Dual-Channel Speech Intelligibility Enhancement Based on the Psychoacoustics

Sang-Hoon Lee
Postech
Electronic and Electrical Engineering
Hoja Dong, Nam Gu, Pohang, Kyungbuk
South Korea

Hong Jeong
Postech
Electronic and Electrical Engineering
Hoja Dong, Nam Gu, Pohang, Kyungbuk
South Korea

Abstract: In this paper, we propose an algorithm which enhances the speech intelligibility using the properties of human auditory system. In previous algorithms related to the speech intelligibility, the improvement in intelligibility has been mostly incorporated in a single-channel environment where the speech and noise signals are mixed together. But the speech enhancement problem of dual channel, in which the speech signal is separated from noise, has been rarely treated. This paper introduces the dual-channel speech enhancement algorithm which enhances the intelligibility by reinforcing speech before it is mixed with noise. To enhance the speech intelligibility, we use the masking phenomenon of the human auditory system. The proposed algorithm has been tested by the subjective experiment which resulted in improved results over the automatic gain control algorithm.

Key–Words: Dual channel speech intelligibility enhancement, Psychoacoustics

1. Introduction

In the environments of ordinary cellular phone communication or wireless communication for military purpose, the dual-channel speech intelligibility enhancement is very important. In most cases of cellular phone communication environments, many kinds of background noise exist. When noise makes it difficult to understand what the speaker is saying, most users are likely to overcome the situation by turning up the volume on the phone. However, this treatment is limited, and is not acceptable in a high level noise environment, where listeners cannot even understand a single word. This paper proposes an algorithm which particularly reinforces the speech signal according to noise in surroundings. The intelligibility of the enhanced speech using this method is significantly elevated in comparison with volume or power boosting. We aim for an algorithm whose purpose is to enhance the speech intelligibility and reduce computations so that it can be used in cellular phones, allowing some decrease in the quality of the speech.

The basic framework of the dual-channel speech enhancement is different from the single-channel speech enhancement, but many important ideas can be shared because both aim to enhance speech signal. Spectral subtraction has been the most frequently used algorithm in the single-channel speech enhancement [1, 2]. In this algorithm, only the averaged spectrum was reduced because the noise is assumed to be stationary. Subtraction parameter determines the tradeoff between the amount of noise reduction, the attenuation of the speech signal and the musical residual noise. These parameters are controlled using masking property of the human auditory system [1]. It also has been used to enhance speech signals based on the result from the measurement of the degree of conspicuousness of speech signals [2]. Amplifying the speech signals by controlling the speech speed has been developed [3]. Elderly listeners have difficulties in understanding speech, especially in the case of rapid utterance. To compensate such deterioration, the speech rate is slowed with invariance in pitch to maintain the timbre. Emphasizing the particular part of the speech signals results in the enhancement of the intelligibility [4, 5]. The second formant is more significant than the first formant for the intelligibility of speech [6, 7], and consonants convey more substantial information for intelligibility than vowels do [8]. In the paper [4], therefore, high pass filtering and amplitude compression are used to emphasize the second formant and consonants respectively. The method which amplifies consonants after dividing the consonants into five classes has been studied [5].

This paper introduces a new method of two-channel speech enhancement that is based on real-

time background noise analysis. We selectively enhance the speech signal which is weak to noise. This can be done by considering the masking effect of the human auditory system.

2. Algorithm

In the proposed algorithm, the speech signal and the noise signal are analyzed and processed mainly in the frequency domain. The processing is conducted on a frame-by-frame basis. We use the 32 ms-size raised cosine window for windowing, and the frame is overlapped every 16 ms. The windowed speech signal and noise signal are transformed using FFT and we can construct the noise masking threshold. Masking effect is a well-known psychoacoustic property of the human auditory system which has already been successfully applied to speech and audio coding in order to partially or totally mask the distortion introduced in the coding process [9]. We only consider the frequency domain masking, or simultaneous masking: a weak signal is made inaudible by a stronger signal that occurs simultaneously. This phenomenon is modeled via a noise masking threshold, below which all components are in audible. We only consider the frequency domain masking, or simultaneous masking: a weak signal is made inaudible by a stronger signal occurring simultaneously. This phenomenon is modeled via a noise masking threshold, below which all components are in audible.

