A Statistical Model-Based Speech Enhancement Using Acoustic Noise Classification for Robust Speech Communication

Jae-Hun CHOI†, Nonmember and Joon-Hyuk CHANG†(a), Member

SUMMARY In this paper, we present a speech enhancement technique based on the ambient noise classification that incorporates 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 set according to the classified context to ensure best performance under each noise. For real-time context 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.

key words: statistical model-based speech enhancement, Gaussian mixture model, noise classification

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 only partly stationary [2]. This due to the major factors which is forced to be the nonlinear smoothing procedure in the “decision-directed” (DD) approach used to obtain a more consistent estimate of the signal-to-noise ratio (SNR). Actually, the a priori SNR determined 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 \( \alpha \) [1], [2].

On the other hand, the speech enhancement technique should realize accurate noise power estimation in adverse environments involving various nonstationary noise components [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, a 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 to estimate the unknown parameters. Note that the long-term smoothing parameter in the SD technique is assumed to be a fixed value regardless of noise type. We note that fixed valued parameters can never be optimal for all noise types since the designer should choose an operating point yielding a reasonable performance across the various noise environments.

In this regard, Krishnamurthy and Hansen proposed the environmental sniffing framework to provide an accurate estimate of the noise update for a given environment [4]. Once the noise type is detected, different noise update is chosen for the given environmental conditions. However, this technique assumes a dual-channel system with one channel that contains only noise and another that contains noisy speech. Thus, it can not be directly applied for the real speech communication circumstances, since this assumption is not possible in the actual acoustic condition. On the other hand, 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 competitive Neyman-Pearson (CNP) VAD. However, the VAD algorithm is designed to improve the detection accuracy, not the speech quality. As observed, few works have been applied to the statistical model-based speech enhancement, which is widely used in the task of single channel noise suppression [1]–[3].

In this paper, 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

$$Y_k(t) = X_k(t) + D_k(t)$$

where $k = 1, 2, \ldots, K$ 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

$$H_0 : \text{speech absent} : Y_k(t) = D_k(t)$$
$$H_1 : \text{speech present} : Y_k(t) = X_k(t) + D_k(t).$$

(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

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

(3)

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

(4)

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

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

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

(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 $k$th frequency can be obtained as follows [3]:

$$\Lambda(Y_k(t)) = \frac{1}{1 + \xi_k(t)} \exp\left\{\frac{\gamma_k(t)\xi_k(t)}{1 + \xi_k(t)}\right\}$$

(6)

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

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

(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:

$$\hat{\xi}_k(t) \equiv \alpha_{\xi} \frac{|\hat{Y}_k(t - 1)|^2}{\lambda_{d,k}(t - 1)} + (1 - \alpha_{\xi})C[\hat{\gamma}_k(t) - 1]$$

(8)

where $\hat{Y}_k(t)$ 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]:

$$\lambda_{d,k}(t+1) = \xi_{d} \lambda_{d,k}(t) + (1 - \xi_{d})E[D_k(t)^2|Y_k(t)]$$

(9)

where $\lambda_{d,k}(t)$ is the estimate for $\lambda_{d,k}(t)$ and $\xi_{d} = 0.99$ as a parameter for smoothing under a general stationary assumption of $D_k(t)$ [3].

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 $\alpha_{\xi}$ in the DD approach and the long-term smoothing parameter $\xi_{d}$ 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 $\alpha_{\xi}$ and $\xi_{d}$ 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 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 $C_{osr}$ is given by combining basic objective measures to form a new measure as following:

$$C_{osr} = 1.594 + 0.805PESQ - 0.512LLR - 0.007WSS$$

(10)

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 $C_{osr}$, we investigated the performance by varying $\alpha_{\xi}$ and $\xi_{d}$
Table 1 Optimal operating points of a priori SNR and noise update for various noise types.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>( \alpha^*_\xi )</th>
<th>( \zeta^*_d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>babble</td>
<td>0.899</td>
<td>0.983</td>
</tr>
<tr>
<td>car1</td>
<td>0.803</td>
<td>0.970</td>
</tr>
<tr>
<td>car2</td>
<td>0.800</td>
<td>0.990</td>
</tr>
<tr>
<td>destroyer-engine</td>
<td>0.810</td>
<td>0.983</td>
</tr>
<tr>
<td>destroyer-operation</td>
<td>0.812</td>
<td>0.983</td>
</tr>
<tr>
<td>factory1</td>
<td>0.937</td>
<td>0.970</td>
</tr>
<tr>
<td>factory2</td>
<td>0.871</td>
<td>0.990</td>
</tr>
<tr>
<td>HF-channel</td>
<td>0.849</td>
<td>0.980</td>
</tr>
<tr>
<td>office</td>
<td>0.865</td>
<td>0.987</td>
</tr>
<tr>
<td>street</td>
<td>0.860</td>
<td>0.990</td>
</tr>
<tr>
<td>white</td>
<td>0.803</td>
<td>0.975</td>
</tr>
<tr>
<td>wind</td>
<td>0.855</td>
<td>0.990</td>
</tr>
</tbody>
</table>

