A Soft Decision-based Speech Enhancement using Acoustic Noise Classification

Jae-Hun Choi, Sang-Kyun Kim, Joon-Hyuk Chang

School of Electronic Engineering
Hanyang University, Seoul, Korea
greatestcjh@naver.com, greenwhity@nate.com, jchang@hanyang.ac.kr

Abstract

In this letter, we present a speech enhancement technique based on the ambient noise classification incorporating the Gaussian mixture model (GMM). The principal parameters of the statistical model-based speech enhancement algorithm such as the weighting parameter in the decision-directed (DD) method and the long-term smoothing parameter of the noise estimation, are chosen as different values according to the classified contexts to ensure best performance for each noise. For the real-time environment awareness, the noise classification is performed on a frame-by-frame basis using the GMM with the soft decision framework. The speech absence probability (SAP) is used in detecting the speech absence periods and updating the likelihood of the GMM.

Index Terms: Speech Enhancement, Noise Classification, Soft Decision, Gaussian Mixture Model

1. Introduction

Speech enhancement is crucial in various communication systems where ambient acoustic noise exists. Spectral subtraction is known to be simple but effective in suppressing stationary background noise. But, since it may introduce musical noise, a number of improvements have been proposed. In particular, the estimation of the uncorrupted signal can be done more accurately using the minimum mean-square error (MMSE) proposed by Ephraim and Malah [1]. This technique is known to be free of the musical noise artifact even if the noise is poorly stationary [2]. This is due to the major factor which is forced to be the non-linear smoothing procedure in the "decision-directed" (DD) approach used to obtain a more consistent estimate the signal-to-noise ratio (SNR). Actually, the *a priori* SNR by the DD rule takes into account the current short-time frame, with a fixed weight $(1-\alpha)$, and the processing in the previous frame, with weight α [1], [2].

On the other hand, the speech enhancement technique should consider the accurate noise power estimation in adverse environments involving various non-stationary noise [1], [3]. An early approach is to average the noisy signal over non-speech section using a first-order recursive scheme. In the soft decision (SD) technique, the well-known noise power estimation algorithm, the long-term smoothed power spectrum of the background noise depending on the probability of speech absence is adopted [3]. The speech absence probability (SAP) is derived from a likelihood ratio test (LRT) by using the DD method for estimation of the unknown parameters. Note that the long-term smoothing parameter in the SD technique is assumed to be a fixed value regardless of noise types. We note that fixed valued parameters can not be always optimal under each noise type since the designer should choose an operating point yielding a reasonable performance across the various noise environments. Actually, Krishnamurthy and Hansen proposed the environmental sniffing framework to provide an accurate estimate of the noise update for a given environment [4]. Also, a voice activity detector (VAD) employing the support vector machine (SVM)-based noise classifier is proposed by Sangwan et al. [5] to set the best operating point in tuning parameters of the VAD.

In this letter, we propose a novel statistical model-based speech enhancement technique using acoustic noise classification. The first step is to find the optimized values of the principal parameters such as the weighting parameter in the DD method and the long-term smoothing parameter in the noise power estimation for each type in the various noises. This is due to the fact that choosing a fixed value for the parameter is clearly suboptimal even though this gives us a reasonably fair performance across a wide variety of noises. As a second step, environmental noise classification is performed on a frame-by-frame basis using a Gaussian mixture model (GMM) [7]. The likelihoods of the GMM for each noise are obtained and updated for the long-term smoothing during speech absence periods only, which can be achieved by the speech absence probability (SAP) within a unified framework.

2. Review of Soft Decision Based-Speech Enhancement

Let $x(n)$ and $d(n)$ denote clean speech and uncorrelated additive noise signals, respectively. The observed noisy speech signal $y(n)$ is the sum of a clean speech signal $x(n)$ and noise $d(n)$, where n is a discrete-time index. By taking a discrete Fourier transform (DFT), we then have

$$Y_k(t) = X_k(t) + D_k(t) \tag{1}$$

where $k(=1,2,\cdots,K)$ is the frequency bin and t is the frame index, respectively. Given two hypotheses, H_0 and H_1 which indicate speech absence and presence, respectively, it is assumed that

$$\begin{aligned} H_0 &: \text{speech absent} : Y_k(t) = D_k(t) \\ H_1 &: \text{speech present} : Y_k(t) = X_k(t) + D_k(t). \end{aligned} \tag{2}$$

