F_0 ESTIMATION USING SRH BASED ON TV-CAR SPEECH ANALYSIS

Keiichi Funaki

funaki@cc.u-ryukyu.ac.jp
Computing & Networking Center
University of the Ryukyus
Senbaru 1, Nishihara, Okinawa, Japan

Takehito Higa

Department of Information Engineering
Graduate School of Engineering
University of the Ryukyus
Senbaru 1, Nishihara, Okinawa, Japan

ABSTRACT

This paper proposes novel robust speech F_0 estimation using SRH (Summation Residual Harmonics) based on TV-CAR (Time-Varying Complex AR) analysis. We have already proposed robust F_0 estimation based on the TV-CAR analysis in which weighted auto-correlation for complex residual signals is used as the criterion. In the SRH method, the criterion is calculated from LP residual signals. The criterion is summation of residual spectrum value for harmonics. In this paper, we propose SRH-based F_0 estimation based on the TV-CAR analysis, in which the criterion is calculated from the complex AR residual. Since complex AR residual provides higher resolution of spectrum, the criterion might be effective for F_0 estimation. The experimental results demonstrate that the proposed method performs better than conventional methods.

Index Terms— F_0 estimation, Summation Residual Harmonics (SRH), complex analysis, analytic signal

1. INTRODUCTION

F_0 estimation has been focused on speech processing since its performance decides performance of speech processing such as speech coding, speech enhancement, and speech recognition. Classic F_0 estimation including auto-correlation[1] or Cepstral method[2] perform well for clean speech in ideal environment. Moreover, YIN[3] has been proposed and it is being used in whole world due to its high accuracy. However these would not perform well for noisy speech in real environment. For this reason, robust F_0 estimation still remains an unsolved and challenging problem. Several robust algorithms have been proposed[4][5][6][7][8]. In [4], auto-correlation function is weighted by a reciprocal of AMDF (Average Magnitude Difference Function) so that the peaks of auto-correlation are emphasized, as a result, it can suppress error estimation of F_0. In [5], EMD (Empirical Mode Decomposition) is applied to auto-correlation function and the EMD-based auto-correlation makes it possible to estimate more accurate F_0. In [6], [7], the Zero Frequency Resonance (ZFR) is introduced to realize more accurate F_0 estimation. In ZFR method, Hilbert Envelope (HE) for LP(Linear Predictive) residual is computed. The HE is defined as magnitude of the analytic signal and it accommodates relatively low frequency components. Then, the pulse components of glottal source are emphasized by Zero Frequency Filtering (ZFF) for the HE and the trend removal. The output signal is ZFR signal of the HE. The ZFR signal is merely periodic signal and accurate Glottal Closure Instance (GCI) can be estimated by its positive zero-crossing. The auto-correlation for the ZFR can improve the performance on F_0 estimation. Moreover, SRH(Summation Residual Harmonics)[8] focuses on residual harmonics. Power spectrum of LP residual presents peaks at the harmonics of the F_0. The SRH is calculated by using the power spectrum. The SRH is summation of the harmonics minus the half-harmonics. By peak-picking of the SRH in certain range of F_0, one can estimate the F_0.

On the other hand, we have already proposed F_0 estimation based on time-varying complex AR (TV-CAR) speech analysis[9][10][11]. In these methods the weighted auto-correlation function is calculated by using complex residual for analytic signal. Analytic signal is complex-valued signal whose real part is observed real speech and whose imaginary part is its Hilbert transform. The complex residual is estimated by time-varying complex AR (TV-CAR) speech analysis based on MMSE and ELS estimation[13], respectively. In [10] it is reported that ELS-based method performs better for speech corrupted by pink noise while in [9] it is reported that MMSE-based method performs better for additive white Gauss noise. [11] shows that the time-varying analysis performs better for strong voiced segments. Moreover, the ZFR-based F_0 estimation based on the TV-CAR analysis has been proposed and evaluated[12].

In this paper, SRH-based F_0 estimation based on the TV-CAR speech analysis is proposed and evaluated. The experimental results demonstrate that the proposed method does perform better than the conventional ones[3][4][8].

This paper is organized as follows. Section 2 describes the TV-CAR speech analysis. In Section 3, the SRH method will be explained. In Section 4, the proposed F_0 estimation will be explained. In Section 5, the experiments will be explained.