After the calculation of noise masking threshold, we selectively enhance the band in which the speech signal is masked by the noise signal. However, if we change the speech signal only using the information of the background noise, then the speech will become a noiselike signal. To treat this problem, we consider the significance of each band, and then selectively enhance the significant part of the speech signal. For example the formant parts of the voiced sound are selectively enhanced.

The proposed enhancement scheme is presented in 1. It is composed of the following main step.

1. Spectral decomposition(FFT)
2. Calculation of the noise masking threshold T(w)
3. Calculation of the significance of the each speech band
4. Calculation of the each band gain based on the noise masking threshold and speech significance information
5. Multiply the band gain to the each speech band and normalize the energy of the speech frame
6. Inverse Fourier transform and overlap add

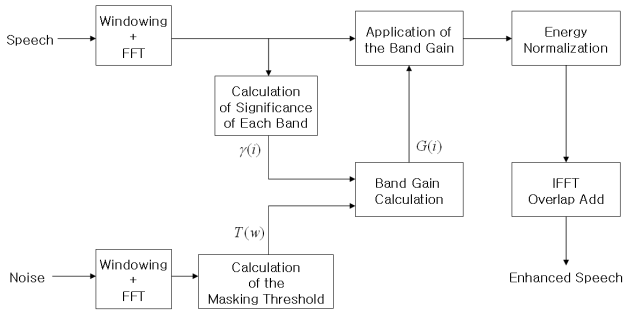

Figure 1. Block Diagram of the Entire System

2.1. Calculation of the Noise Masking Threshold $T(w)$

Let the FFT results of the speech and noise be the followings respectively.

$$S = s(1), s(2), \cdots, s(M) = s(m)_{m=1}^{M} \quad (1)$$

$$N = n(1), n(2), \cdots, n(M) = n(m)_{m=1}^{M}, \quad (2)$$

where M is the frame size.

The noise masking threshold $T(w)$ is obtained by modeling the frequency selectivity of the human ear and its masking property. The different calculation steps are summarized in [9].

1. Frequency analysis along a critical band scale, or Bark scale [10]: This critical band analysis is performed on the Fast Fourier transform (FFT) power spectrum by adding up the energies in each critical band i, according to the values given in [9].
2. Convolution with a spreading function $SF(i)$ to take into account the masking between different critical bands: The function used in this work has been proposed by Schroeder et al. in [11] and is represented in Fig. 2.
3. Subtraction of a relative threshold offset $O(i)$ depending on the noise-like or tone-like nature of the masker. In order to determine the noiselike or tonelike nature of the signal, the Spectral Flatness Measure (SFM) is used. The SFM is defined as the ratio of the geometric mean(G_m) of the power spectrum to the arithmetic mean(A_m)

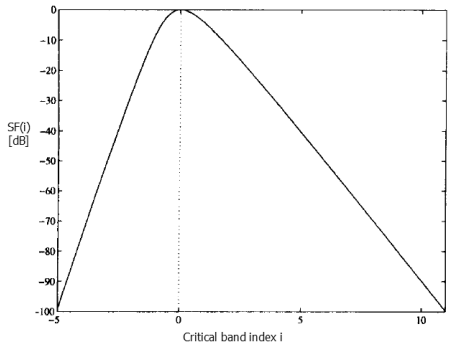

Figure 2. Spreading function used for the noise masking threshold

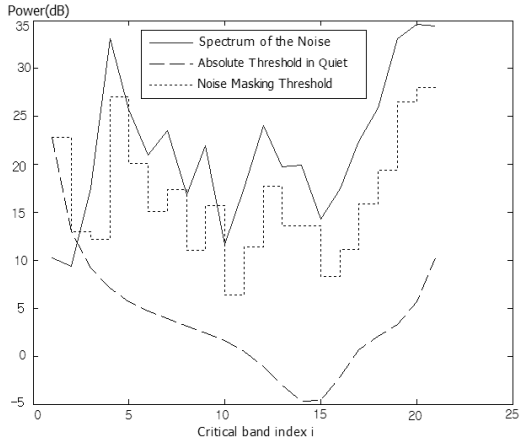

Figure 3. An example of the noise masking threshold

of the power spectrum. In this use, the SFM is converted to decibels, i.e.,

$$SFM_{dB} = 10log_{10}\frac{G_m}{A_m} \tag{3}$$

and further used to generate a coefficient of tonality α as follows :

$$\alpha = min(\frac{SFM_{dB}}{SFM_{dBmax}}, 1) \tag{4}$$

i.e., an SFM of $SFM_{dbmax} = -60dB$ is used to estimate that the signal is entirely tonelike, and SFM of 0dB to indicate a signal that is completely noiselike.

