Speech Reinforcement based on Partial Specific Loudness

Jong Won Shin, Woohyung Lim, Junesig Sung, and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul, Korea
{jwshin, whlim, jssung}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In the presence of background noise, the perceptual loudness of speech signal significantly decreases resulting in the deterioration of intelligibility and clarity. In this paper, we propose a novel approach to enhance the quality of speech signal when the additive noise cannot be directly controlled. Specifically, we propose an approach which reinforces the speech signal so that the partial loudness in each band can be maintained to the level almost the same to that measured without the effect of background noise. To find a suitable reinforcement rule, the loudness perception model proposed by Moore et al. [1] is adopted. Experimental results show that the loudness of the original noise-free speech can be restored by the proposed reinforcement algorithm and the proposed algorithm can enhance the perceived quality of speech signal under various noise environments.

Index Terms: speech reinforcement, partial loudness, loudness perception, speech enhancement

1. Introduction

As the background noise level increases, listening to speech or audio sounds becomes more difficult. The most popular way to resolve this problem is 'speech enhancement', which tries to remove the effect of the additive noise from the noisy signal [2]. A conventional strategy to apply speech enhancement to a speech communication system is to suppress the noise in the noisy input before being transmitted to the far-end [3], as shown in Fig. 1. From this viewpoint, speech enhancement 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 speech quality and this cannot be controlled by means of the traditional speech enhancement algorithms.

In this paper, we propose an approach to reinforce the audio signal in ambient noise environments. When applied to the aforementioned speech communication scenario, instead of processing the near-end noise, the proposed algorithm reinforces the far-end speech so that it can be heard more intelligibly and clearly by the near-end listener. The overall block diagram of the proposed technique is illustrated in Fig. 2 where the background noise can be picked up by either the microphone at the transmitter or a dummy microphone. Basically, the proposed speech reinforcement system boosts the power of the frequency components of the speech. It can be intuitively understood that a better perceptual quality may be achieved by simply amplifying the overall power of the speech, but a simple amplification cannot be an optimal solution when we take the spectral characteristics of the noise into consideration. An alternative idea is to amplify the frequency components of the signal so that the noise level in each critical band becomes lower than the masking threshold created by the signal [4]. However, this method may usually give rise to an excessively loud sound compared with the original one when the noise level is relatively high. Even though one can consider to adjust the frequency components to produce the same signal-to-noise ratio (SNR) for each band [5], [6], SNR is not directly related to the perceptual loudness felt by human auditory system. Moreover, all of these algorithms do not account for the absolute level of the original signal.

In the presence of noise, i.e., signals other than the one of our interest, the perceptual loudness of speech signal is usually diminished [1], [7]. 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. According to the masking effect, a certain weak signal (maskee) cannot be heard especially when a strong signal (masker) exists in a nearby time or frequency region. Even in the case when the maskee is not completely masked, its perceptual loudness decreases to a certain amount. The word partial loudness or partial masked loudness refers to the perceptual loudness of a maskee when a masker is present [1], [7].

Based on this phenomenon, we propose a speech reinforcement algorithm that modifies the partial loudness in each band of the speech signal such that it can be maintained to the level almost the same to that measured without the effect of background noise. This restoration of the loudness may bring on the improvement of the intelligibility and the overall speech quality [8]. The loudness perception model proposed by Moore et al. [1] is adopted in the algorithm. Subjective loudness comparison tests show that the loudness of the speech decreases in the presence of noise and can be restored by the proposed algorithm. In addition, the results of a preference test that measures the overall quality including the intelligibility, clarity, naturalness and pleasantness demonstrate that the proposed algorithm is effective for enhancing the quality of the degraded speech, and outperformed the SNR-based method [5], [6] under various noise environment.

