ON AN IMPROVED F_0 ESTIMATION BASED ON ℓ_2-NORM REGULARIZED TV-CAR SPEECH ANALYSIS

Keiichi Funaki*
* IT Center, University of the Ryukyus, Okinawa, 903-0213, Japan
E-mail: funaki@cc.u-ryukyu.ac.jp

***Abstract*—Spectral estimation performance determines that of speech processing. Linear Prediction (LP) is the most successful speech analysis method commonly introduced worldwide of a smartphone, LINE, Skype to realize CELP speech coding to extract the spectral features with a small amount of computation and fewer parameters. Besides the CELP coding, the LP performs better on the F_0 estimation since the LP residual contains fewer formant structures. We have already proposed a time-varying complex AR (TV-CAR) speech analysis for an analytic signal that estimates the time-varying complex AR parameters from a speech signal. We recently proposed the TV-CAR analysis based on ℓ_2-norm regularized LP and evaluated the effectiveness of the performance on the F_0 estimation using the IRAPT algorithm. On the other hand, bone-conducted (BC) speech is robust against additive noise since it provides a stable harmonic structure in low frequencies and cannot be easily affected by the noise. In this paper, we introduce a pre-filter that simulates BC speech to improve the performance of the F_0 estimation. The experimental results show that the BC filter improves the performance for the high level of noise corrupted female speech.**

I. Introduction

Speech analysis, extracting the spectral features from a speech signal, is a dominant technique in speech processing. The Linear Prediction (LP) proposed in the 1960s[1] is commonly used in CELP speech coding implemented in a smartphone, SKYPE, LINE, ZOOM, or TEAMS. In the CELP coding, the LP residual is predicted using an adaptive codebook, and the resulting residual signals are quantized using the codebook such as a VQ codebook or an Algebraic sparse codebook with several pulses having the value of ±1[2][3]. Furthermore, in ALS(audio lossless coding)[4], the LP is used to compute the residual quantized using entropy coding. The LP is also applied in speech processing, including F_0 estimation, speech enhancement, robust automatic speech recognition (ASR), and speech synthesis. The LP residual is applied to compute the criterion on F_0 estimation. The LP residual is used instead of the speech signal since the LP residual provides fewer formant elements. As a result, it can avoid error estimation such as double pitch, half-pitch and first formant F_1. The auto-correlation of the LP residual is sometimes called the modified auto-correlation method[5]. As a speech enhancement, iterative Wiener Filter(IWF)[6] is being used in which the Wiener filter is designed using the estimated LP spectrum. On the other hand, an augmented Kalman filter (AKF) is applied to suppress the additional noise in which the LP filter is used to estimate the spectrum[7][8][9]. The LP also plays a vital role to improve speech dereverberation[10][11] and Glottal Closure Instant (GCI) detection[12]. In the robust ASR, the IWF reduces the additional noise in ETSI Advanced FrontEnd(AFE)[13]. Although the FFT spectrum is introduced to design the IWF in the AFE, we have already shown that the LP spectrum performs better than the FFT spectrum[14]. Even in the speech synthesis, the LP is embedded. Recently the development of the WaveNet[15] brings a new era to speech synthesis and dramatically improved speech quality. Several improved WaveNet methods have been proposed including GlotNet[16][17], ExcitNet[18], FFTNet[19], LPCNet[20], LP-WaveNet[21] and so on. The GlotNet generates the excitation using a glottal excitation estimated by a glottal inverse filtering. The ExcitNet generates the excitation using the LP residual estimated by the LP inverse filter. The ExcitNet provides more rich excitation, including the noise elements besides glottal excitation, improving speech quality.

The expansion of the LP has been examined in more than a half-century. One approach is to expand the ARMA analysis[22]. These are not so effective since the speech signal does not provide a strong anti-resonance, and the excitation cannot be estimated accurately. The other approach is to expand time-varying analysis[23][24], estimating time-varying spectral features from speech signals by representing the AR coefficients using the basis expansion. The other approach is to expand complex analysis for an analytic signal that can estimate a more accurate speech spectrum due to the nature of the signal. The other approach is to introduce robust criterion instead of the ℓ_2-norm, viz. MMSE estimation. For example, ℓ_0-norm optimization is being introduced in [27]. While ℓ_1-norm optimization is being introduced in [28], compress sensed (CS) ℓ_1-norm optimization is being introduced in [29]. As a ℓ_2 regularized LP, B.Kleijn et.al. proposed RLP (Regularized LP)[30] and P.Alku et.al. proposed TRLP (Time-RLP)[31]. The RLP suppresses rapid changes in the frequency domain to avoid pitch related bias, and The TRLP suppresses rapid changes in the time domain.

We have proposed a Time-Varying Complex AR (TV-CAR) speech analysis based on MMSE (Minimizing Mean Squared Error)[32] that is the combination of time-varying analysis and complex analysis. Moreover, we have proposed GLS(Generated Least Square), ELS(Extended Least Square)[33], LASSO(Least Absolute Shrinkage and Selection Operator)[36], RLP(Regularized LP)[34], TRLP(Time-Regularized LP), and RLP-based & TRLP-based hybrid method[35] are ℓ_2-norm regularized methods while LASSO-based method[36] is the ℓ_1-norm regularized method. We have evaluated the proposed TV-CAR analysis with the IWF[37] in robust ASR[38] and robust F_0 estimation [39]. The IWF is implemented using the ETSI AFE, and F_0 estimation is implemented using the IRAPT (Instantaneous RAPT)[40].

