IMPROVED NOISE REDUCTION WITH PITCH-ENABLED VOICE ACTIVITY DETECTION

Khondaker Abdullah-Al-Mamun and Ghulam Muhammad

Department of Computer Engineering, College of Computer & Information Sciences
King Saud University, P.O. Box: 51178, Riyadh 11543, Saudi Arabia.

ABSTRACT

In this paper, we address the problems in standard noise reduction method, which was designed by ETSI (European Telecommunication Standards Institution) for distributed speech recognition. In ETSI-based procedure, noise spectrum is estimated from noise frames, which are detected by a voice activity detector (VAD). Frame energy from the input signal is calculated by the VAD, and if the frame energy is smaller than a threshold, the corresponding frame is considered as noise frame. In highly corrupted noisy signal, this leads VAD towards false detection of a noise frame as a speech frame. Again in the second stage of ETSI-based procedure, the gain factorization coefficient is set to 0.8 for noise frames. This causes less noise reduction for high noisy signal. In our proposed improvement, pitch information along with frame energy is introduced to detect speech and noise frames, and gain factor is increased for better noise reduction from noise frames. Experimental results on Aurora-2J database show significant achievement in noise reduction using the proposed improvement over the original ETSI-based noise reduction.

1. INTRODUCTION

In many speech processing applications, like audio conferencing and hands-free mobile telephony, the recorded and transmitted speech signals contain a considerable amount of acoustic background noise. This noise comes from various sources depending on the environment. Background noise can stem from stationary noises but most of the time the background noise is non-stationary. Noise causes a signal degradation which can lead to total unintelligibility and which decreases the preference of speech coding, speech synthesis and speech recognition process.

The problem of enhancing speech degraded by additive noise has been widely studied in the past and is still an active field of research. Many noise reduction algorithms have been developed over the last two decades. A subband noise-reduction method for enhancing speech in telephony and teleconferencing is proposed in [1], and a two-step noise reduction technique addressing the problem of single microphone speech enhancement in noisy environments is presented in [2]. Background noise reduction via dual-channel scheme for speech recognition in vehicular environment is discovered by Ahn and Ko [3]. Jitsuhiro et al [4] presents noise suppression using search strategy with multi-model compositions, while noise suppression with high speech quality based on Kalman filter is introduced by Tanabe et al [5]. On the other hand, voice activity detection with noise reduction and long-term spectral divergence estimation is presented by Ramirez et al [6]. Again, Torre et al [7] introduces noise robust model-based voice activity detection, while PS-ZCPA based feature extraction with auditory masking, modulation enhancement and noise reduction for robust automatic speech recognition are discovered by Ghulam et al [8]. Other speech enhancement algorithms include Wiener and power-subtraction methods [9], maximum likelihood (ML) [10] and minimum mean squared error (MMSE) [11, 12], etc. However, improvements of noise reduction in all kinds of noise and at low signal-to-noise ratio (SNR) are still sought.

European Telecommunications Standards Institute (ETSI) developed a standard for speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm, and compression algorithms [13]. In this standard they use Wiener filter in the noise reduction block and it is performed in two stages. The detail of noise reduction algorithm is described in [13], and the corresponding process flow is shown in Figure 1. The input signal is de-noised in the first stage and the output enters into the second stage, where an additional dynamic noise reduction is performed. In the first stage, after framing an input signal, linear spectrum of each frame is estimated. The signal spectrum is then smoothed along time index in Power Spectrum Density (PSD) Mean block. After that, frequency domain Wiener filter coefficients are calculated by using both current frame spectrum estimation and noise spectrum estimation. Noise spectrum is estimated from noise frames, which is detected by a voice activity detector (VAD). VAD is

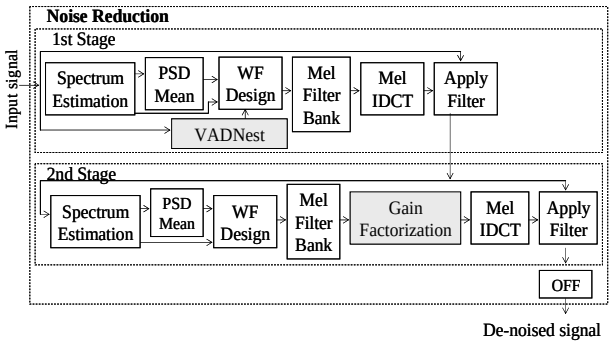

Figure 1. Process flow of noise reduction used in ETSI [13].