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

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

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(Linear Predictive Coding) model is defined by

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

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

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

where $a_i^c(t)$,L,$g_{i,l}^c$ and $f_l^c(t)$ are taken to be i-th complex AR coefficient at time t, 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.

$$Y_{TVCAR}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \tag{4}$$

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

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I} \sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \tag{5}$$

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.(3) parametrizes 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

$$\begin{aligned} \bar{y}_f &= -\bar{\Phi}_f \bar{\theta} + \bar{u}_f \\ \bar{\theta}^T &= [\bar{g}_0^T, \bar{g}_1^T, \cdots, \bar{g}_l^T, \cdots, \bar{g}_{L-1}^T] \\ \bar{g}_l^T &= [g_{1,l}^c, g_{2,l}^c, \cdots, g_{i,l}^c, \cdots, g_{I,l}^c] \\ \bar{y}_f^T &= [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)] \\ \bar{u}_f^T &= [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)] \\ \bar{\Phi}_f &= [\bar{D}_0^f, \bar{D}_1^f, \cdots, \bar{D}_l^f, \cdots, \bar{D}_{L-1}^f] \\ \bar{D}_l^f &= [\bar{d}_{1,l}^f, \cdots, \bar{d}_{i,l}^f, \cdots, \bar{d}_{I,l}^f] \\ \bar{d}_{i,l}^f &= [y^c(I-i) f_l^c(I), y^c(I+1-i) f_l^c(I+1), \\ &\qquad \cdots, y^c(N-1-i) f_l^c(N-1)]^T \end{aligned} \tag{6}$$

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[13]

MSE criterion is defined by

$$\begin{aligned} \bar{r}_f &= [r^c(I), r^c(I+1), \cdots, r^c(N-1)]^T \\ &= \bar{y}_f + \bar{\Phi}_f \hat{\theta} \end{aligned} \tag{7}$$

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

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

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 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.

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

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 $\hat{g}_{i,l}^c$.

3. SRH[8]

The SRH relies on harmonics of LP residual power spectrum that provides less spectrum except F_0. It accounts for effect of not only harmonics but also half-harmonics. The procedure is as follows.

(1)LPC analysis is operated with observed speech signal.
(2)LPC inverse filter is operated with the speech signal to obtain the LP residual.
(3)Power spectrum of the LP residual $E(f)$ is calculated by using FFT.
(4)The SRH is calculated by Eq.(11).
(5)F_0 is estimated by searching maximum value for $SRH(f)$ in certain range of F_0, $[F0_{min}, F0_{max}]$.

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)] \quad (11)$$

The function of $SRH(f)$ is summation of the amplitude corresponding to the harmonics minus that to half-harmonics. The function is not only to emphasize the harmonic elements but also to deemphasize the half-harmonic elements. In [8], two-stage estimation is introduced to avoid estimation error of F_0. In first-stage, average F_0 is estimated. It is denoted as $F0_{mean}$. In second stage, F_0 is estimated by searching maximum value for $SRH(f)$ in limited range of F_0, $[F0_{mean}/2, F0_{mean} * 2]$.

The SRH can be also used for voicing decision by simple local threshoulding. It is important to be noted that SRH searches F_0 in the frequency domain. Consequently, the complex analysis can improve the performance since complex analysis can improve the spectral resolution due to the nature of analytic signal.

4. PROPOSED F_0 ESTIMATION

Auto-correlation function (AUTOC) is defined by

$$f(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} x(t)x(t+\tau) \quad (12)$$

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.(12) within certain range of F_0.

AMDF is defined as follows.

$$p(\tau) = \frac{1}{N} \sum_{t=0}^{N-1} |x(t) - x(t+\tau)| \quad (13)$$

F_0 is selected as notch frequency for Eq.(13) within certain range of F_0.

In Shimamura method [4], the AUTOC is weighted by a reciprocal of the AMDF shown as Eq.(14). 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.

$$G(\tau) = \frac{f(\tau)}{p(\tau) + m} \quad (14)$$

where $f(\tau)$ and $p(\tau)$ are AUTOC shown as in Eq.(12) and AMDF shown as in Eq.(13), respectively.