2. Speech Reinforcement based on Partial Specific Loudness

The block diagram of the loudness perception model employed in our speech reinforcement system is given in Fig. 3 [1]. The aim of using this model is to obtain a mathematical representation of the specific loudness for noise-free signal and the partial specific loudness for noisy signal when both the speech and noise spectra are available. The first and second blocks represent the transfer functions of the sound pressure from the free field to the eardrum and through the middle ear, respectively. It

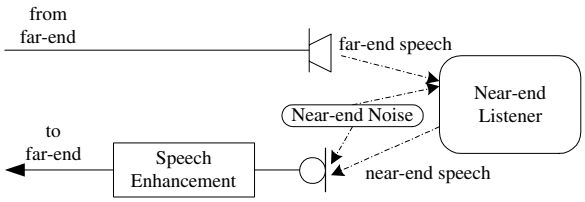

Figure 1: Block diagram of a communication system with speech enhancement.

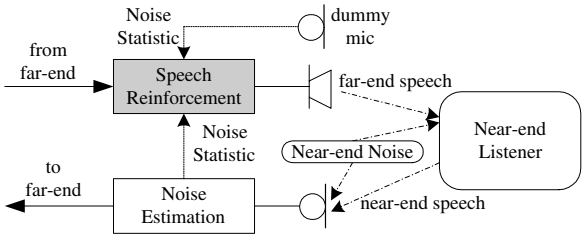

Figure 2: Block diagram of a communication system with speech reinforcement.

is noted that from the aforementioned meaning of each block, the first filter should be omitted if the loudness is to be measured for the sound heard from earphones or headphones. The third block extracts the excitation pattern from the spectrum reaching the cochlea and warps the frequency to the 'equivalent rectangular bandwidth (ERB) scale', which is a refinement of the Bark scale [7].

Given the excitations of speech and noise, we can now compute the specific loudness and partial specific loudness. The specific loudness and partial specific loudness stand for the loudness per ERB and the partial loudness per ERB, respectively. Let N'_Q denote the specific loudness computed in quiet condition when E_{SIG}, the excitation caused by the signal, is larger than or equal to the threshold of hearing in quiet, E_{THRQ}, and also E_{SIG} is less than the saturation limit of the cochlear amplifier. Then, it is given by

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

where C, G, A and α are experimental constants [1]. N'_Q is given in a different form when E_{SIG} is less than E_{THRQ} or larger than the saturation limit of the cochlear amplifier, but these cases are not of much interest since the former corresponds to inaudible sounds, and the latter occurs when the signal level is too high, which can be easily avoided by controlling the volume of the speaker. On the other hand, the partial specific loudness, $N'_{partial}$, is computed in the presence of noise when E_{SIG} is larger than or equal to the masking threshold E_{THRN} and the summation of E_{SIG} and E_{NOISE}, the excitation caused by the noise, is less than the saturation limit of the

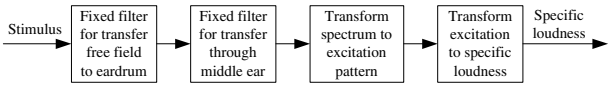

Figure 3: Block diagram of the processing stages in the loudness perception model.

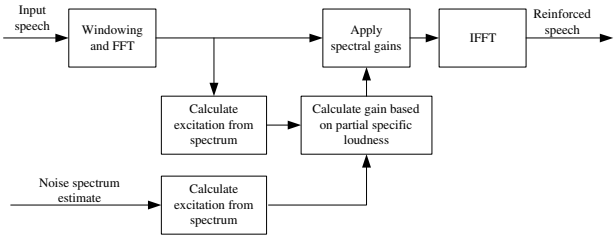

Figure 4: Block diagram of the proposed speech reinforcement system.

cochlear amplifier. In [1], it is shown that

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

or more precisely,

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

where E_{THRN} is modelled as $E_{THRN} = KE_{NOISE} + E_{THRQ}$ with K being an experimentally determined frequency-dependent constant. Again, the cases when $E_{SIG} < E_{THRN}$ or $E_{SIG} + E_{NOISE}$ is large enough to saturate the cochlear amplifier are of little interest. 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].

