NOISE ESTIMATION
AND
NOISE-SUBTRACTED SPEECH QUALITY

M. KAZAMA(1) and M. TOHYAMA(2)
(1)mich@ann.hi-ho.ne.jp, Waseda University, Tokyo, Japan
(2)m_tohyama@waseda.jp, Waseda University and University of York, UK

Abstract
Spectral envelope correlation is a good indicator of the noise dominance in each frame. A frame-wise noise spectrum can be estimated using the spectral envelope correlation between a past estimate of noise and the current frame-signal. The noise spectrum estimation method described in this article includes a frame-by-frame updating process for noise subtraction. We have evaluated the effectiveness of such noise suppression from the viewpoint of temporal signal dynamics and PESQ (following ITU Recommendation P.862) as well as opinion tests (according to a comparison category rating method specified by ITU-T P.800 Annex E). Our evaluation confirmed that the proposed method works well under a wide range of S/N conditions. Both PESQ and opinion rating scores increased for the noise-subtracted speech when the signal-to-noise ratio of the unprocessed noisy speech became higher. The benefit of noise subtraction in terms of the equivalent S/N was estimated to be 6 - 9 dB according to the opinion rating scores and 3 - 4 dB according to PESQ. The equivalent S/N appears to be a good candidate for use as an appropriate measure of signal enhancement.

Keywords
Noise reduction, Speech enhancement, PESQ

1 Introduction

Speech signal enhancement is an important issue in audio telecommunications because both direct and indirect sounds are required for immersive audio communication [1]. Noisy speech is generally not helpful for audio communication, so there is already a long history of research into noise reduction using a single microphone. This article describes a method for noise subtraction based on frame-wise noise estimation. The noise suppression benefit of this method is evaluated according to the signal level distribution and perceptual evaluation of speech quality (PESQ) following ITU Recommendation P.862 [2] as well as through opinion tests conducted according to the comparison category rating (CCR) method specified by ITU-T P.800 Annex E [3].

Some methods that have been investigated extensively are spectrum subtraction [4], estimation of the short-time spectral amplitude (STSA) [5], and a subspace approach that includes spectral subtraction as a special case [6]. These methods are based on the assumption that a speech signal is independent of other noise, but frame-dependent speech or noise-level estimation has also been studied with the goal of reducing the processing noise referred to as musical noise [4].

The noise level is assumed to be constant in spectral subtraction. Since the assumption that there is no correlation between noise and speech signals does not always hold when the frame length is short, an over-subtraction factor that takes into account the frame variances of the noise characteristics has been proposed [4].

Kazama et al. [7] demonstrated that when the analysis and synthesis frame is short (8 - 256 ms) an intelligible utterance can be reconstructed using the speech magnitude with a random phase instead of the speech phase [8]. This can be an ideal case for spectral subtraction when perfect noise estimation and subtraction is performed [9]. This ideal case suggests that the processing noise referred to as musical noise might not be produced when the noise spectrum is accurately estimated.

In this paper we present a procedure for frame-by-frame noise-spectrum estimation based on the dissimilarity of the magnitude-spectrum envelopes of noise and speech. The correlation coefficient between the noise and speech envelopes is used as a measure of their dissimilarity and in every frame can indicate whether noise or speech is dominant. As well as being useful for spectral subtraction, our noise estimation method will be useful for other types of noise reduction since the framework of the spectrum estimation method does not require fixed statistical or theoretical models.

Although noise that is spectrally dissimilar to speech might not severely degrade speech intelligibility in practical situations (except extremely bad cases), it is desirable to reduce the noise level if this can be done without distorting reconstructed speech signals. We therefore evaluated the benefits of noise subtraction through the temporal energy distributions and PESQ [2], as well as through CCR opinion tests [3]. Noise samples were taken in a teleconferencing room.

2 Noise Estimation using Spectral Envelope Correlation

Noise spectrum estimation is a key issue concerning noise reduction, as pointed out above. Here we propose a frame-wise noise estimation method that uses a spectral envelope instead of the spectral fine structure. Suppose we have at the $(l-1)$th frame a noise spectrum estimate $N(k,l-1)$ and have at the lth frame a magnitude spectrum observation $X(k,l)$. When speech is absent (noise is dominant) in the lth frame, we update the noise spectrum estimate for the lth frame by using the rule

$$N(k,l) \equiv aN(k,l-1) + bX(k,l).$$

Otherwise, when speech is dominant, we set

$$N(k,l) \equiv N(k,l-1),$$

where $a+b \equiv 1$ and both a and b are updating parameters (or functions) defined later. In this updating procedure we need a classifier of noise and speech in every frame.