Recently, a bone-conducted (BC) headset has been commonly used because it offers no earphones and noise robustness. The BC speech cannot be affected by additive noise since it provides a stable harmonic structure in low frequencies. The feature can be utilized to improve the performance of the F_0 estimation. A simple AR filter can simulate the BC since it provides low pass filter characteristics. This paper aims to improve the F_0 estimation based on the TV-CAR speech analysis using the IRAPT by introducing the BC filter. The first order of AR filter realizes the BC filter, and it is combined with the pre-emphasis filter. The F_0 estimation is operated by using the pre-filtered speech, and the complex residual is computed with the pre-filtered speech. We conducted the objective evaluation on F_0 estimation using the Keele pitch database[41]. The experimental results show that the BC filter makes it possible to improve the performance for female speech, although it does not make it worse for male speech.

II. Regularized LP

A. LP Analysis

LP analysis is based on an ℓ_2-norm optimization estimating an ith auto-regressive (AR) coefficient $a_i (i = 1, 2, ..., I)$ to minimize the Mean Squared Error (MSE) for the AR model shown in Eq.(1).

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

The power spectrum of the AR model is represented by Eq.(2).

$$S(\omega, \mathbf{a}) = 1/|A(e^{j\omega})|^2 \tag{2}$$

In the LP analysis, the ℓ_2-norm criterion is shown in Eq.(3).

$$\mathcal{D} = E\left[e^2(t)\right] = \mathbf{a}^T\mathbf{R}\mathbf{a} + 2\mathbf{a}^T\mathbf{r} + r_0 \tag{3}$$

where $E[]$ is an expectation, $e(t)$ is the residual signal at time t, $\mathbf{R}$ is the symmetric Toeplitz matrix whose elements are the auto-correlation function $r_i (i = 0, 1, ..., I-1)$, $\mathbf{a}$ is $[a_1, a_2, ..., a_I]^T$, $\mathbf{r}$ is $[r_1, r_2, .., r_I]^T$ and T means Transpose. Minimizing Eq.(3) viz., $\partial\mathcal{D}/\partial\mathbf{a}^T = 0$ results in the following linear equation called Yule-Walker equation.

$$\mathbf{R}\hat{\mathbf{a}} \;\; = \;\; -\mathbf{r} \tag{4}$$

B. Time-Regularized LP(TRLP) Analysis[31]

The TRLP analysis sets ℓ_2-norm for the difference between current parameter vector $\mathbf{a}$ and the previous one $\mathbf{a_{pr}}$, shown as in Eq.(5), as the ℓ_2-norm regularization term.

$$\mathcal{L}_{reg} = \frac{1}{2}\lambda_1 \left(\mathbf{a} - \lambda_2 \mathbf{a}_{\mathrm{pr}}\right)^T \left(\mathbf{a} - \lambda_2 \mathbf{a_{pr}}\right) \tag{5}$$

where λ_1 and λ_2 are regularization factor. $\mathcal{L}_{reg}$ means the penalty term that suppresses rapid spectral changes between adjacent frames. The criterion of the TRLP is $D + \mathcal{L}_{reg}$, thus, from $\partial(\mathcal{D} + \mathcal{L}_{reg})/\partial\mathbf{a}^T = 0$, one can obtain

$$\mathbf{r} + \mathbf{R}\mathbf{a} + \lambda_1\mathbf{a} - \lambda_1\lambda_2\mathbf{a_{pr}} = \mathbf{0}. \tag{6}$$

As a result, we derive the following linear equation shown in Eq.(7).

$$\left(\mathbf{R} + \lambda_1\mathbf{I}\right)\hat{\mathbf{a}} = -\mathbf{r} + \lambda_1\lambda_2\mathbf{a_{pr}} \tag{7}$$

TRLP can be realized by solving Eq.(7). It is worth noting that if λ_2 is 0, the TRLP analysis is equal to Ridge analysis.

C. Regularized LP (RLP) Analysis[30]

LP analysis suffers from pitch-related bias to estimate the unnaturally sharp peak of the F_1 for high pitch speech. To solve the problem, the RLP analysis introduces an ℓ_2-norm regularization term shown in Eq.(8) that means ℓ_2-norm of the AR spectral changes in the frequency domain.

$$\mathcal{R}(S(\omega, \mathbf{a})) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{d}{d\omega}\log S(\omega, \mathbf{a})\right]^2 d\omega \tag{8}$$

The criterion of the RLP is $\mathcal{D} + \lambda_3\mathcal{R}$. λ_3 is called the regularization constant that controls the contribution for the regularized term. The second term means the penalty one that suppresses rapid spectral changes in the frequencies. To estimate the parameter, $\mathbf{a}$, with no iteration, Eq.(8) is approximated to be Eq.(9).

