Noise Estimation Using Negentropy Based Voice-Activity Detector

Rajkishore Prasad, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science,
Nara Institute of Science and Technology, Nara, Japan
{kishor-p, sawatari, shikano}@is.aist-nara.ac.jp

Abstract- This paper presents a noise robust Voice Activity Detector (VAD) based on the chaos measure of the quasi-stationary segment of speech in the frequency domain. The basic idea behind the proposed method is that the addition of noise in the clean speech produces less disorganization in the speech part than the silent/paused part as a result of which spectral contents of the speech part become less chaotic than the silent part in a noisy speech signal. The negentropy can be used as the measure for such a distinction in organization and accordingly noisy speech frames and the noise-only frames are demarcated. The algorithm is almost independent of the SNR level of the noisy speech signal.

I. INTRODUCTION

The task of VAD is related to the automatic segmentation of speech signal into speech and silent parts. Such a technique is of much practical importance and is required in many applications such as in speech enhancement where the noise statistics are estimated from noise-only parts of the degraded signal [1], in speech coding, and in telecommunication e.g. in data and power saving in mobile devices, in the reduction of co-channel interference in mobile telephony [2], for adaptation control in adaptive beamforming. There have been developments of different VAD algorithms that can be broadly categorized into two groups, namely, single channel VAD and multi-channel VAD. The single channel VADs usually exploit temporal or harmonic features of the speech while a multi-channel system uses temporal or harmonic features as well as spatial features such as direction of arrival, source position, etc [3]. Our concern in this paper is with the single channel VAD in the frequency domain. For the single channel VADs conventional methods have been mainly based on the difference of temporal variation in the statistics of speech and interfering noise, as such variations in the noise are much smaller than that in the speech signal. Such methods use information like zero crossing rates, energy of speech and noise, fuzzy rules, etc. [4]. Recently, there have also been developments of VAD algorithms in the different transformed domains that transforms signal into a decorrelated space. A VAD based on the statistical model of Discrete Fourier Transforms (DFT) of the quasi-stationary segments of speech can be found in [5,6] in which Gaussian Probability Distribution Function (PDF) function has been used, in light of [7], to model speech spectral components. In [8], a soft VAD algorithm has been proposed using Laplacian model for the Karhunen-Loeve Transform (KLT) coefficients of speech. Our proposed algorithm is based on the statistical modeling of the magnitude of the DFT coefficients of speech and additive noise by the Generalized Gaussian Distribution (GGD) as it approximate the PDF of spectral components better than a Laplacian or a Gaussian PDF [8]. Since the temporal organization of spectral components is well-patterned in the speech than the noise, the portion of noisy speech signal containing voice will be less chaotic than the silent part e.g. spectrograms of White Gaussian Noise (WGN) and clean speech signal from a female speaker can be observed in Fig.1 to be imbued with such differences in spectral organization. We have used negentropy of the quasi-stationary segment of degraded speech as the measure for such chaos in the signal and by thresholding this measure noisy speech and noise-only frames are discriminated. In contrast to conventional energy based method, the proposed method works well in very low SNR condition.

The rest of the paper is organized as follows. In the next section additive noise model is presented in the frequency domain, Section III deals with the negentropy based voice activity detection and threshold determination for the negentropy based detector. Section IV contains experimental results that are followed by references.

II. SIGNAL MODEL

Speech signal \(y(n)\) picked-up by a single microphone in the presence of additive uncorrelated noise \(d(n)\) is given by

\[
y(n) = x(n) + d(n)
\]

where \(x(n)\) represents clean speech signal, \(n\) is the time-index, and random noise \(d(n)\) is uncorrelated with the clean speech signal. The Short-Time Fourier Transform (STFT) analysis of the above models transforms it into frequency domain in which spectral component in an arbitrary frequency bin \(f\) is given by

\[
y(f) = X(f) + D(f),
\]

where \(y(f),\ X(f),\) and \(D(f)\) represent spectral components of the observed signal, clean signal and noise signal respectively. The same model is used in the enhancement technique to estimate clean signal in which estimation of noise statistics plays important role [7]. Thus accurate estimation of noise statistics is important for the speech enhancement algorithms. In most of the cases, Gaussian models for both the speech and noise spectral components are assumed which is not always true. There are lots of...
of real world noise e.g. clapping, dropping of object, chair crack, speech like babble noise, etc., which do not have Gaussian PDF. Therefore, in the present work speech and noise spectral components have been modeled using a parametric GGD function. The GGD model is defined by mean $\mu$, scale parameter $\alpha$, and shape parameter $\beta$. The GGD PDF for Random Variable (RV) $z$ is given by

$$f_{GG}(z; \mu, \alpha, \beta) = \frac{2^{\alpha/\beta}}{\beta \Gamma(\alpha/\beta)} \left( \frac{1}{\alpha} \right) \exp \left( - \frac{1}{\alpha} \left( |z - \mu| / \alpha \right)^\beta \right)$$

where

$$A = \frac{\beta}{2\Gamma(\alpha/\beta)}, \quad \frac{1}{\alpha} \exp \left( - \frac{1}{\alpha} (|z - \mu| / \alpha)^\beta \right)$$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt = \text{Gamma PDF};$$

$$0 < z < \infty; \quad \alpha > 0; \quad \beta > 0; \quad \sigma = \text{standard deviation (Stdv.)}$$