Assuming that the clean speech $X_k(t)$ and the additive noise $D_k(t)$ are statistically independent and noisy spectral components are characterized by zero-mean complex Gaussian distributions, the probability density functions (PDF's) conditioned

on two hypotheses of H_0 and H_1 are given by

$$p(Y_k(t)|H_0) = \frac{1}{\pi\lambda_{d,k}(t)} \exp\left\{-\frac{|Y_k(t)|^2}{\lambda_{d,k}(t)}\right\} \quad (3)$$

$$p(Y_k(t)|H_1) = \frac{1}{\pi(\lambda_{x,k}(t)+\lambda_{d,k}(t))} \exp\left\{-\frac{|Y_k(t)|^2}{\lambda_{x,k}(t)+\lambda_{d,k}(t)}\right\} \quad (4)$$

where $\lambda_{x,k}(t)$ and $\lambda_{d,k}(t)$ denote the variances of the clean speech and noise for the kth spectral component at the tth frame, respectively [3].

For soft decision, the global SAP (GSAP) $p(H_0|Y(t))$ conditioned on the current observations is derived such that

$$p(H_0|Y(t)) = \frac{p(Y(t)|H_0)p(H_0)}{p(Y(t))} = \frac{1}{1+\frac{P(H_1)}{P(H_0)}\prod_{k=0}^{K-1}\Lambda(Y_k(t))} \quad (5)$$

where $P(H_0) = (1 - P(H_1))$ is the *a posteriori* probability of speech absence. Also, substituting (3) and (4) into (5), the likelihood ratio $\Lambda(Y_k(t))$ at the kth frequency can be obtained as follows [3]:

$$\Lambda(Y_k(t)) = \frac{p(Y_k(t)|H_1)}{p(Y_k(t)|H_0)} = \frac{1}{1+\xi_k(t)} \exp\left\{\frac{\gamma_k(t)\xi_k(t)}{1+\xi_k(t)}\right\} \quad (6)$$

where the *a posteriori* signal-to-noise ratio (SNR) $\gamma_k(t)$ and the *a priori* SNR $\xi_k(t)$ are defined by

$$\gamma_k(t) \equiv \frac{|Y_k(t)|^2}{\lambda_{d,k}(t)},\ \xi_k(t) \equiv \frac{\lambda_{x,k}(t)}{\lambda_{d,k}(t)}. \quad (7)$$

Also, if $\hat{\xi}_k(t)$ and $\hat{\gamma}_k(t)$ are the estimates for $\xi_k(t)$ and $\gamma_k(t)$, $\hat{\xi}_k(t)$ could be estimated using the well-known decision-directed (DD) approach as follows:

$$\hat{\xi}_k(t) \equiv \alpha_\xi \frac{|\hat{X}_k(t-1)|^2}{\hat{\lambda}_{d,k}(t-1)} + (1-\alpha_\xi)C[\hat{\gamma}_k(t)-1] \quad (8)$$

where $\hat{X}_k(t-1)$ represents the estimated clean speech spectrum in the previous frame and $C[x] = x$ if $x \geq 0$, and $C[x] = 0$ otherwise. Here, $\alpha_\xi (0 \leq \alpha_\xi \leq 1)$ is a weighting factor that controls the trade-off between the noise reduction and the transient signal distortion by being chosen very close to 1 (i.e., $\alpha_\xi = 0.99$).

On the other hand, the estimation of the noise power spectrum is a major component in speech enhancement. In particular, the soft decision method adopts a long-term smoothed noise power spectrum of the background noise as the estimate for $\lambda_{d,k}(t)$ as follows [3]:

$$\hat{\lambda}_{d,k}(t+1) = \zeta_d\hat{\lambda}_{d,k}(t) + (1-\zeta_d)E[|D_k(t)|^2|Y_k(t)] \quad (9)$$

where $\hat{\lambda}_{d,k}(t)$ is the estimate for $\lambda_{d,k}(t)$ and $\zeta_d (= 0.99)$ as a parameter for smoothing under a general stationary assumption of $D_k(t)$ [3]. By taking into account the uncertainty for speech absence or presence, the GSAP is applied to the expectation of the power spectrum of noise signal such that

$$E[|D_k(t)|^2|Y_k(t)] = E[|D_k(t)|^2|Y_k(t), H_0]p(H_0|Y(t)) + E[|D_k(t)|^2|Y_k(t), H_1]p(H_1|Y(t)) \quad (10)$$

Given the noisy input signal from the microphone, the clean speech signal $\hat{X}_k(t)$ is obtained by multiplying each spectral component of the noisy speech signal $Y_k(t)$ by a specific spectral gain function $G_k(t)$. Among a number of the spectral gain functions, we follow the MMSE-based noise suppression rule proposed by Ephraim and Malah [1].

