SPEECH REINFORCEMENT BASED ON PARTIAL MASKING EFFECT

Jong Won Shin, Yu Gwang Jin, Seung Seop Park and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
E-mail: {jwshin, ygjin, sspark}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

Perceived quality of the speech signal deteriorates significantly in the presence of ambient noise. In this paper, based on the analysis that the partial masking effect is a main source of the quality degradation when interfering signals are present, we propose a novel approach to enhance the perceived quality of speech signal when the ambient noise cannot be directly controlled by reinforcing it so that it can be heard more clearly. To find a suitable reinforcement rule, the loudness perception model proposed by Moore *et al.* [1] is adopted with the consideration on the prevention of the hearing damage. Experimental results show that the perceived quality and intelligibility can be enhanced under various noise environments.

Index Terms— Speech reinforcement, partial loudness, loudness perception, speech enhancement

1. INTRODUCTION

The perceived quality of the speech signal degrades with the presence of background noises. The algorithms called 'speech enhancement' are the most popular approaches to reduce the effect of the surrounding noise [2], [3]. A typical application of the speech enhancement algorithm is found in the speech communication system. In this scenario, the speech enhancement algorithm in the near-end transmitter is considered to act for reducing the near-end noise perceived by the far-end listener. However, it should be noted that the near-end noise directly arrives at the near-end listener's ears resulting in the deterioration of the quality of the far-end signal.

Instead of processing the surrounding noise, the speech reinforcement approach modifies the speech signal. When applied to the speech communication scenario, this approach reinforces the far-end speech signal to alleviate the effect of the near-end noise to the near-end listener as depicted in Fig. 1. In the figure, the background noise can be picked up by either the microphone at the transmitter or a dummy microphone.

An easy way may be the simple amplification of the overall power of the signal according to the noise level, but this method does not reflect the spectral characteristics of the noise. As an alternative idea, Tzur and Goldin [4] propose to amplify the frequency components of the signal so that the noise level in each critical band becomes lower than the masking threshold. However, this method may usually result in an excessively loud sound compared with the original one when the noise level is relatively high. A simpler approach based on the signal-to-noise ratio (SNR) is proposed independently by Sauert and Vary [5], and Goldin *et al.* [6]. This approach adjusts the frequency components so as to produce the same SNR in each band. Though simple to measure, SNR is not directly related to the perceptual loudness felt by human auditory system. All of these previous algorithms do not account for the absolute power level of the original signal. Furthermore, they do not provide an appropriate analysis on the speech quality deterioration.

This work was supported in part by the Brain Korea 21 Project and the Korea Science and Engineering Foundation (KOSEF) grant funded by Korea government (MEST) (No. R0A-2007-000-10022-0).

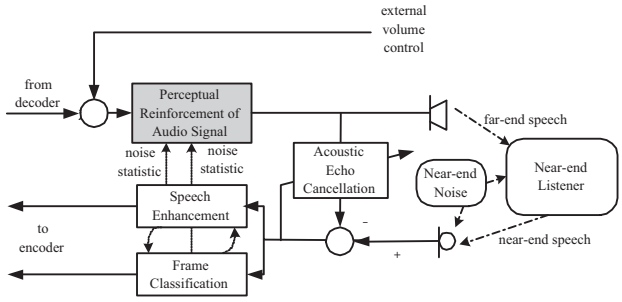

Fig. 1. Block diagram of a communication system in adverse environment.

In our previous work [7], we demonstrate the diminishment of the perceived loudness as the major cause of speech quality degradation in ambient noise condition and propose a reinforcement algorithm based on the human auditory characteristics. In this paper, we provide a full account of the cause of quality degradation, which turns out to be the partial masking effect described in terms of the loudness perception model of Moore *et al.* [1]. We modify our previous algorithm with newly obtained experimental data and practical considerations. Performance evaluation in terms of the ANSI S3.5-1997 Speech Intelligibility Index (SII) [8] and the subjective quality tests demonstrate that the proposed algorithm is effective for enhancing the intelligibility and the subjective quality of the degraded signal.

2. PARTIAL MASKING EFFECT

The properties of the human auditory system are worth investigating to achieve satisfactory performance. The masking effect is one of the most well-known properties of the human audio perception mechanism. The masking effect stands for the phenomenon that a certain weak signal called a maskee cannot be heard, i.e., 'masked' in the presence of a strong signal called a masker in a nearby time or frequency region.

