Speech enhancement based on Gaussian mixture modeling in the sub-band log-power domain

Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura,

Nagoya University, Japan

dat@sp.m.is.nagoya-u.ac.jp

Abstract

We present a speech enhancement system based on Gaussian mixture modeling in the sub-band log-power domain of the observed noisy speech. The basic idea of this method is fitting the actual behaviors of noise and noisy speech powers in terms of their distributions in each sub-band and employing the statistical methods for the noise power estimation and speech activity discrimination. The conventional two components GMM with standard EM algorithm is applied in each sub-band for each segment of half second of the observed noisy speech power. Two statistical methods of maximum a posterior probability (MAP) and cumulative distribution function equalization (CDFE) are developed in this works for the noise estimation. For the voice activity detection, an adaptable decision rule is proposed for the speech recognition application. The noise power and VAD are used in a Wiener filtering system. In an experimental evaluation on AURORA2 database, we compare the proposed to the conventional VAD and noise estimation method. From the experimental results, the proposed VAD method is superior in the non-speech detection rate and the Wiener filtering system based on proposed noise estimation performed better in speech recognition rate, especially in the case when CDFE estimator is employed.

1. Introduction

Noise reduction is an indispensable task of speech technologies, especially in mobile communication, robust speech recognition or hearing aids devices. Among the single channel speech enhancement approaches, the statistical methods in spectral domain is shown to be most effective method [2]-[3]. Considering the additive model of noisy speech

$$\mathbf{X}(n,k) = \mathbf{S}(n,k) + \mathbf{N}(n,k), \quad (1)$$

where $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{N}$ are the complex spectra of noisy speech, clean speech and noise, respectively, these methods are based on statistical estimations for the speech spectral $\mathbf{S}$ using the joint distribution density $p(\mathbf{X}, \mathbf{S})$. Several estimation rules have been studied in literate [1], [2]. This task is separate topic and is not discussed in this work. Here we discuss the modeling of the noise and speech spectrum distributions, which result on the joint distribution $p(\mathbf{X}, \mathbf{S})$. This task is most important but highly difficult due to the dynamic change of the speech and noise behaviors in both time and frequency directions. Conventional methods assume the zero-mean Gaussian distribution of the noise and speech spectrum, which simplifies the mentioned above problem by the noise and signal power (variance) estimation. Among the Gaussian modeling, the signal power is often estimated after noise power estimation, and the decision-directed is well-known as the best method for signal power estimation [2]. Meanwhile, the noise power estimation remains the most difficult problem, especially under low SNR conditions. Several noise estimation methods have been proposed in the literature [2]-[5]. The conventional method uses a voice activity detection [3] to update the noise power during the only noise frames, where the updating information is given by the noisy speech power. The main drawback is that the noise is updated only in the noise frames and therefore, this method is limited for the non-stationary noise environments. Recently, two methods have been proposed in the literature, which do not require VAD and allow to update the noise level even in the speech active duration. The minimum statistic method is proposed by Martin [4], where the updating information is given by tracking the minima from a previous segment of observed noisy powers. Another method uses a q-quantile of the histogram taken from a previous segment to update the local noise power [5]. In both cases on minimum statistic and q-quantile, a segment of 0.5-2 seconds are recommended. The main draw back of these methods is the dependence on control parameter ts, which are difficult to optimized from actual observations. Moreover, this methods is applicable for only a low-dynamic change of the noise level. The basic idea of proposed in this work method is we use the behaviors of the noise and noisy speech powers in term of their probability density functions (PDF) and employ statistical methods to derive the "most like" estimation of the noise power. Furthermore the fitted behaviors of noise and noisy speech in sub-bands are used to discriminate the speech activity. An important point of proposed method is we estimate and upgrade the distributions using a conventional two-components GMM model for short segments of observed noisy speech without any training. Given the noise and noisy speech power distributions, the cumulative distribution equalization (CDFE) estimation method is developed in this work for the noise power estimation. We also compare this method to the conventional maximum a posterior (MAP) estimation. In the VAD for speech recognition application, given the behaviors of noise and noisy speech in sub-bands, an adaptable decision rule is applied to improve the non-speech detection rate and keep the lowest level of distortions. The organization of this paper is follows. Section 2 we describe the Gaussian mixture model used in this work and develop the MAP and CDFE estimators for the noise power estimation.Section 3 proposes a VAD method and reports the experimental evaluation for the VAD performances. In section 4, we implement a Wiener filtering system to compare the proposed noise estimation to the conventional methods and evaluate on AURORA2 database. Section 5 concludes this paper and discusses about the future works.

