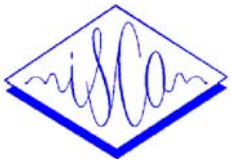


INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Voice activity detection using global soft decision with mixture of Gaussian model

Kiyoung Park, Changkyu Choi, Jeongsu Kim

Human Computer Interaction Laboratory
Samsung Advanced Institute of Technology
San 14-1, Nongseo-ri, Kiheung-eup, Youngin-city Kyounggi-do 449-712, Korea
{kiyoung0.park,changkyu_choi,jeongsu.kim}@samsung.com

Abstract

An improvement on the voice detection algorithm using global soft decision (GSD) is made in this paper. In GSD method, the speech and noise are modelled by the presumed probability density function, e.g. Gaussian pdf. We propose that the estimation and modelling of the signal is done in the domain of filterbank output which widely used in most speech processing applications. Since the output of filterbank is the weighted sum of outputs of several frequency bins, the signals can no longer be estimated using the Gaussian models but mixture of Gaussian models (GMM) in general. It is shown that the estimation of speech absence probability with GMM gives better performance.

1. Introduction

Voice activity detector (VAD) is a system to tell the intervals of sound signals which contain human voice from those of noise only. Such a system is required in many applications including speech coding, enhancement and recognition. Especially in speech enhancement, by knowing the interval including speeches, the characteristics of them can be separated from the noise signal, and noise can be suppressed effectively.

One of the widely used algorithm for VAD is described in [1] and is improved in [2] and [3]. The global soft decision (GSD) algorithm is named from the fact that the decision on the speech absence is made using the global information in all frequency channels in a given time frame and the fact that the decision is made in the form of speech absence probability (SAP).

In this paper, the existing algorithm for the GSD is briefly introduced and a new improvement on that is made to get better SAP estimation performance. Experimental results in simulated environments will be given and conclusion will be followed.

2. GSD for the voice activity detection

Let $Y(t)$ be the spectrum of signals spoken in background noise. Since the speech does not exist all the time, at a given time t, the signal can be represented on the following two hypotheses, H_0, and H_1;

$$H_0 : Y(t) = N(t) \tag{1}$$
$$H_1 : Y(t) = S(t) + N(t), \tag{2}$$

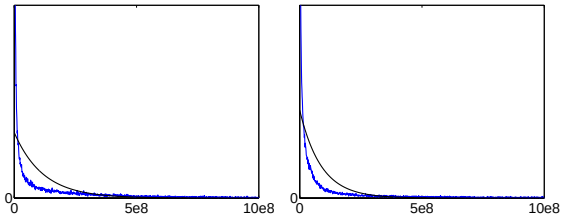
where $S(t)$ and $N(t)$ are the spectrum of clean speech and that of additive noise signal, respectively. It is assumed that the speech and noise are uncorrelated. In many methods in speech enhancement, spectrum of signal $Y(t)$ is denoted as the composition of frequency components $Y(t) = [Y_1(t), Y_2(t), \cdots, Y_M(t)]$, and each component $Y_k(t)$ is regarded as to have independent probability distribution. Usually both the speech and noise spectrums are assumed to have the zero mean complex Gaussian pdfs as follows [4],

$$\begin{aligned} p_S(S_k(t)) &= \frac{1}{\pi\lambda_{s,k}(t)} \exp\left[-\frac{|S_k(t)|^2}{\lambda_{s,k}(t)}\right] \\ p_N(N_k(t)) &= \frac{1}{\pi\lambda_{n,k}(t)} \exp\left[-\frac{|N_k(t)|^2}{\lambda_{n,k}(t)}\right], \end{aligned} \tag{3}$$

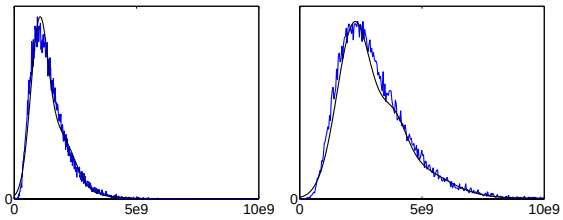
where $\lambda_{s,k}$ and $\lambda_{n,k}$ are the variances of the clean speech and noise signal in the kth frequency bin respectively. Although many of assumptions involved, including independence across channel assumption and pdf model, are far from the reality, many researches with these have shown reasonable preformance[1, 2].

