2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCSLP49672.2021.9362085

On Adaptive LASSO-based Sparse Time-Varying Complex AR Speech Analysis

Keiichi FUNAKI

Computing & Networking Center, University of Nishihara, Okinawa, 903-0213, Japan

funaki@cc.u-ryukyu.ac.jp

Abstract

Linear Prediction (LP) is commonly used in speech processing. In speech coding, the LP is used to remove the formant elements from the speech signal, and the residual is quantized by using the Algebraic code vector after removing pitch elements. In speech synthesis, the LP is also used to generate the glottal or residual excitation for the WaveNet. We have proposed a Time-Varying Complex AR (TV-CAR) speech analysis for an analytic signal to cope with the drawbacks of the LP, such as MMSE, Extended Least Square (ELS), that are the l_2-norm optimization methods. We have already evaluated the performance on F_0 estimation and robust automatic speech recognition. Recently, we have proposed l_2-norm regularized LP-based TV-CAR analysis in the time-domain and the frequency-domain. The regularized TV-CAR method can estimate more accurate formant frequencies, and we have shown that the resulting LP residual makes it possible to estimate a more precise F_0. On the other hand, sparse estimation based on l_1-norm optimization has been focused on image processing that can extract meaningful information from colossal information. LASSO algorithm is an l_1-norm regularized sparse algorithm. In this paper, adaptive LASSO-based TV-CAR analysis is proposed, and the performance is evaluated using the F_0 estimation.

Index Terms: complex speech analysis,analytic signal, LASSO, l_1 regularized sparse analysis, F_0 estimation

1. Introduction

Linear Prediction (LP)[1] proposed in the 1960s is the most commonly and successfully used signal processing algorithm in speech processing that can estimate the Auto-Regressive (AR) spectrum from speech signal by minimizing the l_2-norm of equation error. It is the so-called Least Square (LS) or Minimizing Mean Squared Error (MMSE). The LP provides several advantages; low computational complexity, better performance in terms of quantization of the coefficients, and it is easy to implement the inverse filter. The inverse AR filter can construct with the corresponding MA filter; as a result, one can easily obtain the LP residual. Since the LP residual contains fewer formant structures, a more accurate F_0 can be estimated by using the auto-correlation (AUTOC)[2][3] for the residual than that for the speech signal. This F_0 estimation is called the modified AUTOC. The alternative criterion besides the AUTOC, AMDF[4], and weighted AUTOC[5] can lead to better performance under a noisy condition. The LP residual can also estimate the glottal excitation that is so-called glottal inverse filtering (GIF)[6]. The GIF can help to estimate the glottal model parameter from the speech signal using fitting the glottal excitation to the LP residual. It can obtain not so strange glottal excitation with low computation. However, it cannot perform better than the simultaneous estimation of the glottal and vocal tract model parameters[7][8][9][10][11]. The LP is an essential technique in the CELP speech coding that quantizes the LP residual, computed by the quantized LP coefficients with the Line Spectrum Pair (LSP), using the Vector Quantized (VQ) code vector[12] or Algebraic code vector[13][14], after removing the pitch elements. The LP is also used in Automatic Speech Recognition (ASR). As a front-end of ASR, ETSI (European Telecommunication Standardized Association) standardized an advanced front-end, namely ETSI AFE, that employs an Iterative Wiener Filter (IWF) designed by the estimated speech spectrum and noise one[15]. In the ETSI AFE, the FFT spectrum smoothed in the time and frequency domain is applied to design the filter; however, it was reported that the LPC spectrum performs better than the FFT one[16][17]. Speech synthesis also embeds the LP. Recently the development of the WaveNet[18] brings a new era on speech synthesis and makes it possible to improve the speech quality drastically. Several improved WaveNet methods have been proposed including GlotNet[19][20], ExcitNet[21], FFTNet[22], LPCNet[23], LP-WaveNet[24] and so on. The GlotNet generates the excitation using a glottal excitation estimated by GIF. The ExcitNet generates the excitation using the LP residual estimated by the LP inverse filter. The ExcitNet provides more rich excitation that includes the noise elements besides glottal excitation, making the speech quality improved. The LP is also useful in noise reduction[25][26], in which an adaptive filter for noise reduction introduces the LP to improve the performance since the LP can whiten the speech signal. [27] proposes a restoration scheme for the instantaneous amplitudes and phases in subbands using a Kalman filter with the LP that decides the performance. The LP is also applied in speech dereverberation to improve the performance[28][29]. [29] employs Multi-Step LP (MSLP) to suppress the late reverberation.