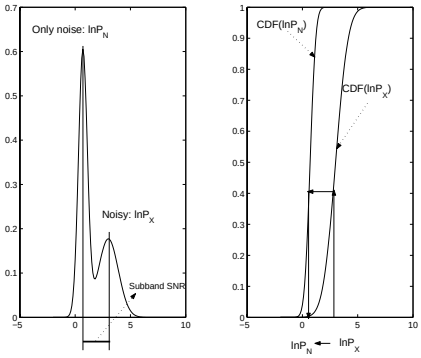

Figure 1: Two components Gaussian mixture model and CDFE in sub-band log-power domain.

2. Noise power estimation by using GMM in the sub-band log power domain

We first describe the GMM fitting is as follows. The observed noisy speech power is smoothed by standard moving average on the noisy periodogram

$$P_X(k,l) = (1-\alpha_X) P_X(k,l-1) + \alpha_X |X(k,l)|^2, \quad (2)$$

where X and P_X denote the observed noisy speech spectrum and the smoothed power, (k,l) denotes the frequency-frame index. The smoothing coefficients should be dependent on the frame length and sampling frequency [4]. In our experiment, we set $\alpha_X = 0.88$. The two components GMM is fitted for each segment of 0.5 second. For simplicity, we assume the diagonal covariance matrix across the sub-bands. The standard EM algorithm is applied where the initial is settled by k-mean algorithm. The estimated parameters in previous segment is used as an initial for the next one. We verified the convergence after just 4-7 iterations. We upgrade the GMM parameter after each 10ms. Note that this GMM model has been applied in full-band log-power domain for the SNR estimation [6].

2.1. MAP estimation of noise power

We first express the noisy speech log-power as follows:

$$\ln P_X(k,l) = \ln P_N(k,l) + \ln SNR(k,l). \quad (3)$$

Recalling two subspace distributions of the GMM in each sub-band, which is fit by the described above procedure

$$\begin{aligned} \ln P_N(k) &\sim \mathrm{N}\left(x, \mu_N(k), \sigma_N^2(k)\right), \\ \ln P_X(k) &\sim \mathrm{N}\left(x, \mu_X(k), \sigma_X^2(k)\right). \end{aligned} \quad (4)$$

The MAP estimation equation for the noise power is denoted as follows:

$$\widehat{\ln P_N} = \max_{P_N} p(\ln P_X | \ln P_N)\, p(\ln P_N). \quad (5)$$

From (3), the condition distribution in (5) is given by the distribution of the instantaneous SNR. This variable is difference of two Gaussian variables and therefore is also assumed to be a Gaussian:

$$\ln SNR(k,l) \sim N\left(\mu_X - \mu_N, \sigma_X^2 - \sigma_N^2\right). \quad (6)$$

For solving (5) we take the derivative to zero. Substituting (4) and (6) into, the estimation of noise power is given as follows:

$$\widehat{lnP_N} = \left[\ \mu_N(k) + \frac{\sigma_N^2(k)}{\sigma_X^2(k)} \left(\ln P_X(l,k) - \mu_X(k)\right)\ \right]. \quad (7)$$

From (7), the MAP estimation of noise power can be viewed as a self-learning linear regression on the logarithmic domain, where the regression parameter is controlled by the fitted distributions.