In the implementation of the proposed speech reinforcement algorithm, all the filters and constants are held fixed to their corresponding values as presented in [1]. Originally, this perceptual model was built to parametrically approximate the results of the measurement experiments performed with a single tone, multiple tones and a uniformly-exciting noise stimuli [7]. For that reason, there may be some modeling mismatch when a signal of broad non-uniform power spectrum is applied.

The block diagram of the proposed speech reinforcement system is given in Fig. 4. We assume that we can obtain an estimate for the noise power spectrum possibly from the near-end transmitter, from a dummy microphone or from any other devices. First, the excitation patterns of the speech and noise are separately derived based on the loudness perception model shown in Fig. 3. 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 is the same as the level of the noise-free signal. Let g denote the gain applied to a band. Then, the partial specific loudness, $N'_{partial}$ derived by (2) or (3) when gE_{SIG} is substituted for E_{SIG} should be equal to N'_Q

in (1). Using (2) as the model for partial specific loudness, we have

$$g = \frac{\frac{[(GE_{SIG}+A)^{\alpha}+f(E_{NOISE})]^{\frac{1}{\alpha}}-A}{G} - E_{NOISE}}{E_{SIG}} \quad (4)$$

where $f(E_{NOISE})$ is given by

$$f(E_{NOISE}) = [(E_{NOISE}(1+K)+E_{THRQ})G+A]^{\alpha} - (E_{THRQ}G+A)^{\alpha}. \quad (5)$$

Note that when E_{NOISE} is constant, the gain given in (4) would be higher for the bands where the speech signal level is lower resulting in the restoration of severely masked components although the reinforced signal level may not exceed the noise level. 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 (4) when $E_{SIG} \gg E_{THRN}$ to approximate it. One of the possible shrinking rules can be described as follows:

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

$$\lambda = \frac{E_{THRN} \times 100}{gE_{SIG}} \quad (6)$$

where $\tilde{g}$ is the modified gain. It prevents an excessive signal amplification when $E_{SIG} \gg E_{THRN}$ and thus makes the gain closer to the one based on (3). Since the process of transforming the power spectrum into the corresponding excitation pattern is well approximated by a linear model for a moderate signal level, the square root of the gain $\tilde{g}$ obtained for each ERB is applied to the associated spectral components resulting in a reinforced signal spectrum.

3. Experimental Results

To show the usefulness of the proposed speech reinforcement algorithm, informal subjective preference tests and loudness comparison tests were performed. Instead of simulating the real environment illustrated in Fig. 2, we simply added the background noises to the original and reinforced speech signals before being played out at the headphone. The test material consisted of eight 7.5 seconds long speech files spoken by 4 male and 4 female speakers. Each file contained two spoken sentences and was sampled at 8 kHz. The noises used in the experiment were the speech babble and white noises extracted from the NOISEX-92 database. Fifteen listeners (9 male and 6 female) whose ages ranged from 19 to 30 participated in the experiment. Six of them were students specialized in speech processing while the others were non-specialists.

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 designed to see 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 based on the near-end microphone input, and compared the quality with that of the unprocessed signal in the presence of background noise. As for the noise power spectrum estimation, we applied the one adopted in the voice activity detection (VAD) algorithm option 2 of the ETSI Adaptive Multi-Rate codec (AMR) [9]. The preference test performed in this experiment was essentially the same as the comparison category rating (CCR) test [10] 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 average test result. The results are summarized in Table 1 where a positive value means that the reinforced speech was preferred. 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.

