TV-CAR speech analysis based on the l_2-norm regularization in the time-domain and frequency domain

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

***Abstract*—Linear Prediction (LP) is a mathematical operation for estimating an all-pole spectrum from the speech signal. It is an essential methodology in speech coding since LP coefficients viz., Auto-Regressive (AR) coefficients can be determined using a small amount of computation and quantized efficiently using Vector Quantization (VQ). Recently, l_2-norm Regularized LP (RLP), and context-aware Time-Regularized LP (TRLP) analysis have been proposed and shown to improve performance. The former suppresses rapid spectral changes in the frequency domain, and the latter suppresses the rapid spectral changes in the time domain. In our previous study, we proposed the MMSE-based Time-Varying Complex AR (TV-CAR) speech analysis that is the complex-valued and time-varying version of the LP and the RLP-based TV-CAR analysis. In this paper, we propose the novel l_2-norm regularized TV-CAR analysis based on not only the TRLP but also the RLP and the objective evaluation using a F_0 estimation applied with the estimated complex residual signals shows that the proposed method performs best.**

I. Introduction

A fundamental methodology in speech processing is using speech analysis to estimate amplitude characteristics of the spectrum since the spectral feature of speech can be represented by the amplitude ones rather than the phase ones. Commonly used speech analysis is Linear Prediction (LP) analysis[1] proposed in the 1960s. It is being applied in speech coding [2][3] for the following reasons. The LP can estimate the all-pole speech spectrum by solving a linear equation, and the LP coefficients can be quantized efficiently using Vector Quantization (VQ) on a Line Spectrum Pair (LSP) domain. Also, in the MPEG-4 Audio LosslesS coding(ALS)[4], the LP is applied to obtain the LP residual, and the residual is quantized by an entropy coding to realize the lossless coding. The LP is also applicable to robust speech recognition (ASR) to implement a speech enhancement using iterative Wiener filter (IWF)[5] as a front end of ASR[6][7] and F_0 estimation of speech in which LP residual is exploited to compute a criterion, including auto-correlation (AUTOC), AMDF[8][9][10], weighted AUTOC[11], instantaneous frequencies (IF)[12] or so on. The LP is also utilized to estimate glottal excitation by using the glottal inverse filtering[13]. The LP method estimates the parameter in the analysis frame and does not account for frame boundaries. As a result, it cannot handle noise-corrupted speech. Context-aware speech analysis methods have been proposed to solve this difficulty, Time-Varying LP(TVLP)[14], and Frequency Domain LP(FDLP)[15]. However, the context-aware method requires a long frame that results in not only a long delay but also a large amount of computation. Recently, P.Alku et al. proposed the context-aware method that uses only short frame signals, namely, Time-Regularized LP(TRLP)[16]. In the TRLP, the LP parameters are estimated by minimizing the summation of the LP criterion and the l_2-regularized penalty term that is the norm of the difference between current and previous frame parameters; meanwhile, the Regularized LP (RLP)[17] introduces the regularized penalty term that is the l_2-norm of the spectral changes in frequencies. It can cope with the pitch-related bias that is overestimating the first formant (F_1) bandwidth. The TRLP provides the regularized term in the time-domain. In contrast, the RLP provides one in the frequency domain.

On the other hand, we have been studying the Time-Varying Complex AR (TV-CAR) speech analysis method for an analytic signal. The TV-CAR analysis is a time-varying and complex-valued LP method that can estimate complex AR coefficients of every sample and can estimate a more accurate spectrum due to the nature of the analytic signal. We have already proposed MMSE-based[18], a robust ELS-based method[19], and so on[20][21][22], and we have already evaluated the performance on speech processing such as F_0 estimation of speech and robust Automatic Speech Recognition (ASR). Since a complex analysis for an analytic signal can estimate a more accurate spectrum, the TV-CAR analysis makes it possible to have the performance improved on the F_0 estimation and robust ASR. In the former case, the complex residual signals are utilized to compute the criterion, including the weighted auto-correlation[23][24], the summation of spectral harmonics[25], and instantaneous frequencies[26]. They can perform better since a complex analysis can separate F_0 and F_1 due to the improved spectral resolution of the analytic signal so that the formant elements can be eliminated in the residual signals. In the latter case, the IWF is designed by using the estimated spectrum as a front-end of the ASR[7][27][28]. The TV-CAR analyzes the wide-band signal, and only the power spectrum in the lower narrow-band frequencies is extracted to design the IWF since it results in avoiding the spectrum distortion in higher frequencies.