2.2. Cumulative distribution function mapping

A weak point of MAP estimation in previous section is that the assumption (6) is exactly hold only if two components in right side of (5) are statistically independent and it might get some errors under low SNR conditions. An alternative statistical estimation method which overcomes this problem is cumulative distribution equalization (CDFE) . This estimation method finds a best non linear transform from observed noisy speech to clean speech in log-power domain to match the its CDF

$$\widehat{\ln P_N(k,l)} = g\left\{\ln P_X(k,l)\right\}, \quad (8)$$

$$F_{g(\ln P_X)}\left[g(\ln P_X)\right] = F_{\ln P_N}(\ln P_N). \quad (9)$$

Here $g(.)$ denotes a nonlinear function and $F(.)$ denotes the cumulative distribution function. The key point of CDFE estimation is the invariant property of CDF which is noted as follows

$$F_{g(x)}\left[g(x)\right] = F_x(x). \quad (10)$$

From (9) and (10), the noise power estimation is given by mapping from CDF of observed noisy speech to the CDF of noise power as follows,

$$g(\ln P_X) = F^{-1}_{\ln P_N}\left[F_{\ln P_X}(\ln P_X)\right]. \quad (11)$$

The principle of CDFE estimation using GMM fitting is shown in Figure1. Note that, the Gaussian CDFs have tractable forms expression and therefore, this transform is carried out without any difficulties. Figure 2 show examples of the CDFE for two sub-bands. This method implies a non-linear regression, where the non-linear function is controlled by the cumulative distribution functions.

2.3. Speech spectral magnitude estimation

Given the noise estimation in power domain, the clean speech power is estimated using decision-directed method

$$P_S = \alpha_S \widehat{S^2} + (1-\alpha_S)\max(P_X - P_N, 0) \quad (12)$$

To investigate performance of the proposed noise estimation method, we implement a simple Wiener filter, the gain function of which is denoted by

$$G = \frac{\hat{S}}{X} = \max\left(\frac{P_S}{P_X}, \beta\right), \quad (13)$$

where $\beta = 0.1$ is a spectral floor, which is used to mask the residual noise effect. The proposed Wiener filter will be combined to a proposed in next section voice activity detection.

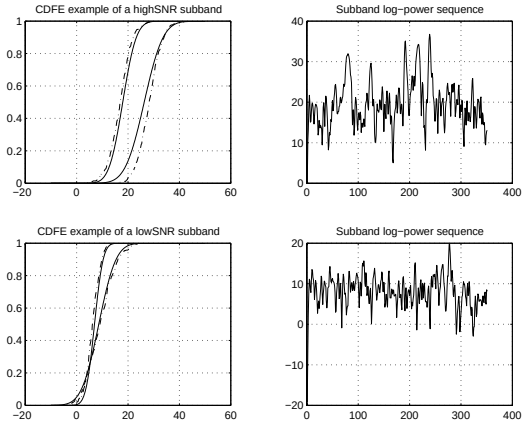

Figure 2: Example of fitted cumulative distribution functions of the noise and noisy speech in sub-band log-power domain. (Dotted lines draw the actual histograms)

3. Voice activity detection based on GMM in sub-band log power domain

The second approach presented in this work is a voice activity detection. This VAD also uses the fitted modeling distributions of the noise and noisy speech power subspaces. One important point is we set an adaptable decision rule to control the false error rate and avoid the high level of distortions.

3.1. GMM in Mel-frequency banks

Unlike in the noise estimation, where the linear frequency scale should be used in order to reproduce the sounds, the sub-band GMM model for the VAD is applied in the Mel-frequency filter banks. We employ a 24 Mel-filter banks system. For saving computational cost, the covariance matrix is assumed to be a diagonal. Assuming two hypothesis: H1(speech present) and H0 (only noise present), for each frame index, the decision rule is formulated by averaging the log-likelihood ratio and noted as follows:

$$\frac{1}{K_{Mel}} \sum_{k=0}^{K_{Mel}} \log \frac{p\left(\ln P_{ob}\left(k\right)|H1\right)}{p\left(\ln P_{ob}\left(k\right)|H0\right)} \begin{array}{c} > \\ \leq \end{array} \eta, \qquad (14)$$

where P_{ob} denotes the observed Mel-bank power at a particular frame. the hypotheses H1 and H0 are denoted as follows:

$$\begin{array}{l} H1 \sim N\left(\mu_X, \sigma_X^2\right) \\ H0 \sim N\left(\mu_N, \sigma_N^2\right) \end{array} \qquad (15)$$