If the level of one signal is much higher than that of other interfering signals called noise, the perceived loudness of the signal remains almost the same since the noise is masked by the signal. Conversely, if the level of the signal is much lower than that of the

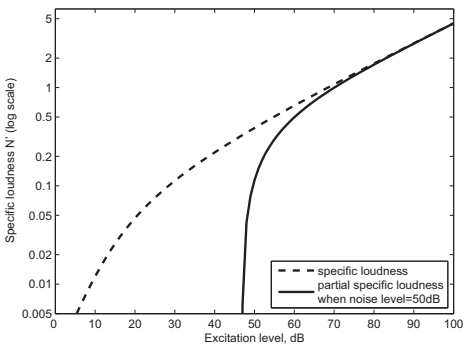

Fig. 2. Moore's models for specific loudness and the partial specific loudness of the 1 kHz tone when the noise excitation level is 50 dB as functions of the signal excitation level (Eqs. (1) and (3)).

noise, the perceived loudness of the signal decreases dramatically to zero. Now the question is how the perceived loudness of the signal becomes if the level of the signal lies in between these two extreme conditions. An intuitive anticipation will be that the loudness of the signal diminishes for a certain amount. This phenomenon is actually called the partial masking effect and the reduced loudness of a signal when other signals are present is referred to as the partial loudness [1]. This phenomenon is illustrated in Fig. 2 where the loudness of the 1 kHz tone in quiet environment and that under the presence of noise are displayed along the signal excitation level according to the Moore's loudness perception model [1]. A detailed description of the Moore's loudness perception model will be presented in the next section.

In fact, almost everybody has experienced such phenomena in daily life, for example, when one listens to music or has a conversation over the mobile phone in the presence of surrounding noises. The decrease in the perceived loudness caused by surrounding noise, known as the partial masking effect, requires audio players or cell phones to provide a very wide range of volume control. Moreover, the partial masking effect modifies the tone color of the speech signal unless the level of the signal is much higher than the noise level since the amount of partial masking in each band is not the same.

Although the partial masking effect can be felt in everyday life, the mathematical modeling of the partial loudness attracts less interest. Recently, Moore *et al.* proposed a fairly systematic mathematical model for the partial loudness [1].

In this paper, we attribute the speech quality degradation under the presence of noise to the partial masking effect and propose a speech reinforcement algorithm that restores the loudness of the degraded signal. Specifically, the proposed approach reinforces the signal under noisy environment in such a way that the partial loudness in each band is maintained to the same level as that of the original noise-free signal. To calculate the appropriate gain for each band, the mathematical models of the loudness and the partial loudness for each band proposed in [1] are employed. A detailed description of the algorithm is given in the next section.

3. SPEECH REINFORCEMENT BASED ON PARTIAL SPECIFIC LOUDNESS

To get into the details of the proposed algorithm, some psychoacoustical terms should be introduced in advance. The equivalent rectangular bandwidth (ERB) scale is a warped frequency scale which can be considered as a refinement of the well-known Bark scale [1]. The loudness is an intensive attribute of an auditory sensation, in terms of which sounds may be ordered on a scale extending from quiet to loud [10]. The partial loudness refers to the reduced loudness of a target signal when other interfering signals exist [1]. The specific loudness is the loudness per ERB. In other words, the loudness can be calculated by integrating the specific loudness over all ERBs. The partial specific loudness means the partial loudness per ERB.

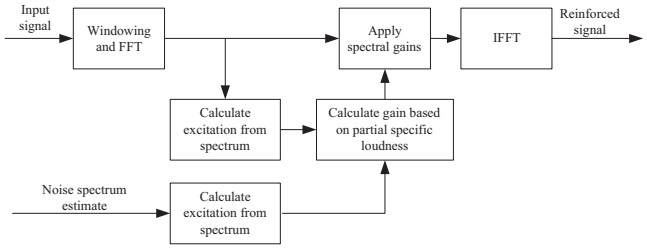

Fig. 3. Block diagram of proposed speech reinforcement system

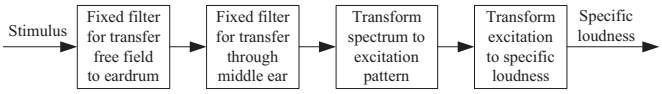

Fig. 4. Block diagram of the processing stages in the loudness perception model

The overall speech reinforcement algorithm based on partial masking effect is shown in Fig. 3 where we do not include the module for estimating the noise power spectra. They can be obtained from a conventional noise power spectra estimation algorithm. First, the excitation patterns of the signal and noise are separately derived based on the loudness perception model shown in Fig. 4. Next, an appropriate gain is computed for each band so that the signal when multiplied by the gain, will yield the partial specific loudness which becomes the same as the level of the noise-free signal. Then, the resulting gain is applied to the corresponding discrete Fourier transform (DFT) coefficients to produce the reinforced spectrum.