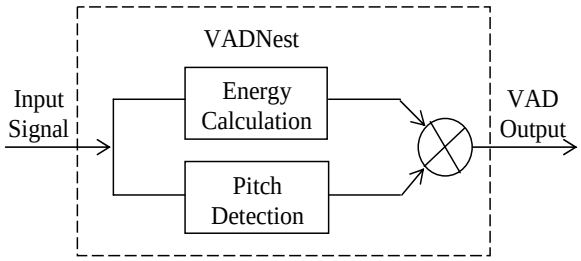

Figure 2. The proposed improved VAD.

calculated based on signal energy which consists of noise signal energy and speech signal energy. Linear Wiener filter coefficients are smoothed along frequency axis by using a Mel-filter bank, resulting in a Mel-warped frequency domain Wiener filter. The impulse response is obtained by applying Mel-warped inverse DCT. Finally, input signal of each stage is filtered in the Apply Filter. At the end of Noise Reduction, DC offset of noise-reduced signal is removed in the OFF block. Additionally, in the second stage, the aggression of noise reduction is controlled by gain factorization. In our proposed improvement we mainly concentrate in VADNest and Gain Factorization block of the above standard.

The problem of the VAD used by ETSI is that it uses only frame energy from the input signal, and if the frame energy is smaller than a threshold, the corresponding frame is considered as noise frame. This leads VAD in highly corrupted noisy signal towards false detection of noise frame as a speech frame. Again in the second stage of Wiener filter, the gain factorization coefficient is set to 80% for low SNR frames. This causes less noise reduction for high noisy signal.

In this paper, we introduce some improvements in ETSI-based noise reduction procedure. In our proposed improvements, pitch information along with frame energy is introduced to detect speech and noise frames, and gain factor is increased for better noise reduction from noise frames (low SNR frames).

The organization of the paper is as follows. Section 2 describes the proposed improvement in ETSI-based noise reduction method. Section 3 gives the experimental results and discussion. Finally, Section 4 draws some conclusions.

2. PROPOSED IMPROVEMENT

In the proposed improvement, we focus on how to enhance the performance of ETSI-based noise reduction method. At high SNR speech signal ETSI perform well, but at low SNR the performance is not optimum. In ETSI, noise spectrum is estimated from noise frames, which is detected by a VAD. The VAD is calculated based on frame energy, which consists of speech signal energy and noise signal energy of an input signal. Later Wiener filter works depending on VAD. However, frame energy cannot provide accurate differentiation between noise and speech especially in low SNR condition. In this condition VAD leads towards false detection of a noise frame as a speech frame. To overcome this situation, we propose an improved VAD based on pitch information. An existence of pitch corresponds to voiced speech frame, while an absence of pitch refers to unvoiced speech / silent frame. Pitch depends on periodicity of a signal, not on signal energy. Therefore, pitch is independent of intensity of noise in a signal. We apply pitch detection algorithm as well as energy for calculating VAD at the first stage of the ETSI noise reduction process.

Figure 2 shows the proposed pitch enabled VAD for ETSI. In energy calculation block, energy is calculated as same as ETSI. In pitch detection block, pitch is estimated for each analysis segment by using the RAPT algorithm [15] which uses normalized cross correlation function (NCCF) and dynamic programming. The NCCF defined as follows:

Given a frame of speech sampled, $s(n)$ $0 \le n \le N-1$

$$NCCF(k) = \frac{\sum_{n=0}^{N-k} s(n)s(n+k)}{\sqrt{e_0 e_k}}$$

$$\text{Where } e_k = \sum_{n=k}^{n=k+N-K} s^2(n), \qquad 0 \le k \le K-1$$

Again in the second stage of Wiener filter in ETSI, the gain factorization coefficients are set to 10% for noisy speech fames and 80% for noise frames. This causes less noise reduction for low SNR signal. In the proposed improvement, we increase the upper limit of gain factorization coefficient and vary it from 85% to 99% and remain the lower limit same as 10%. This provides better noise reduction from noise frames.