The LP provides three shortcomings: 1) It cannot extract time-varying spectral features or cannot account for the frame boundaries. 2) It cannot estimate anti-resonance. 3) It cannot handle non-Gaussian excitation. A considerable number of improved LP methods have been investigated to address the shortcoming. Time-varying analysis methods have been proposed[30][31], and context-aware analysis methods have been proposed[32][33] to cope with 1). Progress has been achieved to cope with 2) by ARMA (Auto-Regressive and Moving Average) analysis method[34][35]. A further effort has been made to solve 3) by simultaneous estimation of not only glottal but also vocal tract model parameters [7][8][9][10][11], as shown above. Moreover, recently, sparse estimation based on l_0 or l_1-norm optimization has been focused since it can extract meaningful information from big data. Sparse LP methods based on l_1-norm optimization have been proposed to handle non-Gaussian excitation[36][37][38].

On the other hand, we have already studied Time-Varying Complex AR (TV-CAR) speech analysis for an analytic signal, MMSE[39], Extended Least Square (ELS)[40], or so on, that are based on l_2-norm optimization. These can estimate the time-

varying spectrum within the frame to cope with 1, and these can also estimate a more accurate speech spectrum due to the nature of the analytic signal. We have proposed l_2 regularized LP-based TV-CAR analysis methods proposed[41][42] that can perform better than the MMSE methods. This paper proposes two types of sparse TV-CAR analysis methods based on LASSO(Least absolute shrinkage and selection operator)[43] and adaptive LASSO[44]. The LASSO, l_1-norm regularized method, is realized by using IRLS (Iterative Reweighted Least Square)[45], and the performance is compared by using that of F_0 estimation with the estimated complex LP residual. The IRAPT[46], Instantaneous RAPT(Robust Algorithm for Pitch Tracking)[47] is introduced to implement the F_0 estimation and Keele Pitch database[48] corrupted by white Gaussian, and Pink noise are used for evaluation.

2. TV-CAR model

2.1. Analytic signal

The target signal of the TV-CAR analysis is the analysis signal, which is a complex signal defined as

$$z^c(t) = x(t) + jx_H(t) \tag{1}$$

where $z^c(t)$, $x(t)$, and $x_H(t)$ represent the analysis signal at time t, the observation signal at time t, and the Hilbert-transformed observation signal. Note that the superscript c represents a complex number in this paper. The analytic signal retains only the spectrum in positive frequencies. So, it can be thinned out by a factor of 2. The signal $y^c(t)$, which is $z^c(t)$ thinned out by 2, is used as the analysis signal. The term $1/\sqrt{2}$ is multiplied to adjust the power of $y^c(t)$ with that of $y(t)$.

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

Conventional LP model is defined by

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

where a_i and I are i-th order LP coefficient and LP order, respectively. Since the conventional LP model cannot express the time-varying spectrum, the LP 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, a finite order of complex basis expansion, the complex parameter, and a complex-valued basis function, respectively. By substituting Eq.(3) into Eq.(2), one can obtain the following transfer function of the TV-CAR model.

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

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 $u^c(t)$ is taken to be a complex-valued input signal. 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 the continuous time-varying spectrum. Also, the complex-valued analysis facilitates accurate spectral estimation in low frequencies. As a result, this feature allows for more appropriate inverse filtering to obtain the residual with fewer formant components. For the sake of a more convenient expression, we represent Eq.(5) using vector-matrix notation as follows.