The TV-CAR analysis methods are not the context-aware

methods that estimate the time-varying complex parameters within the current frame by representing the complex AR coefficient with a complex basis expansion, leading to being affected easily by environmental noise. We further proposed an RLP-based TV-CAR analysis[29]. In this paper, we propose context-aware TV-CAR analysis based on not only the TRLP but also the RLP that seems to be robust against additive noise than the RLP-based method[29]. F_0 estimation in a noisy environment is carried out to evaluate the performance by using the complex residual. IRAPT[12] method is introduced to implement the estimation since IRAPT is an improved version of well-known and commonly used RAPT[30], and the analytic signal is employed to compute the instantaneous frequencies instead of cross-correlation.

II. Regularized TV-CAR method

A. TV-CAR model

The TV-CAR model can be defined by Eq.(1).

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

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. The input-output relationship for Eq.(1) is shown in Eq.(2).

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \quad (2)$$

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 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 to estimate the parameters in every sample.

Alternatively, Eq.(2) can be formulated by the following vector-matrix representation.

$$\begin{aligned} \mathbf{y}_f &= -\mathbf{\Phi}_f \theta + \mathbf{u}_f \quad (3) \\ \bar{\theta}^T &= [\mathbf{g}_0^T, \mathbf{g}_1^T, \cdots, \mathbf{g}_l^T, \cdots, \mathbf{g}_{L-1}^T] \\ \mathbf{g}_l^T &= [g_{1,l}^c, g_{2,l}^c, \cdots, g_{i,l}^c, \cdots, g_{I,l}^c] \\ \mathbf{y}_f^T &= [y^c(I), y^c(I+1), y^c(I+2), \cdots, y^c(N-1)] \\ \mathbf{u}_f^T &= [u^c(I), u^c(I+1), u^c(I+2), \cdots, u^c(N-1)] \\ \mathbf{\Phi}_f &= [\mathbf{S}_0^f, \mathbf{S}_1^f, \cdots, \mathbf{S}_l^f, \cdots, \mathbf{S}_{L-1}^f] \\ \mathbf{S}_l^f &= [\mathbf{s}_{1,l}^f, \mathbf{s}_{2,l}^f, \cdots, \mathbf{s}_{i,l}^f, \cdots, \mathbf{s}_{I,l}^f] \\ \mathbf{s}_{i,l}^f &= [y^c(I-i) f_l^c(I), y^c(I+1-i) f_l^c(I+1), \\ &\qquad \cdots, y^c(N-1-i) f_l^c(N-1)]^T \end{aligned}$$

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

$$\hat{\theta} = \arg\min_{\bar{\theta}} \|\mathbf{y}_f + \mathbf{\Phi}_f \bar{\theta}\|_2^2 \quad (4)$$

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

$$\left(\mathbf{\Phi}_f^H \mathbf{\Phi}_f\right) \hat{\theta} = -\mathbf{\Phi}_f^H \mathbf{y}_f \quad (5)$$

where H is an Hermite operator. It is the time-varying, complex and covariance analysis version of the conventional LP.

B. RLP-based TV-CAR analysis[29]

Since the TV-CAR analysis is the complex, time-varying and covariance type of LP analysis, Eq.(6) can be derived by integrating the RLP onto the TV-CAR analysis. As the l_2-norm regularized term, the power spectrum at the center sample of the frame, $N/2$, is applied.

$$\left(\mathbf{\Phi}_f^H \mathbf{\Phi}_f + \lambda_3 \mathbf{D}_{tv}^H \mathbf{F} \mathbf{D}_{tv}\right) \hat{\theta} = -\mathbf{\Phi}_f^H \mathbf{y}_f \quad (6)$$