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{d}{d\omega}\log A(e^{j\omega})\right|^2 d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{A'\left(e^{j\omega}\right)}{A\left(e^{j\omega}\right)}\right|^2 d\omega \tag{9}$$

By using Eq.(9), Eq.(8) turns to be Eq.(10).

$$\hat{\mathcal{R}}(S(\omega, \mathbf{a})) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{A'\left(e^{j\omega}\right)}{W(\omega)}\right|^2 d\omega \tag{10}$$

where $|W(\omega)|^2$ is a crude estimation of $|A(\omega)|^2$.

$$A'(e^{j\omega}) = -\sum_{k=0}^{M} jka_k e^{jk\omega} \tag{11}$$

As a result, Eq.(10) reduces to Eq.(12).

$$\sum_{k=0}^{I}\sum_{m=0}^{I} ka_k ma_m \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{-j\omega(k-m)}}{|W(\omega)|^2} d\omega \tag{12}$$

Since the integral term in Eq.(12) is an inverse discrete transform of $|1/W(\omega)|^2$, Eq.(10) reduces to Eq.(13).

$$\hat{\mathcal{R}}(S(\omega,\mathbf{a})) = \sum_{k=0}^{I}\sum_{m=0}^{I} k a_k m a_m h(m-k) \tag{13}$$

where

$$h(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{j\omega x}}{|W(\omega)|^2}d\omega \tag{14}$$

that is the inverse Fourier transform of the power spectrum so that it is the auto-correlation function. Finally, Eq.(13) reduces to Eq.(15).

$$\hat{\mathcal{R}}(S(\omega,\mathbf{a})) = \mathbf{a}^T\mathbf{D}^T\mathbf{F}\mathbf{D}\mathbf{a} \tag{15}$$

where $\mathbf{D}$ is a diagonal matrix whose element is $d(i,i) = i$, $\mathbf{F}$ is a Toeplitz auto-covariance matrix. From Eq.(3) and Eq.(15), the criterion of RLP, $\mathcal{D} + \lambda_3\mathcal{R}$ is as follows.

$$\mathbf{a}^T(\mathbf{R} + \lambda_3\mathbf{D}^T\mathbf{F}\mathbf{D})\mathbf{a} + 2\mathbf{a}^T\mathbf{r} + r_0 \tag{16}$$

Minimizing Eq.(16), $\partial(\mathcal{D} + \lambda_3\mathcal{R})/\partial\mathbf{a}^T = 0$ results in the following linear equation.

$$(\mathbf{R} + \lambda_3\mathbf{D}^T\mathbf{F}\mathbf{D})\hat{\mathbf{a}} = -\mathbf{r} \tag{17}$$

The RLP analysis can be realized by solving Eq.(17). It is worth noting that if λ_3 is 0, the RLP analysis is the same as the LP analysis in Eq.(4).

III. Regularized TV-CAR method

A. TV-CAR model

Eq.(18) defines the TV-CAR model.

$$\begin{aligned} Y_{TVCAR}(z^{-1}) &= \frac{1}{A(z^{-1})} = \frac{1}{1+\sum_{i=1}^{I} a_i^c(t) z^{-i}} \\ &= \frac{1}{1+\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \end{aligned} \tag{18}$$

where $a_i^c(t)$, L, $g_{i,l}^c$ and $f_l^c(t)$ are ith complex AR coefficient at time t, an order of complex basis expansion, complex parameter and complex basis function, respectively. Eq.(19) denotes the input-output relationship for Eq.(18).

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \tag{19}$$

where $y^c(t)$ is the target analytic signal at time t and $u^c(t)$ is a complex input signal at time t. The analytic signal is a complex-valued signal whose real part is the speech signal, and the imaginary part is the Hilbert transformed signal of the real one. Since the analytic signal yields the spectrum only over positive frequencies, the signal can be decimated by a factor of two; consequently, the complex analysis can estimate a more accurate spectrum in low frequencies. Moreover, the TV-CAR analysis is a time-varying analysis that introduces complex basis expansion of the AR parameter to represent the parameter as a function of time, enabling the parameter estimation in every sample.

Alternatively, Eq.(19) can be formulated by the following vector-matrix representation.

$$\begin{aligned} \mathbf{y}_f &= -\mathbf{\Phi}_f\theta + \mathbf{u}_f \\ \bar{\theta}^T &= [\mathbf{g}_0^T, \mathbf{g}_1^T, \cdots, \mathbf{g}_l^T, \cdots, \mathbf{g}_{L-1}^T] \\ \mathbf{g}_l^T &= [g_{1,l}^c, g_{2,l}^c, \cdots, g_{i,l}^c, \cdots, g_{I,l}^c] \\ \mathbf{y}_f^T &= [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)] \\ \mathbf{u}_f^T &= [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)] \\ \mathbf{\Phi}_f &= [\mathbf{S}_0^f, \mathbf{S}_1^f, \cdots, \mathbf{S}_l^f, \cdots, \mathbf{S}_{L-1}^f] \\ \mathbf{S}_l^f &= [\mathbf{s}_{1,l}^f, \mathbf{s}_{2,l}^f, \cdots, \mathbf{s}_{i,l}^f, \cdots, \mathbf{s}_{I,l}^f] \\ \mathbf{s}_{i,l}^f &= [y^c(I-i)f_l^c(I), y^c(I+1-i)f_l^c(I+1), \\ &\qquad \cdots, y^c(N-1-i)f_l^c(N-1)]^T \end{aligned} \tag{20}$$

where N is analysis length, $\mathbf{y}_f$ is $(N-I,1)$ column vector whose element is the analytic signal, $\bar{\theta}$ is $(L\cdot I,1)$ column vector whose element is the complex parameter, $\mathbf{\Phi}_f$ is $(N-I, L\cdot I)$ matrix whose element is the weighted analytic signal by a complex basis.

