On a robust ASR based on complex AR speech analysis

Keita HIGA* and Keiichi FUNAKI†
* Graduate School of Science and Engineering, University of the Ryukyus, Okinawa, JAPAN
† C & N Center, University of the Ryukyus, Okinawa, JAPAN
E-mail: funaki@cc.u-ryukyu.ac.jp Tel/Fax: +81-98-8946/8963

***Abstract*—The advanced front-end (AFE) for automatic speech recognition (ASR) was standardized by the European Telecommunications Standards Institute (ETSI). The AFE provides speech enhancement realized by an iterative Wiener filter (IWF) in which a smoothed FFT spectrum over adjacent frames is used to design the filter. We have previously proposed robust time-varying complex AR (TV-CAR) speech analysis and evaluated the performance of speech processing such as F_0 estimation and speech enhancement. TV-CAR analysis can estimate more accurate spectrum than FFT, especially in low frequencies because of the nature of the analytic signal. In addition, the TV-CAR can estimate more accurate speech spectrum against additive noise. In this paper, the time-invariant version of wide-band TV-CAR analysis is introduced to the IWF in the AFE and is evaluated using the CENSREC-2 database.**

I. Introduction

Since speech recognition is commonly used in realistic noisy environment, performance can be seriously degraded by additive noise or reverberation. Speech recognition that is robust against additive noise can be classified into the following two approaches. The first uses speech enhancement to reduce additive noise from noise-corrupted speech. The other uses a robust feature vector against additive noise. The former methods include spectral subtraction (SS)[1], the iterative Wiener filter(IWF)[2][3], the minimizing mean square error (MMSE) approach[4], and the Vector Taylor series(VTS)[5]. The latter methods include RASTA[6], Power-normalized Cepstral Coefficients (PNCC)[7], GMM adaptation[8], double auto-correlation spectrum expression [9], extended weighted linear prediction (XLP) analysis[10], phase MFCC[11], phase spectrum-based group delay spectrum[12], and q-LSMN[13].

SS is a very simple method in which a noise spectrum is estimated and then reduced from observed speech in the frequency domain. Although the SS method suffers from musical noise generation, it is commonly used since it can be easily implemented due to its simple structure. In addition, it is suitable for speech recognition applications that do not have to re-synthesize speech. The VTS method modifies a spectrum using a non-linear mismatch function between a model in a realistic environment and a model in an ideal environment. The PNCC method is based on auditory processing that introduces power-based non-linearity and reduces noise using non-linear filtering and temporal masking. P.Alku,et.al. reported experimental results for large vocabulary speech recognition obtained using the MFCC converted from a spectrum based on weighted linear prediction (WLP) or extended weighted linear prediction (XLP)[10]. While MFCC and RASTA are spectral parameters based on amplitude characteristics, studies focused on phase characteristics are currently being conducted. For example, Paliwal, et. al. reported that phase characteristics can be possibly used for speech recognition. They demonstrated experimental results with MFCC based on synthetic speech by long term phase characteristics[11]. Yamamoto, et. al. performed speech recognition using group delay spectral features based on a phase spectrum[12].

Furthermore, the European Telecommunications Standards Institute (ETSI) has standardized the advanced front-end (AFE)[14] for automatic speech recognition (ASR) in which a smoothed FFT spectrum-based IWF[3] is adopted. The method is regarded as the reference for the front-end of ASR.

We have proposed an IWF on the basis of MMSE-based complex LPC analysis[15]. In [15], the complex LPC analysis for the analytic signal was introduced to estimate the power spectra rather than LPC analysis. Complex LPC analysis can estimate more accurate spectrum in low frequencies than the real one because of the nature of the analytic signal. We have also proposed MMSE-based time-varying complex auto-regressive (AR) speech analysis[16] that can estimate time-varying complex AR spectrum since AR coefficients are represented by complex basis expansion as a function of time. Complex LPC analysis is realized by setting the basis expansion order to 1[15]. We have proposed robust time-varying complex AR (TV-CAR) analysis based on an extended least square (ELS)[17] in which an additional whitening filter is introduced to realize unbiased estimation. The robust ELS method can estimate more accurate speech spectrum against additive noise.