Moore *et al.* proposed a mathematical model for the loudness perception consisting of four parts as shown in Fig. 4 [1]. The first and second blocks are the fixed filters representing the transmission through the outer ear and the middle ear, respectively. In this paper, the revised middle ear transfer function proposed in [9] is adopted instead of the one in [1] which was used in our previous paper [7]. The third block calculates the excitation pattern and warps the frequency scale into the ERB scale. Finally in the last block, the specific loudness and the partial specific loudness is computed using the excitations for speech signal and noise.

According to [1], the specific loudness computed in quiet condition, N'_Q, can be described in terms of the excitation caused by the signal, E_{SIG}, and the threshold of hearing in quiet, E_{THRQ}, under certain mild conditions as

$$N'_Q = C[(GE_{SIG} + A)^\alpha - A^\alpha] \quad (1)$$

where C, G, A and α are experimentally determined constants. On the other hand, the partial specific loudness, $N'_{partial}$ under the pres-

ence of noise is modelled under some mild conditions as

$$\begin{aligned} N'_{partial} &= C\{[(E_{SIG}+E_{NOISE})G+A]^{\alpha}-A^{\alpha}\} \\ &- C\{[(E_{NOISE}(1+K)+E_{THRQ})G+A]^{\alpha} \\ &- (E_{THRQ}G+A)^{\alpha}\}, \end{aligned} \quad (2)$$

or more precisely,

$$\begin{aligned} N'_{partial} &= C\{[(E_{SIG}+E_{NOISE})G+A]^{\alpha}-A^{\alpha}\} \\ &- C\{[(E_{NOISE}(1+K)+E_{THRQ})G+A]^{\alpha} \\ &- (E_{THRQ}G+A)^{\alpha}\}(\frac{E_{THRN}}{E_{SIG}})^{0.3} \end{aligned} \quad (3)$$

where E_{NOISE} is the excitation caused by the noise and the masking threshold E_{THRN} is given by

$$E_{THRN} = KE_{NOISE} + E_{THRQ} \quad (4)$$

with K being an experimentally determined frequency-dependent constant [1]. The difference between (2) and (3) lies in the scaling factor $(E_{THRN}/E_{SIG})^{0.3}$, which is adopted to reflect the evidence that when $E_{SIG} \gg E_{THRN}$, the partial loudness of the signal remains the same as that of the noise-free signal [1].

Given the excitations for signal and noise, we can compute the appropriate reinforcement gain for each band. Let g denote the gain applied to a band. Then, the partial specific loudness, $N'_{partial}$ derived from (2) or (3) when gE_{SIG} is substituted for E_{SIG} should become equal to N'_Q in (1). Using (2) as the model for partial specific loudness, it turns out that the optimal gain is given by

$$g = \frac{\frac{[(GE_{SIG}+A)^{\alpha}+f(E_{NOISE})]^{\frac{1}{\alpha}}-A}{G} - E_{NOISE}}{E_{SIG}} \quad (5)$$

in which

$$\begin{aligned} f(E_{NOISE}) = [(E_{NOISE}(1+K)+E_{THRQ})G+A]^{\alpha} \\ -(E_{THRQ}G+A)^{\alpha}. \end{aligned} \quad (6)$$

As for the more precise model given by (3), it is not easy to describe the gain in a closed form. Instead, we simply shrink the gain in (5) when $E_{SIG} \gg E_{THRN}$ to approximate it. One of the possible shrinking rules can be described as follows:

$$\bar{g} = \lambda g + (1-\lambda)\times 1.0 \quad \text{if } gE_{SIG} > E_{THRN}\times 100 \quad (7)$$

where $\bar{g}$ is the modified gain and $\lambda = (E_{THRN}\times 100)/gE_{SIG}$. The shrunk gain prevents an excessive signal amplification when $E_{SIG} \gg E_{THRN}$ and thus makes the gain closer to the one provided by (3). On the other hand, to avoid the hearing damage, the gain should not be large when the total excitation for the band $gE_{SIG}+E_{NOISE}$ is huge. A possible approach can be diminishing the gain g close to one when the excitation is very high as follows:

$$\tilde{g} = \lambda\bar{g} + (1-\lambda)\times 1.0 \quad \text{if } \bar{g}E_{SIG}+E_{NOISE} > 10^{10} \quad (8)$$

where $\tilde{g}$ denotes the final gain and $\lambda = 10^{10}/(\bar{g}E_{SIG}+E_{NOISE})$. Although the specific loudness and the partial specific loudness are expressed by highly nonlinear functions of excitation, the process of transforming the power spectrum into the corresponding excitation pattern is well approximated by a linear model at a moderate signal level. Hence, the square root of the gain $\tilde{g}$ obtained for each ERB can be applied to the associated spectral components resulting in a reinforced signal spectrum.

4. EXPERIMENTAL RESULTS

It is obvious that the subjective quality test is the most appropriate way to exhibit the superiority of the algorithm since it adopts a psychoacoustic model for the loudness perception. For that reason, we performed a set of informal subjective preference tests to show that the proposed algorithm can enhance the perceived quality. In addition, we measured the ANSI S3.5-1997 Speech Intelligibility Indices (SII's) [8] of the original and reinforced signals for an objective performance evaluation. In the subjective tests, instead of simulating the real environment as illustrated in Fig. 1, we simply added the background noises to the original and reinforced speech signals before being played out at the headphone. The test material used in the subjective quality tests consisted of eight 7.5 seconds long speech files spoken by 4 male and 4 female speakers, while the test material used for measuring the SII was composed of thirty two 7.5 seconds long speech files spoken by the same 8 speakers. Each file was sampled at 8 kHz. The noises applied in the subjective tests were the speech babble, factory floor and white noises extracted from the NOISEX-92 database. Twenty one listeners (8 male and 13 female) whose ages ranged from 20 to 30 participated in the experiments.

Firstly, an informal subjective preference test was performed to compare the perceived quality of the reinforced signal with that of the unprocessed signal in the presence of background noise. This experiment was to show how efficient the proposed algorithm could be in enhancing the quality of the speech under various noise conditions. At first, the quality of the signal reinforced by the proposed algorithm where the true value of the noise power in each band was utilized (denoted as 'SRPSLt') was compared with that of the unprocessed signal in noisy condition. This experiment can provide the performance bound of the proposed reinforcement algorithm. We also implemented a reinforcement algorithm (denoted as 'SRPSLe') in which the noise power spectrum was estimated by the voice activity detection (VAD) algorithm option 2 of the ETSI Adaptive Multi-Rate codec (AMR) [11], and compared the quality with that of the unprocessed noisy signal. The preference test performed in this experiment was essentially the same as the ITU-T P.800 comparison category rating (CCR) test [12] except that the original clean speech signal was provided to the listeners as a reference. Each participant gave his/her opinion on the perceptual preference with a score from -3 to 3. All the scores from the listeners were then averaged to yield the overall test result. The results are summarized in Table 1 where a positive value indicates that the reinforced speech was preferred to the unprocessed signal. The average score was higher at lower SNR since the unprocessed speech would be severely masked by the noise. We can also see that the score was lower for the babble noise, which has a spectral tilt similar to that of the speech signal resulting in mild partial masking for every spectral component. When the noise power spectrum was estimated by a practical technique, the performance gain was slightly reduced but still meaningful. From the result, we can conclude that the proposed reinforcement algorithm enhances the perceived quality of the speech signal in noisy environments.

Secondly, the quality of the signal reinforced by the proposed algorithm was compared with that of the signal reinforced by the SNR-based algorithm [5], [6] which was found to be superior to the simple power amplification [6]. The signal reinforced by the SNR-based algorithm was produced by amplifying the spectral components so as to retain the same SNR for all bands. The target SNR value of the output signal was set to make the power of the output equal to that of the signal reinforced by the proposed method. Two different versions of the SNR-based method were also implemented. The SNR-based

Table 1. Result of subjective preference test: reinforced speech vs. unprocessed speech under noise conditions.

	SRPSLt - unprocessed			SRPSLe - unprocessed		
noise	babble	factory	white	babble	factory	white
-5 dB	1.76	2.13	2.01	1.67	2.07	2.03
0 dB	0.94	1.45	1.38	1.02	1.41	1.48
5 dB	0.38	0.72	1.13	0.35	0.70	1.19
10 dB	0.10	0.20	0.82	0.08	0.21	0.78
avg	0.80	1.13	1.33	0.78	1.10	1.37

Table 2. Result of subjective preference test: proposed reinforcement algorithm vs. SNR-based method.

