Robust Entropy Based Voice Activity Detection Using Parameter Reconstruction in Noisy Environment

Hag-Yong Han, Kwang-Seok Lee, Si-Young Koh and **Kang-In Hur**, *Member, KIMICS*

Abstract—Voice activity detection is a important problem in the speech recognition and speech communication. This paper introduces new feature parameter which are reconstructed by spectral entropy of information theory for robust voice activity detection in the noise environment, then analyzes and compares it with energy method of voice activity detection and performance. In experiments, we confirmed that spectral entropy and its reconstructed parameter are superior than the energy method for robust voice activity detection in the various noise environment.

Index Terms—Noisy Speech, Voice Activity Detection, Entropy

I. INTRODUCTION

VAD(Voice Activity Detection), which can exert decisive effect on the speech recognition rate, is the important pre-processing of speech recognition and speech communication. VAD detects real speech section among various noise and sounds. However, implementation of VAD with independent and stable characteristics in every sounds is difficult work.

VAD is usually based on energy method[1][3]. In this method, it is only suitable in good environment without noise. For example, VAD will need more complicated analysis and method at start point of the speech section and in consonants having low energy. Furthermore, cough sound of outside and additional noise of breathing at start and end point of speech section should be removed. For those problems, when sound section is longer than threshold holding time of speech, exceeding threshold value of short time average energy, speech activity, is detected and start point of speech is located ahead of particular time from detected point of energy threshold value. Moreover, this method is used with zero crossing for more trusty VAD[4]. Another VAD method which use the spectral analysis detect voice activity by using difference between input signal and reference noise spectrum.

In this paper, it is used spectral entropy which is based on entropy, the principle concept of information theory. This method is firstly used by J.L. Shen in speech processing for the first time. Through his experiment Shen showed a lot of difference between speech and non-speech of spectral entropy[5].

In order to complement spectral entropy, this paper suggests new feature parameters which are reconstructed by spectral entropy, analyzes and compares suggested new feature parameters with energy method, and confirms application possibility of VAD parameter in various noise environment.

Manuscript received received November 27, 2003.

Hag-Yong Han (phone: +82-51-200-6961, email: hyhan@donga.ac.kr and Kang-In Hur (email: kihur@mail.donga.ac.kr) are with the Department of Electronic Engineering, Dong-A University, Busan, Korea.

Kwang-Seok Lee (phone: +82-55-751-3333, email: kslee@jinju.ac.kr) is with the Department of Electronic Engineering, Jinju National University, Jinju, Korea.

Si-Young Koh (phone: +82-53-850-7164, email: kohsy@kiu.ac.kr) is with the School of Electronic Information and Communicati on Engineering, Kyungil University, Kyungsan, Korea.

II. PARAMETERS FOR VAD

A. Entropy

Entropy based on the Shannon's information theory is the scale measuring the amount of information. Information derivable from outcome x_i depends on its probability according to information theory. If probability P(x_i) is small, we can derive a large degree of information, because the outcome that has occurred is very rare. On the other hand, if probability is large, information derived will be small, because the outcome is well expected. Thus, the amount of information is defined as follows:

$$I(x_i) = \log\frac{1}{P(x_i)} \tag{1}$$

Suppose X is a discrete random variable taking value x_i (referred to as a symbol) from a finite or countable infinite sample space $S = \{x_1, x_2, ..., x_i, ...\}$ (referred to as a symbol). The symbol is produced from an information source with alphabet S, according to the probability distribution of the random variable X. One of the most important properties of an information source is the entropy H(S) of the random variable X, defined as the average amount of information (expected information):

$$\begin{aligned} H(X) &= E[I(X)] \\ &= \sum_S P(x_i) I(x_i) \\ &= \sum_S P(x_i) \log\frac{1}{P(x_i)} \\ &= E[-\log P(X)] \end{aligned} \tag{2}$$

B. Spectral entropy

Spectral entropy process consists of calculating FFT of input signal, probability density of power spectrum in

the band limited speech signal, and entropy. Probability density of spectrum is estimated in the method that has the normalization effect about frequency components.

