SPEECH ENHANCEMENT BASED ON THE DECOMPOSITION OF SPEECH INTO DETERMINISTIC AND STOCHASTIC COMPONENTS AND PSYCHOACOUSTIC MODEL

Seokhwan Jo, Chang D. Yoo

Div. of EE, Dept. of EECS, KAIST,
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
antiland00@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

A novel speech enhancement algorithm based on both a decomposition of speech into coexisting deterministic and stochastic components and a psychoacoustic model is proposed. Noisy speech is first decomposed into deterministic and stochastic components, and then each component is enhanced preserving its individual characteristics. A psychoacoustic model is taken into account when enhancing the stochastic component which usually has much lower energy than the deterministic component. Simulation results show that the proposed algorithm performs better than some of the more popular algorithms in terms of segmental signal-to-noise ratio (SNR) and speech recognition rate.

Index Terms— Speech enhancement, deterministic component, stochastic component, masking threshold

1. INTRODUCTION

It is understood that the presence of background noise can degrade the performance of many speech communication and recognition systems. Most important of all, noise induces fatigue on a listener. For these reasons, a variety of enhancement algorithms have been proposed [1], [2], [3], albeit with limited success. Part of the reason for their limitation relates to the simplistic model assumption which the past enhancement algorithms are based on. Often, these algorithms are derived under the assumption that speech signal or speech spectral coefficients follow either deterministic model or stochastic model. In [2] and [4], speech enhancement algorithms are derived under the deterministic speech model. In [3] and [5], speech spectral coefficients or speech spectral amplitudes are stochastically estimated assuming the coefficients follow a certain probability distribution. In general, both deterministic and stochastic components coexist in speech. Observation of the short-time spectra often indicates that speech is composed of a mixture of harmonically modulated sinusoids and some random noise. Therefore, enhancement algorithms derived under either only stochastic model or deterministic model are limited in performance since it does not capture the full speech characteristics.

In this paper, a novel algorithm that is based on coexisting deterministic and stochastic models is proposed. In terms of speech production, the deterministic component can be thought of as the voiced speech and the stochastic component as the unvoiced speech. Noisy speech is first decomposed into deterministic and stochastic components, and then each component is enhanced preserving its individual characteristics. Generally, it is more difficult to estimate the stochastic component than the deterministic component since the former is of lower energy than the latter. In this work, a psychoacoustic masking threshold is used to adaptively enhance the stochastic component.

This paper is organized as follows. Section 2 presents a coexisting deterministic and stochastic model. Section 2.1 presents enhancement of deterministic component. Section 2.2 presents enhancement of stochastic component using psychoacoustic model. Section 3 shows the simulation results, and Section 4 concludes the paper.

2. SPEECH ENHANCEMENT BASED ON THE COEXISTING DETERMINISTIC AND STOCHASTIC MODEL

Speech consists of deterministic and stochastic components, and both windowed components are denoted respectively as $v_w[n]$ and $u_w[n]$. Thus, a windowed speech segment $s_w[n]$ can be represented as

$$s_w[n] = v_w[n] + u_w[n] \tag{1}$$

where subscript signifies that each term is a short-time segment which is obtained by applying a window function $w[n]$. In the Fourier domain, it is represented as

$$S(k,l) = V(k,l) + U(k,l) \tag{2}$$

where $S(k,l)$, $V(k,l)$ and $U(k,l)$ are Fourier coefficients for frequency bin k and time frame l of clean speech, the deterministic component, and the stochastic component, respectively.

In many speech applications of today, speech is considered as a response of a linear time varying filter driven by excitation which is a sum of periodic impulse train and white noise sequence. The deterministic component represents the response due to the periodic impulse train and the stochastic component represents the response due to the white noise sequence.

When speech is corrupted by additive noise as follows

$$y[n] = s[n] + z[n], \tag{3}$$

the Fourier coefficients satisfy the following equation

$$\begin{aligned} Y(k,l) &= S(k,l) + Z(k,l) \\ &= V(k,l) + U(k,l) + Z(k,l) \\ &= V(k,l) + X(k,l) \end{aligned} \tag{4}$$

where $Y(k,l)$ and $Z(k,l)$ are Fourier coefficients of noisy speech $y[n]$ and noise $z[n]$, respectively. Further, we assume that speech and noise are uncorrelated, and the Fourier coefficients of both $U(k,l)$ and $Z(k,l)$ follow a zero-mean Gaussian distribution.