In our F_0 estimation[9][10][11], the weighted auto-correlation for complex AR residual is used as the criterion. Needless to say, in SRH method[8], the SRH shown as Eq.(11) is calculated using LP residual to estimate F_0. In the proposed method, the SRH shown as Eq.(11) is calculated using complex AR residual shown as Eq.(7) to estimate F_0.

The procedure is summarized as follows.

(1)The complex AR residual is computed by Eq.(7) using the TV-CAR analysis for analytic signal.
(2)The power spectrum $E(f)$ is computed for the complex AR residual.
(3)The SRH is computed by Eq.(11).
(4)F_0 search is carried out by peak-picking for the criterion SRH.

The $E(f)$ is non-symmetric one-size spectrum and the resolution is twice larger than that for real-valued analysis. It allows for better performance on the estimation. Note that $F0_{mean}$ is set by the estimated F_0 using Shimamura method[4] in which speech signal is used to calculate the criterion shown as Eq.(14).

5. EXPERIMENTS

In order to compare the performance, the experiments were carried out with Keele Pitch Database[14] corrupted by white Gauss or Pink noise[15] whose noise level was -5, 0, 5, 10, 20, 30[dB]. The noise corrupted speech is filtered by the IRS filter[16] for speech coding application. The proposed method was compared with conventional methods as follows.

(1) Weighted auto-correlation for Speech signal[4]
(2) SRH of time-varying real LP residual
(3) SRH of time-invariant complex AR residual
(4) SRH of time-varying complex AR residual

Experimental conditions are summarized in Table 1. AR order I is 14 for real analysis, 7 for complex analysis. Basis expansion order L is 2 and first order polynomial $(1, t)$ is selected as a basis function. The performance is evaluated by using 10 % of GPE (Gross Pitch Error) and FPE (Fine Pitch

Error). Figures 1 and 2 show the GPEs and FPEs corresponding each method in which black line with black square means GPEs/FPEs for method (1), black line with black diamond means GPEs/FPEs for method (2), red line means GPEs/FPEs for method (3) and blue line means GPEs/FPEs for proposed method (4). Figures 1 and 2 demonstrate that the SRH for LP residual and complex AR residual perform better than Shimamura method[4] in terms of GPE as well as FPE. Moreover, the proposed method can perform better than time-invariant complex-valued residual and real-valued residual in terms of GPE as well as FPE. The reason why the SRH based on complex AR residual perform better is that spectrum resolution is improved due to the nature of analytic signal. The reason why there is a little difference between time-varying and time-invariant speech analysis is that power spectrum is calculated over several pitch periods to compute the SRH, as a result, time-varying analysis cannot always be effective. Figures 3 and 4 show the results for proposed method and YIN[3]. The proposed method performs better than YIN in terms of GPE.

Table 1: Experimental conditions

Speech data	Keele Pitch database [14] Male 5 long sentences Female 5 long sentences
IRS filter	64-th FIR
Sampling	10kHz/16bit
Analysis window	Window Length: 25.6[ms] Shift Length: 10.0[ms]
F_0 search range	$50 - 400$[Hz]
Complex-valued AR Target signal	I=7, L=2 (time-varying) complex AR residual
Real-valued AR Target signal	I=14, L=2 (time-varying) real AR residual
Noise Noise Level	(1)white Gauss noise (2)pink noise[15] 30,20,10,5,0,-5[dB]
SRH	$N_{harm} = 5$

6. CONCLUSIONS

This paper has proposed the SRH-based F_0 estimation using the TV-CAR speech analysis. The SRH is summation of harmonics minus half-harmonics for complex AR residual calculated by the TV-CAR speech analysis. The performance comparison was carried out using five kinds of methods. The experimental results demonstrate that the proposed TV-CAR residual-based SRH performs best. As a future work, we are going to propose the SRH voicing detection based on the TV-CAR analysis and to evaluate it.

7. REFERENCES

[1] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust., vol. 20(5), pp.367-377, Dec. 1972.

[2] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Amer., vol. 41(2), pp. 293-309, Feb. 1967.

[3] Alain de Cheveigne and H.Kawahara, YIN, "A fundamental frequency estimator for speech and music," Journal of the Acoustical Society of America, Vol.111, No.4, pp.1917-1930. 2002.

[4] T.Shimamura and H.Kobayashi,"Weighted Autocorrelation for Pitch Extraction of Noisy Speech," IEEE Trans. Speech and Audio Processing, vol. 9, no. 7, pp. 727-730, 2001.

