I.J. Wireless and Microwave Technologies, **2011, 2, 55-61**
Published Online April 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijwmt.2011.02.08

Available online at http://www.mecs-press.net/ijwmt

A Speech Enhancement Method Based on Kalman Filtering

Chaogang Wu, Bo Li, Jin Zheng

Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Abstract

The enhancement of speech degraded by non-stationary interferers is a highly relevant and difficult task for many signal processing applications. In this study, we present a monaural speech enhancement method based on spectral subtraction and Kalman filtering (KF) by extracting the Liljencrants–Fant (LF) excitation during voiced speech, in which the nature of glottal flow can be maintained. Therefore, the approach could preserve the glottal pulse's nature characteristic in Kalman filtering and thus achieve significant improvements on objective quality. The quality of the enhanced speech has been evaluated by perceptual evaluation of speech quality (PESQ) score. The results indicate that the proposed algorithm could improve the output speech quality compared with the conventional KF algorithm and sub-band spectral subtraction.

Index Terms: Speech enhancement; LF glottal flow; source separation

1. Introduction

Enhancing speech degraded by interference is an important task for many signal processing applications, including speech coding and speech recognition. The challenge in the enhancement of speech is to raise the signal to noise ratio (SNR) and to reduce the effect of reverberation. The main problem in processing the speech for enhancement is the non-stationary nature of speech production process and the potentially speech-like real-world interferers, so there is a significant and variable spectral overlap between speech and interferer. Therefore, the speech enhancement is very important, which can improve both the intelligibility and the quality of speech by attenuating the interferer without substantially degrading the speech [1-3].

The temporal and spectra characteristics of the speech change continuously, both in the wave shape and in the energy issues. Thus, the SNR is not only a function of time but also a function of frequency. Hence, it is difficult to estimate the characteristics of the degraded speech signal. The methods based on spectral subtraction for noise reduction and deconvolution of room response for dereverberation are not enough [1] to figure out this problem.

* Corresponding author.
E-mail address: wuchaogang@gmail.com

In this paper, we propose a method for enhancing speech by focusing on the characteristics of speech production. The enhancement technique is based on the property of the excitation source for voiced speech, in which the strength of excitation is the maximum around the instant of glottal closure. It is possible to enhance speech by exploiting the characteristics of excitation source in speech production. Therefore, the approach could preserve the glottal pulse's original characteristic in Kalman filtering and thus achieve significant improvements on objective quality. The PESQ tests show that the proposed algorithm could give out better output speech quality compared with the conventional KF algorithm and sub-band spectral subtraction.

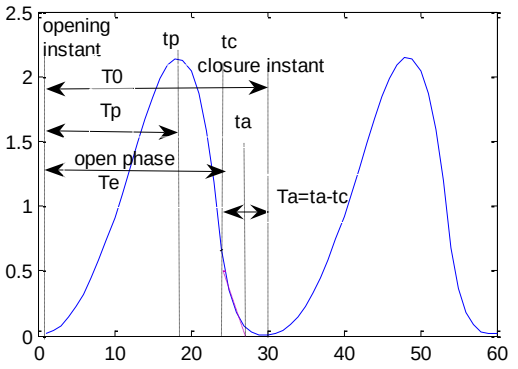
2. The LF-Filter Model

In speech processing, a source-filter model is widely used to represent the speech production mechanism. For voiced speech segments, the source component is indicated by the glottal flow derivative (GFD) [4], which incorporates the derivative effect caused by the lips radiation to the signal observed at the glottis. For unvoiced speech segments, the source component is referred to Gauss-like noise signal. Moreover, the Liljencrants–Fant (LF) model is regarded as a reasonable approximation of the GFD, which enables the characterization of the glottal source signal with five parameters: one is for the cycle interval of the glottal source T0, one is for the amplitude and other three are to define the shape of the glottal flow. A typical LF waveform is depicted on Fig. 1. Among the possible parameter sets to define the shape, the vector θ = [Qq, am, Qa] has been chosen, in which Qq corresponds to the open quotient (Qa = Te/T0), am is the asymmetry coefficient and Qa is the return phase quotient (Qa = Ta/((1−Oq)T0)). θ presents the space of shape parameters. The explicit expression of the model for one fundamental period is given by:

$$E(t)=\begin{cases} E_1(t)=E_0e^{\alpha t}\cos(\omega_g t), 0<t<t_e \\ E_2(t)=-\dfrac{E_e[e^{-\varepsilon(t-t_e)}-e^{-\varepsilon(t_c-t_e)}]}{\varepsilon t}, t_e<t<t_c \end{cases} \tag{1}$$

where the parameters a, ε and ωg are implicitly related to θ. Normally, the glottal closure instant tc is chosen for the reference of one cycle interval of voiced speech signal. The vector [T0, Te, Tp, Ta] is another parameter set to define the shape of glottal flow and can be transformed to parameter set of θ.