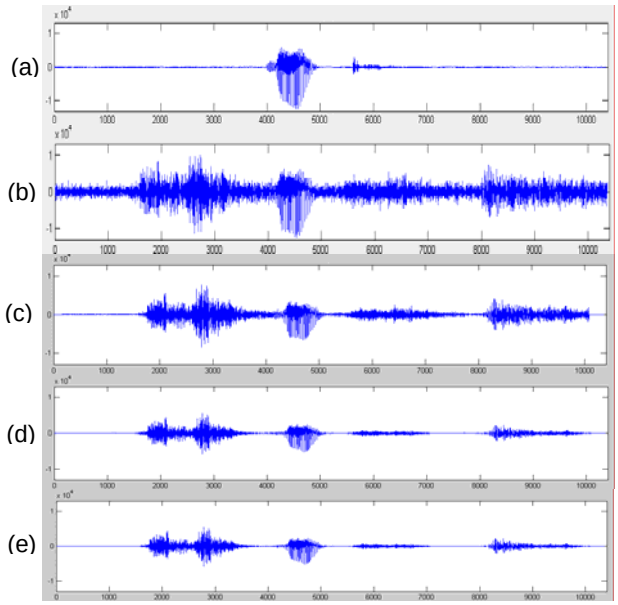

Figure 3. Speech enhancement results: utterance is [roku] (six) for a) clean signal, b) noisy version (SNR=0db), c) noise reduction using ETSI, d) noise reduction using the proposed improvement with pitch enabled VAD, and e) that using the proposed improvement with pitch enabled VAD and increased gain factor.

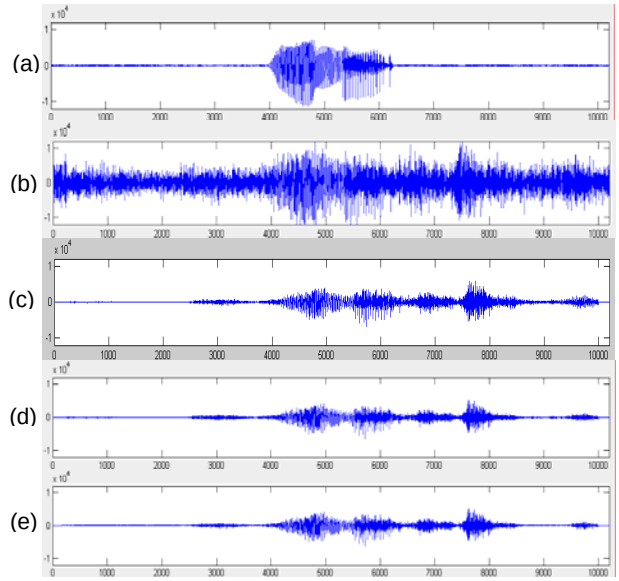

Figure 4. Noise reduction results: utterance is [nana] (seven) for a) clean signal, b) noisy version (SNR=0db), c) noise reduction using ETSI, d) noise reduction using the proposed improvement with pitch enabled VAD, and e) that using the proposed improvement with pitch enabled VAD and increased gain factor.

Table 1. Comparative SNR improvements between the procedures.

Input Noisy Speech: SNR (dB)	Original ETSI output: SNR (dB)	Improved ETSI with Pitch Enabled VAD: SNR (dB)	Improved ETSI with Pitch Enabled VAD and increased Gain Factorization: SNR (dB)
20	23.4	23.8	23.8
15	23.7	24.5	24.6
10	21.7	23.3	23.3
5	20.3	22.3	22.4
0	17.0	19.0	19.2

3. EXPERIMENTS

3.1. Database

Aurora-2J database [14] is used in the experiments. The utterances are connected Japanese digit strings and sampling rate is 8 kHz. Selections of 8 different real-world noises have been added to the speech over a range of signal-to-noise ratios (SNRs: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, clean). In the Aurora-2J digit recognition task, the evaluation focuses on robustness against additive noise and distortion by an unknown transmission channel. Eight different real noises are divided into two groups for testing. Data in Test A are added to by noises of Subway, Babble, Car, and Exhibition. Data in Test B are added to by noises of Restaurant, Street, Airport, and Station. In Test C, besides the additive noise, channel distortion is also included. In our experiments, we use 20 different utterances with different SNRs (from 0 dB to 20 dB) from Data in Test A and Test B.

3.2. Experimental setup

The parameters for the experiments are same as with ETSI [13] except for gain factor. Gain factor is varied from 0.8 to 0.99 with a step size of 0.05. The optimum result was found using 0.99 though some noisy signals showed better performances at 0.90. In the experimental result, we fix it for 0.99.

The KTH's WaveSurfer implementation of a noise robust algorithm for pitch tracking [15] is used in the proposed improvement. This algorithm is based on normalized cross-correlation and dynamic programming.

The proposed improvement (pitch enabled VAD, and increased gain factorization) is compared with the original ETSI-based noise reduction procedure. We use segmental SNR as a measure for comparison.

