SNR AND LOCAL NOISE POWER ESTIMATIONS BASED ON GAUSSIAN MIXTURE MODELING ON THE LOG-POWER DOMAIN

Kazuya Takeda, Tran Huy Dat, Hiroshi Fujimura, and Fumitada Itakura
Nagoya University, Japan

ABSTRACT

We propose a flexible and robust SNR estimation method for the real conditions, when neither clean reference signal nor speech activity is available. This method is based on Gaussian mixture modeling on the log-power domain of the noisy speech and use the estimated subspace distribution parameters to derive the SNR measures. The experimental results show better performances in estimating both the segmental and global SNR compared to conventional method based on VAD. The second application presented in this work is local noise power estimation, where the same model is applied to each frequency bin. Furthermore, an empirical MAP solution using second order statistics is applied to estimate the local noise powers in order to implement a Wiener filtering system. The evaluation experiments show the improvements of the proposed speech enhancement method in both segmental SNR and ASR performances.

1. INTRODUCTION

The SNR and noise powers estimations are important problems in speech processing. The SNR is a main measure of the speech quality index which is frequently used in the data collection and classification task and the local noise powers estimation is used in speech enhancement systems. In many applications, SNR and local noise power estimations of noisy speech are highly difficult because neither clean reference signal nor speech activity is given. Conventionally, the SNR and local noise power estimations are based on voice activity detection (VAD). However, these methods work well only at high SNR conditions and therefore, its application is limited. The basic idea of this work is using the natural property of the speech signal, which always contains the silent duration to model and estimate them via a probabilistic mixture model. Furthermore the estimated distribution parameters are being used in the SNR and local noise power estimation and it is carried out without VAD. The organization of this paper is as follows. In section 2 we discuss the stochastic view of SNR definitions and Gaussian mixture modeling (GMM) on the log-power domain, and propose the statistical estimation method for the segmental and global SNR indexes. In section 3 the model is applied on each frequency bin and to estimate the local noise powers with application to a Wiener filtering. Section 4 summarizes the work.

2. SNR ESTIMATIONS

2.1 Stochastic view of SNR definitions

Several measures of SNR are used as the speech quality indexes. Since a speech signal is short-time stationary, a segmental SNR is advantageous [1]. Originally, the segmental SNR is calculated in speech active frames:

$$SNR_{seg} = \frac{1}{L}\sum_{i=1}^{L} 10\log_{10}\frac{P_S(i)}{P_N(i)}, \qquad (1)$$

where: $P_S(i)$ and $P_N(i)$ are respectively the clean reference speech and the noise power at the i-th speech active frame, and thus are called frame powers. Alternatively the global SNR is noted as

$$SNR_{global} = 10\log_{10}\frac{P_S}{P_N} = 10\log_{10}\frac{\frac{1}{L}\sum_{i=1}^{L} P_S(i)}{\frac{1}{L}\sum_{i=1}^{L} P_N(i)}. \qquad (2)$$

When clean speech is not available the total (noisy speech) to noise ratio (TNR), is more suitable to use and noted as:

$$TNR_{seg} = \frac{1}{L}\sum_{i=1}^{L} 10\log_{10}\frac{P_X(i)}{P_N(i)}, \qquad (3)$$

where $P_X(i)$ is the frame power in the i-th speech active frame. Due to the independence of speech and noise we suppose that:

$$P_X(i) = P_S(i) + P_N(i). \qquad (4)$$

At very high SNR conditions, the TNR approximates well the SNR and therefore can be used as the SNR index [2]. When the frame number L is large, the averaging in (1) converges to the expectation. The segmental SNR can be denoted in a stochastic form of expectations as follows:

$$SNR_{seg} = \left\langle 10\log_{10}\frac{P_S}{P_N}\right\rangle = \left\langle 10\log_{10} P_S\right\rangle - \left\langle 10\log_{10} P_N\right\rangle, \qquad (5)$$