	SRPSLt - SNRt			SRPSLe - SNRe		
noise	babble	factory	white	babble	factory	white
-5 dB	0.65	0.52	0.79	0.10	0.91	1.53
0 dB	0.60	0.63	0.82	0.37	0.91	1.16
5 dB	0.48	0.37	0.69	0.36	0.67	0.76
10 dB	0.36	0.21	0.29	0.10	0.20	0.32
avg	0.52	0.43	0.65	0.23	0.67	0.94

algorithm utilizing the actual noise power spectrum is denoted as 'SNRt', and the other one where the noise power spectrum is estimated by the AMR VAD option 2 is denoted as 'SNRe'. The result of these tests is shown in Table 2 where a positive number means that the signal reinforced by the proposed algorithm was preferred. From the result, it can be seen that the partial loudness-based technique outperformed the SNR-based method. It was also observed that the tone color of the speech signal was altered when we applied the SNR-based algorithm since the modified spectral shape of the speech generally tended to approach the background noise spectrum [5] and the relative perceived loudness for each band varied due to the difference in the amount of partial masking.

Finally, we compared the ANSI S3.5-1997 SIIs [8] of the unprocessed noisy signal with those of the signal reinforced by the proposed algorithm and the SNR-based method. SII provides an objective measure of speech intelligibility although it does not well reflect human auditory characteristics but focuses more on the average signal and noise power than their temporal variations. We tested all the 15 noises from the NOISEX-92 DB. The values of the SII averaged over 15 noises are given in Table 3. From the result, we can conclude that the proposed algorithm enhances the intelligibility of the speech signal. It is interesting to see that the SII for 'SNRe' is larger than that for 'SNRt' though the average power is smaller. It may be due to the fact that the noise power spectrum estimation algorithm relatively overestimates the high frequency components than the low frequency components resulting in high SNRs in the high frequency bins.

Table 3. Speech intelligibility index averaged over 15 noises.

	unprocessed	SNRt	SRPSLt	SNRe	SRPSLe
-10 dB	0.254	0.387	0.438	0.415	0.411
-5 dB	0.381	0.477	0.517	0.509	0.497
0 dB	0.513	0.573	0.599	0.601	0.582
5 dB	0.640	0.671	0.684	0.690	0.673
10 dB	0.758	0.773	0.777	0.783	0.771
avg	0.509	0.576	0.603	0.600	0.587

5. CONCLUSIONS

In this paper, we have proposed a novel approach to enhance the quality of speech signal in adverse environment when the noise cannot be directly controlled. The proposed approach reinforces speech signal under noise to have the partial specific loudness for each band almost the same as that of the original noise-free signal. The loudness perception model proposed by Moore *et al.* [1] has been adopted to calculate the specific loudness and partial specific loudness. Experimental results have shown that the reinforced speech enhances the perceived quality of the noisy speech signal.

6. REFERENCES

[1] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of Audio Engineering Society*, vol. 45, no. 4, pp. 224-240, Apr. 1997.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[3] N. S. Kim and J. -H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[4] M. Tzur (Zibulski) and A. A. Goldin, "Sound equalization in a noisy environment," *Audio Engineering Society 110th Convention*, Preprint No. 5364, May 2001.

[5] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. I-493-I-496, 2006.

[6] A. A. Goldin, A. Budkin and S. Kib, "Automatic volume and equalization control in mobile devices," *Audio Engineering Society 121th Convention*, Preprint No. 6960, Oct. 2006.

[7] J. W. Shin and N. S. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," *IEEE Signal Processing Letters*, vol. 14, issue 11, pp. 887-890, Nov. 2007.

[8] ANSI S3.5-1997 (R2002), *American National Standard Method for Calculation of the Speech Intelligibility Index*.

[9] B. R. Glasberg and B. C. J. Moore, "Prediction of absolute thresholds and equal-loudness contours using a modified loudness model," *Journal of Acoustical Society of America*, vol. 120, no. 2, pp. 585-588, Aug. 2006.

[10] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, fifth edition, London: Elsevier, 2004.

[11] 3GPP Document ETSI TS 26.094, *Voice Activity Detector for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, v. 6.1.0, Jun. 2006.

[12] ITU-T P.800, *Methods for Subjective Determination of Transmission Quality*, Aug. 1996.