3. Proposed speech enhancement using acoustic noise classification

From the previous section, it is discovered that two key parameters of the speech enhancement technique as in [3], such as the weight α_ξ in the DD approach and the long-term smoothing parameter ζ_d in the noise power estimation, are set to fixed values. Since, however, those parameters should be differently set according to the noise type to ensure best performance, we organize the environment knowledge associated with noise to adaptive selection of the parameter in speech enhancement.

3.1. Finding Optimal Operating Points For Given Noises

The operating points about α_ξ and ζ_d according to specific noises should be built on a relevant criterion in terms of speech quality. The most accurate way to evaluate speech quality can be achieved through exhaustive the subjective listening test. But, since it is very costly and time consuming, we adopt the relevant method such as the composite measure in [6] to check overall speech quality. Specifically, the composite measure for overall quality C_{ovl} is given by combining basic objective measures to form a new measure as following:

$$C_{ovl} = 1.594 + 0.805PESQ - 0.512LLR - 0.007WSS \quad (11)$$

where PESQ is the perceptual evaluation of speech quality (PESQ) in the ITU-T P.862, the LLR denotes log-likelihood ratio (LLR) and the WSS is weighted-slope spectral distance. In [6], it is known that the composite measure has a significant correlation with the overall perceptual speech quality such as the mean opinion score (MOS). In terms of C_{ovl}, we investigated the performance by varying α_ξ and ζ_d and plotted the graphical curve. For this, we prepared the NTT database that consists of a number of speech material. In order to create noisy environments, we added twelve different noise such as babble, car1, car2, destroyer-engine, destroyer-operation, factory1, factory2, HF-channel, office, street, white, wind noises to the clean speech data at 5, 10, and15 dB SNR. For each noise type, we obtained the 3D mesh curve as a function of the various values of α_ξ and ζ_d as plotted in Fig. 1. From Fig. 1, it is discovered that the point directed by arrow becomes the optimal point in the case of office noise. In a similar way, we obtained the optimal points $(\alpha_\xi^*, \zeta_d^*)$ for various noise types as shown in Table I.

3.2. Acoustic Noise Classification-Based Speech Enhancement

As given in the previous subsection, optimal operating points for various noise types are achieved. For the real-time implementation in choosing the optimal point according to the given

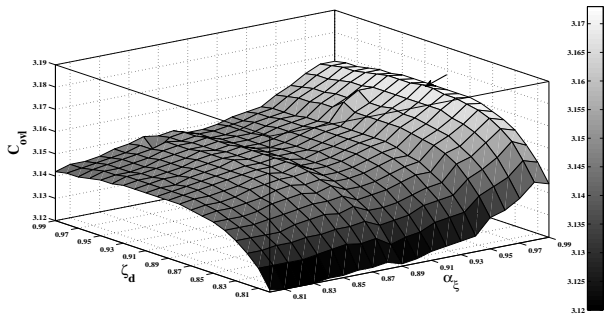

Figure 1: 3D mesh curve of the optimal operating point for the office noise at 10 dB SNR

noise condition, we should classify the noise signal on a frame-by-frame basis during speech absence. To achieve a successful classification, a feature vector that effectively characterize the discrimination among the various noise environments must be chosen. From [7], we selected 10 linear predictive coding (LPC) coefficients, the energy, the partial residual energy, the running mean of energy and the running mean of the partial residual energy due to their superior classification performance.

Given the feature vector $\vec{x} = \{x_1, x_2, \cdots, x_D\}$, the likelihood for the GMM of a weight sum of M mixture components is denoted as follows:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} \alpha_i p_i(\vec{x}) \tag{12}$$

where $p_i(\vec{x})$ is a Gaussian distribution and α_i is the weight of ith Gaussian mixture. Based on this, each noise is modeled by the GMM parameter (λ).