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

$$\mathbf{D}_{tv} = [\mathbf{d_0}, \mathbf{d_1}, ..., \mathbf{d_l}, ..., \mathbf{d_{L-1}}] \quad (7)$$

$$\mathbf{d_l} = \mathbf{diag}[f_l^c(N/2), 2f_l^c(N/2), ..., I f_l^c(N/2)] \quad (8)$$

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

C. TRLP-based TV-CAR method

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

$$\hat{\theta} = \arg\min_{\bar{\theta}} \|\mathbf{y}_f + \mathbf{\Phi}_f \bar{\theta}\|_2^2 + \frac{1}{2}\lambda_1 \|\bar{\theta} - \lambda_2 \hat{\theta}_{\mathrm{pr}}\|_2^2 \quad (9)$$

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.

$$\left(\mathbf{\Phi}_f^H \mathbf{\Phi}_f + \lambda_1 \mathbf{I}\right) \hat{\theta} = -\mathbf{\Phi}_f^H \mathbf{y}_f + \lambda_1 \lambda_2 \hat{\theta}_{\mathbf{pr}} \quad (10)$$

D. Proposed RLP and TRLP-based Hybrid TV-CAR analysis

Furthermore, by combining Eq.(6) and Eq.(10), we can easily derive the following hybrid approach of the RLP and TRLP.

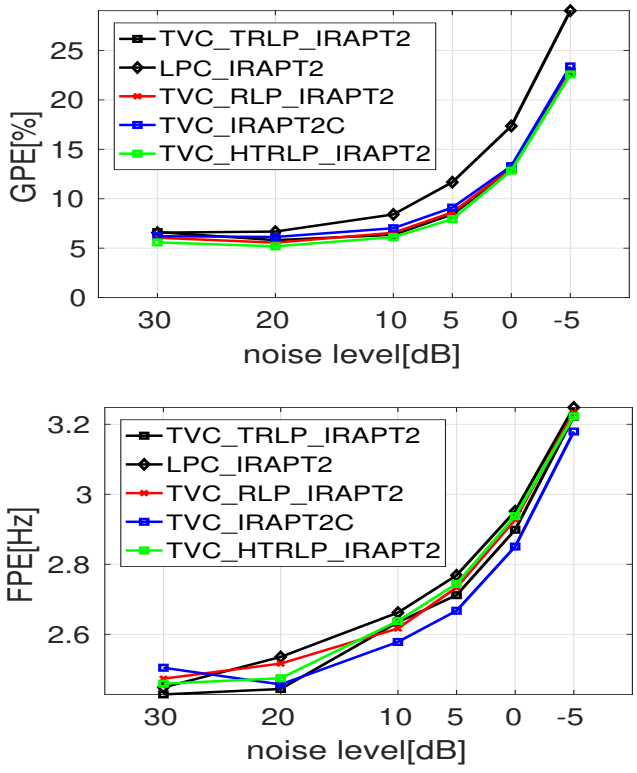
$$\left(\mathbf{\Phi}_f^H\mathbf{\Phi}_f + \lambda_1\mathbf{I} + \lambda_3\mathbf{D}_{tv}^H\mathbf{F}\mathbf{D}_{tv}\right)\hat{\theta} = -\mathbf{\Phi}_f^H\mathbf{y}_f + \lambda_1\lambda_2\hat{\theta}_{\mathbf{pr}} \quad (11)$$

III. Experiments

The proposed TRLP-based TV-CAR methods are evaluated by comparing with the conventional ones using the F_0 estimation in noisy environments. F_0 estimation is implemented by the IRAPT[12] in which the instantaneous frequency is computed by the analytic signal. Thus, the complex residual estimated by the TV-CAR methods can be applied. The IRAPT is an improved, well-known, and commonly used RAPT[30] method, and it performs better than the original RAPT. The following signals are applied in the performance comparison,