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{M} S(f_k)} \tag{3}$$

where, $s(f_i)$ is power spectrum of frequency component f_i, p_i is corresponding probability density, and N is the total number of frequency components in FFT.

Next step is calculating entropy. However, the above process emphasize entropy of the noise and non-speech section, therefore, we can emphasize speech section through entropy conversion, and then, last step, estimated entropy is reconstructed.

C. Energy

Energy is represented by eq.(4),(5) generally as follow. To obtain the stable margin for the VAD, we use more eq.(4) rather than eq.(5) average square logarithm.

$$E_n = \frac{1}{N}\sum_{m=0}^{N-1} x^2(m) \tag{4}$$

$$E_n(dB) = 10\log E_n \tag{5}$$

D. ZCR & LCR

ZCR(Zero Crossing Rate) is the one of the most useful parameter for the Voiced/Unvoiced Detection generally. But ZCR is based on zero point of horizontal level, so this parameter is capable of being affected the background noise.

LCR(Level Crossing Rate) is the parameter to get the only speech crossing rate in order to remove the background noise by using the threshold.

$$L_n = \frac{1}{2}\sum_{m=0}^{N-1} |\,\mathrm{sgn}[x(n-m)-TH] - \mathrm{sgn}[x(n-m-1) - TH\,|\,| \tag{6}$$

where, $\mathrm{sgn}[s(n)] = 1, s(n) \geq 0$ and in case of ZCR, TH=0.

E. Reconstruction of Spectral Entropy for VAD

Reconstruction of spectral entropy gives margin setting threshold value and stresses speech section. In this paper, there are various feature parameters for VAD as follows.

Feature 1 : En
Feature 2 : En×LogEn
Feature 3 : En × (MAX_ZCR - ZCR)
Feature 4 : En × LogEn × (MAX_ZCR - ZCR)
Feature 5 : En × GDF of Speech Entropy
Feature 6 : En×GDF of Speech Entropy × LogEn

En : Entropy
LogEn : Log Entropy
GDF : Gasussian Distribution Function
MAX_ZCR : Maximum Zero Crossing Rate

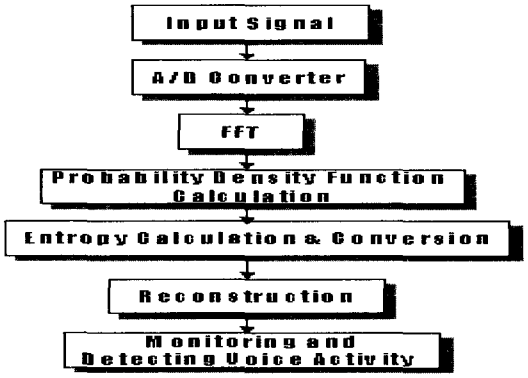


Fig. 1 The process of spectral entropy

Gaussian distribution function of Feature 5, 6 which is applied by gaussian distribution function of spectral entropy and made by speech sample previously is to emphasize speech section. The applied gaussian distribution function is followed.

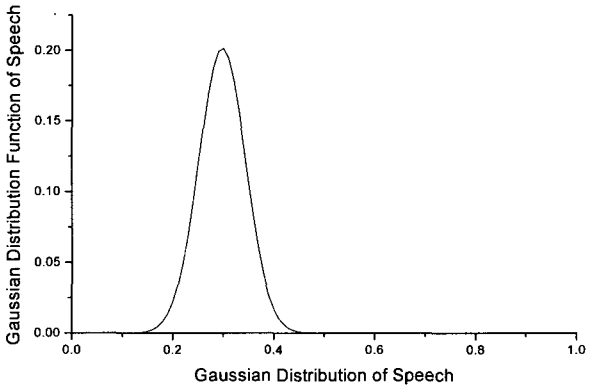


Fig. 2 Gaussian distribution of speech entropy

III. EXPERIMENTAL RESULTS

A. Data Base

In experiment this paper uses NOISE-92[8] database which has various noise, 19.98 KHz - 16 bit and filtering anti-aliasing, and that is changed to 16 KHz - 16 bit.