When speech degraded by additive random noise is decomposed into coexisting deterministic and stochastic components, the bulk of the noise energy appears in the stochastic component [6]. Thus, the deterministic component $V(k,l)$ is relatively less corrupted by the noise process. In this paper, the Fourier coefficients of noisy speech is assumed to follow a non-zero mean complex Gaussian distribution, which is given by

$$f(Y(k,l)) = \frac{1}{\pi\sigma_x^2(k,l)} \exp\left\{ -\frac{|Y(k,l) - V(k,l)|^2}{\sigma_x^2(k,l)} \right\} \quad (5)$$

where $\sigma_x^2(k,l)$ is the variance the Fourier coefficient $X(k,l)$ and equals the sum of the variance of the stochastic component and the noise, that is, $\sigma_x^2(k,l) = \sigma_u^2(k,l) + \sigma_z^2(k,l)$. In terms of speech production, the deterministic component is constructed to have the properties of a voiced speech (a sum of harmonically modulated sinusoids) and the stochastic component to have the properties of a unvoiced speech (autoregressive (AR) signal).

The scheme of the proposed enhancement algorithm is shown in Fig. 1. The system enhances the deterministic component and the stochastic component preserving its individual characteristic, respectively. The stochastic component is characterized as AR random noise and is estimated using a Wiener filter which is adaptively controlled by a masking threshold of the stochastic component. The deterministic component is characterized as a sum of harmonically modulated sinusoids and is estimated by synthesizing the sinusoids using the estimated parameters. Since the masking threshold of the stochastic component is difficult to calculate from noisy speech, speech is roughly estimated, and it is used to compute the masking threshold of the stochastic component. The masking threshold of the stochastic component as well as stochastic model parameters are computed iteratively.

2.1. Enhancement of deterministic component

As previously mentioned, the deterministic component can be viewed as the voiced component of speech and is only slightly degraded by the noise process compared to the stochastic component. The deterministic component is modeled mathematically as a sum of harmonically modulated sinusoids over the time duration of the window. The pitch period in lth frame, $P_0(l)$, can be used to form a harmonic series representation for the deterministic component as follows

$$V(k,l) = \sum_{m=-M(l)}^{M(l)} A_m(l) W((k - mk_0(l)) \bmod N) \quad (6)$$

where $W(k)$ is the Fourier coefficients of the window function $w[n]$ and N is the number of DFT points. The parameter $A_m(l)$ represents the amplitude of mth harmonic in lth frame. The parameter $k_0(l)$ represents the fundamental frequency in lth frame which is related to the pitch period $P_0(l)$ as

$$k_0(l) = \frac{N}{P_0(l)}. \quad (7)$$

The number of harmonics in the lth frame $M(l)$ is a function of the fundamental frequency and is given by

$$M(l) = \left\lfloor \frac{N}{2k_0(l)} \right\rfloor \quad (8)$$

where $\lfloor \cdot \rfloor$ denotes the smallest integer less than or equal to the argument.

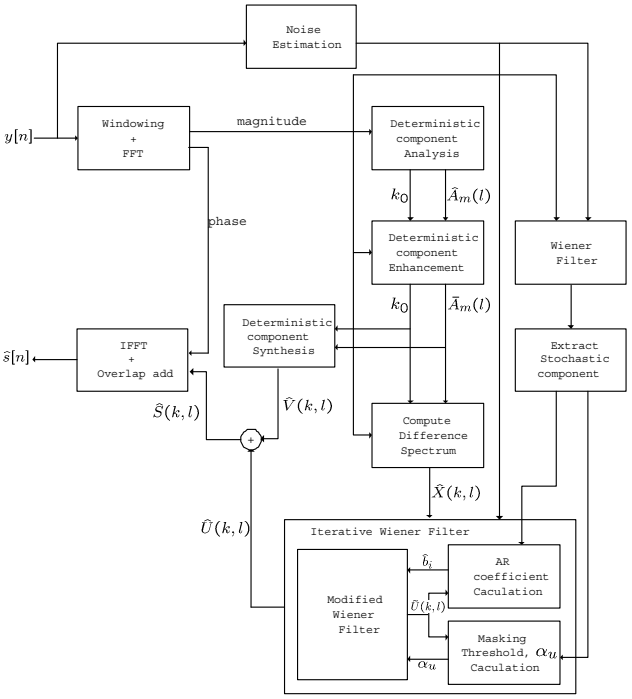

Fig. 1. Block diagram for dual excitation speech enhancement using psychoacoustic model.