In this paper, a time-invariant version of wide-band TV-CAR analysis for an analytic signal is introduced rather than FFT onto the AFE and performance is evaluated. MMSE-based real-valued AR and wide-band MMSE-based complex-valued AR analysis are evaluated. CENSREC-2[18][19] is used for evaluation and the Hidden Markov Model (HMM) Tool Kit (HTK)[20] is used to realize a HMM speech recognizer. The CENSREC-2 database includes a task for continuous digit recognition in real-car-driving environments. In-car speech data is collected in a specially equipped vehicle under 11 environmental conditions. The speech recorded by a microphone attached to the ceiling above the driver's seat is used for

evaluation. There are four evaluation environments for speech recognition depending on whether the recording environments and microphones used between training and testing data match.

II. ETSI AFE

A. Wiener Filter

The key factor of the ETSI AFE is the Wiener filter. In this subsection, the Wiener filter is briefly explained. We assume that a clean speech signal $s(t)$ can be estimated by filtering an observed noise-corrupted signal $x(t) = s(t) + n(t)$ with optimal filter $h(\tau)$, where $n(t)$ is additive noise and $x(t)$ is the observed speech signal. The estimated speech using filter $h(\tau)$ is expressed as follows.

$$\hat{s}(t) = \int_0^\infty h(\tau)x(t-\tau)d\tau \tag{1}$$

The filter is designed so as to minimize the means squared error (MSE) between the estimated signal and the clean speech $s(t)$. $h_{opt}(\tau)$ is designed so as to minimize the following criterion.

$$V[h(\tau)] = E[(s(t) - \hat{s}(t))^2] \tag{2}$$

The Wiener filter $h_{opt}(\tau)$ is designed as follows.

$$H(\omega) = \frac{S_{ss}(\omega)}{S_{ss}(\omega) + S_{ww}(\omega)} \tag{3}$$

In Eq(3), $S_{ss}(\omega)$ and $S_{ww}(\omega)$ are spectrum of clean speech and additive noise, respectively. These spectra are commonly estimated by LPC analysis. Since clean speech cannot be observed, the spectrum $S_{ss}(\omega)$ is estimated using the observed signal, and then the filter is designed and the enhanced speech $\hat{s}(t)$ is used to re-estimate $S_{ss}(\omega)$. The iteration indicates the number of stages (commonly two) in the Wiener filter.

B. ETSI AFE

In this subsection, the ETSI AFE is explained. The procedure consists of speech enhancement, waveform processing, feature extraction, blind equalization, compression, and encoding. In the AFE, a two-stage Wiener filter is performed in every frame for speech enhancement. In the first stage, speech and noise power spectrum are estimated by FFT using voice active detection (VAD) and the Wiener filter is designed according to Eq.(3) using the estimated spectra. While FFT is introduced by estimating the speech spectrum in the AFE, TV-CAR analysis is introduced in the proposed method. The filter coefficients are converted into mel-frequencies, and convolution is performed to generate enhanced speech. In the second stage, the output signal obtained in the first stage is set to the input signal and speech enhancement is performed again. Finally, DC offset is removed from the enhanced speech.

III. TV-CAR Speech Analysis

In this paper, time-invariant real and complex-valued analysis are introduced as a spectrum estimator for the Wiener Filtering. Since the TV-CAR analysis includes the real-valued and time-invariant analysis, the TV-CAR analysis is explained for convenience.

A. Analytic speech signal

The target signal of the TV-CAR method is an analytic signal which is a complex-valued signal defined as follows.

$$y^c(t) = \frac{x(2t) + j \cdot x_H(2t)}{\sqrt{2}} \tag{4}$$

where $y^c(t)$, $x(t)$, and $x_H(t)$ denote an analytic signal at time t, an observed signal at time t, and a Hilbert transformed signal for the observed signal, respectively. Note that the superscript c denotes a complex value in this paper. Analytic signals provide the spectra over the range $(0, \pi)$; thus they can be decimated by a factor of two. $2t$ indicates this decimation. The term $1/\sqrt{2}$ is multiplied to adjust the power of an analytic signal to that of the observed signal.