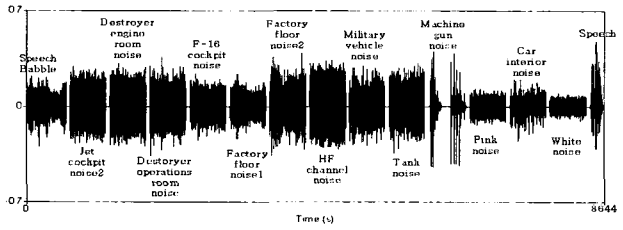
Table 1 NOISEX-92

No.	Noise Types
1	Speech babble
2	Jet cockpit noise2
3	Destroyer engine room noise
4	Destroyer operators room noise
5	F-16 cockpit noise
6	Factory floor noise1
7	Factory floor noise2
8	HF channel noise
9	Military vehicle noise
10	Tank noise
11	Machine gun noise
12	Pink noise
13	Car interior noise
14	White noise

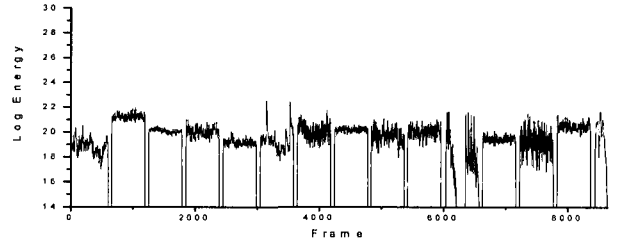
B. Experimental results and inquiry

Figure 3 is sample data for estimating, which compares above-mentioned features, section 2.2, with energy.

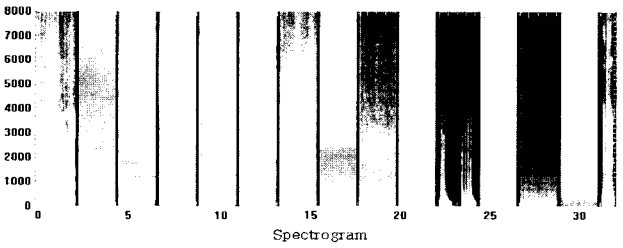
Figure 4 shows averages of reconstructed features in each sample data during five second. Features and index of horizontal axis are section 2.2 and Table 1, respectively. All feature are normalized as speech for comparing energy. We verified that energy is difficult to set threshold value but reconstructed features are easily to set threshold value. Index 15 is speech and feature 5 is excellent.