Given the above assumptions, the speech signal s(n) can be represented by means of the LF-filter model:



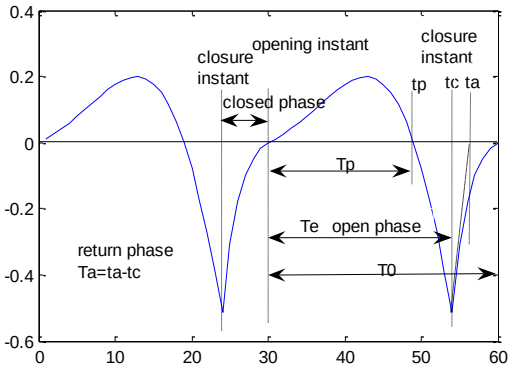


Figure. 1 Glottal flow and GFD of the LF model.

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + b_0 g_{LF}(n) + e(n)$$

Where ak(n) is the time-varying coefficient of the auto regression (AR) model which characterizes the vocal tract and gLF(n) denotes the LF glottal flow derivative. Coefficient b0 is related to the LF waveform amplitude, e(n) is a residual signal which contains the information that is not explicitly captured by the linear LF-Filter model. e(n) can not be ignored because of the mismatch between the deterministic part of glottal waveform and the LF model, the nonlinear effects such as ripple which results from source-vocal tract interactions and the noise components of the glottal waveform. For unvoiced segments, the speech signal s(n) can be represented as

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + e(n) \tag{3}$$

where the excitation signal *e(n)* is modeled by Gauss noise.

3. The Kalman Fitering Speech Enhancement Based on LF-Filter Model

In order to improve the quality of the speech signal with noise using Kalman filter, the parameter of LF-Filter model must be estimated from the speech degraded by interferers. Therefore, the speech enhancement (LF-KF) in this paper has been divided into two main steps: 1) estimating the LF-ARX model parameters and 2) enhancing speech signal by Kalman filtering.

A. Estimation of LF-filter parameters

Linear prediction (LP) analysis can be used to derive the source characteristics and the inverse filtering based on LP analysis, which is the common technique to get the glottal flow from speech signal. But the filter parameters can not be computed exactly by LP analysis for the interference of noise [5]. To get the better estimation of LF-Filter model, a new method including two steps is proposed. In this method, the improved inverse filtering is used to get LF parameters and the approximate LP coefficients firstly, and then the LP filter parameters are optimized by an iterative process. We reduce the noise by using of the inter-pith smoothing before LP analysis:

$$\begin{aligned} S_n &= \begin{bmatrix} S_1 + n_1 & S_2 + n_2 & S_3 + n_3 & S_4 + n_4 \end{bmatrix} \\ S_{ns} &= (S_1 + S_2 + S_3 + S_4)/4 + (n_1 + n_2 + n_3 + n_4)/4 \end{aligned} \tag{4}$$

Additionally, the Wl-Jaroudi-Makhoul arithmetic is embedded for optimizing the LP coefficients.

1) LF parameter estimation

The estimation of the LF model parameters can be formulated as the minimization of the energy of the residual signal. However, it can be seen that this optimization is highly nonlinear and thus rather intricate (2). The method proposed in this paper modified and developed these previous methods and can be completed through the following steps:

a) Estimation of T_0 and the pitch synchronization.

b) LP analysis and the estimation of approximate estimation of vocal tract filter.

In particular, voiced speech is produced as a result of excitation of the vocal tract system with quasi-periodic glottal pulses. The significant excitation in each glottal cycle takes place at the instant of glottal closure. The strength of excitation of the vocal tract system depends on the rate at the instant of glottal closure. In fact, it is this strength that enables us to perceive speech in spite of degradation in speech. Due to the high strength of excitation, the SNR of speech is high around these instants compared to other regions. Thus, it is possible to enhance speech by exploiting the characteristics of excitation source in speech production.

c) Estimation of LF model using the simple pitch synchronous LP (PSLP) inverse filtering.