The offset $O(i)$ in decibels for the masking threshold in each band i is then set as

$$O(i) = \alpha(14.5 + i) + (1 - \alpha)5.5 \tag{5}$$

4. Renormalization and comparison with the absolute threshold of hearing : Since the energy estimates in each critical band are increased due to the effects of convolution in step 2, the renormalization need be applied as described in [9], i.e., multiply each threshold estimate by the inverse of the energy gain, assuming a uniform energy of 1 in each band.

An example of the noise masking threshold for a given speech frame is represented in Fig. 3.

2.2. Calculation of the significance of the each speech band

The speech will become a noiselike signal if we only change the speech signal using only the information of background noise. Hence, we have to introduce the concept of significance in enhancing the speech signal. For this purpose, we define an estimation for the presence of speech cues in each critical band. The estimation is determined as a function of the energy portion of the each speech band, i.e.,

$$\gamma(i) = \Lambda(\frac{\sum_{b=1}^{B_i} s^2(b)}{\sum_{m=1}^{M} s^2(m)}), \tag{6}$$

where i is the critical band index and B_i is the size of critical band i. Thus, $\frac{\sum_{b=1}^{B_i} s^2(b)}{\sum_{m=1}^{M} s^2(m)}$ is the energy portion of the band i. $\Lambda(k)$ is the soft decision function for smooth estimation using an exponential form.

$$\Lambda(k) = \frac{1 + exp(k)}{2} \tag{7}$$

Since the input range of $\Lambda(k)$ is [0-1], the range of the output $\Lambda(k)$ is [1-1.85].

2.3. Calculation of the each band gain

To make the speech signal robust to noise, the gains of each speech band are determined using the equation below.

$$G(i) = \gamma(i) \times \left(\alpha \times R + \beta \times \frac{\sum_{b=1}^{B_i} T^2(b)}{\sum_{m=1}^{M} T^2(m)} \right), \quad i = 1, 2, \cdots, 21 \tag{8}$$

where i is the critical band index and $T(w)$ is noise masking threshold which has been derived in section 2.1. Here, $\gamma(i)$ is multiplied to reflect the significance of each speech band. The parameters α and β have the following meanings.

1. SNR reference factor α

   No process will be carried out if R is too large to be influenced by noise. For a large α, each band gain has no difference between each other and it will become unity under the influence of the frame gain adjustment which will be discussed later. This results in a small change in the speech signal. Under the condition, where speech signal is comparatively larger than background noise, the distortion of the speech signal will be little if α is adjusted from 4 to 6. On the contrary, when the speech signal is required to be changed significantly due to the heavy noise, a value from 1 to 3 is adequate for α.

2. Noise distribution coefficient β

   In Eq. (8), $\sum_{b=1}^{B_i} T^2(b)$ and $\sum_{m=1}^{M} T^2(m)$ represent the noise energy of the i_th band and the total energy of the noise frame respectively. That is, $G(i)$ becomes larger when the noise energy of the i_th band is high. In the bands with more noise, the noise masking effect is suppressed by raising the corresponding band gain of speech. Therefore as β becomes large, the difference between each speech band gain becomes high and the speech timbre is changed considerably. When the noise level is low, a value from 5 to 9 is appropriate for β. But if the noise level is high enough to disturb the speech, then β should be adjusted from 10 to 15. When β is less than 5, the difference between each band gain is small, which means no change of timbre occurs. If β is particularly high, then the band gain decision excessively depends upon noise so that the original speech signal is heavily distorted.

After many experiments, the parameters are set to $\alpha = 3,\ \beta = 10$.