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

where N is the analysis interval, $\mathbf{y}_f$ is a $(N-I, 1)$ column vector whose elements are analytic speech signals, θ is a $(L \cdot I, 1)$ column vector whose elements are complex parameters, and $\mathbf{\Phi}_f$ is a $(N-I, L \cdot I)$ matrix whose elements are weighted analytic speech signals by the complex basis. The superscript T denotes transposition.

3. Proposed LASSO-based sparse TV-CAR analysis

This Section explains the proposed sparse methods. In general, l_p-norm criterion is defined as follows.

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{y}_f + \mathbf{\Phi}_f \theta\|_p^p + \gamma \|\theta\|_k^k \tag{7}$$

l_p-norm $\|\cdot\|_p$ is defined as

$$\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{1/p}. \tag{8}$$

LASSO(Least Absolute Shrinkage and Selection Operator)[43] introduces the absolute parameter as a penalty term, and it can estimate a sparse solution with less redundancy.

3.1. l_2-norm/MMSE algorithm

In l_2-norm/MMSE, $\gamma = 0$ and $p = 2$, thus, the criterion is as follows.

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{y}_f + \mathbf{\Phi}_f \theta\|_2^2 \tag{9}$$

Minimizing the MSE criterion of Eq.(9) with respect to the complex parameter leads to the following MMSE algorithm.

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

where, the superscript H denotes an Hermitian transposition. After solving the linear equation of Eq.(10), we obtain the complex AR parameter ($a_i^c(t)$) at time t with the estimated complex parameter $\hat{g}_{i,l}^c$.

3.2. Adaptive LASSO algorithm

In LASSO, $k = 1$, thus, the criterion of LASSO is as follows.

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{y}_f + \mathbf{\Phi}_f\theta\|_2^2 + \gamma\|\theta\|_1 \tag{11}$$

γ is coefficient for Lagrange's multiplier method that permits the difference for the first term of Eq.(11). If the noise level is high, and the input signal of $y(t)$ cannot be reliable, γ is set to be a high value. Otherwise, γ is set to be small. Thus, it can be thought that the LASSO is more robust against additive noise. Adaptive LASSO[44] is formulated as follows.

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{y}_f + \mathbf{\Phi}_f\theta\|_2^2 + \gamma\hat{\omega}\|\theta\|_1 \tag{12}$$

$$\hat{\omega} = 1/(\|\theta\|_1)^{\eta} \tag{13}$$

$\hat{\omega}$ is a weighting factor for the parameter. LASSO can be solved approximately by using Iterative Reweighted Least Square (IRLS)[45]. If $\mathbf{\Theta}$ is set to be $diag(|\theta|)$, $\|\theta\|_1$ can be expressed by $\theta^H\mathbf{\Theta}^{-1}\theta$. As a result, l_1-norm optimization can be simplified as a weighted l_2-norm optimization problem. If current approximate solution θ_{k-1} is given, $\mathbf{\Theta}_{k-1}$ is set to be $diag(|\theta_{k-1}|)$ and the following equation is solved.

$$\hat{\theta} = \arg\min_{\theta} \|\mathbf{y}_f + \mathbf{\Phi}_f\theta\|_2^2 + \gamma\hat{\omega}\theta^H\mathbf{\Theta}_{k-1}\theta \tag{14}$$

Eq.(14) can be solved by a linear equation. In the IRLS, the initialize parameters of $\hat{\theta}_0$ and $\mathbf{\Theta}_0$ are set and the following equation is solved.

$$(\mathbf{\Phi}_f^H\mathbf{\Phi}_f + \gamma\hat{\omega}\mathbf{\Theta}_{k-1}^{-1})\hat{\theta}_k = -\mathbf{\Phi}_f^H\mathbf{y}_f \tag{15}$$

In Eq.(15), second term ($\gamma\hat{\omega}\mathbf{\Theta}_{k-1}^{-1}$) is real-valued diagonal matrix and the diagonal elements are added to the first term ($\mathbf{\Phi}_f^H\mathbf{\Phi}_f$) and then the linear equation is solved. Next, $\mathbf{\Theta}_k$ is updated by using $\hat{\theta}_k$. The estimation is iterated until the $\hat{\theta}_k$ estimation gets to its convergence. Table 1 summarizes the IRLS algorithm. Note that the initial estimation is equal to be that of the Ridge.