B. MMSE algorithm

The MMSE algorithm is an ℓ_2-norm optimization realized by Minimizing the MSE for the equation error.

$$\hat{\theta} = \arg\min_{\bar{\theta}} \|\mathbf{y}_f + \mathbf{\Phi}_f\bar{\theta}\|_2^2 \tag{21}$$

Minimizing the MSE for the equation error leads to the following MMSE algorithm.

$$\left(\mathbf{\Phi}_f^H\mathbf{\Phi}_f\right)\hat{\theta} = -\mathbf{\Phi}_f^H\mathbf{y}_f \tag{22}$$

where H is an Hermite operator, it is the time-varying, complex and covariance analysis version of the conventional LP, Eq.(4).

C. RLP-based TV-CAR analysis[34]

Since the TV-CAR analysis is the complex, time-varying and covariance type of LP analysis, Eq.(23) can be derived by integrating the RLP onto the TV-CAR analysis. The ℓ_2-norm regularized term, the power spectrum at the center sample of the frame, $N/2$, is applied.

$$\left(\mathbf{\Phi}_f^H\mathbf{\Phi}_f + \lambda_3\mathbf{D}_{tv}^H\mathbf{F}\mathbf{D}_{tv}\right)\hat{\theta} = -\mathbf{\Phi}_f^H\mathbf{y}_f \tag{23}$$

where λ_3 is the regularization factor that controls the contribution for the regularized term, and $\mathbf{D}_{tv}$ is defined as follows.

$$\mathbf{D}_{tv} = [\mathbf{d_0}, \mathbf{d_1}, ..., \mathbf{d_l}, ..., \mathbf{d_{L-1}}] \tag{24}$$

$$\mathbf{d_l} = \mathbf{diag}[f_l^c(N/2), 2f_l^c(N/2), ..., If_l^c(N/2)] \tag{25}$$

$\mathbf{d_l}$ is (I,I) diagonal matrix and $\mathbf{D_{tv}}$ is $(I, L\cdot I)$ matrix that is generated by aligning L number of $\mathbf{d_l}(l = 0,1,...,L-1)$.

D. TRLP-based TV-CAR method

The TRLP-based TV-CAR algorithm is realized as follows.

$$\hat{\theta} = \arg\min_{\bar{\theta}} \|\mathbf{y}_f + \mathbf{\Phi}_f \bar{\theta}\|_2^2 + \frac{1}{2}\lambda_1 \|\bar{\theta} - \lambda_2 \hat{\theta}_{\mathrm{pr}}\|_2^2 \tag{26}$$

where $\hat{\theta}_{\mathbf{pr}}$ is the parameter estimated in the previous frame. The linear equation can be easily derived as the TRLP-based TV-CAR method.

$$\left(\mathbf{\Phi}_f^H \mathbf{\Phi}_f + \lambda_1 \mathbf{I}\right) \hat{\theta} = -\mathbf{\Phi}_f^H \mathbf{y}_f + \lambda_1 \lambda_2 \hat{\theta}_{\mathbf{pr}} \tag{27}$$

E. RLP and TRLP-based Hybrid TV-CAR analysis

Furthermore, by combining Eq.(23) and Eq.(27), we can easily derive the following hybrid approach of the RLP and TRLP.

$$\left(\mathbf{\Phi}_f^H \mathbf{\Phi}_f + \lambda_1 \mathbf{I} + \lambda_3 \mathbf{D}_{tv}^H \mathbf{F} \mathbf{D}_{tv}\right) \hat{\theta} = -\mathbf{\Phi}_f^H \mathbf{y}_f + \lambda_1 \lambda_2 \hat{\theta}_{\mathbf{pr}} \tag{28}$$

IV. Pre-operation

In speech processing, the input speech from a microphone is air-conducted (AC) sound. The sound we hear with our ears contains many bone-conducted (BC) components transmitted through the skull. Unlike AC sound, the BC component is not easily affected by noise, so it is thought that utilizing the BC characteristics lead to improve the noise reduction performance. As a filter with BC characteristics, we introduce the ARMA filter in Eq(29).

$$H(z) = \frac{\beta(1-\gamma z^{-1})}{1-\alpha z^{-1}} \tag{29}$$

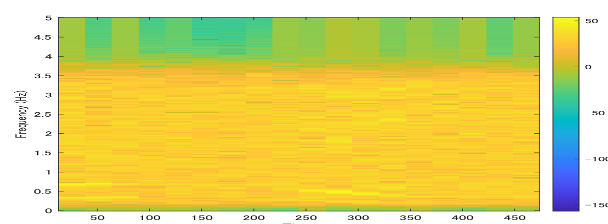
The ARMA filter represented by Eq.(29) is applied as pre-processing, TV-CAR analysis is performed, an inverse filter calculates complex residuals, and IRAPT performs F_0 estimation using the residuals. Fig.1 shows the spectrogram for AC and BC filtered female speech corrupted by -5[dB] Pink noise. According to informal listening, the BC filtered speech contains fewer noise components. Fig.1 demonstrates that the lower spectral components are emphasized in the BC filtered speech compared to the AC speech. It can be thought that the BC filter is effective for the F_0 estimation.