It can be solved via sequential unconstrained minimization method. By uniformly sampling the Qa values and solving the problem at each sample, the best estimation can be obtained when it has minimum error.

2) Optimization of vocal tract filter parameters

We can estimate the vocal tract filter and LF model parameter using LP analysis, but the filter parameters are not accurate. To estimate the accurate vocal tract filter without the effect of the glottal flow derivative on the output speech is very important. A(k,a) is the frequency response of the accurate vocal tract filter, S(k) is the FFT of speech signal s(n) and the G(k) is the FFT of glottal flow signal gLF. Suppose that U(k)=S(k)/G(k), then we can get:

$$f_U(U(k)|a)=\frac{1}{\pi A(k,a)}\exp\left(-\frac{|U(k)|^2}{A(k,a)}\right) \tag{5}$$

It has the same form as the formula in the discrete all-pole model. So the Wl-Jaroudi-Makhoul algorithm can be utilized to optimize the filter parameters:

- Computing the responded $\hat{h}(-i)$ of the estimated $a_k, k=1,2,\cdots,M$, $\hat{h}(-i)$ is the time reserved form of the response signal of the all pole filter $a_k, k=1,2,\cdots,M$.

$$\hat{h}(-i)=\frac{1}{N}\sum_{m=1}^{N}\frac{e^{-jmiw_0}}{\sum_{k=0}^{M}e^{-jkmw_0}a_k} \tag{6}$$

- Updating the $a_k, k=1,2,\cdots,M$ according to the $\hat{h}(-i)$ computed in (6) by the following equation.

$$\sum_{k=0}^{M}a_k r(i-k)=\hat{h}(-i), 0\le i\le M \tag{7}$$

By using the Wl-Jaroudi-Makhoul algorithm, an accurate estimation of vocal tract filter has been achieved.

B. Speech enhancement using Kalman filter

With the estimated LF model mentioned above and the vocal tract filter, we can construct the Kalman filter to enhance the degraded speech signal as follows.

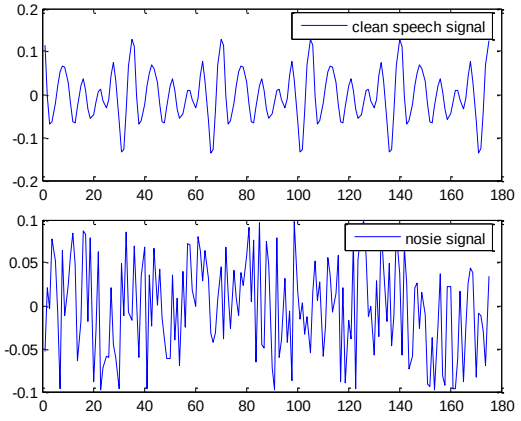
$$\begin{aligned}
&\vec{e}(n) = \vec{y}(n) - \hat{x}(n|n-1) \\
&K(n) = P(n|n-1)[R + P(n|n-1)]^{-1} \\
&\hat{x}(n|n) = \hat{x}(n|n-1) + K(n)\vec{e}(n) \\
&P(n|n) = [I - K(n)]P(n|n-1) \\
&P(n|n-1) = FP(n-1)F^T + H\sigma^2 \\
&P(n|n-1) = E\{[\vec{x}(n) - F\hat{x}(n-1)][x(n) - F\hat{x}(n-1)]^T\}
\end{aligned} \tag{8}$$

where k(n) is the Kalman gain vector, P(n|n) is an error covariance matrix, P(n|n-1) is an a priori error covariance matrix. By averaging the interferer magnitude spectrum during a speech pause, we can estimate the average interferer magnitude. The constructed Kalman filter is used to filter every speech frame, which could reduce the noise while keeping the speech characteristics.

4. Experiments and Results

Compared to the established speech enhancement approaches, such as the conventional Kalman filtering and multi-band spectral subtraction (MBSB), the approach in this paper (LF-KF) could keep the glottal pulse's nature characteristic in Kalman filtering and thus can achieve significant improvements on the speech quality. The multi-band spectral subtraction achieves equal or better subjective listener ratings than many other approaches. However, only estimating the average interferer magnitude would limit the enhancement performance, because that the interferer contribution to the mixture at some points in time domain can deviate significantly from the average.