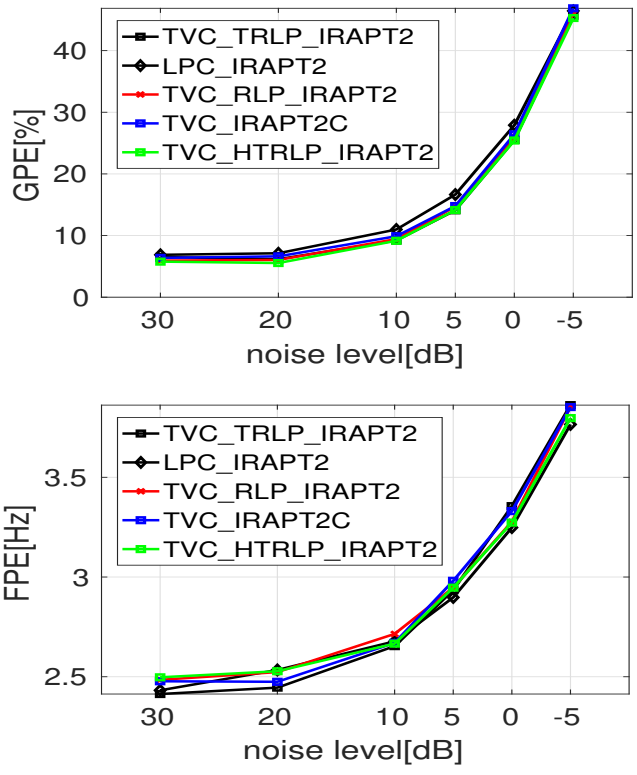
(1)The real residual computed by the LP.
(2)The complex residual computed by the MMSE-based TV-CAR has shown in Eq.(5)[18].
(3)The complex residual computed by the RLP-based TV-CAR has shown in Eq.(6)[29].
(4)The complex residual computed by the TRLP-based TV-CAR has shown in Eq.(10).
(5)The complex residual computed by the hybrid approach of the TRLP and RLP-based TV-CAR shown in Eq.(11).

Keele pitch database[31] added by white Gauss or Pink noise[32] is applied for evaluation. The noise-corrupted signal is filtered by the IRS filter[33] for speech coding applications. Gross Pitch Error(GPE) and Fine Pitch Error(FPE) are adopted as the objective criterion. The experimental conditions are shown in Table 1. Figures 1 and 2 show the experimental results for additive white Gauss and Pink noise, respectively. In the figures, (a) means 10[%] of GPEs and (b) means 10[%] of FPEs. The five lines indicate as follows. The black solid line with lozenge means (1)**LPC_IRAPT2**. The blue line means (2)**TVC_IRAPT2C**. The red line means (3)**TVC_RLP_IRAPT2**. The solid black line with square means (4)**TVC_TRLP_IRAPT2**. The green line means (5)**TVC_HTRLP_IRAPT2**. The figures are as follows. The complex-valued methods perform better than the LP. The proposed TRLP-based and RLP-based methods perform better than the MMSE-based TV-CAR method in terms of GPE that shows a fatal error such as double pitch or half pitch leading to serious low performance so that GPE is more important index than FPE. The proposed hybrid approach of the TRLP and RLP performs best. The reason why the GPEs are considerable value is that the IRS filtered speech is used for speech coding applications. 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[26][29].



(b)FPE(10%)

Fig.1: F_0 estimation performance(Gauss)



(b)FPE(10%)

Fig.2: F_0 estimation performance (Pink)

Table 1: Experimental Conditions

Speech data	Keele Pitch Database[31]
	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	$I = 7, L = 2$(Time-Varying)
Basis	$f_l^c(t) = t^l/l!$
Pre-emphasis	$1 - z^{-1}$
TRLP/RLP	$\lambda_1 = 0.02, \lambda_2 = 0.99$ / $\lambda_3 = 0.0001$
Noise	White Gauss or Pink noise[32]
Noise Level	30,20,10,5,0,-5[dB]

IV. Conclusions

In this paper, we proposed two types of regularized LP-based TV-CAR speech analysis methods, TRLP introduced l_2-norm regularized LP in the time domain and A hybrid approach of the TRLP and RLP[29] that introduces l_2-norm regularized LP in the time and frequency domain. The former method penalizes rapid changes of the estimated spectrum among adjacent frames that can take into account a broader macro context and seems to be robust against background noise. The latter one penalizes the rapid changes of the estimated spectrum not only in the time-domain but also in the frequency domain that makes it possible to suppress pitch-related bias. Furthermore, the objective evaluation is accomplished to compare 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 IRAPT implements the F_0 estimation. The experimental results show that the proposed regularization methods perform better than the MMSE method in terms of GPE, and the hybrid approach of the TRLP and RLP realizes the improved performance in comparison to the TRLP-based, RLP-based, MMSE-based methods. We aim to evaluate the proposed methods on a front-end of robust ASR[7][34]. Besides, sparse TV-CAR analysis based on the Least Absolute Shrinkage and Selection Operator (LASSO)[35][36][37] or Elastic Net will be proposed and be evaluated on speech processing.