V. Experiments

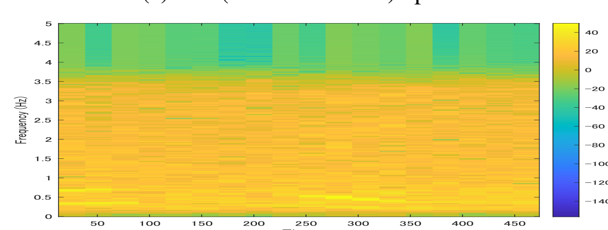
The proposed RLP and TRLP-based hybrid TV-CAR method with the pre-filter is compared with conventional methods using the F_0 estimation in noisy environments. The following signals are applied in the performance comparison,

(1)The real residual computed by the LP with the IRAPT.
(2)The complex residual computed by the MMSE-based TV-CAR has shown in Eq.(22)[32] with the IRAPT.
(3)The complex residual computed by the RLP-based TV-CAR has shown in Eq.(23)[34] with the IRAPT.
(4)The complex residual computed by the hybrid approach of the RLP and TRLP-based TV-CAR shown in Eq.(28) with the IRAPT.
(5)Proposed method. The complex residual computed by the hybrid approach of the RLP and TRLP-based TV-CAR shown in Eq.(28) for the BC speech resulting from the pre-filtering using Eq.(29).

Keele pitch database[41] added by white Gauss or Pink noise[42] is applied for evaluation. The noise-corrupted signal is filtered by the IRS filter[43] for speech coding applications. Gross Pitch Error(GPE) and Fine Pitch Error(FPE) are adopted as the objective criterion. The pitch database provides the true F_0. If the estimation error is smaller than p-percent of the true F0, the estimation is regarded as SUCCEED. Otherwise, the estimation is regarded as FAILURE. The GPE is a percentage of FAILURE frames, and the FPE is a variance of the estimation error at the SUCCEED frames. The experimental conditions are shown in Table 1. Figures 2 and 3 show the experimental results for Male and Female speech, only Female speech, respectively. In the figures, (a) and (c) mean 10[%] of GPEs and (b) and (d) mean 10[%] of FPEs. The five lines indicate as follows. The solid black line with lozenge means (1)**LPC_IRAPT2** with the IRAPT. The blue line means (2)**TVC_IRAPT2C** with the IRAPT. The red line means (3)**TVC_RLP_IRAPT2** with the IRAPT. The green line means (4)**TVC_HTRLP_IRAPT2** with the IRAPT. The solid black line with square means (5)Proposed **TVC_HTRLPBC_IRAPT2** for BC speech. Fig.3 demonstrates that the BC improves a high level of noise corrupted female speech on GPE. Fig.2 demonstrates that the BC improves a high level of white Gauss noise corrupted speech on GPE. The GPE is more critical than FPE since the GPE means fatal estimation error such as double pitch or half-pitch, leading to low performance on speech processing. Fig.2 also demonstrates that the performance is down for the pink noise corrupted speech for male and female speech. The reason is that the performance for male speech is down by introducing the BC pre-filter. It is worth noting that the original IRAPT for speech signal is omitted since the performance is much lower than real and complex residual signals[34][39].



(a) AC (non BC filtered) speech



(b) BC filtered speech

Fig.1: spectrogram of the BC filtered speech

Table 1: Experimental Conditions

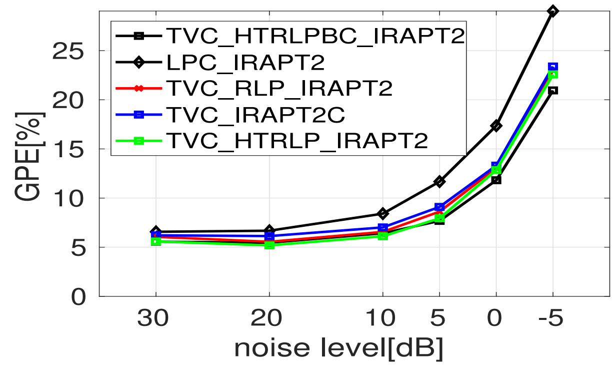
Speech data	Keele Pitch Database[41] 5 long Male sentence 5 long Female sentence
Sampling	10kHz/16bit
Analysis window	Window Length: 25.6[ms] Shift Length: 10.0[ms]
TV-CAR Basis Pre-emphasis	$I = 7, L = 2$(Time-Varying) $f_l^c(t) = t^l/l!$ Eq.(29)
TRLP/RLP	$\lambda_1 = 0.02, \lambda_2 = 0.99$ / $\lambda_3 = 0.0001$
Noise Noise Level	White Gauss or Pink noise[42] 30,20,10,5,0,-5[dB]