The shape parameter $\beta$ determines the shape of the distribution. For $\beta=1$ and $\beta=2$ distributions are Laplacian and Gaussian respectively and distribution tends to become uniform as $\beta \rightarrow \infty$. The shapes of distribution for the different values of $\beta$ are shown in Fig.2. GGD with different parameters will be used to model the underlying PDF of the spectral components of the noise as well as speech. Thus use of GGD as a statistical model for the spectral components of speech and noise can provide flexible model. The GGD parameters will be estimated from the data samples of $Y(f)$ and $D(f)$ using Maximum Likelihood (ML) approach [9].

### III VAD Based on Negentropy

The conventional energy based VAD is based on the energy-level difference of the speech and noise signal. The energy of the spectral component in any frequency bin $f$ is given by

$$E(L) = \frac{1}{N} \sum_{n=1}^{N} |Y(f_i)|^2.$$  

Using such energy based method detection of noise-only frames and noisy speech frames is difficult, especially, in very low SNR conditions or when the spectral energy of noise in any frame is more than that of the speech. Under such adverse conditions the performance of conventional energy based VAD detector worsen. We propose here a statistical VAD detectors based on the chaos measure of the spectral magnitude of quasi-stationary segments. Since in the observed noisy speech data, speech containing segments are less chaotic than the noise-only frames, chaos-based measure can discriminate noise-only and noisy speech frames.

For doing voice activity detection based on such chaotic characteristic, we have used negentropy as a measure [10]. The benefit of using negentropy over others such as entropy [11] is that it is always positive, scale invariant, and can be computed only in terms of shape parameters of the used GGD model as is shown below.

The negentropy of each frame in DFT domain is obtained in terms of Differential Entropy (DE) $\Delta H$ of the magnitude of spectral components. The DE of the any frame data $U = \{Y_1, Y_2, \ldots, Y_N\}$ is given by

$$\Delta H(U) = - \int p(U) \log p(U) dU,$$

where $p(U)$ represents PDF of the frame data $U$. The PDF of magnitude of spectral components of each frame is represented by GGD with mean $\mu$, scale parameter $\alpha$, and shape parameter $\beta$. estimated from the data. Using GGD model for $p(U)$ in Eq.(6) it can be integrated to give

$$\Delta H(U) = \int f_G(\mu, \alpha, \beta) = \log \left( \frac{2\alpha \Gamma(1/\beta)}{\beta} \right) + \frac{1}{\beta},$$

which depends on the scale and shape parameters. The negentropy $H(\beta)$ is computed as the difference of DE of Gaussian RV, with same variance as of that of spectral components of speech, and DE of speech spectral components modeled by the GGD $f_G(0, \alpha, \beta)$. Accordingly, negentropy $H(\beta)$ is given by

$$H(\beta) = \Delta H(\alpha_s, \beta_s) = 2 - \Delta H(\alpha_s, \beta_s)$$

$$= \log \left( \frac{\beta_s}{2} \right) \frac{6.29 \Gamma \left( \frac{3}{\beta_s} \right)}{\Gamma \left( \frac{1}{\beta_s} \right)} + \left( 0.5 - \frac{1}{\beta_s} \right).$$

The theoretical variation of negentropy of GGD with shape parameter is shown in the Fig.3. It is obvious from Fig.3 that the negentropy is zero for the Gaussian distribution and goes up in the positive direction for the spiky distribution. Since the speech frames are more parsimonious than and noise frames the noise-only frames will have lower negentropy while for the noisy speech frames negentropy will be relatively high and thus a threshold value of the negentropy can be chosen to separate noisy speech frames and noise-only frames.