3. GSD with Gaussian mixture model

In many speech processing applications including recognition and coding, the speech signals are dealt with on the 'band of frequency bins' basis, e.g. the output of Mel-scaled filterbank in speech feature extraction methods. By introducing the frequency bands of a few tens rather than using all frequency bins, the number of which usually ranges upto a few hundreds, the amount of computation can be greatly reduced, and moreover the amount of



(a) Clean Speech, the 9th band (b) Clean Speech, the 11th band



(c) White noise, the 9th band (d) White noise, the 11th band

Figure 1: The output of frequency band. Dim lines real data, bold lines estimated pdf.

data for each set of parameters increases, which in turn improves the reliability of parameter estimation.

When two or more frequency bins in (3) are added to yield a band of frequency, the band output cannot be assumed to have Gaussian pdf anymore, since each of them is assumed to have Gaussian pdf by itself. In this work, the characteristics of band output are analyzed and experimental results with a new pdf model are proposed.

The outputs of frequency band are computed according to the feature extraction methods suggested by the ETSI standard [5];

$$\mathcal{Y}_j(t) = \sum_{k=j_l}^{j_h} c(j,k) Y_k(t), \tag{4}$$

where $\mathcal{Y}_j(t)$ is the output of jth band at time t, $Y_k(t)$ is the output of kth frequency bin and j_l, j_h and $c(j,k)$ are the Mel-scaled filterbank coefficients as in [5].

Fig.1(a)–1(d) show the distributions of the output of the 9th and the 11th band for the clean speech and white noise respectively. Peaky lines indicate the real data histogram. The distribution of clean speech is quite sparse and the Laplacian model seems to be valid. That of noise, however, is far from Gaussian distribution and rather resembles Rayleigh distribution.

In this paper, we propose the band output for the clean speech has the Laplacian pdf and distribution of band output for the noise can be modelled by the mixture of Gaussians (GMM) with positive means and variances. In the Fig.1(a)–1(d), smooth lines indicate the estimated distribution for Laplacian and GMM, respectively. Now we propose that each has the following pdfs, respectively

$$p_{\mathcal{S}_j}(\mathcal{S}_j(t)) = \frac{1}{2a_j} e^{-\frac{|\mathcal{S}_j(t)|}{a_j}} \tag{5}$$

$$p_{\mathcal{N}_j}(\mathcal{N}_j(t)) = \sum_i w_{j,i} \frac{1}{\sqrt{2\pi\sigma_{j,i}^2}} e^{-\frac{(\mathcal{N}_j(t)-m_{j,i})^2}{\sigma_{j,i}^2}}, \tag{6}$$

where a_j is Laplacian parameter, and $w_{j,i}$, $m_{j,i}$ and $\sigma_{j,i}^2$ are mixing coefficients, means, variance of ith mixture of GMM respectively for the jth band output.

Then the conditional probability densities given speech absence and presence hypotheses are as follows [3].

$$p(\mathcal{Y}|H_0) = \frac{1}{2a} e^{-|\frac{\mathcal{Y}|}{a}} \tag{7}$$

$$\begin{aligned} p(\mathcal{Y}|H_1) &= \int_{-\infty}^{\infty} f_{\mathcal{N}}(\mathcal{Y}-\mathcal{S}) f_{\mathcal{S}}(\mathcal{S}) \mathrm{d}\mathcal{S} \\ &= \sum_i \frac{w_i}{4a} e^{\frac{\sigma_i^2}{a^2}} \left[e^{\frac{\mathcal{Y}}{a}} \mathrm{erfc}\left(\frac{a\mathcal{Y}+\sigma_i^2}{\sqrt{2}a\sigma_i} \right) \right. \\ &\quad \left. + e^{-\frac{\mathcal{Y}}{a}} \mathrm{erfc}\left(\frac{-a\mathcal{Y}+\sigma_i^2}{\sqrt{2}a\sigma_i} \right) \right] \end{aligned} \tag{8}$$

where time index t and band index j are omitted for the simplicity, and $\mathrm{erfc}(\cdot)$ is an error function.