Table 1: IRLS algorithm

Algorithm: LASSO
Initialize:
$k = 0$,
$\hat{\theta}_k = (1, 1, 1, ..., 1)^T$,
$\bar{\mathbf{\Theta}}_k = \boldsymbol{I}$(Unit Matrix)
Iterations:
$k \leftarrow k + 1$
Eq.(15) is solved to get $\hat{\theta}_k$
$\bar{\mathbf{\Theta}}$ is updated by $\mathbf{\Theta}_k(j, j) = \|\theta_k(j)\| + \epsilon$.
if $\|\hat{\theta}_k - \hat{\theta}_{k-1}\|_2^2 < thr$ then
break iterations;
end if
Output:
Solution $\hat{\theta}_k$

4. Experiments

As the source signal to the IRAPT, the following four signals are evaluated. a) speech signal (original IRAPT[46]), b) the complex residual by adaptive LASSO-based TV-CAR analysis, c) the complex residual by LASSO-based TV-CAR analysis, d) the complex residual by MMSE-based TV-CAR[51]. b), c) and d) are obtained by the TV-CAR analysis and its inverse filtering for the analytic signal. It is worthwhile to note that if the F_0 estimation accuracy is improved by using the complex residual, we can take into account that the formant estimation performance of the TV-CAR analysis is high since the formant elements are removed in the residual. Table 2 describes the experimental conditions. The experiments were carried out using Keele Pitch Database[48] corrupted by white Gauss or Pink noise[49]. Note that the Keele Pitch database contains five long sentences uttered by different five male speakers and five long sentences uttered by five female speakers. Each sentence is sufficiently long, more than 30 seconds long, and the total length is around 6 minutes. The noise-corrupted speech is filtered by an IRS(Intermediate Reference System) filter[50] for speech coding application. F_0 estimation performance is evaluated by using GPE(Gross Pitch Error)and FPE(Fine Pitch Error). Figure 1 (1-1) and (1-2) show the result of additive white Gauss noise and Figure 1 (2-1) and (2-2) show the result of Pink additive noise. In figures, (1-1) and (2-1) mean a graph of GPE of 10[%], and (1-2) and (2-2) means that of FPE of 10[%]. X-axis denotes noise level [dB], Y-axis denotes GPE[%] and FPE[Hz]. Four lines of the figures are as follows.

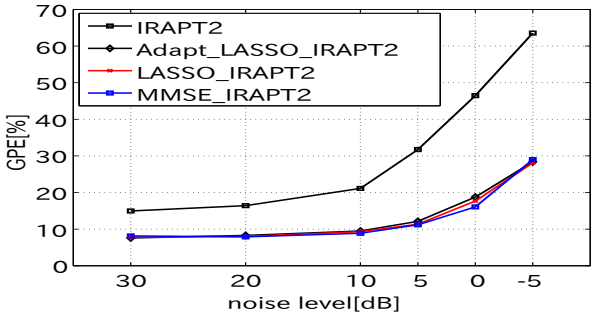
(a) IRAPT2 (square solid line) is GPE/FPE for the speech of IRAPT2[46].
(b) Adapt_LASSO_IRAPT2 (diamond solid line) is GPE/FPE of IRAPT2 for the complex AR residual estimated by the adaptive LASSO-based TV-CAR analysis.
(c) LASSO_IRAPT2 (red line) is GPE/FPE of IRAPT2 for the complex AR residual estimated by the LASSO-based TV-CAR analysis($\hat{\omega} = 1$ in Eq.(12)).
(d) MMSE_IRAPT2 (blue line) is GPE /FPE of IRAPT2 for the complex AR residual estimated by the MMSE-based TV-CAR analysis[51].