For smoothing the VAD, the spike frame firstly detected as speech frame will be reclassified. An important thing is the setting of threshold decision η which highly affects to the performances of the VAD system. Naturally, the optimal, in statistical meaning, threshold for the log-like lihood ratio is zero. However for the speech recognition system, where the less distortion level is to more important than the noise reduction, the VAD system should keep a high level of speech detection rate, while is working in improvement of non-speech detection rate. From figure 1, we can see that, when the subspace distributions are closed, the optimal solution (which in fact is the intersection point of two distributions) yields a high level of false alarm and consequently possible high level of distortions. For that case, we should move the decision threshold on left side. Following this consideration, we propose a adaptable threshold, which is dependent on the average distance between subspace distributions in each sub-band. This measurement is defined below somehow like the segmental SNR

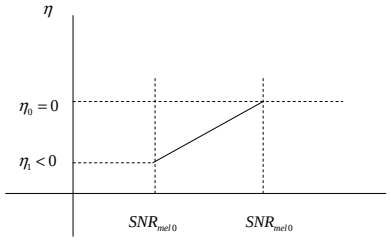

Figure 3: Adaptive threshold to SNR_{mel}

$$S_{mel} = \frac{1}{K_{mel}} \sum_{k=0}^{K_{Mel}} \left(\mu_X\left(k\right) - \mu_N\left(k\right)\right), \qquad (16)$$

where the number of processing filter banks is $K_{mel} = 24$. Figure 3 shows the linear tuning of threshold using in this work. We set $SNR_{mel0} = 30$ and $SNR_{mel1} = 5$ which are approximately equivalent to 0dB and 20dB levels of the segmental SNR consequently. The left boundary $\eta_1 = -1$.

3.2. Experimental evaluation

The proposed VAD is evaluated in terms of the speech/non-speech discrimination analysis using AURORA2 database and is compared to the most representative methods. Two measurements of speech hit rate HR1 (i.e. the average rate of all speech frames that are correctly detected as speech) and non-speech hit rate HR0 (i.e. the average rate of all non-speech frames that are correctly detected as noise) are evaluated. The VAD performances as a function of the SNR are evaluated on the AURORA2 database and are shown in Figure 4 and 5. The proposed method (GMM) is compared to the standard G729 [8] and most recent method KL-FBE based on Kullback-Leiber distance [9]. Table 1 reports the overall results of three methods. We can see that, the proposed GMM method is compatible to the KL-FBE at the HR1 rate but greatly overcomes KL-FBE at HR0 rate. Note that,the setting decision rule is critically important, since from our experiments, the GMM method with non adaptable decision rule yields the performances even worse than G729.

Table 1: *Overall HR1 and HR0 evaluation on AURORA*

Rates	G729	FL-FBE	GMM
HR1	93.00	96.96	96.24
HR0	31.77	46.83	59.18

Table 2: *WER performances of WF systems*

Method	MS	GMMMAP	GMMCDFE
Multi-condition	12.32	12.18	10.02
Clean	25.20	24.78	20.34