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

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

[7] 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.

[8] W.J.Hess, "Pitch and voicing determination,"in Advances in Speech Signal Processing, ed. S.Furui and M.Sondhi, Marcel Dekker, 1992.

[9] L.R. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust. Speech Signal Process., Vol.24, No.5, pp.399-417, 1976.

[10] M.J. Ross, H.L. ShaiTer, A. Cahen, R. Freudberg, and. H.J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech Signal Process., Vol.22, No.5, pp.353-362, 1974.

[11] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. Speech Audio Process., Vol.9, No.7, pp.727-730, Oct. 2001.

[12] E.Azarov, M.Vashkevich, A.Petrovsky, "Instantaneous pitch estimation based on RAPT framework," Proc. EUSIPCO-2012, Bucharest, Romania, Aug., 2012.

[13] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," IEEE/ACM TASL., Vol.22, No.3, Mar. 2014.

[14] 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.

[15] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," IEEE/ACM Transactions on Audio, Speech, and Language Processing , Vol. 22, No. 8, pp. 1285-1295, 2014.

[16] 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.

[17] L. A. Ekman, W. B. Kleijn, and M. N. Murthi,"Regularized linear prediction of speech," IEEE Trans. ASLP., Vol.16, No.1, 2008.

[18] K.Funaki, Y.Miyanaga and K.Tochinai, "On a Time-varying Complex Speech Analysis," Proc. EUSIPCO-98, Rhodes, Greece, Sep.,1998.

[19] K.Funaki, "A time-varying complex AR speech analysis based on GLS and ELS method," Proc. Eurospeech2001, Aalborg, Denmark, Sep. 2001.

[20] K.Funaki, Y.Miyanaga and K.Tochinai,"On Robust speech analysis based on time-varying complex AR model,"ICSLP-98, Sydney, Australia, Dec., 1998.

[21] K.Funaki, "A time-varying complex speech analysis based on IV method," Proc. ICSLP-2000, Beijing, China, Oct.,2000.

[22] K.Funaki,"WLP-based TV-CAR speech analysis and its evaluation for F0 estimation," Proc MAVEBA2013, Firenze, Italy, Dec. 2013.

[23] T.Kinjo and K.Funaki, "Robust F0 Estimation Based on Complex LPC Analysis for IRS Filtered Noisy Speech," IEICE Trans., E90-A, No.8, 2007.

[24] K.Funaki and T.Kinjo, "Robust F0 Estimation Using ELS-Based Robust Complex Speech Analysis," IEICE Trans. Vol. E91-A, 2008.

[25] K.Funaki and T.Higa,"F_0 Estimation using SRH based on TV-CAR Speech Analysis," Proc.EUSIPCO-2012. Bucharest, Romania, Aug., 2012.

[26] 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.

[27] K.Funaki,"Speech Enhancement based on Iterative Wiener Filter using Complex Speech Analysis," EUSIPCO-2008, Lausanne, Switzerland, Aug.2008.

[28] 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.

[29] K.Funaki, "TV-CAR Speech Analysis Based on Regularized LP," Proc. EUSIPCO-2019, Spain, Sep.2019.

[30] D.Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in Speech Coding and Synthesis, W.B.Kleijn and K. K.Palatal (eds), pp.497-518, Elsevier Science B.V., 1995.

[31] F.Plante, G.F.Meyer, W.A.Ainsworth, "A Pitch Extraction Reference Database," Proc.EUROSPEECH-95, 1995.

[32] NOISE-X92, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

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

[34] 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.

[35] R.Tibshirani, "Regression shrinkage and selection via the lasso," J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288, 1996.

[36] H.Zou, "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, Vol.101, 2006.

[37] 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.