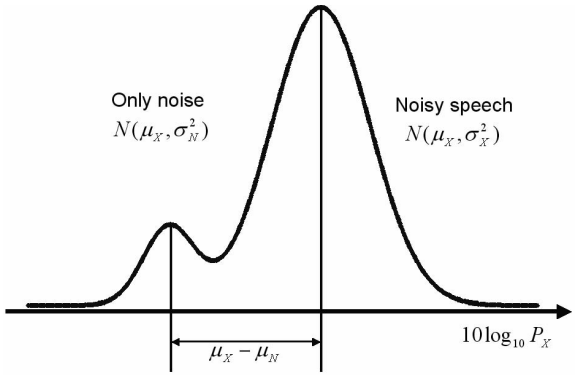

Figure1. GMM on log-powers domain

Analogously, we denote other measures in stochastic form:

$$TNR_{seg} = \left\langle 10\log_{10}\frac{P_X}{P_N}\right\rangle = \left\langle 10\log_{10} P_X\right\rangle - \left\langle 10\log_{10} P_N\right\rangle, \quad (6)$$

$$SNR_{seg} = 10\log_{10}\frac{\langle P_S\rangle}{\langle P_N\rangle} = 10\log_{10}\langle P_S\rangle - 10\log_{10}\langle P_N\rangle, \quad (7)$$

$$TNR_{global} = 10\log_{10}\frac{\langle P_X\rangle}{\langle P_N\rangle} = 10\log_{10}\langle P_X\rangle - 10\log_{10}\langle P_N\rangle. \quad (8)$$

2.2 Gaussian mixture modeling

As was mentioned above, the basic concept of this work is using the probabilistic mixture modeling of the observed noisy speech. The two questions are which model and on which domain the modeling should be assumed? In this work we use the two-component Gaussian mixture on the log-power domain and noted by

$$p(x) = \alpha_0 \mathcal{N}\left(x, \mu_N, \sigma_N{}^2\right) + \alpha_1 \mathcal{N}\left(x, \mu_X, \sigma_X{}^2\right), \quad (9)$$

where: $x = 10\log_{10} P_{observed}$. Note that, the logarithm plays a role of a compressed operator which reduces the dynamic range (or variance) of speech subspace and therefore should provide better the parameter estimation. The simple two-component Gaussian mixture model is chosen due the requirement of the consistency of the parameter estimation, where only a single noisy speech is available. Our procedure for processing is as follows. The power of the observed noisy speech is estimated using a frame length of 4ms and a frame shift of 2ms. Next, the EM algorithm is applied to fit the two mixture Gaussian model to the log frame-power sequence. The initial parameters are chosen by the standard K-means method [3, 4]. We verified the EM convergence after 5-7 iterations and the required minimum length of the noisy speech is approximately 0.5 second with 8 kHz sampling frequency.

2.3 Segmental SNR estimation

On basic of the definition given in (6), the segmental TNR is considered equal to the difference between two estimated means,

$$TNR_{seg} = \mu_X - \mu_N, \quad (10)$$

where: $\mu_X > \mu_N$. Note that the segmental TNR approximates the SNR well at high SNR conditions and can be used as a segmental SNR estimation. However the segmental SNR can be derived more accurately by using estimated distribution parameters. Substituting (4) into (5), the segmental SNR is denoted by

$$SNR_{seg} = \left\langle 10\log_{10}\left(\frac{P_X}{P_N} - 1\right)\right\rangle. \quad (11)$$

Recalling the Gaussian mixture model, the two random variables in (11) have Gaussian distributions:

$$\begin{aligned} 10\log_{10} P_N &\sim \mathcal{N}\left(x, \mu_N, \sigma_N^2\right), \\ 10\log_{10} P_X &\sim \mathcal{N}\left(x, \mu_X, \sigma_X^2\right), \end{aligned} \quad (12)$$

and therefore their difference is also Gaussian distributed:

$$10\log_{10}\frac{P_X}{P_N} \sim \mathcal{N}\left(x, \mu_X - \mu_N, \sigma_X^2 - \sigma_N^2\right). \quad (13)$$