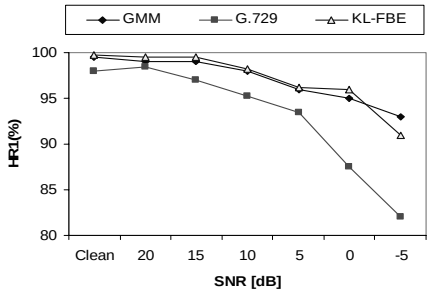

Figure 4: Speech hit rate as a function of SNR

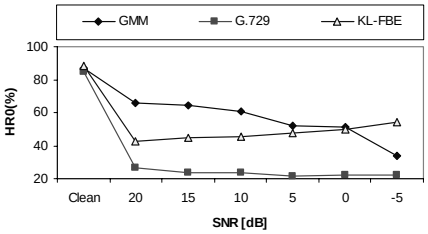

Figure 5: Non-speech hit rate as function of SNR

4. Application to speech enhancement

The proposed noise power estimation and VAD methods are used for a Wiener filtering speech enhancement system. The GMM fitting is same as was described in section 2. The MAP or CDFE estimation of noise power according to (7) and(11) consequently. The gain function is calculated following (13). Finally the phase adding, IFFT and overlap and add are combined to the waveform reconstruction For reference, the analogous WF system based on the minimum statistic method (MS) is also implemented. The standard Aurora2 speech data set is used for evaluation. Speech recognition experiments are performed on the Aurora 2 connected to the digit recognition task [9]. The digit HMMs are the standard complex back-end models of 16 states , and each state has a 20 components Gaussian mixture with diagonal covariance matrix. The training process is carried out at each front-end before training. The feature vector has 39 components of 12 MFCC coefficients together with C0, their first and second derivatives. Table 2 compare the noise power estimation methods in overall performance of speech recognition at the word errors rate. The WF-GMMCDFE system overcomes conventional WF-MS method in WER with approximately 2 percents in multi-condition training, and 4 percents at clean training, while the WF-GMMMAP yields approximately same results compared to WF-MS method. Moreover, from the hearing test, we verify that, the WF-GMMCDFE output speeches have best speech intelligibility with small musical noise level compatible to WF-MS method. The WF-GMMMAP provides better noise reduction, however some musical noise is remained. In next, the WFGMMCDFE system is combined to the proposed VAD. The results of WF+FD systems are in table 4. The system using GMMVAD is performed best for the clean training with about 2.5 percents improvement in WER. However the multi-condition training, the improvements are very small.

5. Conclusion

We propose a speech enhancement based on GMM fitting in the log-power domain of observed noisy speech. Given the distributions of noise and noisy speech subspaces, the statistical methods are employed for the noise estimation and speech activity discrimination. The main point here is that the two-components GMM model in a short segment is able to estimate well the subspace distributions of noise and noisy speech. Furthermore, the employment of statistical method improve the performance of both noise estimation and voice activity detection tasks. The experiment results show the advantage use of proposed method compared to conventional noise estimation methods.

Table 3: *WER performances of WFGMMCDFE+FD systems*

Method	KL-FBE	GMM	G729
Multicondition	9.32	13.25	9.89
Clean	19.35	26.54	17.75

6. Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for CC Society, Course Management System under Ubiquitous Computing Environment, 2004.

7. References

[1] R. Martin “Statistical methods for enhancement of noisy speech,” *Proc. IWAENC*, Kyoto, 2003.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. ASSP*, vol.ASSP-32, pp.1109-1121,1984.

[3] D. Pearce, “Developing the ESTI AURORA advance distributed speech recognition front-end and what next”, *ASRU 2001, pp. 131-134*,2001.

[4] R. Martin, “Noise power spectral estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. ASSP*, Vol. 9, No.5, pp.504-512, 2001.

[5] V. Stahl, A. Fischer and R. Bippus, “Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering,”, *Proc. ICSLP* 2000.

[6] TH. Dat, K.Takeda and F. Itakura ”Robust SNR estimation of noisy speech based on Gaussian mixture modeling on log-power domain”, *Proc. ISCA ITRW Robust*, 2004.

[7] ITU, ITU-T Rec. G729 (AnnexB),1996

[8] J. Ramirez, J. Segura, C. Benitez ”A New Kullback-Leiber VAD for speech Recognition in Noise”, *IEEE SPL*, vol.11, No.2,2004.

[9] H. Hirsch, D. Pearce, ”The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” *Proc. ISCA ITRW ASR*, 2000.