Table 2: Experimental Conditions

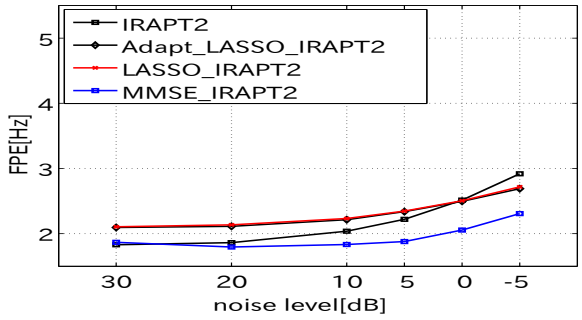
Speech data	Keele Pitch Database[48]
	5 long Male sentence
Sampling	10kHz/16bit
Analysis window	Window Length: 25.6[ms]
	Shift Length: 10.0[ms]
MMSE	$I = 7, L = 2$
Basis	$f_l^c(t) = t^l/l!$
Pre-emphasis	$1 - z^{-1}$
LASSO	$I = 14, L = 2$
Basis	$f_l^c(t) = t^l/l!$
γ	0.1
ϵ	0.01
Pre-emphasis	$1 - z^{-1}$
Noise	White Gauss or Pink noise[49]
Noise Level	30,20,10,5,0,-5[dB]

Figure 1 demonstrates as follows. The l_1-norm regularization based LASSO methods do perform better for a higher level of additive pink noise in terms of GPE , although there is no significant difference among them for other cases. In this case, the adaptive LASSO does perform better than the LASSO method.

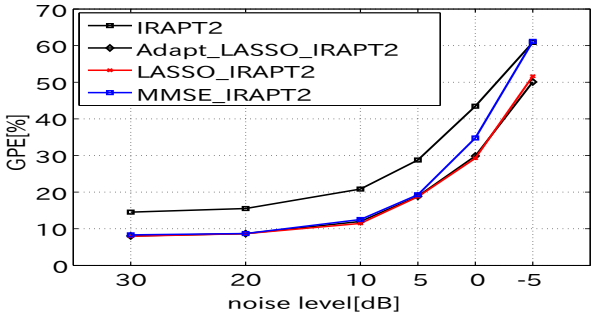
In terms of FPE, it is evident that the MMSE performs best. GPE means fatal estimation error, half-pitch, or double pitch, while FPE means subtle estimation error, so we can consider GPE is more critical than FPE. Consequently, the LASSO methods outperform the MMSE. It can be thought that we can use the LASSO method as the pre-search, and then we can use the MMSE method as the last search with the neighbor of the pre-selected value to avoid fatal estimation. The reason why the l_1-norm regularized LASSO methods do not perform so well is that analysis order or parameters are not optimized. Moreover, the proposed method introduces the IRLS algorithm that sometimes fails to a local minimum.



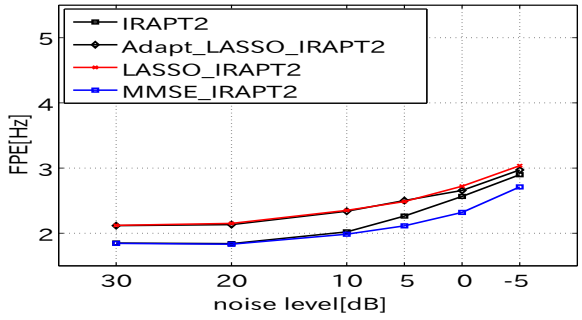
(1-1)GPEs for additive white Gauss noise



(1-2)FPEs for additive white Gauss noise



(2-1)GPEs for additive Pink noise



(2-2)FPEs for additive additive Pink noise

Figure 1: F_0 estimation performance(10 % GEP)