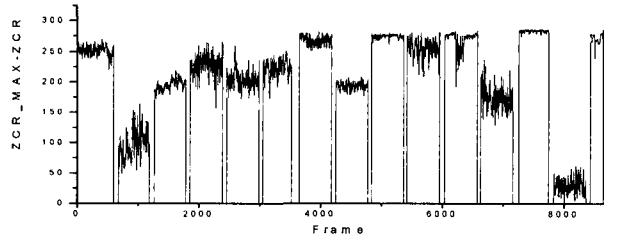
(a) Source signal



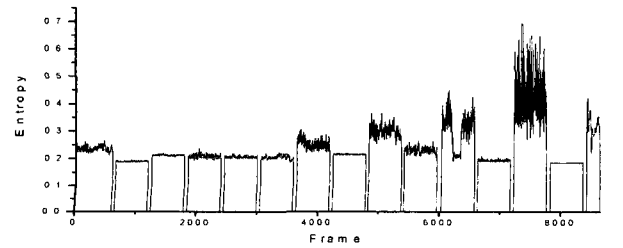
(b) Log energy



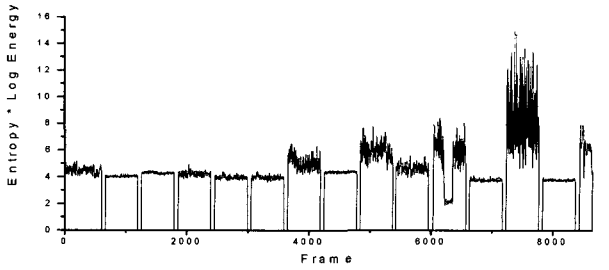
(c) Spectrogram



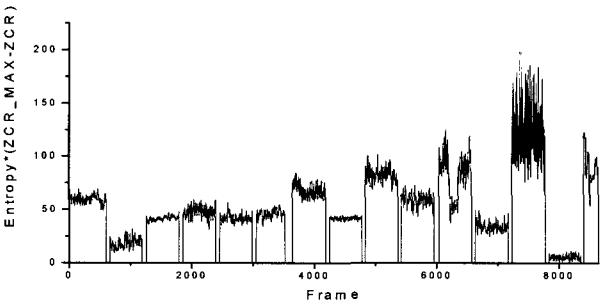
(d) ZCR_{MAX} - ZCR



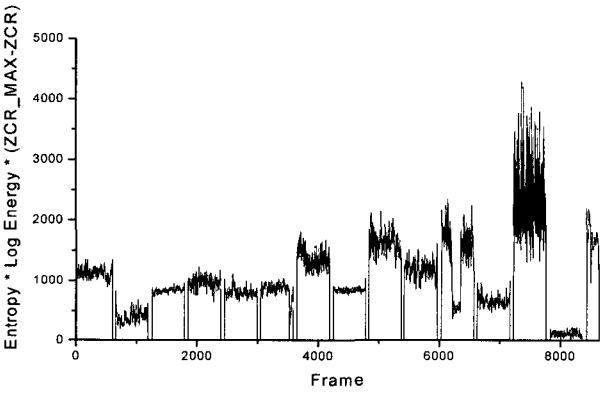
(e) Entropy



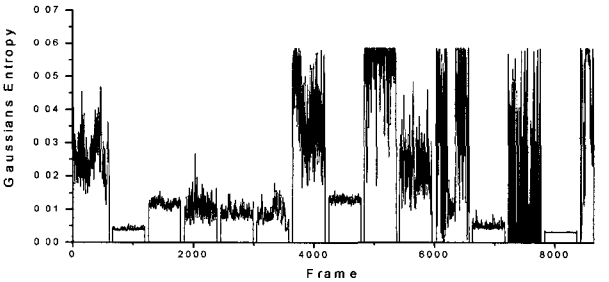
(f) Entropy × Log energy



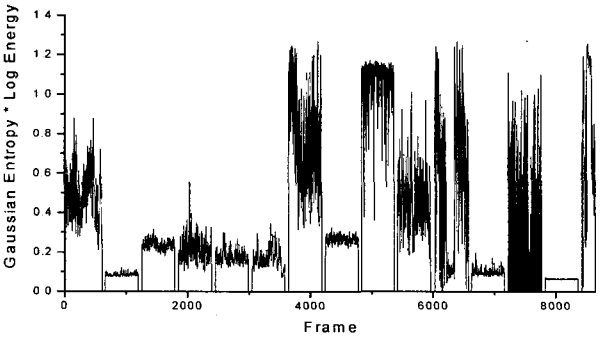
(g) Entropy × (ZCR_{MAX} - ZCR)



(h) Entropy × Log energy × (ZCRMAX - ZCR)



(i) Gaussian entropy



(j) Gaussian entropy × Log energy

Fig. 3 Comparison of log energy and reconstruction feature

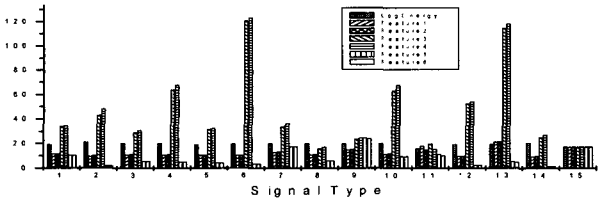


Fig 4. Average of log energy and reconstructed features

IV. CONCLUSIONS

Recently, Diverse research is studied for more trusty and robust VAD based on not only energy but also various noise and audio sound classification because of help by development of hardware process. In this paper both spectral entropy and several reconstructed features for robust VAD in noise are used, compared and analyzed with cases in energy method. Experiments showed that the reconstructed spectral entropy features were more effectual than energy based on algorithms, specially, it is confirmed that applying gaussian distribution function of spectral entropy in speech has good performance.