We introduce a probability $P(N)$ that represents the noise dominance and combine the above equations as

$$\begin{aligned} N(k,l) &\equiv P(N)\left[aN(k,l-1) + bX(k,l)\right] \\ &+ (1-P(N))\,N(k,l-1). \end{aligned}$$

Here we express the probability using the magnitude-envelope correlation between the noise estimate $N(k,l-1)$ and the lth frame spectrum $X(k,l)$.

Suppose that the lth frame spectrum-envelope $X_E(k,l)$ is composed of

$$X_E(k,l) \equiv N_E(k,l) + S_E(k,l),$$

where $N_E(k,l)$ and $S_E(k,l)$ respectively denote the noise and speech envelope in the frame. If we assume that

$$N_E(k,l-1) \cong N_E(k,l),$$

where $N_E(k,l-1)$ is the envelope of the noise estimate in the $(l-1)$th frame, then the cross-correlation coefficients $\rho(l)$ between $X_E(k,l)$ and $N_E(k,l-1)$ can be written as

$$\rho^2(l) \cong \frac{\overline{N_E^2(k,l)}}{\overline{X_E^2(k,l)}} \equiv P(N)$$

where $\overline{*}$ are the frequency averages and the noise and speech envelopes are assumed not to be correlated with each other.

Consequently we set

$$\begin{aligned} N(k,l) &\cong \left[1 - b\rho^2(l)\right] N(k,l-1) + b\rho^2(l)X(k,l) \\ &\equiv \left[1 - \alpha\rho^q(l)\right] N(k,l-1) + \alpha\rho^q(l)X(k,l), \end{aligned}$$

where q can be defined so that

$$\rho^q(l) \to 1/2 \quad \text{when} \quad \rho(l) \to 1.$$

Thus, we can update the noise estimates according to the temporal characteristics of noise and experimentally determine the parameter α which depends on the frame length conditions.

3 Noise Subtraction Using the Proposed Method

In this numerical study we used speech signals and noise from a projector fan recorded in a teleconferencing room, but synthesized the noisy speech samples through superposition. The noise estimation and subtraction procedure is summarized schematically in Fig. 1. The process proceeds frame-by-frame with a short frame-hop size because we need to track the temporal properties of noise signals. Spectrum analysis is done using a conventional short-time Fourier transform (STFT).

According to the noise estimation process, the noise subtraction performed in the lth frame is

$$\hat{S}(k,l) \equiv X(k,l) - N(k,l)$$

when the estimated speech magnitude spectrum $\hat{S}(k,l)$ is non-negative; otherwise we set

$$\hat{S}(k,l) \equiv 0.$$

The noise-suppressed speech signal can be synthesized through the inverse STFT of the estimated speech magnitude with the observed phase of the noisy signal in every frame.

The signals were sampled at 16 kHz, and every 16 ms we took a 32-ms frame by using a rectangular window after zero padding so that the total frame-length was 256 ms. We could get sinusoidal components of a signal without windowing artifacts from the STFT spectrum through a rectangular window if this was necessary [10]. The frame length stated above might be the upper limit at which an intelligible speech signal can be synthesized with random phase [7][8]. A triangular window was used to synthesize the signal according to an overlapping add method after subtraction.

A longer frame length (256 ms without zero-padding) was taken to obtain noise estimates with sufficient frequency resolution. Noise subtraction based on these two different frame lengths can be done using a moving-average of the past noise-spectrum estimates. This moving-average-based noise subtraction generally could be done on an over-subtraction basis, and we also discarded the noise levels below the average so that the moving average traced the minimum levels.

This could cause aggressive noise subtraction. Therefore, we will normally need post-processing in the frequency domain just before the inverse STFT for speech reconstruction so that severe spectral distortion will not be perceived. The details regarding this processing are still under investigation.