The expectation of the non-linear function (11) has no closed form but can be approximated using an asymptotic expansion:

$$\ln\left(e^r - 1\right) \approx r - 0.7e^{-r} - 0.9e^{-2r} - e^{-3r}. \quad (14)$$

Note that the approximation error is less than 1% at $r > 0.12$, i.e.

$$10\log_{10}\frac{P_S}{P_N} > -9.78 \text{ dB}. \quad (15)$$

In this work our interest is the signals with SNR from 0 to 20 dB and thus the error of approximation (14) can be neglected. The expectation of approximation (14) when r is a Gaussian random variable can easily be calculated. A closed form of SNR_{seg} is derived as follows:

$$SNR_{seg} = \frac{10}{\ln 10}\left\{\begin{array}{l} m - 0.7\exp\left[-\left(m - \dfrac{d}{2}\right)\right] \\ -0.9\exp\left[-2\left(m - d\right)\right] \\ -\exp\left[-3\left(m - \dfrac{3d}{2}\right)\right] \end{array}\right\}, \quad (16)$$

where

$$m = \frac{\ln 10}{10}\left(\mu_X - \mu_N\right),\ d = \left(\frac{\ln 10}{10}\right)^2\left(\sigma_X^2 - \sigma_N^2\right). \quad (17)$$

A simulation experiment is performed to verify the effectiveness of the proposed estimation. The clean speech signal is taken from the JNAS database (Japanese Newspaper Article Sentences Speech corpus). Twenty sentences of 5 male and 5 female speakers are randomly

chosen. Noise is artificially added to the speech signals so that the noisy speech reaches the given level of segmental SNR from 0dB to 20dB. Babble noise is used for the simulation. The segmental SNR and segmental TNR are estimated by the above described algorithms. For reference, the segmental SNR based on VAD [2] is also implemented. Figure 2 shows the results obtained in the babble noise condition. The segmental TNR using GMM gives more accurate estimation than that using VAD [2]. However, the errors in the segmental TNR estimation using GMM are seen at low SNR conditions. The

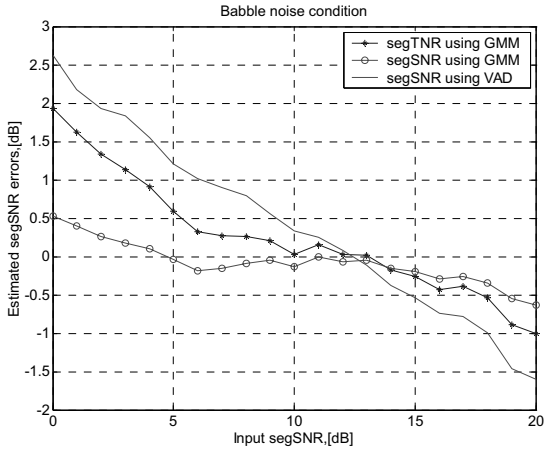

Figure2**:** Segmental SNR estimation errors

segmental SNR estimation using GMM (16) overcomes this problem and is more accurate. Some errors at very high SNR conditions can be explained as being due to the influences of speech pause durations being added to the noise subspace. These come to have an influence on the estimation of noise power subspace when the noise level is low and comparable to them.

2.4. Global SNR estimation

Gaussian mixture modeling can also be used for global SNR estimation. Note that global TNR and SNR are directly related:

$$SNR_{global} = 10\log_{10}\frac{\langle P_S\rangle}{\langle P_N\rangle} = 10\log_{10}\left(10^{\frac{TNR_{global}}{10\log_{10}e}} - 1\right), \quad (18)$$

and therefore the estimation of one of them is sufficient as a basic for calculating the other one. The GMM on the log-power domain is equivalent to the log-normal mixture on the power domain. Given the Gaussian distribution of log-power, the power expectation is given by

$$10\log_{10} X \sim \mathcal{N}\left(\mu,\sigma^2\right) \Rightarrow \langle X\rangle = 10^{\frac{\mu}{10}+\frac{\sigma^2\ln 10}{200}}.$$