With these hypotheses and global soft decision rule, SAP can be obtained using the improved GSD (IGSD) [2] by

$$\begin{aligned} p(H_0|\mathcal{Y}) &= \frac{p(H_0,\mathcal{Y})}{p(\mathcal{Y})} \\ &= \frac{\prod_j [P(H_0)p(\mathcal{Y}_j|H_0)]}{\prod_j [P(H_0)p(\mathcal{Y}_j|H_0) + P(H_1)p(\mathcal{Y}_j|H_1)]}, \\ &= \frac{1}{\prod_j [1+q\Lambda_j]} \end{aligned} \tag{9}$$

where $P(H_0)$ is *a priori* SAP $P(H_1) = 1 - P(H_0)$, $q = P(H_1)/P(H_0)$, and likelihood ratio $\Lambda_j = p(\mathcal{Y}_j|H_1)/p(\mathcal{Y}_j|H_0)$. The output of each band is assumed to be statistically independent as well.

4. Experimental results

To compare the performance of GSD with Gaussian and GMM to estimate SAP from noisy speech, the artificially generated white noise signals are added to 50 clean speeches. Each of speech is a sequence of isolated words, and lasts about 19.2 seconds in average. Speeches are spoken by a male in a quiet place and recorded in 16 kHz sampling rate. As in [5], 23 Mel-scaled bands of frequency bins are used, and for each of them the parameters for speech and noise signal model are estimated using Gaussian-Gaussian and Laplacian-GMM pdfs respectively. To clarify the effects of the underlying pdfs and to exclude effects of

wrong parameter estimation, the model parameters for both speech and noise models are not estimated adaptively but estimated using all the speeches before computing the SAP and the fixed values are used through all the experiments. At first the variances in Gaussian model is estimated as

$$\hat{\sigma}^2_{s,k} = E[|S_k|^2] \tag{10}$$

$$\hat{\sigma}^2_{n,k} = E[|Y_k|^2|H_0] \tag{11}$$

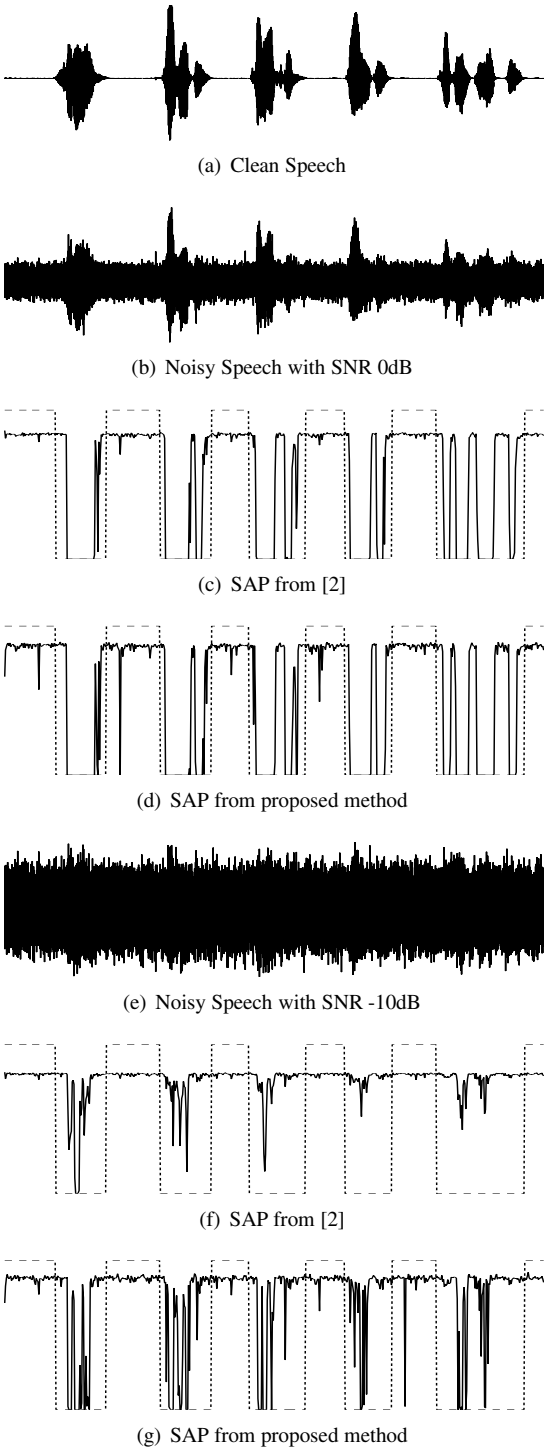
In 10 it is assumed that we know the clean speech, and 11 can be comuted since the noise intervals are marked by human as exact as possible beforehand. The parameter in Laplacian is estimated by the mean of the band output as,

$$\hat{a}_j = E[\mathcal{S}_j]. \tag{12}$$

The estimation of GMM parameters in the noise only interval, greedy EM algorithm in [6] is used. The maximum number of mixture for each band is limited not to exceed 4, and after parameter estimation 3.5 mixtures for each band are used in average.