Spectral envelopes were obtained by smoothing the magnitude spectrum every frame before taking the moving average. If we describe this smoothing process of the sequence in the frequency domain in terms of time-sequence processing sampled every 1/16 ms, it corresponds to low-pass filtering with a cutoff frequency of 1,000 Hz.

Figure 2 shows a noise-subtraction example. Panels (a) and (b) illustrate samples of clean (a) and noisy (b) speech waveforms, where the average segmental signal-to-noise ratio (S/N) was about 0 dB. The samples of speech and noise from the projector-fan were separately recorded in a teleconferencing room so that we could control S/N in the noisy speech signals. Panel (c) shows the spectral envelope correlation on a frame-by-frame basis. We can see that it was low in the frames where speech signals were likely to be dominant. This indicates that the spectral envelope correlation should be a good estimator for speech dominance even in short frames. Panel (d) is a resultant waveform obtained through the noise-subtraction process. We analyze the effect of noise reduction on the enhancement of noisy speech in the following sections.

4 Noise-Reduction Effect

The effect of spectrum subtraction on noise reduction can be determined through signal dynamics in the temporal domain. In general, we would expect a signal level increase as the noise level rises according to the hypothesis that noise and speech are independent of each other. Figure 3 compares clean and noisy speech frame energy (with and without subtraction) for every 32 ms of frames. We can see that the frame energy for the noisy speech was above that for the clean speech, while the frame energy for the reconstructed speech was lower than that for the clean speech. This was due to the over-subtraction. If we consider the noise reduction benefit as including the over-subtraction effect, the noise levels were reduced by about 15 dB.

5 Noise-Subtracted Speech Quality Evaluated by PESQ

The noise-subtracted samples, such as those shown in Fig. 2, were intelligible as was the original noisy speech. We evaluated speech quality by PESQ following ITU Recommendation P.862. Figure 4 shows the evaluation results, which indicate an improved PESQ after noise subtraction. Here, PESQ was averaged over 10 sentences, where five sentences were spoken by a male speaker and the other five were spoken by a female speaker, for each of the S/N conditions. Figure 4 shows PESQ for both noisy and noise-subtracted samples as the averaged segmental S/N changed. We found that PESQ increased almost linearly for the noisy samples used in this experiment as the S/N increased. Interestingly, PESQ for the noise-subtracted speech also rose as the S/N increased. The benefit of noise subtraction in terms of PESQ was about 0.4, and this value was almost independent of S/N for the original samples (except for extreme cases). This increase in PESQ corresponded to an S/N improvement of 3 - 4 dB. Thus, our subtraction scheme appears to work well when the S/N is between -6 and +18 dB.

6 Quality Evaluation through an Opinion Test

We also evaluated the noise-subtraction effect on speech quality through an opinion test based on the CCR method specified by ITU-T P.800 Annex E [3]. Seven male subjects and one female subject participated in the opinion test. The subjects were asked to evaluate sample pairs by giving rating scores on a seven-point scale according to the guidelines of the CCR method [3]. The subjects were PhD students and researchers (including the authors) who had previously participated in various types of listening test. Each subject evaluated 100 pairs of samples by listening through a headphone under a diotic condition. The samples consisted of both unprocessed and processed noisy speech signals. The processed samples were created through the noise-subtraction procedure described above and might have included both residual noise and distortion due to the processing. Every subject was therefore instructed to give an over-all rating without attempting to decompose the factors causing sound deterioration. Each rating score was assigned to the second entry of each pair based on the CCR method.

Figure 5 shows the opinion ratings on the seven-point scale for (a) unprocessed and (b) processed sample-pairs. Figure 6 shows scores for pairs consisting of unprocessed and processed samples. There were no significant differences between Figs. 5(a) and (b), which both show that the opinion scores increased as the absolute difference in the S/N between the compared entries became larger. The results shown in Figs. 3 to 5 suggest that the equivalent S/N might be a suitable measure for evaluating the effect of noise subtraction on speech quality. In other words, the noise subtraction algorithm might be acceptable for practical use, if the processing benefit can be evaluated in the equivalent S/N sense, because the signal processing does not produce severe distortion.