For the enhancement of the deterministic component, the pitch is estimated using a robust algorithm developed by Griffin *et al.* [7]. The harmonic amplitude $A_m(l)$ is adjusted as follows

$$\bar{A}_m(l) = \begin{cases} \hat{A}_m(l), & \text{if } |\hat{A}_m(l)| > \alpha_v \sqrt{\sigma_z^2(mk_0, l)} \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

where $\hat{A}_m(l)$ is a harmonic amplitude estimate from noisy speech. Here, α_v is an oversubtraction factor for deterministic component enhancement, and $\sigma_z^2(mk_0, l)$ is a noise power spectrum density estimate at mth harmonic.

2.2. Enhancement of stochastic component

According to a speech production model, unvoiced speech can be represented as the response of a linear quasi-stationary system to a noise-like excitation. For this reason, the stochastic component is modelled as an output of an AR process of order p. This is expressed mathematically as

$$\sum_{i=0}^{p} b_i u_w[n-i] + G \cdot d[n] = 1, \quad b_0 = -1 \quad (10)$$

where G and $d[n]$ represent the gain and white Gaussian noise with zero mean and unit variance, respectively.

The minimum mean square error (MMSE) estimate of the stochastic component is the conditional mean and is obtained using the

Wiener filter. The stochastic component estimate is given by

$$\begin{aligned}\hat{U}(k,l) &= E[U(k,l)|Y(k,l),\hat{V}(k,l)] \\ &= \frac{\sigma_u^2(k,l)}{\sigma_u^2(k,l)+\sigma_z^2(k,l)}(Y(k,l)-\hat{V}(k,l)) \\ &= \frac{\sigma_u^2(k,l)}{\sigma_u^2(k,l)+\sigma_z^2(k,l)}\hat{X}(k,l). \end{aligned} \quad (11)$$

Here, $\sigma_u^2(k,l)$ is represented by b_i and G as follows,

$$\sigma_u^2(k,l) = \frac{G^2}{\left|1-\sum_{i=1}^{p} b_i e^{-ji\frac{2\pi}{N}k}\right|^2}. \quad (12)$$

Since the energy of the stochastic component is generally much lower than that of the deterministic component, it is difficult to estimate the stochastic component from noisy observation. Using the masking property of a psychoacoustic model, the stochastic component is estimated as proposed in [8]. The oversubtraction factor of the Wiener filter is controlled by the masking threshold of stochastic component.

The new MMSE estimator is given by

$$\hat{U}(k,l) = \frac{\sigma_u^2(k,l)}{\sigma_u^2(k,l)+\alpha_u(k,l)\sigma_z^2(k,l)}\hat{X}(k,l) \quad (13)$$

where $\alpha_u(k,l)$ is the oversubtraction factor of stochastic component and is adjusted by the masking threshold $T_u(k,l)$. The calculation of the masking threshold is summarized in a number of literatures [8] [9]. The steps involved in determining the masking threshold are as follows:

1. *Critical band analysis* : sum up the power spectrum in each critical band (Bark), where the power spectrum is obtained by magnitude squaring the Fourier coefficient.
2. *Spreading* : convolve with a spreading function to take into account the effect of adjacent critical bands.
3. *Offset* : subtract the offset by considering the tone-like or noise-like nature of the speech.
4. *Re-normalization* : convert the spread spectrum back to Bark domain.
5. *Absolute threshold* : compare with the absolute threshold and choose the maximum between them.

The adjustment $\alpha_u(k,l)$ is obtained by the following

$$\alpha_u(k,l) = F_{\alpha_u}[\alpha_{min}, \alpha_{max}(\bar{\xi}_u(l)), T_u(k,l)] \quad (14)$$

where α_{min} is the minimal value of the oversubtraction factor and typically equals to 1, and $\alpha_{max}(\bar{\xi}_u(l))$ is the maximal value and a function of $\bar{\xi}_u(l)$. It will be explained in more detail soon.