Table 1: Subjective preference test result: The reinforced speech vs. the unprocessed speech under noise conditions.

test set	SRPSLt - unprocessed		SRPSLe - unprocessed	
noise	babble	white	babble	white
-5 dB	1.81	1.84	1.57	1.73
0 dB	1.24	1.43	0.79	1.37
5 dB	0.50	1.15	0.28	1.08
10 dB	0.04	0.96	0.02	0.88
average	0.90	1.34	0.66	1.26

Table 2: Subjective preference test result: The proposed reinforcement algorithm vs. the SNR-based method.

test set	SRPSLt - SNRt		SRPSLe - SNRe	
noise	babble	white	babble	white
-5 dB	1.02	1.08	0.69	1.07
0 dB	1.07	0.89	0.80	0.63
5 dB	0.76	0.53	0.50	0.53
10 dB	0.55	0.32	0.34	0.21
average	0.85	0.70	0.58	0.61

Next, 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 of the output signal was set to make the power of the output equal to that of the signal reinforced by the proposed method. In much the same way to our previous experiment, two different versions of the SNR-based method were implemented according to the manner of obtaining noise power spectrum. The SNR-based 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

Table 3: Perceived partial loudness comparison test result: A positive score means clean speech is perceived louder.

test set	clean - noisy			clean - SRPSLt			clean - SRPSLe		
noise	babble	white	average	babble	white	average	babble	white	average
-5 dB	1.93	2.21	2.07	-0.43	-0.93	-0.68	0.85	0.09	0.47
0 dB	1.13	1.67	1.40	0.11	-0.67	-0.28	0.63	0.18	0.40
5 dB	0.73	1.20	0.97	0.28	-0.11	0.09	0.57	0.23	0.40
10 dB	0.38	0.99	0.68	0.25	0.18	0.21	0.28	0.45	0.37
average	1.04	1.52	1.28	0.05	-0.38	-0.16	0.58	0.24	0.41

2 is denoted as 'SNRe'. The overall performance of 'SRPSLt' was compared with that of 'SNRt' to show the superiority of the proposed method. We also compared 'SRPSLe' with 'SNRe' to show the practical applicability of the reinforcement algorithm. 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 observed that the tone color of the speech signal was altered by applying the SNR-based algorithm since the modified spectral shape of the speech generally followed that of the background noise [5] and the relative perceived loudness for each band varied even if the same gain was applied to all bands due to the difference in the amount of partial masking.

Finally, to demonstrate that the loudness of speech signal decreases in the presence of noise and can be restored by the proposed reinforcement scheme, subjective listening tests were conducted with emphasis on the perceived loudness. The procedure of the test was almost the same as that of the previous preference test. But, this time, each listener gave his/her opinion on the perceived loudness, not the quality, with a score which indicates how much the former speech is felt 'louder' than the latter one. The result of test comparing the partial loudness of unprocessed and reinforced speech signals in noisy condition with clean speech signal is given in Table 3 where a positive score means that the clean speech is perceived louder. It is evident that the perceived loudness of the unprocessed noisy speech decreased as the SNR got lower. The test results confirm that the reinforcement algorithm proposed in this paper effectively restores the partial loudness of the noisy signal. In the case of employing a practical noise power spectrum estimation algorithm, the partial loudness restoration was found to be somewhat imperfect since the noise power spectrum might be slightly underestimated but was still considered to be tolerable on average.

4. Conclusions

In this paper, we have proposed a novel approach to enhance the quality of speech signal in adverse environment where the noise cannot be directly controlled but the power spectrum of it can be estimated. The proposed approach reinforces speech signal under noise to have the same partial specific loudness as the specific loudness that the speech would provide when noise is absent. The loudness perception model proposed by Moore et al. [1] has been adopted to calculate the specific loudness and partial specific loudness of the signal. Experimental results have shown that the proposed reinforcement algorithm restores the perceptual loudness of the degraded signal and enhances the quality of the signal under various noise environments.

5. Acknowledgements

This work was supported in part by ETRI SoC Industry Promotion Center and the Korea Science and Engineering Foundation (KOSEF) grant funded by Korea government (MOST) (No. R01-2007-000-10818-0).

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] 3GPP2 Document C.S0014-0 v1.0, *Enhanced Variable Rate Codec (EVRC)*, Dec. 1999.

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

[8] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *Journal of Acoustical Society of America*, vol. 94, no. 4, pp. 2050-2062, Oct. 1993.

[9] 3GPP Document ETSI TS 26.094, *Voice Activity Detector for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, v. 6.1.0, Jun. 2006.

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