VI. Conclusions

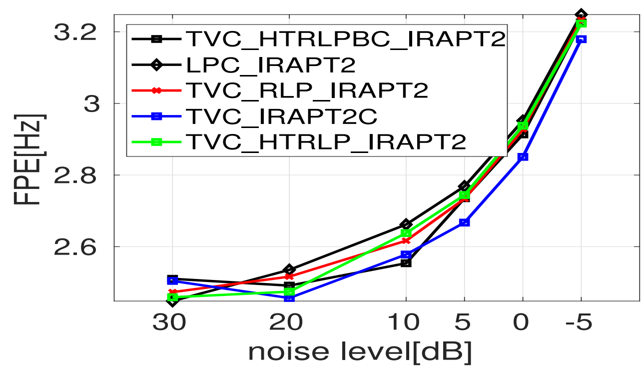
We have proposed the improved F_0 estimation based on the regularized LP-based TV-CAR speech analysis method, a hybrid approach of the TRLP and RLP[34] that introduces ℓ_2-norm regularized LP in the time and frequency domain. The ℓ_2-norm regularized terms penalize the rapid changes of the estimated spectrum in the time-domain and frequency-domain, making it possible to suppress pitch-related bias, overestimation of the first formant. We have already evaluated the speech analysis on the F_0 estimation and have shown that it leads to better performance. This paper introduces the BC filter as the pre-operation combined with the pre-emphasis filter to improve the performance. The BC components provide more stable harmonics in low frequencies so that it can be expected that it is robust against additional noise. The first order AR filter realizes the BC characteristics, and it improves the F_0 estimation performance. The objective evaluation is compared with the conventional methods employing F_0 estimation using the estimated complex residual for IRS filtered Keele pitch database added by white Gauss or Pink noise. The experimental results illustrate that the BC performs better than the conventional method, especially for female speech. Although the performance for male speech is down, we found out that the performance is not so bad for the other pre-filter coefficients—the poor performance results from the simplest AR filter. We are convinced that a more appropriate pre-filter can bring better performance. The investigation of more complicated and precise pre-filter is a continuous way. PEFAC[44][45] is more popular and more accurate F_0 estimation and it is open sourced on VOICEBOX[46]. We intend to adopt PEFAC instead of IRAPT as the F_0 estimation. Moreover, we aim to evaluate the proposed methods on a front-end of robust ASR[14][47]. Besides, sparse TV-CAR analysis based on the LASSO[36][48][49], or Elastic Net will be proposed and be evaluated on speech processing.

VII. Acknowledgement

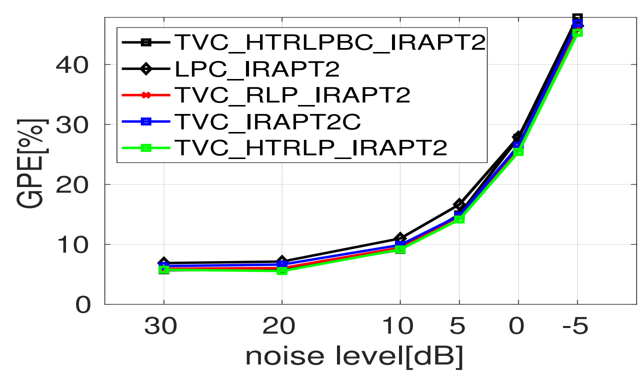
We wish to thank two professors and several staff members in the IT centre of the University of the Ryukyus for their support to achieve this research.



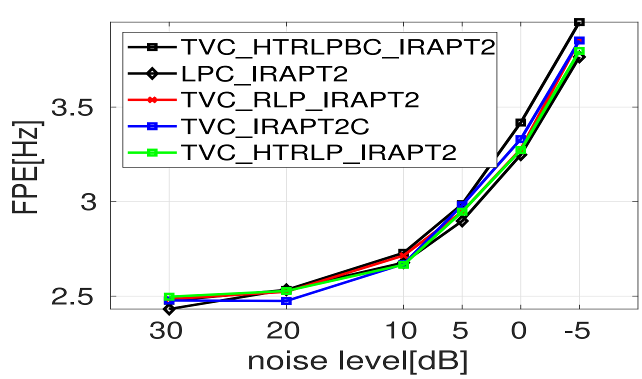
(a)GPE(10%) for additive white Gauss noise