[5] S.K.Roy, Md.Khademul I.Molla, K.Hirose, M.K.Hasan, "Pitch estimation of noisy speech signals using EMD-fourier based hybrid algorithm," Proc. ISCAS2010, 2010.

[6] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," IEEE Trans. Audio, Speech, Lang. Process., vol. 17(4), pp.614-624, May 2009.

[7] B. Yegnanarayana, S.R.M. Prasanna, and S.Guruprasad, "Study of robustness of zero frequency resonator method for extraction of fundamental frequency," Proc. ICASSP2011

[8] T.Drugman and A.Alwan,"Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," Proc. of Interspeech2011, Firenze, Italy, Sep. 2011.

[9] K.Funaki,et.al., "Robust F0 Estimation Based on Complex LPC Analysis for IRS Filtered Noisy Speech," IEICE Trans. on Fundamentals, Vol. E90-A, No.8.,1579-1586, Aug. 2007.

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

[11] K.Funaki, "On Evaluation of the F_0 estimation based on time-varying complex speech analysis," Proc. Interspeech2010, Makuhari, Japan, Sep. 2010.

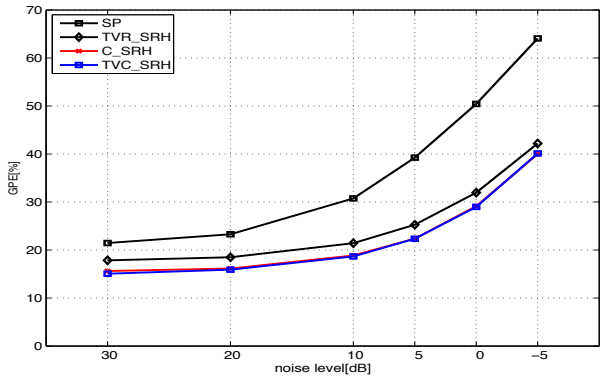
[12] K.Funaki and T.Higa, "Evaluation of F0 estimation using ZFR based on time-varying speech analysis," Proc. of ISCAS2012, Seoul, May, 2012.

[13] K.Funaki, "A Time-Varying Complex AR Speech Analysis Based on GLS and ELS Method," EUROSPEECH-2001,Aalborg, Denmark, Sep.7,2001.

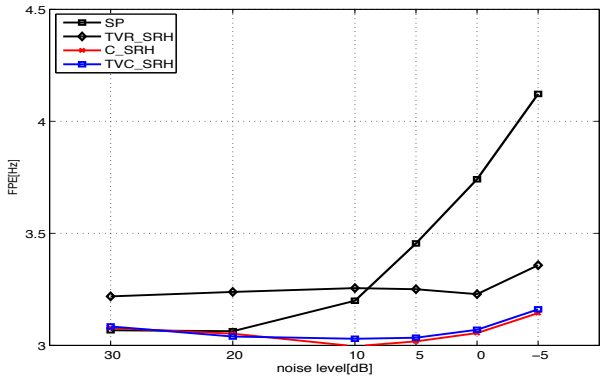
[14] Keele Pitch Database, University of Liverpool, http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html

[15] NOISE-X92, http://spib.rice.edu/spib/select_noise.html

[16] ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, Nov. 2000.