![Fig. 3 Shape parameter versus negentropy of the GGD. It is zero for Gaussian distribution and positive for the spiky distribution (0<β<1).](image-url)
The threshold value of negentropy can be decided on the basis of the global statistics of the negentropy. The negentropy of the frames itself is a random variable and its PDF represents joint probability of occurrence of noisy speech frame and noise-only frames. The PDF of negentropy of each frame can also be modeled by GGD with mean $\mu_h$, scale parameter $\alpha_h$ and shape parameter $\beta_h$ as follows

$$p_r(H) = \frac{1}{2\alpha_h \Gamma(\frac{1}{\beta_h})} \exp\left(-\frac{\alpha_h}{\beta_h}\left(\frac{|H - \mu_h|^2}{\sigma^2} \right)\right)$$

(9)

where $\nu = \{speech, noise\}$ to represent noisy speech frame and noise-only frames respectively. Since the occurrence of noise and speech frame is independent

$$p_r(H) = p(noise|H)p(speech|H)$$

(10)

The threshold value $H_{TH}$ of the negentropy is estimated under assumption that the conditional probabilities of noise-only frame and noisy speech frames are same (say $p_r$) at the threshold.

Accordingly, the threshold $H_{TH}$ is given by

$$H_{TH} = \mu_h \pm \left[\alpha_h \log \left(\frac{2p_r(1-p_r)\Gamma(1/\beta)}{\beta}\right)\right]^{1/\beta_h}$$

(11)

The similarity between the spectral bands of the estimated noise and original noise can be measured by measuring the Kullback Leibler Divergence between the PDF of their spectral bands. Since the PDF of the spectral bands are parameterized by the GGD, KLD between them can be measured in terms of GGD parameters. The KLD between two GGD functions defined by scale parameters $\alpha_1, \alpha_2$ and shape parameters $\beta_1, \beta_2$ is given by

$$D_{KL} = \log \left(\frac{\beta_1 \alpha_1 \Gamma(1/\beta_1)}{\beta_2 \alpha_2 \Gamma(1/\beta_2)}\right) \frac{\alpha_1}{\alpha_2} \Gamma(1/\beta_2) \frac{\Gamma(1/\beta_1)}{\beta_2}$$

(12)

Further DCT coefficients in each frequency bin are independent, so the overall distance between the original and estimated spectrums of noise can be given by averaging the distances calculated in Eq.(12) for each pair of the spectral bands.

IV EXPERIMENTS AND RESULTS

In the experiments we have used two sentences, of the time length 3 sec. and spoken by a Japanese male and female speakers, from the ASJ continuous speech corpus for the research [12] and noise data from the NOISEX-92 database available at http://mi.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html. The speech signals were degraded under the additive model of Eq.(1) to different SNR levels. The speech signal and noise signal are STFTed using hanning analysis window of 20 ms with 50% overlap. The used DFT size was 512. The GGD parameters for the spectral components of each frame were estimated using ML approach which were further used in Eq.(8) to calculate negentropy of spectral components in each frame. The energy of each frames were computed using Eq.(5). The result of negentropy based VAD for the speech signal degraded to -5dB SNR by clapping noise (self recorded) and WGN are shown in Fig.4. The theoretical values of threshold, as derived in the Eq.(12), are shown in the Fig.7 as a function of the probability of occurrence of the speech segments. The threshold values were estimated for the speech signal (female speaker) degraded to 0 dB and 15 dB by WGN and clapping noises. There is a little variation in the negentropy value where the probability of each is assumed to be equal (0.5). Also, if the probability of occurrence of speech frame is increased, threshold goes down and chances of taking larger number of frames as speech and less number of frames as a noise-only frame increases. The GGD parameters estimated for the magnitude of spectral components of the WGN, clap, and babble noise are shown in the Fig.5. It is evident from there that Gaussian distribution for the spectral components of all noise is not a suitable model e.g. clapping noise and babble noise do not have Gaussian distribution as $\beta$ is less than 2 for them. For the speech enhancement algorithms, estimated statistics of the noise are important. In order to estimate noise
Fig. 6. Voice detection for speech signal (female speaker) degraded to -5 dB SNR by WGN (from NOISEX database). The upper figure shows degraded speech, second subplot from top shows WGN, third from top shows clean speech and subplot in bottom shows spectrogram of noisy speech. On each subplot smoothed plot of negentropy and energy for each noisy frame are also shown. Legend indication is same for all subplots as shown in the bottom subplot. It reflects negentropy tracks speech and noise segments better than that of energy-based measure.

statistics, it is very common practice to use few frames from the starting of the observed data as it is silent part. We compare here the GGD parameters of the spectral components of the original noise signal and estimated noise signal discriminated by the negentropy based VAD, and from the beginning’s Few Frames of the Signal (FFS) of the noisy speech data. In Fig. 8 mean, scale parameter, shape parameter and KLD between the spectral bands of original and estimated noise frequency bins are shown for the different SNR conditions. These values are averaged over the number of frequency bins. The estimated spectrum depends on the noise type. As shape parameter of the noise becomes very near to that of the speech, VAD’s performance degrades and difference in the estimated and original noise spectrum increases.

Fig. 7 Variation of threshold for VAD based on the negentropy. It is evident from figure that there is very slight variation in the threshold level for the large variation in the SNR of the degraded speech.

REFERENCES