5. Conclusion

This paper presented l_1-norm regularized sparse TV-CAR analysis that introduces the adaptive LASSO whose criterion is the l_2 norm for the equation error added by the l_1-norm of the parameter. We have explained the proposed adaptive LASSO algorithm with the MMSE algorithm. The IRLS is used to obtain the approximate estimation for the adaptive LASSO. The performance is evaluated by F_0 estimation using the IRAPT since the performance is improved if the LP residual signal contained fewer formant elements. The experimental results show that the proposed LASSO-based TV-CAR methods perform better for a high level of Pink additive noise, although the MMSE-based TV-CAR analysis cannot perform well. The performance for this case is almost equal to that for original IRAPT, and a high level of Pink noise is considered a weak point of the MMSE-based l_2-norm analysis and the proposed l_1-norm analysis can overcome the difficulties. This paper used a fixed value of γ. It is, however, natural that the analysis order I, L, and value of γ have to be estimated to be optimum. To solve this problem, we are going to introduce LARS (Least Angle Regression)[52]. The IRLS sometimes fail to non-optimal estimation so that ADMM[53] is going to be introduced to solve the LASSO. We are going to evaluate the proposed analysis for robust speech recognition[16][17][54]. Besides, we are going to propose the Elastic net-based TV-CAR analysis combined with l_2-norm regularization [41][42].

6. References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] H.Fujisaki and M.Ljungqvist, "Estimation of Voice Source and Vocal Tract Parameters Based on ARMA Analysis and a Model for the Glottal Source Waveform," Proc. ICASSP-87, 1987.

[8] T.Ohtsuka and H.Kasuya, "Robust ARX-based speech analysis method taking voicing source pulse train into account," The Journal of the Acoustical Society of Japan, Vol.58, 2002.(in Japanese)

[9] K.Funaki, Y.Miyanaga and K.Tochinai, "Recursive ARMAX speech analysis based on a glottal source model with phase compensation," Signal Processing, Vol. 74, 1999.

[10] K.Funaki, Y.Miyanaga and K.Tochinai, "On Subband analysis based on Glottal-ARMAX speech model," ESCA 3rd International Workshop on Speech Synthesis, Bluemountain, Australia, Nov., 1998.

[11] Y.Li, K.Sakakibara and M.Akagi, "Estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model," Proc. ISCSLP-2018, Taipei, Nov., 2018.

[12] W.B.Kleijn, "Principles of Speech Coding," Springer Handbook of Speech Processing pp.283-306.2008.

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

[14] "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.

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

[16] Keita Higa and Keiichi Funaki, "Robust ASR Based on ETSI Advanced Front-End Using Complex Speech Analysis." IEICE Trans.E98-A,2015.

[17] Keita Higa and Keiichi Funaki, "Improved ETSI advanced front-end for ASR based on robust complex speech analysis," Proc. APSIPA-2016, Jeju, Korea, Dec. 2016.

[18] A.v.d.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.

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

[20] 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.

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

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

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

[24] 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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] H.Morikawa and H.Fujisaki, "Adaptive analysis of speech based on a pole-zero representation," IEEE Trans. ASSP, Vol.30, Feb 1982.

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

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

[37] T.L.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.

[38] D.Giacobello, M.G.Christensen, M.N. Murthi, S.Jensen and M.Moonen, "Sparse Linear Prediction and its Applications to Speech Processing," IEEE Trans. ASLP., Vol.20, No.5, 2012.

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

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

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

[42] K.Funaki, "TV-CAR speech analysis based on the l_2-norm regularization in the time-domain and frequency domain," Proc. EUSIPCO-2020, AMS, Jan.2021.(submitted)

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

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

[45] M.Elad,"Sparse and Redundant Representations From Theory to Applications in Signal and Image Processing," Springer; 2010.

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

[47] David 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.

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

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

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

[51] K.Funaki and K.Hotta,"On a Robust F_0 Estimation of Speech based on IRAPT using Robust TV-CAR Analysis," Proc.APSIPA2014, Siem Reap, Cambodia, Dec.2014.

[52] B.Efron, T.Hastie, I.Johnstone and R.Tibshirani,"Least Angle Regression," The Annals of Statistics, Vol. 32, No. 2, 2004, Institute of Mathematical Statistics, 2004

[53] S.Boyd, N.Parikh, E.Chu, B.Peleato and J.Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," Foundations and Trends in Machine Learning, Vol.3, No.1, pp.1-122, 2010

[54] S-T.Niu, J.Du, L.Chai, C-H.Lee "A Maximum Likelihood Approach to Multi-Objective Learning using Generalized Gaussian Distributions for DNN-based Speech Enhancement," Proc. ICASSP-2020, Barcelona, Spain, May, 2020.