2.4. Energy normalization

The amplified signal according to the each band gain in Eq. (8), is processed again by following the comparison between the speech frame power and the noise frame power. The speech frame power and noise frame power are defined as follows.

$$P_s = \sum_{m=1}^{M} s^2(m) \tag{9}$$

$$P_n = \sum_{m=1}^{M} T^2(m) \tag{10}$$

The frame power of the speech signal changed by the band gain adjustment is defined as follows.

$$P'_s = \sum_{i=1}^{I} \sum_{m \in 1,2,\cdots,B_i} \Big(G(i) \times s(m) \Big), \tag{11}$$

where i is the critical band index.
For the frame power adjustment of the speech signal, band gains are modified using next equations.

$$G'(i) = \frac{\sqrt{P_n}}{\sqrt{P'_s}} \times G(i) \quad \text{if } P_n > P_s \tag{12}$$

$$G'(i) = \frac{\sqrt{P_s}}{\sqrt{P'_s}} \times G(i) \quad \text{if } P_n < P_s \tag{13}$$

When the speech power is greater than the noise power in the current frame, the band gain is modified to maintain the original speech power as in Eq. (12) Otherwise, masking effect occurs when the noise power is greater than the speech power. To avoid this effect, the frame power of the speech is amplified because the masking effect decreases as the speech power increases. Therefore, each band gain is adjusted as Eq. (13) to enhance the speech band by band and to avoid the masking effect.

Fig. 4. shows the result to which each algorithm is applied. (a), (b), (c), and (d) are the waveforms of the input speech, input noise, enhanced speech (our algorithm), and volume controlled speech, respectively. For an experimental comparison, in (d), the volume is adjusted so that the power of the controlled speech has the same magnitude with the power of speech enhance by our algorithm. Fig. 5. presents spectrograms of each algorithm. As can be seen in Fig. 4, the proposed algorithm functions as an emphasis on the regions of the signal that contains acoustic cues [5]. In addition, it selectively amplifies the formant of the voiced sounds, which is also shown in Fig. 5.

3. Experiment

3.1. Test material

Four speaker participated in constructing speech database for experiment. Each speaker recorded 77 pair monosyllabic CVC(consonant-vowel-consonant) words. The difference between members of a pair is a phoneme. For example, [kal] and [pal] were selected, or [bim] and [bam] were selected to construct a pair. Eight types of noise were chosen from Noisex-92 database for the experiment, which have different time-frequency distributions.
1) Buccaneer jet traveling at 190 knots; cockpit noise

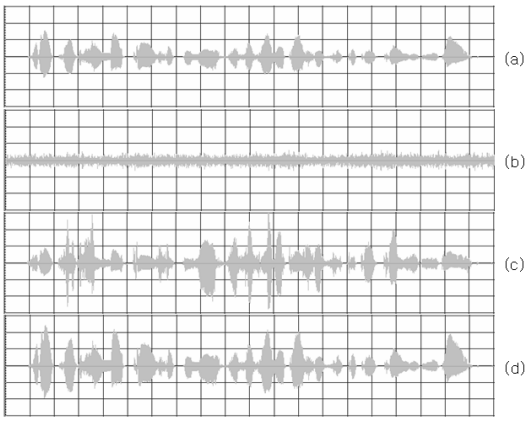

Figure 4. Waveforms of the (a) Input speech, (b) Input noise, (c) Enhanced speech(our algorithm) (d) Volume controlled speech

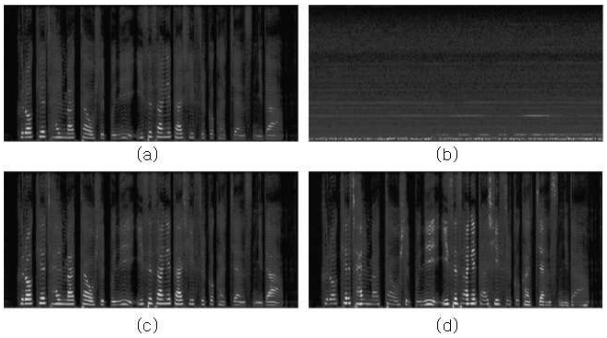

Figure 5. Spectrograms of the (a) Input speech, (b) Input noise, (c) Enhanced speech(our algorithm) (d) Volume controlled speech

2) Destroyer engine room noise
3) Destroyer operations room background noise
4) F-16 cockpit noise
5) Factory floor noise
6) Leopard : Military vehicle noise
7) Pink noise
8) White noise

3.2. Test method

In the experiment, a monosyllabic word and noise were played at the same time, and then corresponding pair words were displayed in the monitor. For examples, if [kal] had been played, then [kal] and [pal] were displayed in the monitor together. Next the tester decided which one was played and selected one of them. The played speech was randomly selected from the followings.
1) no processed speech
2) volume controlled speech
3) enhanced speech (our algorithm)

The noise was randomly selected from one of 8 types and so were the speakers. In case of 2), the volume of speech is adjusted so as to have the same power with 3).