Here, $F_{\alpha_u} = \alpha_{max}(\bar{\xi}_u(l))$ when $T_u(k,l) = T_u(k,l)_{min}$ and $F_{\alpha_u} = \alpha_{min}$ when $T_u(k,l) = T_u(k,l)_{max}$. When the masking threshold is high, speech signal is strong and residual noise is naturally inaudible. Hence, the noise reduction should be low in order to reduce speech distortion. On the other hand, when the masking threshold is low, speech signal is weak and residual noise may be annoying to listener. So, the noise reduction should be high in order to reduce noise. Other values between two extreme cases are interpolated based on the logarithmic values of $T_u(k,l)$.

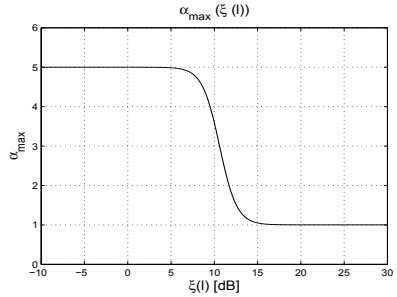

Fig. 2. Adjustment of α_{max}. The equation is $\alpha_{max} = \frac{1}{0.25+e^{10\log_{10}\bar{\xi}(l)-12}} + 1$ which is decided based on the value of SegSNR.

When signal-to-noise ration (SNR) is high, noise reduction needs to be reduced. For this reason, α_{max} is a function of SNR. Here, $\bar{\xi}_u(l)$ is the average value of *a priori* SNR in lth frame,

$$\bar{\xi}_u(l) = \frac{1}{N}\sum_{k=1}^{N}\frac{\sigma_u^2(k,l)}{\sigma_z^2(k,l)}. \quad (15)$$

Hence, when $\bar{\xi}_u(l)$ is low, α_{max} is a large value and the noise reduction is strengthened. The converse happens when $\bar{\xi}_u(l)$ is high.

To estimate the stochastic component more accurately, an iterative method is used. In other words, α_u, b_i and G are obtained iteratively. Thus, the stochastic component estimator is iterative Wiener filter.

3. PERFORMANCE EVALUATION

In this section, the performance of the proposed algorithm was evaluated and compared to other speech enhancement algorithms. The test sentences were selected from TIMIT database. The Kaiser window with $\beta = 7$ was used with frame size $N = 512$ (32ms) with 50% overlap. The performance of the proposed algorithm was evaluated in terms of segmental signal-to-noise ratio (SegSNR) and speech recognition rate. The SegSNR is defined as

$$\text{SegSNR} = \frac{1}{T}\sum_{m=0}^{T-1} 10\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1} s^2[n+Nm]}{\frac{1}{N}\sum_{n=0}^{N-1}(s[n+Nm]-\hat{s}[n+Nm])^2}\right),$$

where $s[n]$ and $\hat{s}[n]$ are the original clean speech samples and the estimated speech samples. The upper and lower bound of the frame SNR were set to 35 dB and -10 dB respectively. All SegSNR results were averaged over 20 different speech signals.

For speech recognition test, we created new database by adding white Gaussian noise and f16 cockpit noise to the test set of TIMIT database, respectively. The input SNR level was set equal to 5 dB. In the speech recognition test, the mono-phone hidden Markov models (HMMs) of three states with 16 Gaussian mixtures and 39 dimension MFCC features were used. The HMMs were trained on clean speech training set and tested on noisy and enhanced versions of the testing set with free phone grammar. The speech recognition rate is measured in terms of phone correction rate (PCR) and phone accuracy rate (PAR).