Figure 7 shows the equivalent S/N estimated from the opinion scores or PESQ for processed speech signals according to the results shown in Fig. 6. The equivalent S/N is defined as the S/N for unprocessed noisy speech that gives the same opinion score or PESQ as for the corresponding processed speech. We can see the realized processing benefit in terms of the equivalent S/N shown in Fig. 7. The benefit was 6 - 9 dB according to the opinion scores and 3 - 4 dB according to PESQ. The difference in these benefit ranges was probably because all of the speech samples were basically intelligible, so the subjects participating in the opinion tests mainly made their decision according to the signal dynamics (as shown in Fig. 3) while PESQ gives scores that include waveform distortion.

7 Summary

We have proposed a method for frame-wise noise estimation and subtraction. A frame-by-frame updating equation was derived using the spectral envelope correlation between the last estimate of noise and a current frame-signal. The correlation reflects the dissimilarity of the speech and noise spectra and thus should be a good indicator of noise dominance in each frame.

A long frame-length with sufficient frequency resolution was used for noise estimation, while a short one removing the phase effect on intelligible speech representation was used for speech reconstruction. These different frame lengths enabled effective noise subtraction on an over-subtraction basis, and these conditions were integrated by taking a moving average for the noise estimates using the long frame and zero-padding for the short frame. Finally post-processing in the frequency domain was added to reduce the distortion caused by aggressive over-subtraction.

The noise reduction benefits for noise-subtracted samples, which were simulated using room noise recorded in a teleconferencing room, were evaluated through PESQ and opinion tests. The results indicate that our proposed method works well under a wide range of S/N conditions. We found that PESQ increased when the signal-to-noise ratio of the original noisy speech rose. We could evaluate the benefit of noise subtraction based on the equivalent S/N, which might be an appropriate measure of signal enhancement. The estimated benefit in terms of the equivalent S/N was 6 - 9 dB based on the opinion rating scores and 3 - 4 dB based on PESQ. We have not yet explained this difference in the estimated benefit, or how to discriminate between the deterioration factors due to remaining noise and those due to the produced distortion. We thank Professor T. Houtgast for his suggestions on the signal dynamics representation, YAMAHA Corporation for providing a facility including the teleconferencing rooms, and Professor Y. Yamasaki for his constant encouragement.

References

[1] C.Kyriakakis, et al., Surrounded by Sound, IEEE Signal Processing Magazine, 1, pp.55-66, 1999.

[2] ITU-T Rec. P.862, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, 2001 (Feb).

[3] ITU-T Rec. P.800, SERIES P: TELEPHONE TRANSMISSION QUALITY, Methods for objective and subjective assessment of quality, Methods for subjective determination of transmission quality, 1996 (Aug).

[4] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of Speech Corrupted by Acoustic Noise, ICASSP 79, pp.208-211, 1979.

[5] O. Cappe, Elimination of Musical Noise Phenomenon with Ephraim and Malah Noise Suppressor, IEEE SAP 2(2), pp.345-349, 1994.

[6] Y. Hu and P.C. Loizou, A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise, IEEE SAP 11(4), pp.334-341, 2003.

[7] M.Kazama, M.Tohyama, and T. Houtgast, Speech Reconstruction by Using only Its Magnitude Spectrum or Only Its Phase, 17th Int. Cong. Acoust. 7, pp.51, 2001.

[8] L. Liu, J. He, and G. Palm, Effects of phase on the perception of intervocalic stop consonants, Speech Communication 22, pp.403-417, 1997.

[9] Vary, P., Noise suppression by spectral magnitude estimation-mechanism and theoretical limits, Signal Processing, 8, pp.387-400, 1985.

[10] M. Kazama, K. Yoshida, and M. Tohyama, Signal Analysis by Clustered Line-Spectrum Modeling, J. Audio Eng. Soc. 51(3), pp.123-137, 2003.

