Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 85-88.

Voice Activity Detection based on Combination of Weighted Sub-band Features using Auto-Correlation Function

Kun-Ching Wang[1], *Chiun-Li Chin*[2], *Yi-Hsing Tsai*[3]

[1] Department of Information Technology & Communication, Shin Chien University
[2] Department of Applied Information Sciences, Chung Shan Medical University
[3] Information & Communications Research Laboratories, Industrial Technology Research Institute
wkc0224@seed.net.tw[1]

Abstract

This paper shows the voice activity detection (VAD) based on combination of weighted sub-band features using auto-correlation function. According to the fact that the noise corruption on each sub-band is different from each other, so the estimated signal to noise ratio (SNR) is employed to weight utility rate of each frequency sub-band. Furthermore, a strategy of sub-band features combination is used to integrate all of weighted sub-band auto-correlation function feature parameter and to develop the combined feature parameter. Experimental results demonstrate that the proposed VAD achieves better performance than existing standard VADs at any noise level.

Index Terms: voice activity detection, auto-correlation, wavelet packet transform, sub-band weighting, feature combination

1. Introduction

Determination of the instances of presence of speech periods in a given signal is an important problem in many fields of speech processing, usually known as Voice Activity Detection (VAD). Voice Activity Detection, in fact, is becoming increasingly important and relevant in modem telecommunication and speech enhancement systems such as speech coding, speech recognition, hands-free telephony, audio conferencing and echo cancellation [1]-[3].

Traditionally, VAD algorithms use short-term energy, zero-crossing rate and LPC coefficients [4] as feature parameters. Cepstral features [5], formant shape [6], and least-square periodicity measure [7] are also some of the more recent metrics used in VAD designs. In addition, a time-frequency parameter is introduced in [8] and spectral entropy in [9]. Temporal power envelope [10] and spectral divergence [11] are inspired to develop novel features. For spectral divergence, it uses a long-term speech window instead of instantaneous values of the spectrum to track the spectral envelope. The estimation of the so-called long-term spectral envelope (LTSE) is used to classify the difference between speech and non-speech segments. Recently, some sub-band approaches based on wavelet have been proposed in [12-15]. In [14], Chen et al. employed the wavelet-packet transform to decompose the input speech signal into critical sub-band signals.

This paper shows a novel voice activity detection (VAD) algorithm using weighted auto-correlation function in sub-band feature combination. The sub-band feature is developed by the auto-correlation function (ACF) defined in the wavelet packet domain. Three-level wavelet decomposition is further divided into four non-uniform sub-bands. Besides, the voiced or vowel speech sounds have a stronger periodic property than unvoiced sounds and noise signals, and this property is concentrated in low frequency sub-bands. The low frequency sub-bands have high resolution in order to enhance the periodic property by decomposing only the low sub-band in each level. In fact, the noise corruption on each sub-band is different from each other. To develop a VAD scheme that is robust against various kinds of noise, a combination of sub-band feature is include. So, each sub-band weighting coefficient is adjusted by the signal to noise ratio (SNR) through a sigmoid function. Finally, the combined feature parameter is developed through sub-band features combination where the sub-band weighting coefficient adjusting sub-band ACF.

2. The proposed VAD algorithm

The architecture of the proposed VAD method is shown in Figure 1. The input noisy speech is first decomposed into four sub-bands through three-level wavelet decomposition. Observing Figure 1, the posterior sub-band SNR, $SNR_{pot}(\xi, m)$, is derived from sub-band noise estimate. The sub-band signal ACF (called as SSACF) is determined from ACF defined in the sub-band domain. Through Mean-Delta processing, the Mean-Delta SSACF (called as MDSSACF) is achieved on each sub-band. In addition, these parameters were also proposed in [24]. The weighting coefficient, $w^{\xi}(m)$, is adapted by sigmoid function against $SNR_{pot}(\xi, m)$. The output is summation of the values of four weighted MDSSACF feature parameters. After a dynamic thresholding, the VAD output will show whether speech or noise-dominated frame.

2.1. Sub-band feature parameter

The well-known "Auto-Correlation Function", $R(k)$, is used to measure the self-periodic intensity of sub-band signal sequences [17] and is shown below:

$$R(k) = \sum_{n=0}^{p-k} s(n)s(n+k), \quad k = 0,1,.....p \,, \tag{1}$$

where $s(n)$ is a discrete-time signal. p is the length of ACF and k denotes the shift of the sample.

In this subsection, the ACF will be defined in the sub-band domain and be called the "Sub-band Signal Auto-Correlation Function (SSACF)." Figure 2 shows that sub-band decomposition by employing the structure of three-level wavelet decomposition (WD). By using 3-level WD, we can divide the speech signal into four non-uniform sub-bands. To determine the value of the periodic intensity of sub-band

signals, a method of Mean-Delta [18] is applied here to SSACF on each sub-band.