Fig. 2(a) shows an example of clean speech recording and Fig. 2(e) and Fig. 2(e) show noisy speeches of SNR 0dB and -10dB. It is very hard to detect voice interval just by tracking the frame energy. While Fig. 2(c) and 2(f) illustrate the results of VAD based on the Gaussian and Gaussian model as in [2], Fig. 2(d) and 2(g) show those based on the Laplacian and GMM. Dashed lines indicate the target SAP which are marked by human with clean speeches. As shown in the figures, the result with GMM shows more accurate SAP in noise interval while similar SAP in voice interval, and this is more evident under severe noise environments. This is because GMM has the capability to model the signals of large variation.

The average SAP in the voice interval (given H_1) and in the silence interval (given H_0) for all speeches are summarized in Table. 1. The value of *a priori* likelihood ratio q is set to 0.01 empirically. Since SAP is a monotonic function of q, the value itself doesn't affect relative SAP. With the same value of q, the proposed method gives low SAP in voice interval, and high SAP in silence interval, both of which are what we want. To make fair comparison, the SAP in voice interval is adjusted by changing the value of q in proposed method. When the SAPs in voice interval is equal to each other, the SAPs in silence interval of proposed method is higher about 0.07 and 0.11 than those of GSD with Gaussian models. With proper hard decision rule e.g. described in [3], the discrimination between voice and silence interval becomes possible even in excessive noise environments with GSD using GMM, which was not possible with Gaussian model.



(a) Clean Speech

(b) Noisy Speech with SNR 0dB

(c) SAP from [2]

(d) SAP from proposed method

(e) Noisy Speech with SNR -10dB

(f) SAP from [2]

(g) SAP from proposed method

Figure 2: Comparison of SAP estimation between Gaussian and mixture of Gaussian modelling.

5. Discussion

In this paper, the signal model is considered for the computation of SAP. SAP is widely used for many speech

Table 1: Average SAP in voice interval and in silence interval.

(a) SNR 0dB

	$\frac{P(H_1)}{P(H_0)}$	SAP in Voice Interval	SAP in Silence
IGSD	0.0100	0.3801	0.8330
Proposed Method	0.0100	0.3501	0.8506
	0.0057	0.3802	0.9102

(b) SNR -10dB

	$\frac{P(H_1)}{P(H_0)}$	SAP in Voice Interval	SAP in Silence
IGSD	0.0100	0.7183	0.8008
Proposed Method	0.0100	0.6792	0.8748
	0.0068	0.7188	0.9116

applications including speech enhancement, coding and recognition. The Gaussian model which is usually adopted in most applications is not suitable for the sum of several frequency bins. Instead, the GMM is used and is shown to represent the noisy speech signal better than the Gaussian model.

As further works, the algorithm for the speech enhancement can be developed using the various algorithm e.g. one proposed by Ephraim and Malah [4], since the signal and noise power can be estimated from the model parameters. More elaborately the enhancement method should be developed based on the pdf model and this remains as a further research.

6. References

[1] 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.

[2] V. I. Shin, D.-S. Kim, M. Y. Kim, and J. Kim, "Enhancement of noisy speech by using improved global soft decision," in *Proc. 7th European Conference on Speech Communication and Technology*, vol. 3, 2001, pp. 1929–1932.

[3] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, September 2003.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, December 1984.

[5] *ETSI ES 202 212 v1.1.1, Speech processing, transmission and quality aspect (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm*, European Telecommunications Standards Institute, http://www.etsi.org, November 2003.

[6] N. Vlassis and A. Likas, "A greedy em algorithm for gaussian mixture learning," *Neural Processing Letters*, vol. 15, no. 1, pp. 77–87, February 2002.