3.3. Test result

Fig. 6 shows the entire experimental result. Mean intelligibility over the four speakers improved from 74.3% in the natural(none processing) speech to 96.5% in the speech enhanced by our algorithm. The difference is 22.2%. The intelligibility enhancement by our algorithm is 10.1% higher than the enhancement by volume control. Fig. 7 shows the result for each speaker. The difference in intelligibility between the least and most intelligible speaker was 16.4% for the natural speech but only 5.9% for the enhanced speech as a result of a much greater effect of enhancement for the originally less intelligible speaker.

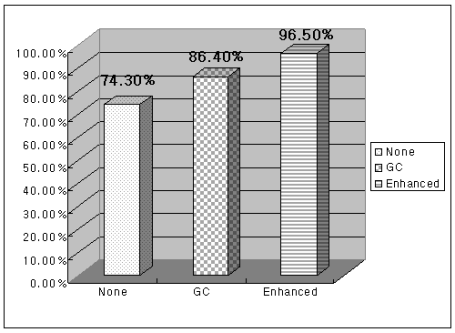

Figure 6. Intelligibility score of the overall experiments

4. Conclusions

In the cellular phone communication environment, we are exposed to various kinds of noise. In previous papers, speech was enhanced without any consideration of noise or even if it was considered, the enhancement systems have only treated the single-

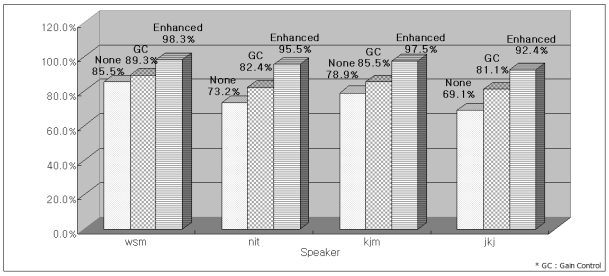

Figure 7. Intelligibility score of the each speaker

channel case. We proposed an algorithm which enhances speech based on the real-time noise analysis under dual-channel condition. The main advantages of the proposed algorithm are the followings.

1. It is computationally efficient. (The most computative part is the FFT part. The computation amount of the rest parts is trivial.)

2. Speech is enhanced by considering the significance of the each critical band. For example, the formants which include much information are particularly emphasized.

3. It is adaptive to noise. We slightly changed the speech when the background noise is negligible, and considerably changed the speech when we could not ignore the noise. In this way, the intelligibility was enhanced.

The proposed algorithm has been tested and compared to the natural speech and gain controlled speech. The subjective evaluation has been completed by five testers and the results show that there is a significant improvement in the enhanced speech than natural and gain controlled speech.

References:

[1] Nathalie Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, 7(2):126–137, March 1999.

[2] Rongqiang Hu and David V. Anderson. Improved perceptually inspired speech enhancement using a psychoacoustic model. In *IEEE Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 440–444, 2004.

[3] Akira Nakamura, Nobumasa Seiyama, Atsushi Imai, Tohru Takagi, and Eiichi Miyasaka. A new approach to compensate degeneration of speech intelligibility for elderly listeners. *IEEE Transactions on Broadcasting*, 42(3):285–293, September 1996.

[4] Russell J. Niederjohn and James H. Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):277–282, August 1976.

[5] Valerie Hazan, Andrew Simpson, and Mark Huckvale. Enhancement techniques to improve the intelligibility of consonants in noise : Speaker and listener effects. In *Proceedings of International Conference of Speech and Language Processing*, volume 5, pages 2163–2167, 1998.

[6] Thomas I. B. The second formant and speech intelligibility. In *Proc. Nut. Electronics Conference*, volume 23, pages 544–548, 1967.

[7] Russell J. Niederjohn and James H. Grotelueschen. Tthe influence of first and second formants on the intelligibility. *Journal of the Audio Engineering Society*, 16(2):182–185, April 1968.

[8] G. A. Miller. *Language and Communication*. McGraw-Hill, New York, 1963.

[9] James D. Johnston. Transform coding of audio signal using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323, February 1988.

[10] E. Zwicker and H. Fastl. *Psychoacoustics : Facts and Models*. Springer-Verlag, Germany, 1990.

[11] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *The Journal of the Acoustical Society of America*, 66:1647–1652, August 1979.