(b)FPE(10%) for additive white Gauss noise

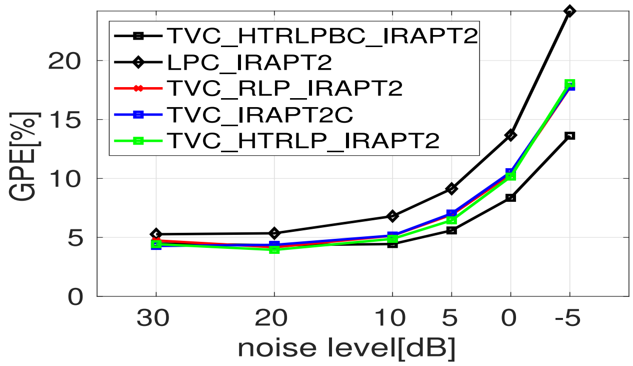


(c)GPE(10%) for additive Pink noise

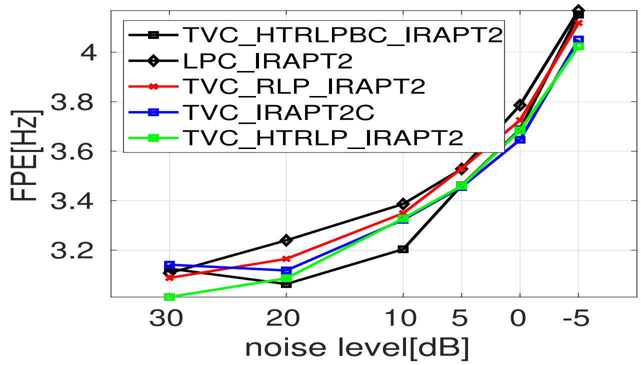


(d)FPE(10%) for additive Pink noise

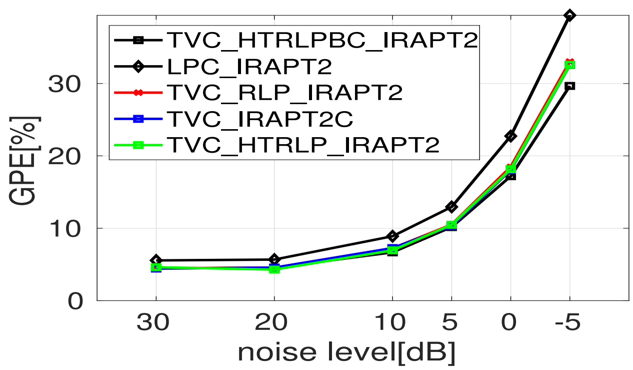
Fig.2: F_0 estimation performance



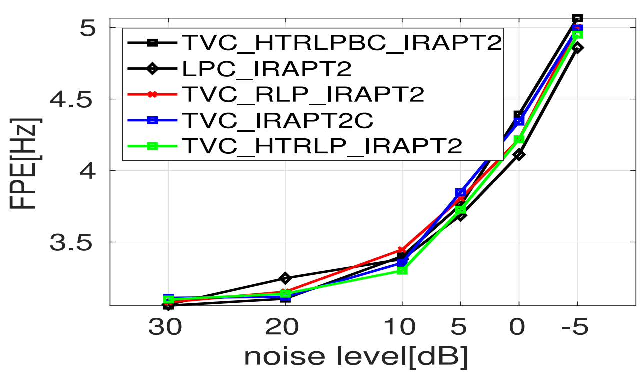
(a)GPE(10%) for additive white Gauss noise



(b)FPE(10%) for additive white Gauss noise



(c)GPE(10%) for additive Pink noise



(d)FPE(10%) for additive Pink noise

Fig.3: F_0 estimation performance for only Female speech

REFERENCES

[1] J.Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, Vol. 63, No. 4, pp. 561-580, Apr. 1975.

[2] ITU-T G.729: "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)," Mar., 1996.

[3] "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Service Option 62 for Spread Spectrum Systems," 3GPP2. C.S0052-0 Version 1.0, pp.73-85, 3GPP2, June, 2004.

[4] T.Liebchen, T.Moriya, N.Harada, Y.Kamamoto, and Y.A.Reznik, "The MPEG-4 Audio Lossless Coding (ALS) Standard - Technology and Applications," Audio Engineering Society, Convention Paper 119th Convention, Oct., New York, NY, USA, 2005. http://elvera.nue.tu-berlin.de/files/0737Liebchen2005.pdf

[5] Sadaoki Furui, "Digital Speech Processing, Synthesis, and Recognition Synthesis, and Recognition, Second Edition," CRC Press, 2001.

[6] J.S.Lim and A.Oppenheim, "All-pole Modeling of Degraded Speech," IEEE Tran. ASSP, 1978.

[7] A. Kawamura, K. Fujii, Y. Itoh, and Y. Fukui, "A new noise reduction method using estimated noise spectrum," IEICE Trans. Fundamentals, vol.E85-A, no.4, pp.784-789, April 2002.

[8] A. Kawamura, K. Fujii, Y. Itoh, and Y. Fukui, "A new noise reduction method using linear prediction error filter and adaptive digital filter," Proc. ISCAS-2002 May 2002.

[9] N.Nower, Y.Liu, and M.Unoki "Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement," Speech Communication, June 2015.

[10] Y.Liu, S.Morita, M.Unoki, "MTF-Based Kalman Filtering with Linear Prediction for Power Envelope Restoration in Noisy Reverberant Environments," IEICE Trans. E99-A, 2016.

[11] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of Late Reverberation Effect on Speech Signal Using Long-term Multiple-step Linear Prediction," IEEE Trans. ASLP., Vol.17, No.4, pp.534-545, 2009.

[12] T. Drugman and T. Dutoit, "Glottal Closure and Opening Instant Detection from Speech Signals," Proc. INTERSPEECH, 2009.

[13] ETSI Advanced Front-End, ES 202 050 v1.1.5(2007-01), Jan.2007.

[14] K.Higa and K.Funaki, "Robust ASR Based on ETSI Advanced Front-End Using Complex Speech Analysis," IEICE Trans. Vol.E98-A, No.11, 2015.

[15] A.van den Oord, S.Dieleman, H.Zen, K.Simonyan, O.Vinyals, A.Graves, N.Kalchbrenner, A.Senior, K.Kavukcuoglu, "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016.

[16] L.Juvela, V.Tsiaras, B.Bollepalli, M.Airaksinen, J.Yamagishi, P.Alku, "Speaker-independent raw waveform model for glottal excitation," Proc. Interspeech-2018, 2018.

[17] L.Juvela, B.Bollepalli, V.Tsiaras, and P.Alku, "GlotNet-A Raw Waveform Model for the Glottal Excitation in Statistical Parametric Speech Synthesis," IEEE/ACM Trans. on ASLP., 2019.

[18] E.Song, K.Byun, H-G.Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," Proc. EUSIPCO-2019, Spain, Sep.2019.