REFERENCES

[1] Sadaoki Furui : *"Digital Speech Processing Synthesis, and Recognition"*, MAECEL DEKKER, INC. 2001 pp. 248-249.

[2] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon: *"SPOKEN LANGUAGE PROCESSING"*, Prentice Hall 2001. pp. 120-130.

[3] Nikos Doukas, Patrick Naylor and Tania Stathaki : "Voice Activity Detection Using Source Separation Techniques", Signal Processing Section, Proc. Eurospeech '97.

[4] L.R.Rabiner, R.W.Schafer : *"Digital Processing of Speech Signals"*, PRENTICE HALL.

[5] J.L. Shen, J.Hung, L.S.Lee : "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments", Proceeding of ICLP-98, 1998.

[6] S. McClellan and J.D. Gibson : "Variable-rate celp based on subband flatness", in IEEE Transactions on Speech and Audio-Processing, vol. 5, pp. 120-130, 1997.

[7] J.Sohn and W.Sung : "A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation", in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 356-368, 1998.

[8] J.D. Hoyt and H. Wechsler : "Detection of human speech in structured noise" in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 237-240, 1994.

[9] http://spib.rice.edu/spib/select_noise.html

[10] Jianping Zhang, Wayne Ward, Bryan Pellom, "Phone Based Voice Activity Detection Using Online Bayesian Adaptation with Conjugate Normal Distributions", ICASSP'2002, Orlando Florida, May 2002.

[11] E.Neimer, R.Goubran and S.Mahmould, "Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain", IEEE Trans. Speech and Audio Proc., 9(3), pp. 217-231, 2001.

[12] R. Sarikaya and J. Hansen, Robust Speech Activity Detection in the Presence of Noise, ICSLP, Sydney, 1998.

[13] J.Sohn, N. Kim and W. Sung, "A Statistical Model-Based Voice Activity Detection", IEEE Signal Proc. Lett., 6(1), pp. 1-3, 1999.

[14] S. Tanyer and H. Ozer, "Voice Activity Detection in Nonstationary Noise", IEEE Trans. Speech and Audio Proc.,8(4), pp. 478-482, 2000.

[15] Mak, B., Jungua, J.-C., and Reaves, B., "A robust speech/non speech detection algorithm using time and frequency-based features", IEEE ICASSP, vol. I, pp. 269-272, 1992

Hag-Yong Han
Received B.S and M.S degrees of electronic engineering in 1994 and 1998 respectively, from Dong-A University. And He is a Ph. D. candidate currently at Dong-A University. In 2001 he joined the Easy Harmony CO., LTD. in Busan, Korea, where is currently an Chief of R&D. His research interests are in Biometrics, Pattern Recognition and DSP Applications.

Kwang-Seok Lee
Received B.S. and M.S. degrees of electronic engineering in 1983 and 1985 respectively, from Dong-A University. And Ph.D. from Dong-A University, in 1992. In 1995, He joined the JinJu National University in JinJu, Korea, where is currently an Professor and Industrial-University Cooperation Director. His research interests are in Intelligent System, DSP, Speech Recognition, Synthesis and Biometrics.

Si-Young Koh
Received the B.S and M.S degrees of electronic engineering in 1979 and 1983 respectively, from Young Nam University. And Ph.D. from Dong-A University, in 1992. In 1986 he joined the Kyungil University in Kyungsan, Korea, where is currently an Professor. His research interests are in Bio Signal Processing, Speech Signal Processing.

Kang-In Hur
Received the B.S and M.S degrees of electronic engineering in 1980 and 1982 respectively, from Dong-A University. And Ph.D. from Kyung Hee University, in 1990. In 1984 he joined the Dong-A University in Busan, Korea, where is currently an Professor. His research interests are in DSP, Speech Recognition, Synthesis and Neural Networks.