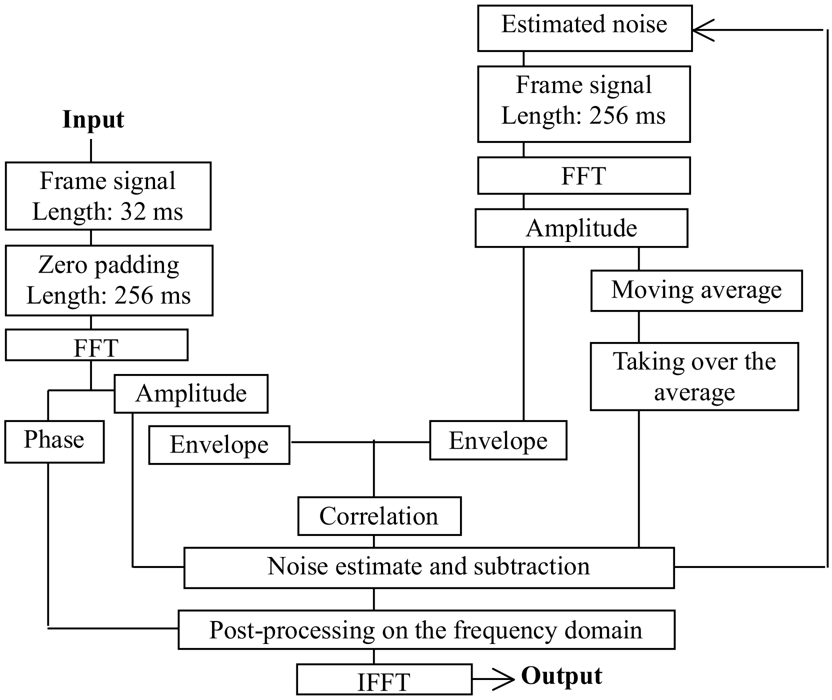

Fig. 1 Schematic procedure for noise estimation and subtraction

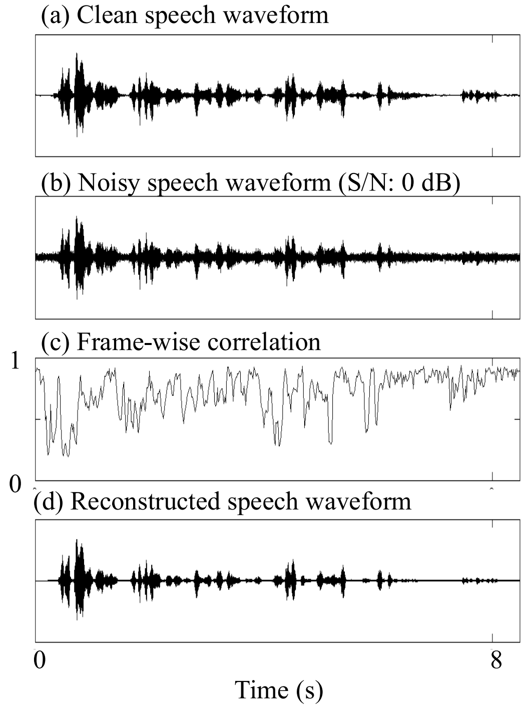

Fig. 2 Examples of signal waveforms and frame-wise correlation between spectral envelopes for present noisy speech and past-estimated noise

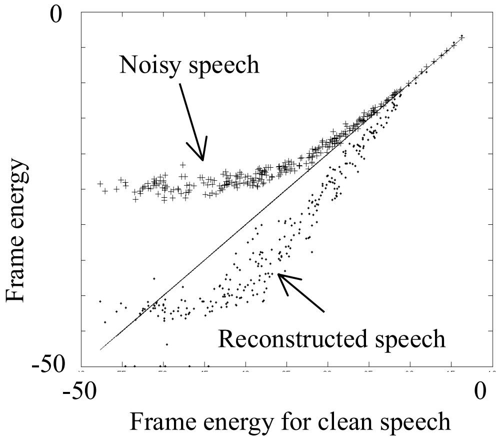

Fig. 3 Distributions of clean speech and noisy (or reconstructed) speech levels.
S/N for the noisy speech: 0 dB