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, and 15 dB SNR. For each noise type, we obtained the 3D mesh curve as a function of the various values of \( \alpha \) and \( \zeta \). According to simulation results, it is discovered that highest point becomes the optimal point in each noise. Indeed, we obtained the optimal points \((\alpha^*_\xi, \zeta^*_d)\) for various noise types as shown in Table 1.

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 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 \( \tilde{x} = \{x_1, x_2, \ldots, x_D\} \), the likelihood for the GMM of a weight sum of \( M \) mixture components is denoted as follows:

\[
p(\tilde{x} | \lambda) = \sum_{i=1}^{M} \alpha_i p_i(\tilde{x})
\]  

(11)

where \( p_i(\tilde{x}) \) is a Gaussian distribution and \( \alpha_i \) is the weight of \( i \)th Gaussian mixture. Based on this, each noise is modeled by the GMM parameter \( \lambda \).

In this noise classification, each noise characterized by a GMM, i.e., \( \lambda_j \) 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

\[
\log \hat{p}(\tilde{x}(t)|\lambda_t) = p(H_0|Y(t))[\beta \log \hat{p}(\tilde{x}(t-1)|\lambda_t) \\
+ (1 - \beta) \log \hat{p}(\tilde{x}(t)|\lambda_t)] \\
+ (1 - p(H_0|Y(t)) \log \hat{p}(\tilde{x}(t-1)|\lambda_t)
\]

(12)

Based on this, we then determine the noise model \( \hat{s}(t) \) with the maximum \textit{a posteriori} probability on a current frame assuming equally likely noises such that

\[
s(t) = \arg \max_{s=1,2,\ldots,13} \log \hat{p}(\hat{s}(t)|\lambda_s(t)).
\]

(13)

Using the classified noise information \( s(t) \) on the current frame, the two key parameters \( \alpha^*_\xi \) and \( \zeta^*_d \) are substituted with \( \alpha^*_\xi \) and \( \zeta^*_d \) using Table 1 every frame. As a result, the proposed \( \hat{\xi}(t, k) \) becomes

\[
\hat{\xi}_{p,b}(t) = \hat{\alpha}^*_\xi(t) \frac{|\hat{X}_k(t-1)|^2}{\hat{\lambda}_{d,k}(t-1)} + (1 - \hat{\alpha}^*_\xi(t))C[\hat{Y}_k(t) - 1].
\]

(14)

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

\[
\hat{\alpha}^*_\xi(t) = \kappa_\alpha \hat{\alpha}^*_\xi(t-1) + (1 - \kappa_\alpha)\hat{\alpha}^*_\xi(t)
\]

(15)

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

Also, the estimation of the noise power is then changed using \( \zeta^*_d \) such that

\[
\hat{\lambda}_{d,k}(t) = \hat{\xi}^*_\zeta(t)\hat{\lambda}_{d,k}(t-1) \\
+ (1 - \hat{\xi}^*_\zeta(t))E[|D_k(t)|^2|Y_k(t)]
\]

(16)

in which

\[
\hat{\xi}^*_\zeta(t) = \kappa_\zeta \hat{\xi}^*_\zeta(t-1) + (1 - \kappa_\zeta)\hat{\xi}^*_\zeta(t)
\]

(17)

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 investigated the PESQ scores between the previous method in [3] and the proposed algorithm as in Table 2. From the table, we see that the proposed method outperformed the previous approach in all the tested conditions. The second assessment is based on the composite measure providing higher correlations with objective speech quality. From the results in Table 3, it can be seen that the proposed algorithm effectively enhance the speech quality compared to the previous method.

5. Conclusion

In this paper, we proposed a novel speech enhancement technique using the environment-awareness provided by frame-by-frame noise classification. The principal contribution of this work is a finding the optimal values for the principal parameters in a statistical model-based speech enhancement for further performance improvement. In order to realize frame-by-frame 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 objective quality tests.

Acknowledgement

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (MEST) (NRF-2011-0009182). And, this research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-C1090-1121-0007). And, this work was supported by the research fund of Hanyang University (HY-2011-20110000000210)

References