B. Time-varying complex AR model

The TV-CAR model is defined as follows.

$$Y_{TVCAR}(z^{-1}) = \frac{1}{1 + \sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) z^{-i}} \tag{5}$$

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

$$\begin{aligned} y^c(t) &= -\sum_{i=1}^{I} a_i^c(t) y^c(t-i) + u^c(t) \\ &= -\sum_{i=1}^{I}\sum_{l=0}^{L-1} g_{i,l}^c f_l^c(t) y^c(t-i) + u^c(t) \end{aligned} \tag{6}$$

where $u^c(t)$ and $y^c(t)$ are taken as a complex-valued input and an 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.(6) parameterizes the AR coefficient trajectories that continuously change as a function of time so that the time-varying analysis is feasible to estimate continuous time-varying speech spectrum. In addition, the complex-valued analysis facilitates accurate spectral estimation in low frequencies. Therefore, this feature allows more appropriate Wiener filtering. Eq.(6) can be represented in vector-matrix notation as follows.

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

where N is the analysis interval, $\bar{y}_f$ is an $(N-I, 1)$ column vector whose elements are analytic speech signals, $\bar{\theta}$ is an $(L \cdot I, 1)$ column vector whose elements are complex parameters,

and $\bar{\Phi}_f$ is an $(N-I, L\cdot I)$ matrix whose elements are weighted analytic speech signals by the complex basis. The superscript T denotes transposition.

C. MMSE-based algorithm

MSE criterion is defined as follows.

$$\begin{aligned} \bar{r}_f &= [r^c(I), r^c(I+1), \cdots, r^c(N-1)]^T \\ &= \bar{y}_f + \bar{\Phi}_f \hat{\theta} \qquad (7) \\ r^c(t) &= y^c(t) + \sum_{i=1}^{I} \sum_{l=0}^{L-1} \hat{g}^c_{i,l} f^c_l(t) y^c(t-i) \qquad (8) \\ E &= \bar{r}_f^H \bar{r}_f = (\bar{y}_f + \bar{\Phi}\hat{\theta})^H (\bar{y}_f + \bar{\Phi}\hat{\theta}) \qquad (9) \end{aligned}$$

where $\hat{g}^c_{i,l}$ is the estimated complex parameter, $r^c(t)$ is an equation error, i.e., complex AR residual, and E is the MSE for the equation. 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.

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

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

IV. Proposed AFE

In this paper, the Wiener filter is modified using the spectra in Eq.(3) estimated by MMSE-based TV-CAR analysis. The estimated spectrum is obtained by Eq.(5) with the estimated complex parameter $g^c_{i,l}$. The gain of the spectrum is determined by the corresponding power of the AR residual[15].

Although the complex analysis can estimate more accurate speech spectrum in low frequencies, it suffers from low estimation in high frequencies because of the aliasing of an analytic signal. In the AFE, 16kHz speech is input and uniformly divided into two sub-band signals by QMF whose frequency response is shown in Figure 1. Note that the lower band signal is Wiener filtered and the higher band signal is not. To remove the influence of the aliasing, complex analysis is performed to the input 16kHz wide-band speech, and the spectrum for the Wiener filter is extracted. The Wiener filtering is carried out by the wide-band analysis as in Figure 2. The procedure is as follows.

(1-w)Wide-band analysis is carried out with the 16KHz wide-band signals.
(2-w)Narrow band speech power spectrum is extracted.
(3)The Wiener Filter (WF) is designed by using the narrow band power spectrum by Eq.(3),
(4)The WF is operated.

The extraction (2-w) is realized as follows. The estimated power spectrum from 0-4kHz is extracted and multiplied by the frequency response of the low pass filter of the QMF, as shown by the blue line in Figure 1. Note that less distorted spectrum can be obtained by wide-band complex analysis