Denote the expectations of noise and noisy speech:

$$\begin{aligned}\langle P_N\rangle &= 10^{\frac{\mu_N}{10}+\frac{\sigma_N^2\ln 10}{200}}\\ \langle P_X\rangle &= 10^{\frac{\mu_X}{10}+\frac{\sigma_X^2\ln 10}{200}}\end{aligned}. \quad (19)$$

Substituting (19) into (8) yields:

$$TNR_{global} = \left(\mu_X - \mu_N\right) + \frac{\ln 10}{20}\left(\sigma_X^2 - \sigma_N^2\right). \quad (20)$$

An experiment using the database described in section 2.1 is performed. The global SNR using VAD is also implemented. Figure 3 plots the average global SNR estimation errors for the babble noise estimation errors for the babble noise condition, as it was expected the

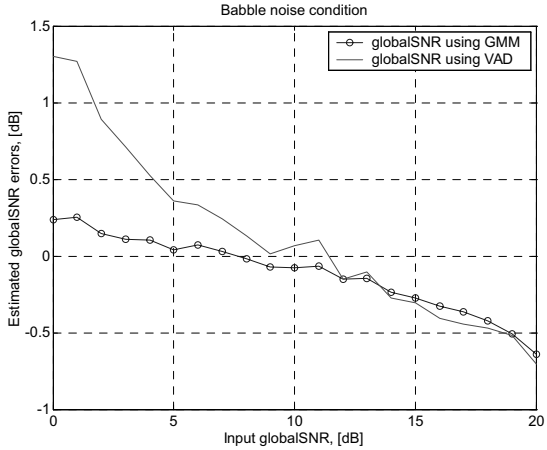

Figure 3**:** Global SNR estimation errors

proposed method estimates the global SNR more accurately than the conventional method especially at the low SNR conditions.

3. LOCAL NOISE POWER ESTIMATION AND APPLICATION TO SPEECH ENHANCEMENT

Second application presented in this work is the local noise power estimation. The conventional methods for the local noise power estimation are mainly based on a heuristic manner: -the moving average method estimates the local noise powers recursively [6]: -the minimum statistic (MS) method estimates the local noise powers on basic of the minimum value of the preceding N frames [5]. A common problem of these methods is that, their effectiveness always depends on some control parameters which is in general difficult to define. In contrast, in this work we present a statistical approach for this problem by applying the mixture modeling on each frequency bin k for the log-power sequences.

$$\begin{aligned}\ln P_N(m,k) &\sim \mathcal{N}\left(x,\mu_N(k),\sigma_N^2(k)\right),\\ \ln P_X(m,k) &\sim \mathcal{N}\left(x,\mu_X(k),\sigma_X^2(k)\right).\end{aligned} \quad (21)$$

As same as in (19) the variances of local noise and noisy speech powers can be denoted

$$\begin{aligned}\mathrm{var}\left\{P_N(k)\right\} &= \exp\left[2\mu_N(k)+\sigma_N^2(k)\right]\exp\left(\sigma_N^2(k)-1\right)\\ \mathrm{var}\left\{P_X(k)\right\} &= \exp\left[2\mu_X(k)+\sigma_X^2(k)\right]\exp\left(\sigma_X^2(k)-1\right)\end{aligned}. \quad (22)$$

The maximum a posterior (MAP) estimation for local noise powers is denoted by

$$\hat{P}_N = \arg\max_{P_N}\left[\log\left(p\left(P_X \mid P_N\right)p\left(P_N\right)\right)\right]. \quad (23)$$