The AR order of stochastic component model was set to $p = 30$. In the simulation, the parameters of algorithm were set to the following values: $\alpha_v = 3$, $\alpha_{min} = 1$, and $\alpha_{max} = \frac{1}{0.25+e^{10\log_{10}\bar{\xi}(l)-12}} + 1$ (shown on Fig. 2). Their parameters were heuristically chosen based on the performance of SegSNR. The proposed algorithm was

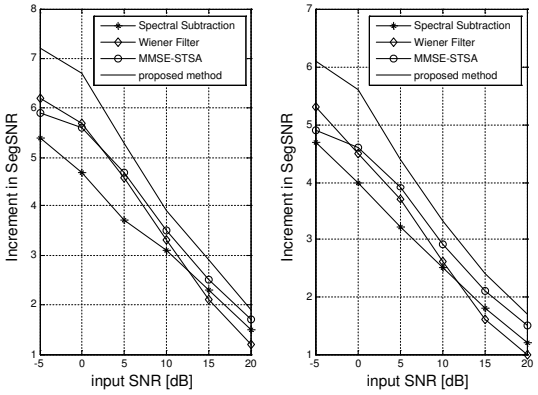

Fig. 3. (a) SegSNR improvement of proposed algorithm and other algorithms in white Gaussian noise. (b) SegSNR improvement in f16 cockpit noise.

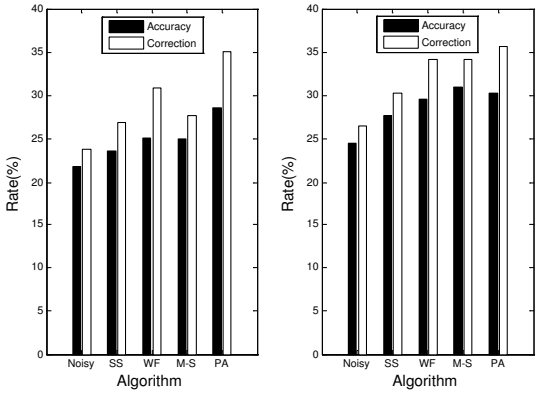

Fig. 4. SS: Spectral Subtraction, WF: Wiener filter, M-S: MMSE-STSA estimator, PA: the proposed algorithm. (a) Improvement of recognition accuracy and correction rate in white Gaussian noise. (b)Improvement of recognition accuracy and correction rate in f16 cockpit noise.

compared to spectral subtraction (SS) [1], Wiener filter (WF), and MMSE-STSA estimator (M-S) [3]. For WF and M-S, *a priori* SNR was estimated with the decision directed method with $\alpha = 0.98$ as proposed in [3].

In Fig. 3 (a), the average SegSNR improvement using the proposed algorithm in various white Gaussian noise level is shown. In this result, the proposed method performed better than the other methods, and it performed better with lower SNR. Fig. 3 (b) shows SegSNR improvement in various f16 cockpit noise level.

In the speech recognition test, the proposed algorithm performed better than others. The PCR and PAR using clean test set were respectively 67% and 64%. In Fig. 4 (a), the PCR and PAR of speech recognition are shown in white Gaussian noise. Both PCR and PAR of the proposed algorithm were higher than those of other algorithms. In Fig. 4 (b), the recognition result when f16 cockpit noise was used is shown. The PCR of the proposed algorithm is higher than that of other algorithms. The PAR improvement of the proposed algorithm was higher than those of any other algorithms except the M-S.

4. CONCLUSION

A novel speech enhancement algorithm based on both coexisting deterministic and stochastic models and a psychoacoustic masking property is proposed. When noisy speech is decomposed into deterministic and stochastic components, most of the noise energy appears in the stochastic component. Since the energy of the stochastic component is generally lower than that of the deterministic component, it is difficult to estimate the stochastic component. In the proposed method, a psychoacoustic masking property obtained from the stochastic component is used to adaptively enhance the stochastic component. Under various simulation conditions, the proposed algorithm performed better than other baseline algorithms.

5. ACKNOWLEDGMENTS

This work was supported in part by MIC & IITA through IT Leading R&D Support Project.

6. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, April 1979.

[2] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, pp. 137–145, 1980.

[3] Y. Ephraim, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, September December 1984.

[4] J. Jensen and J.H.L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, October 2001.

[5] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 845–856, 2005.

[6] C.D. Yoo J. Hardwick and J.S. Lim, "Speech enhancement using the dual excitation model," *Proc. IEEE Int. Conf. Acoustic, Speechm and Signal Processing*, vol. 2, pp. 367–370, 1993.

[7] D.W. Griffin and J.S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-36, August 1988.

[8] N. Virag, "Single channel speech enhancement based on masking properties of human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 126–137, 1999.

[9] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. of IEEE*, vol. 88, April 2000.