In this noise classification, each noise characterized by a GMM, i.e., λ_s where $s = 1$ (babble), 2 (car1), 3 (car2), 4 (destroyer-engine), 5 (destroyer-operation), 6 (factory1), 7 (factory2), 8 (HF-channel), 9 (office), 10 (street), 11 (white), 12 (wind), 13 (universal background model). Based on the established model, the current input frame is classified into one of the noise classes. Actually, we first use the update routine incorporating the long-term smoothed likelihood based on the soft decision to prevent the update during speech periods such that

$$\begin{aligned} \log \hat{p}(\vec{x}(t)|\lambda_s) &= p(H_0|Y(t)[\beta \log p(\vec{x}(t-1)|\lambda_s) \\ &+ (1-\beta) \log\ p(\vec{x}(t)|\lambda_s)] \\ &+ (1 - p(H_0|Y(t)) \log p(\vec{x}(t-1)|\lambda_s) \end{aligned} \tag{13}$$

Based on this, we then determine the noise model (s) with the maximum *a posteriori* probability on a current frame assuming equally likely noises such that

$$s(t) = \arg \max_{s=1,2,\cdots,13} \log \hat{p}(\lambda_s|\vec{x}_s(t)). \tag{14}$$

Using the classified noise information $s(t)$ on the current frame, the two key parameters α_ξ and ζ_d are substituted with α_ξ^* and ζ_d^* using Tab. 1 every frame. As a result, the proposed

Table 1: Optimal operating points of *a priori* SNR and noise update for various noise types.

Noise Type	optimal points	
	α_ξ^*	ζ_d^*
babble	0.899	0.983
car1	0.803	0.970
car2	0.800	0.990
destroyer-engine	0.810	0.983
destroyer-operation	0.812	0.983
factory1	0.937	0.970
factory2	0.871	0.990
HF-channel	0.814	0.980
office	0.863	0.987
street	0.860	0.990
white	0.803	0.975
wind	0.855	0.990

$\hat{\xi}(t,k)$ becomes

$$\hat{\xi}_{p,k}(t) = \hat{\alpha}_\xi^*(t) \frac{|\hat{X}_k(t-1)|^2}{\hat{\lambda_{d,k}}(t-1)} + (1 - \hat{\alpha}_\xi^*(t)) C[\hat{\gamma}_k(t) - 1]. \tag{15}$$

This time, $\hat{\alpha}_\xi^*(t)$ is obtained using the long-term smoothing to prevent abrupt change of α_ξ^* for ensuring robust performance as follows:

$$\hat{\alpha}_\xi^*(t) = \kappa_\alpha \hat{\alpha}_\xi^*(t-1) + (1-\kappa_\alpha)\hat{\alpha}_\xi^*(t) \tag{16}$$

where $\kappa_\alpha (= 0.9)$ is a smoothing parameter.

Also, the estimation of the noise power is then changed using ζ_d^* such that

$$\begin{aligned} \hat{\lambda}_{d,k}(t) &= \hat{\zeta}_d^*(t)\hat{\lambda}_{d,k}(t-1) \\ &+ (1 - \hat{\zeta}_d^*(t)) E[|D_k(t)|^2|Y_k(t)] \end{aligned} \tag{17}$$

in which

$$\hat{\zeta}_d^*(t) = \kappa_\zeta \hat{\zeta}_d^*(t-1) + (1-\kappa_\zeta)\hat{\zeta}_d^*(t) \tag{18}$$

with a smoothing parameter $\kappa_\zeta (= 0.9)$. As a result, the statistical model-based speech enhancement using noise classification is finally achieved using (15) and (17).

4. Experiments and Results

The proposed statistical model-based speech enhancement technique using noise classification was evaluated with a objective speech quality measures. Test data, not used in data training, which consisted of the one hundred phrases from the NTT database, spoken by four male and four female speakers, were used. Each phrase included two different meaningful sentences and the whole length of each file lasted 8 sec. 10 ms input signal was sampled at 8 kHz and transformed to the DFT domain. We added the aforementioned various noises to the clean speech signal at different SNRs of 5, 10, 15 dB.