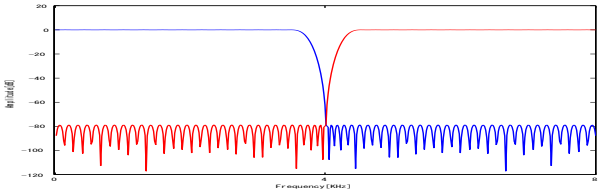

Figure 1: QMF frequency response

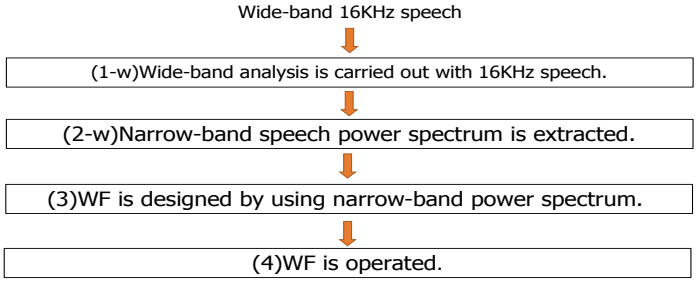

Figure 2: Blockdiagram of wide-band analysis

using the low pass filter. The wide-band signal cannot be obtained by the first stage of the Wiener filter since the higher band is not emphasized; thus, the wide-band analysis is introduced only in the first stage and TV-CAR analysis is performed for the 8kHz sampled emphasized speech in the second stage. Since the time-varying analysis is difficult to introduce in the frequency-domain Wiener filtering framework, time-invariant analysis is implemented by setting the complex basis expansion order L to 1. The corresponding real-valued analysis can be easily implemented by setting L to 1 and input signal as the input speech signal rather than the analytic signal.

V. Experiments

The CENSREC-2 speech database[18] was used for evaluation. It consists of four types of data from various internal car environments (Table 1) as follows. Record environment indicates a difference between the learning data and evaluation data.

Table 1: Types of data in CENSREC-2

	Microphone	Record environment
Condition1	Matched	Matched
Condition2	Matched	not Matched
Condition3	not Matched	Matched
Condition4	not Matched	not Matched

Table 2 shows the analysis conditions for the MMSE-based speech analysis. AR order I was 7 for complex analysis and was 14 for real-valued analysis. In the first stage of the IWF, the 16kHz signal was analyzed with I at 14. In the second stage of the IWF, the 8kHz signal was analyzed with I at 7. The basis expansion order L was 1 (i.e., time-invariant analysis).

The experiments are carried out by using the CENSREC-2 baseline script[18][19]. Remainder of the conditions including HMM configuration are the same as those in the script. The following 39 order parameters estimated by HCopy in the

HTK[20] was used as the HMM feature vector; MFCC(12th) +Δ MFCC(12th) +ΔΔ MFCC(12th) +log power(1st) +Δ log power(1st) +ΔΔ log power(1st). The experimental results are shown in Figures 3 and 4. In Figures, **CENSREC-2 Baseline** means HMM speech recognition without the front-end based on IWF, **AFE Baseline** means the original ESTI AFE that is based on the FFT-based IWF. **AFE Baseline + RASTA** means that RASTA filtering[6] is carried out for the MFCC estimated by the FFT-based IWF. It is the other conventional method. **Real_8k+Real_8k** means the real-valued AR analysis shown in Table 2 is carried out on first and second stages. **Complex_16k+Complex_8k** means the complex-valued AR analysis shown in Table 2 is carried out. Figures 3 and 4 demonstrate the performance (recognition rate and relative improvement) based on the ETSI AFE(FFT), the real-valued MMSE, and the proposed wide-band complex-valued MMSE analysis. The results demonstrate that the proposed AR analysis methods outperform the original FFT-based AFE. The proposed wide-band complex-valued MMSE analysis performs best since complex analysis can estimate more accurate spectrum than the conventional analysis, especially in low frequencies, and the wide-band analysis can eliminate spectral distortion in high frequencies. In addition, the AR analysis methods outperform the original AFE since AR analysis can estimate a more accurate speech spectrum without a fine harmonics structure than FFT analysis. It is important to be noted that the proposed method outperforms the other methods in the case of condition 4 that is common environments for ASR. All front-end methods do not work well in the case of condition 2. It may indicate a limitation of MMSE-based analysis method.

Table 2: Experimental conditions

Analysis window	Window Length: 25.6[ms] Shift Length: 10.0[ms]
Complex-valued AR	
First stage (16kHz)	I=14, L=1 (time-invariant)
Second stage (8kHz)	I= 7, L=1 (time-invariant)
Pre-emphasis	None
Real-valued AR	I=14, L=1 (time-invariant)
Pre-emphasis	None