3.3. Experimental results and discussion

Table 1 shows comparative evaluation between the procedures. The proposed improvement is evaluated from two points: (i) using pitch enabled VAD, and (ii) using pitch enabled VAD and increased gain factor for low

SNR. Table 1 justifies the use of pitch enabled VAD into ETSI to increase the performance in noise reduction. For example, the original ETSI gained 17.0 dB for SNR 0 dB input signal while the improved pitch enabled VAD increased it to 19.0 dB. The improved pitch enabled VAD with increased gain factorization achieved further 19.2 dB SNR for the same input signal. It is also shown in the Table 1 that the improved approaches performed better for all other SNR input signals and the pitch enabled VAD had more positive impact than the increased gain factorization.

Figure 3 and 4 graphically represent the achievement of the proposed improvement on ETSI for two utterances, [roku] (six) and [nana] (seven), respectively. Top panels (a) of these figures show clean speech, (b) show 0 dB signals with subway (Fig. 3) and exhibition (Fig. 4) noises.

Figure 3(c) and 4(c) show the result of ETSI based noise reduction method, while Figure 3(d), 3(e) and 4(d), 4(e) show the result of the proposed improvements using (i) ETSI with pitch enabled VAD and (ii) ETSI with pitch enabled VAD and increased gain factorization coefficient, respectively. Using pitch enabled VAD, more noise is reduced comparing to the original ETSI based noise reduction method. Again, applying increased gain factorization coefficient at the second stage in ETSI results in further improvement.

4. CONCLUSION

We have proposed an improvement of ETSI standard noise reduction method. The performance is increased by using a pitch enabled VAD in VADNest block and increased gain factor for low SNR frames in Gain Factorization block. In future, we try to design the whole noise reduction procedure in one stage instead of two stages used by ETSI to reduce the time complexity.

5. ACKNOWLEDGEMENT

This work has been supported by Prince Sultan Bin Abdulaziz International Program for Distinguished Research Scholarships, King Saud University.

6. REFERENCES

[1] E. J. Diethorn, “A subband noise-reduction method for enhancing speech in telephony and teleconferencing,” IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, USA, 19-22 Oct. 1997.

[2] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, “A two-step noise reduction technique,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 04), Canada, 17-21 May 2004, vol. 1, pp. 289-292.

[3] Ahn Sungjoo, and Ko Hanseok, “Background noise reduction via dual-channel scheme for speech recognition in vehicular environment,” IEEE Transactions on Consumer Electronics, vol. 51, no. 1, pp. 22 – 27, Feb. 2005.

[4] Takatoshi Jitsuhiro, Tomoji Toriyama, and Kiyoshi Kogure, “Noise Suppression Using Search Strategy with Multi-Model Compositions,” in Proc. INTERSPEECH2007, Belgium, 26-31 Aug. 2007. pp. 1078-1081.

[5] N. Tanabe, T. Furukawa, H. Matsue, and S. Tsujii, “Noise Suppression with High Speech Quality Based on Kalman Filter,” in Proc. International Symposium on Intelligent Signal Processing and Communications (ISPACS '06), Japan, 12-15 Dec. 2006, pp. 315-318.

[6] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. J. Rubio, “Voice activity detection with noise reduction and long-term spectral divergence estimation,” in Proc. IEEE ICASSP 04, Canada, 17-21, May 2004, vol. 2, pp. 1093-1096.

[7] A. de la Torre, J. Ramirez, C. Benitez, J. C. Segura, L. Garcia, and A. J. Rubio, “Noise robust model-based Voice Activity Detection,” in Proc. INTERSPEECH2006, USA, 17-21 Sep. 2006, pp. 1954-1957.

[8] Ghulam Muhammad, Takashi Fukuda, Kouichi Katsurada, Junsei Horikawa, Tsuneo Nitta, “PS-ZCPA based feature extraction with auditory masking, modulation enhancement and noise reduction for robust ASR,” IEICE transactions on information and systems, vol. E89-D, no. 3, pp. 1015-1023, Nov. 2006.

[9] J. S. Lim and A. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” in Proc. of the IEEE, vol. 67, no. 2, pp. 1586–1604, Dec. 1979.

[10] R. McAulay and M. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

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

[12] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[13] ETSI ES 202 050 V1.1.5, “Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms,” 2007.

[14] S. Nakamura, K. Yamamoto, K. Takeda, S. Kuroiwa, N. Kitaoka, T. Yamada, M. Mizumachi, T. Nishiura, M. Fujimoto, A. Sasou and T. Endo, "Data Collection and Evaluation of AURORA-2 Japanese Corpus," IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 619-623, 2003.

[15] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *Speech Coding and Synthesis*, Elsevier Science, pp. 495-518, 1995.