Firstly, we evaluated the performance of the acoustic noise classification for the proposed method under various SNR conditions. Among the experimental results, a confusion matrix at SNR=5dB is presented as shown in Tab. 2. From the table, it can be seen that the proposed approach yielded the reliable classification accuracy. Secondly, we investigated the PESQ scores between the previous method in [3] and the proposed algorithm as in Tab. 3. From the table, we see that the proposed method

Table 2: Result of the noise classification through a confusion matrix (SNR=5dB)

Accuracy	bab	car1	car2	des-eng	des-ops	fac-1	fac-2	HF-ch	office	street	white	wind	ubm
bab	97.10	0	0	0	0	0	0	0	2.90	0	0	0	0
car1	0	92.63	0	0	0	0	5.26	0	0	2.11	0	0	0
car2	0	0.14	99.86	0	0	0	0	0	0	0	0	0	0
des-eng	0	0	0	100.00	0	0	0	0	0	0	0	0	0
des-ops	0	0	0	0	99.96	0	0	0	0	0.04	0	0	0
fac-1	0	0	0	0	0	96.56	0	0	3.44	0	0	0	0
fac-2	0.34	0	0	0	0	0.68	90.51	0	8.48	0	0	0	0
HF-ch	0	0	0	0	0	0	0	100.00	0	0	0	0	0
office	5.97	0	0	0	0	0.24	1.00	0	92.79	0	0	0	0
street	0	0	0	0	0	1.75	1.08	0	5.80	91.37	0	0	0
white	0	0	0	0	0	0	0	0	0	0	100.00	0	0
wind	0	0	0	0	0	0	0	0	0	0	0	100.00	0
Average accuracy : **96.73** %													

Table 3: The PESQ results obtained from the proposed algorithm and the soft decision method.

Environments		PESQ	
Noise	SNR (dB)	[3]-based	proposed
car1	5	3.450	**3.521**
	10	3.740	**3.770**
	15	3.985	**4.000**
destroyer-engine	5	2.330	**2.384**
	10	2.623	**2.659**
	15	2.918	**2.942**
HF-channel	5	2.035	**2.109**
	10	2.379	**2.435**
	15	2.715	**2.754**
white	5	2.036	**2.109**
	10	2.373	**2.441**
	15	2.744	**2.795**

Table 4: The overall quality results [6] obtained from the proposed algorithm and the soft decision method.

Environments		C_{ovl}	
Noise	SNR (dB)	[3]-based	proposed
car1	5	3.750	**3.834**
	10	4.069	**4.110**
	15	4.342	**4.365**
destroyer-engine	5	2.530	**2.594**
	10	2.873	**2.922**
	15	3.210	**3.253**
HF-channel	5	1.994	**2.077**
	10	2.415	**2.483**
	15	2.818	**2.874**
white	5	2.241	**2.327**
	10	2.634	**2.715**
	15	3.042	**3.111**

outperformed the previous approach in all the tested conditions. The final assessment is based on the composite measure providing higher correlations with objective speech quality. From the results in Tab. 4, it can be seen that the proposed algorithm effectively enhances the speech quality compared to the previous method.

5. Conclusions

In this letter, we propose a novel speech enhancement technique using the environment-awareness provided by the noise classification. The principal contribution of this work is a finding the optimal points for the principal parameters in a statistical model-based speech enhancement for further performance improvement. In order to take a frame-by-frame basis implementation of the noise classification, the GMM-based likelihood is used. The performance of the proposed approach has been found superior to that of the conventional technique through the objective quality tests.

6. Acknowledgements

This work was partly supported by the IT R&D program of MKE/KEIT [KI001824] and This work was supported by National Research Foundation of Korea(NRF) grant funded by the Korean Government(MEST) (NRF-2009-0085162)

7. References

[1] Y. Ephraim and D. Malah, “Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[2] O. Cappé, “Elimination of musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech Audio Processing*, vol. 2, pp. 345-349, Apr. 1994.

[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] N. Krishnamurthy and J. Hansen, “Noise update modeling for speech enhancement: when do we do enough?," in*Interspeech*, pp. 1431-1434, Sept. 2006.

[5] A. Sangwan, N. Krishnamurthy, and J. H. L. Hansen, “Environmentally aware voice activity detector ," in *Proc. Interspeech*, pp. 2929-2932, Aug. 2007.

[6] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229-238, Jan. 2008.

[7] K.-H. Lee and J.-H. Chang, “Acoustic environment classification based on SMV speech codec parameters for context-aware mobile phone," *IEICE Trans. on Information and Systems*, vol. E92-D, no. 7, pp. 1491-1495, July 2009.