However, since $p(P_S)$ has no closed form, the MAP estimation (23) is not realized. Here, we use an empirical MAP solution by approximating the distributions based on first and second order statistics [7]. This yields an empirical solution denoted by:

$$\hat{P}_N(m,k) \approx \mu_{P_N}(k) + \frac{\mathrm{var}\left\{P_N(k)\right\}}{\mathrm{var}\left\{P_X(k)\right\}}\left(P_X(m,k)-\mu_{P_X}(k)\right), \quad (24)$$

where: the means and variances of the noise and noisy speech powers are estimated by (19) and (22) respectively. The estimated local noise powers (24) can be used for a noise suppression filter which is noted by:

$$\hat{S}(m,k) = G(m,k)\,X(m,k) \quad (25)$$

For the Wiener filtering, the gain factor is estimated via the local SNR:

$$G = \frac{P_S}{P_X} = 1 - \frac{\mu_{P_N} + \frac{\mathrm{var}_{P_N}}{\mathrm{var}_{P_X}}\left(P_X - \mu_{P_X}\right)}{P_X}. \quad (26)$$

The wave form of enhanced speech can be reproduced by the phase adding, inverse IFFT, overlap and add technique [1]. In the evaluation, we implement the Wiener filter with minimum statistics and the proposed GMM method for local noise power estimation. The Aurora 2 data is used for the evaluation. Figures 8-9 show the results of average segmental SNR and ASR improvements over moderate SNR and noise conditions. The GMM noise estimation based Wiener filter are shown to be comparable to the MS method at low and middle SNR conditions and even better at high SNR conditions.

4. CONLUSIONS

We propose a flexible and robust SNR method for the real case when neither clean reference signal nor speech activity is available. This method uses the natural property of speech signal which always contains the silent durations to model them as a probabilistic mixture model. The simple two-component GMM on the log-power domain is shown to be simple but effective to estimate the SNR indexes. This model can also be extended to apply on each frequency bin in order to noise power estimation. The performance of proposed noise estimation method is shown to be at least comparable to the conventional method, when is free from any control setting parameter.

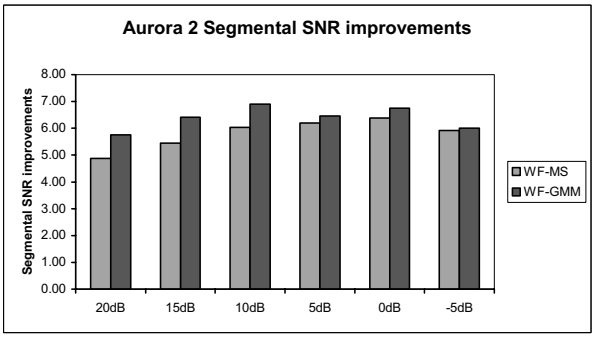

Figure 8**:** Segmental SNR improvements

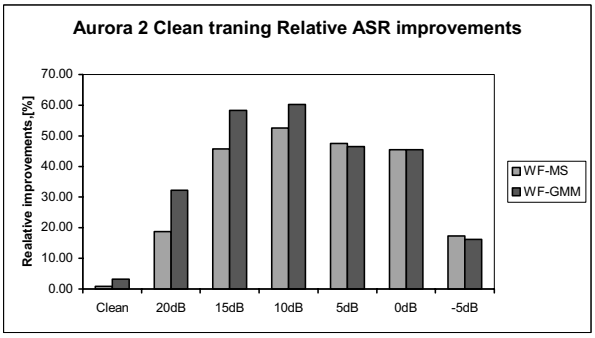

Figure 9**:** Relative ASR improvements with clean training.

5. ACKNOWLEGDGMENTS

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.

6. REFERENCES

[1] J.Deller, J. Proakis, and J. Hansen, Discrete time processing of speech signals, Prentice Hall, 1987

[2] A.Korthauer, "Robust estimation of the SNR of noisy speech for the quality evaluation of speech databases," in Proc. IEEE RMSRAC, 1999

[3] S. Dasgusta, "Learning Gaussian mixtures," in Proc. IEEE SFCF 1999.

[4] G.J. McLachlan, and D.Peel, Finite mixture models, Wiley, 2000.

[5] R. Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics,'' IEEE Trans. ASSP, Vol. 9, No5, 2001, pp.504-512.

[6] I. Cohen, "Noise spectrum estimation in adverse environments," IEEE Trans. SAP Vol. 20, 2002.

[7] V.I Tikhonov, Statistical radio technique, Moscow, Soviet Radio, 1983.