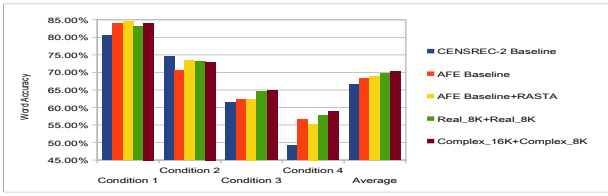

Figure 3: Recognition rates

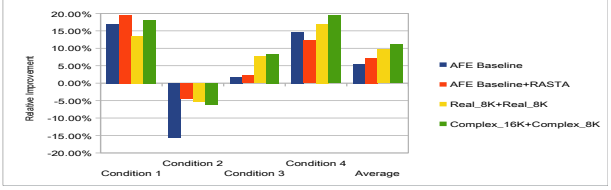

Figure 4: Relative improvements

VI. Conclusions

We have evaluated time-invariant real/complex-valued AR analysis based on the ETSI AFE for ASR. While FFT is used to estimate the spectra for the Wiener filter in the standard AFE, in the proposed method, spectra are estimated by MMSE-based complex-valued AR analysis. Performance was evaluated using the CENSREC-2 in-car noise-corrupted speech database and its baseline script. The complex-valued and real-valued MMSE-based analysis methods were implemented using MMSE-based TV-CAR speech analysis. To suppress the influence of aliasing, complex analysis was performed for a 16kHz signal, and the lower band spectrum was extracted by the low pass filter of the . The experimental results demonstrate that the wide-band complex-valued analysis outperforms the compared methods. In future, robust and time-varying analyses will be investigated.

References

[1] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans., ASSP-27, pp.113-120,1979.
[2] J.S.Lim and A.V.Oppenheim, "All-pole modeling of degraded speech," IEEE Trans., ASSP-26, pp.197-210, 1978.
[3] H.L.Hansen and M.A.Clements, "Constrained iterative speech enhancement with application to speech recognition," IEEE Trans. Signal Processing, vol.39, pp.795-805, April 1991.
[4] Y.Ephraim and D.Malah, "Speech enhancement using minimum mean-square error log-spectral amplitude estimator," IEEE Trans., ASSP-33, pp.443-445, 1985.
[5] Pedro J. Moreno, Bhiksha Raj and Richard M. Stern, "A vector TAYLOR series approach for environment-independent speech recognition," Proc.ICASSP96, 1996.
[6] H.Hermansky and N.Morgan, "RASTA processing of speech," IEEE. Trans. Speech Audio Process., vol. 2, no. 4, pp. 578-589, Oct. 1994.
[7] C.Kim and R.M.Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," Proc. ICASSP2012, Kyoto, 2012.
[8] W.Kim and J.H. L.Hansen, "Feature compensation employing online GMM adaptation for speech recognition in unknown severely adverse environments," Proc. ICASSP2012, Kyoto, 2012.
[9] T.Shimamura, N.D.Nguyen, "Autocorrelation and double autocorrelation based spectral representations for a noisy word recognition system," Proc. Interspeech-2010, Makuhari, Japan, Sep. 2010.
[10] S.Keronen, J.Pohjalainen, P.Alku, M.Kurimo, "Noise Robust Feature Extraction Based on Extended Weighted Linear Prediction in LVCSR," Proc. INTERSPEECH2011, Firenze, 2011.
[11] K.Paliwal,et.al., "Usefulness of Phase spectrum in human speech perception," Proc. EUROSPEECH2003
[12] K. Yamamoto, E. Sueyoshi, S. Nakagawa, "Speech recognition using long-term phase information," Proc. Interspeech-2010, Makuhari, Japan, Sep. 2010.
[13] H.F.Pardede, et.al., "Feature normalization based on non-extensive statistics for speech recognition," Speech Communication. Vol. 55, pp. 587-599, Mar, 2013.
[14] ETSI Advanced Front-End,
ES 202 050 v1.1.5(2007-01)
http://www.etsi.org/deliver/etsi_es/202000_202099/202050/01.01.05_60/es_202050v010105p.pdf
[15] K.Funaki, "Speech Enhancement based on Iterative Wiener Filter using Complex Speech Analysis," EUSIPCO-2008, Lausanne, Switzerland, Aug.2008.
[16] K.Funaki,et al., "On a Time-varying Complex Speech Analysis," EUSIPCO-98, Rhodes, Greece, Sep.9-11,1998.
[17] K.Funaki, "A time-varying complex AR speech analysis based on GLS and ELS method," Eurospeech2001, Aalborg, Denmark, Sep.2001.
[18] CENSREC-2, http://www.slp.cs.tut.ac.jp/CENSREC/
[19] S. Nakamura, M. Fujimoto, and K. Takeda, "CENSREC2: corpus and evaluation environments for in car continuous digit speech recognition," Proc. INTERSPEECH, 2006.
[20] HTK Web site, http://htk.eng.cam.ac.uk