[19] Z.Jin, A.Finkelstein, G.J.Mysore, J.Lu, "FFTnet: A Real-Time Speaker-Dependent Neural Vocoder," Proc. ICASSP-2018, 2018.

[20] J-M.Valin, J.Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," https://arxiv.org/abs/1810.11846

[21] M-J.Hwang, F.Soong, F.Xie, X.Wang, H-G.Kang, "LP-WaveNet: Linear Prediction-based WaveNet Speech Synthesis," https://arxiv.org/abs/1811.11913

[22] Y.Miyanaga, N.Miki and N.Nagai, "Adaptive identification of a time-varying ARMA speech model," IEEE Trans. ASSP-34, 423-433, 1986.

[23] M.G.Hall, A.V.Oppenheim, and A.S.Willsky, "Time-varying parametric modeling of speech," Signal Processing , Vol. 5, No. 3, pp. 267-285, 1983.

[24] Y.Grenier,"Time-dependent ARMA modeling of nonstationary signals," IEEE Trans. on ASSP., Vol.31, No.4, 1983.

[25] S.Kay, "Maximum entropy spectral estimation using the analytic signal," IEEE Trans. ASSP-26, 1980.

[26] T.Shimamura and S.Takahashi, "Complex linear predictiton method based on positive frequency domain," Trans. IEICE A-72, 1989. (in Japanese)

[27] E. Denoel and J-P.Solvay, "Linear Prediction of Speech with a Least Absolute Error Criterion," IEEE Trans. ASSP., Vol.33, No.6, 1985.

[28] T.Jensen, D.Giacobello, M.G.Christensen, S.H.Jensen, M.Moonen, "Real-Time Implementations of Sparse Linear Prediction for Speech Processing," Proc. ICASSP-2013, 2013. IEEE Trans. ASSP., Vol.33, No.6, 1985.
[29] D.Giacobello, M.G.Christensen, M.N.Murthi, S.H.Jensen, M.Moonen, "Sparse Linear Prediction and its Applications to Speech Processing," IEEE Trans. ASLP., Vol.20, No.5, 2012.
[30] L. A. Ekman, W. B. Kleijn, and M. N. Murthi,"Regularized linear prediction of speech," IEEE Trans. ASLP., Vol.16, No.1, 2008.
[31] M.Airaksinen, L.Juvela, O.Rasanen, P.Alku, "Time-regularized Linear Prediction for Noise-robust Extraction of the Spectral Envelope of Speech," Proc. Interspeech-2018, India, 2018.
[32] K.Funaki, Y.Miyanaga and K.Tochinai, "On a Time-varying Complex Speech Analysis," Proc. EUSIPCO-98, Rhodes, Greece, Sep.,1998.
[33] K.Funaki, "A time-varying complex AR speech analysis based on GLS and ELS method," Proc. Eurospeech2001, Aalborg, Denmark, Sep. 2001.
[34] K.Funaki, "TV-CAR Speech Analysis Based on Regularized LP," Proc. EUSIPCO-2019, Spain, Sep.2019.
[35] Keiichi Funaki, "TV-CAR speech analysis based on the l2-norm regularization in the time-domain and frequency domain." Proc.APSIPA2020, pp.568-571, Auckland, New Zealand, Dec.2020.
[36] K.Funaki, "Sparse Time-Varying Complex AR(TV-CAR) speech analysis based on Adaptive LASSO," IEICE, Trans. on Fundamentals, Vol.E102-A, No.12, Dec.2019.
[37] K.Funaki,"Speech Enhancement based on Iterative Wiener Filter using Complex Speech Analysis," EUSIPCO-2008, Lausanne, Switzerland, Aug.2008.
[38] K.Higa and K.Funaki, "Improved ETSI advanced front-end for ASR based on robust complex speech analysis," Proc. APSIPA-2016, Jeju, Korea, Dec. 2016.
[39] K.Hotta and K.Funaki "On a robust F0 estimation of speech based on IRAPT using robust TV-CAR analysis," Proc. APSIPA-2014, Dec. 2014.
[40] E.Azarov, M.Vashkevich, A.Petrovsky, "Instantaneous pitch estimation based on RAPT framework," Proc. EUSIPCO-2012, Bucharest, Romania, Aug., 2012.
[41] F.Plante, G.F.Meyer, W.A.Ainsworth, "A Pitch Extraction Reference Database," Proc.EUROSPEECH-95, 1995.
[42] NOISE-X92, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
[43] ITU-T Recommendation G.191, Software tools for speech and audio coding standardization, Nov. 2000.
[44] S. Gonzalez and M. Brookes. "PEFAC - a pitch estimation algorithm robust to high levels of noise," IEEE Trans. Audio, Speech, Language Processing, pp.518-530, Feb. 2014.
[45] S.Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," Proc.EUSIPCO, Aug 2011.
[46] "VOICEBOX: Speech Processing Toolbox for MATLAB" written by Mike Brookes, Department of Electrical & Electronic Engineering, Imperial College, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[47] Y-H.Tu, J.Du, and C-H.Lee, "DNN Training Based on Classic Gain Function for Single-Channel Speech Enhancement and Recognition," Proc. ICASSP-2019, 2019.
[48] R.Tibshirani, "Regression shrinkage and selection via the lasso," J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288, 1996.
[49] H.Zou, "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, Vol.101, 2006.