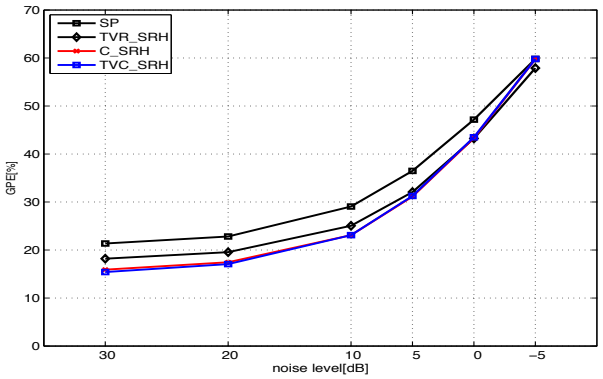


(1)GPEs for additive white Gauss noise

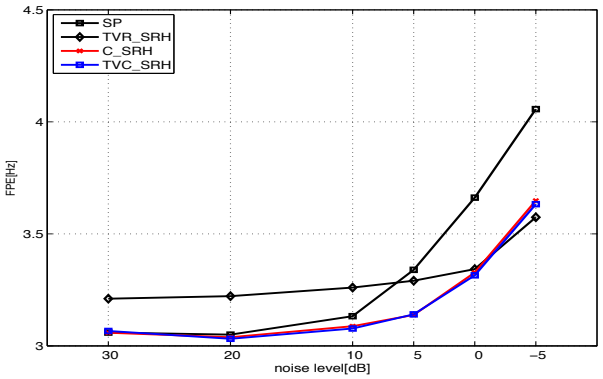


(2)FPEs for additive white Gauss noise

Figure 1: Experimental results (Gaussian noise)

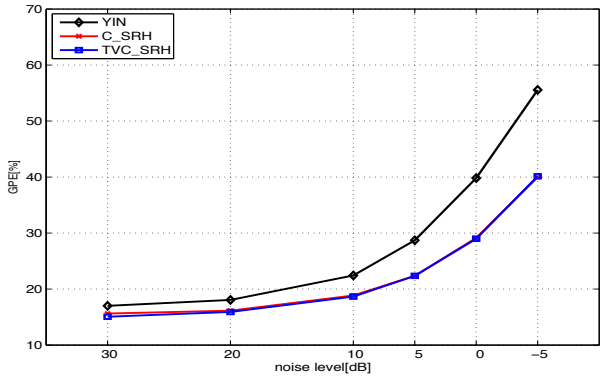


(1)GPEs for additive pink noise

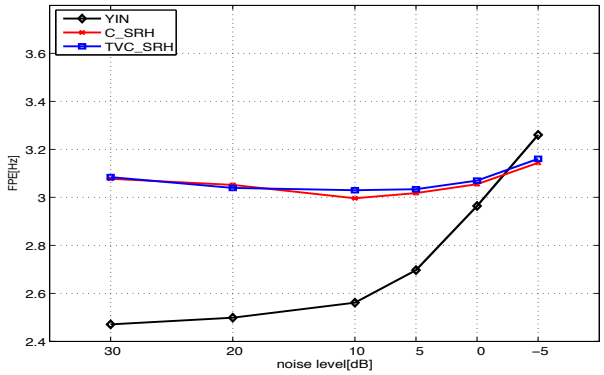


(2)FPEs for additive pink noise

Figure 2: Experimental results (Pink noise)

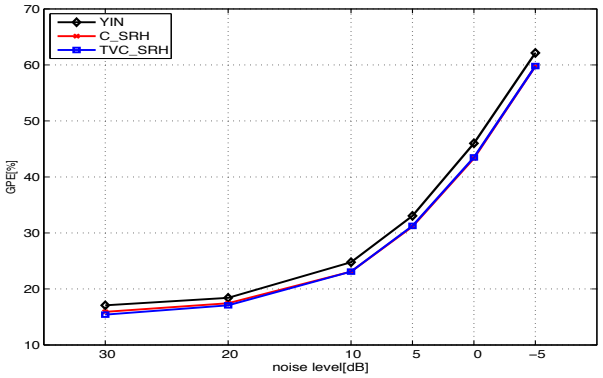


(1)GPEs for additive white Gauss noise

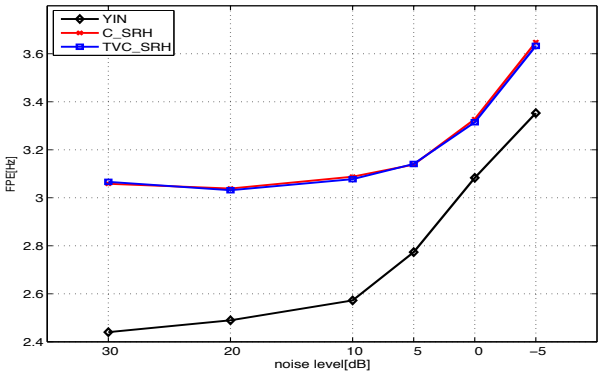


(2)FPEs for additive white Gauss noise

Figure 3: Experimental results (Gaussian noise)



(1)GPEs for additive pink noise



(2)FPEs for additive pink noise

Figure 4: Experimental results (Pink noise)