In our experiment, the speech signals are collected in a live room simultaneously from three different spatially located microphones. The data was sampled at 8 kHz and stored as 16 bit integers. The quality of the enhanced speech was evaluated by SNR and PESQ score. The results show that the LF-KF can achieve higher SNR and PESQ score (Table 1 and 2). This improvement is achieved for the LF model which is close to the true glottal pulse's nature characteristic. Fig. 2 demonstrates the filtered signals in the enhancement process.



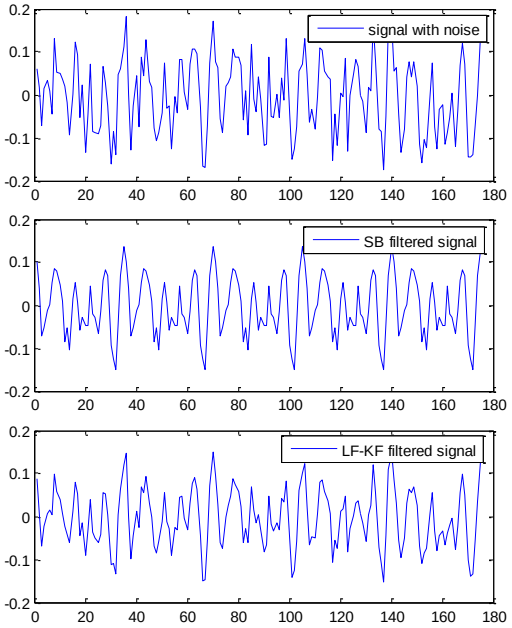


Figure 2. Enhancement of the signal using LF-KF.

TABLE 1. THE PESQ SCORES OF SPEECH ENHANCEMENT METHODS

PESQ	MBSB	KF	LF-KF
Gauss_0dB	2.111	2.838	2.976
Gauss_5dB	2.361	3.100	3.361
Gauss_10dB	2.895	3.650	3.702
Gauss_15dB	3.226	3.978	3.958

TABLE 2. THE SNR OF SPEECH ENHANCEMENT METHODS

SNR/dB	MBSB	KF	LF-KF
Gauss_0dB	5.1	11.1	12.0
Gauss_5dB	4.9	8.5	8.9
Gauss_10dB	2.0	5.4	5.7
Gauss_15dB	2.1	3.3	4.1

The enhanced speech is much better compared to the coherently added speech signal because better approximate all-pole filter can be derived from the degraded speech by processing the speech around the instants of glottal closure and the accurate filter parameters can be achieved by Wl-Jaroudi-Makhoul algorithm.

5. Conclusion

This paper describes a method for enhancing speech by focusing on the characteristics of speech production. Based on the property of the voiced speech's excitation source, the approach can preserve the glottal pulse's original characteristic in Kalman filtering and thus achieve significant improvements on speech quality. It achieves the enhanced signal to noise ratio and perceptual evaluation of speech quality score compared with

methods such as the multi-band spectral subtraction and the conventional approach based on Kalman filtering during the quality tests. This result would provide high potential for the application in the speech enhancement.

Acknowledgment

This work was partially supported by the 973 program (2010CB327900) and the Defense Industrial Technology Development Program (BXX2011XXX8)

References

[1] D. Vincent, O. Rosec, and T. Chonavel, "A New Method for Speech Synthesis and Transformation Based On an ARX-LF Source-Filter Decomposition and HNM Modeling," ICASSP. pp. 525–528, 2007.

[2] H. Zhao, and X. Zou, "A Speech Enhancement Preprocessor for Low Bit Rate Speech Coding," Pacific-Asia Conference on Circuits, Communications and System. pp. 443–445, 2009.

[3] Christian D. Sigg, Tomas Dikk, and Joachim M. Buhmann, "Speech Enhancement with Sparse Coding in Learning," Dictionaries in Proc. ICASSP. pp. 4758–4761, 2010.

[4] B. Yegnanarayana, S. R. Mahadeva Prasanna and K. Sreenivasa Rao, "Speech Enhancement Using Excitation Source Information," ICASSP. vol. 1, pp. 541–544, 2002.

[5] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF glottal source parameters based on ARX model," Interspeech. pp. 333–336, 2005.