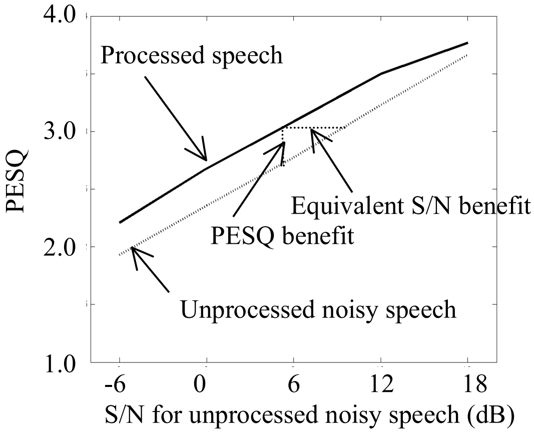

Fig. 4 Noise-reduction benefits evaluated using PESQ

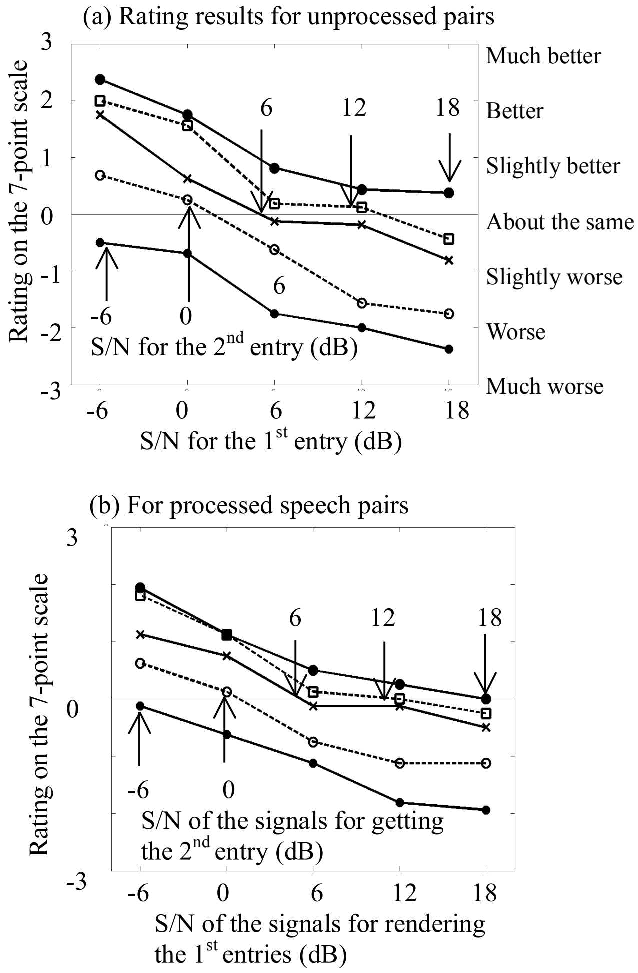

Fig. 5 Opinion scores for (a) unprocessed and (b) processed speech pairs

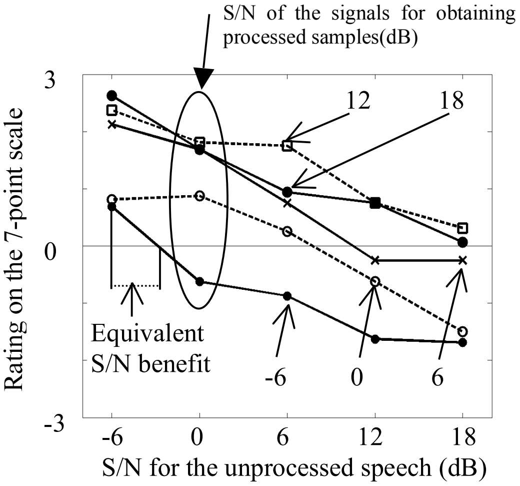

Fig. 6 Opinion scores for pairs of unprocessed (1st entry) and processed (2nd entry) speech samples

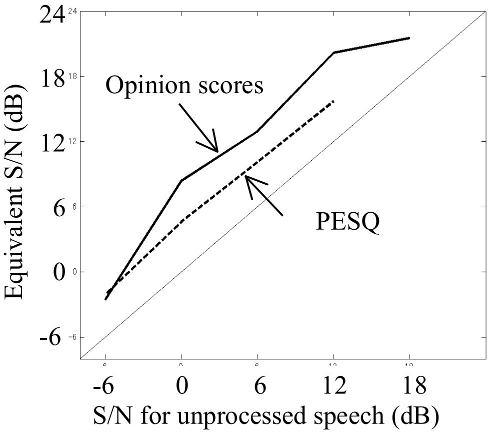

Fig. 7 Equivalent S/N estimated from opinion rating scores and PESQ for noise-subtraction processing