First, a measure similar to delta cepstrum evaluation is used to estimate the periodic intensity of the SSACF, namely, the "Delta Sub-band Signal Auto-Correlation Function (DSSACF)," as shown below:

$$\dot{R}_M(k) = \frac{\sum_{m=-M}^{M} mR(k+m)}{\sum_{m=-M}^{M} m^2}, \quad (2)$$

where $\dot{R}_M$ is the DSSACF over an M-sample neighborhood.

For a particular frame, it is computed by using only the frame's SSACF (intra-frame processing), while the delta cepstrum is computed by using cepstrum coefficients from neighboring frames (inter-frame processing). It is observed that the DSSACF value is almost similar to the local variation over the SSACF.

Second, the delta of the SSACF is averaged over an M-sample neighborhood $\bar{R}_M$, where the mean of the absolute values of the DSSACF (MDSSACF) is given by

$$\bar{R}_M = \frac{1}{N_b}\sum_{k=0}^{N_b-1}\left|\dot{R}_M(k)\right|, \quad (3)$$

where N_b indicates the length of the sub-band signal.

2.2. Sub-band weighting coefficient

In order to determine the utility rate of each sub-band, the estimated SNR estimation is required. A posterior SNR, $SNR_{pot}(\xi,m)$, is formulated as:

$$SNR_{pot}(\xi,m) = 10\cdot\log_{10}\frac{WE(\xi,m)}{\tilde{\sigma}_w^2(\xi,m)}, \quad (4)$$

where $WE(\xi,m)$ means the wavelet energy of the observed noisy speech signal. $\tilde{\sigma}_w^2(\xi,m)$ is the estimated noise power for current frame.

Observing the Eq.(4), we know that the sub-band noise power spectrum has to be es-timated while determining the value of a posterior SNR. In order to estimate the noise-level quickly and accurately, various methods [19] were proposed for tracking the minimum of the noisy speech power spectrum energy over a fixed search window length. To speed up the determination of local minimum of noisy speech spectrum over a search window size, Doblinger's efficient method [20] is used here, which is not constrained by any window length to update noise spectrum estimate.

$$\begin{aligned}&\text{If } WE_{\min}(\xi,m-1) < WE(\xi,m),\\ &\text{then } WE_{\min}(\xi,m) = \gamma\cdot WE_{\min}(\xi,m-1)\\ &\quad + \frac{1-\gamma}{1-\beta}\left[WE(\xi,m) - \beta\cdot WE(\xi,m-1)\right],\\ &\text{else } WE_{\min}(\xi,m) = WE(\xi,m),\end{aligned} \quad (5)$$

where $WE_{\min}(\xi,m) = \tilde{\sigma}_w^2(\xi,m)$ denotes the local minimum of wavelet energy of the noisy speech. β and γ are constants determined experimentally.

Consequently, the $SNR_{pot}(\xi,m)$ parameter will help us sense how much the current sub-band is corrupted by noise. After the value of a posterior SNR obtained, the sub-band weight coefficient, $w^{\xi}(m)$, is calculated by applying a sigmoid function to sub-band SNR as

$$w^{\xi}(m) = \frac{1}{1+\exp\left[-0.5\cdot\left(SNR_{pot}(\xi,m) - \eta^{\xi}(m)\right)\right]}. \quad (6)$$

Therefore, we will use this information to weight each sub-band. Figure 2 shows the plots of the weighting coefficients from Eq.(6) depending on η.

In the Fig. 2, η is a center-offset of the sigmoid function. Under the fixed value of a posterior SNR, the weighting coefficient decrease toward to zero when η is increasing. In addition, the values of the all parameter η are determined by experimental test. According the fact that the speech components almost focus in low-frequency sub-band, let the sigmoid function with largest η (such as $\eta = 20$) locate to highest frequency sub-band (such as D1 sub-band). On the contrary, let the sigmoid function with smallest (such as $\eta = 5$) locate to lowest frequency sub-band (such as A3 sub-band).

2.3. Dynamic thresholding used in VAD decision

In order to determine the boundary of voice-activity accurately, a scheme of dynamic thresholding is used in the VAD decision. An adaptive threshold value is derived from the statistics of the combined MDSSACF parameter during a noise-only frame, and the VAD decision process recursively updates the threshold by using the mean and variance of the values of the parameter.

Assuming the first five frames as noise-only frame, we compute the initial noise mean and variance within those frames. Then, the thresholds for the speech and noise are given as follows [21]:

$$Th_s = \mu_n + \alpha_s\cdot\sigma_n, \quad (7)$$

$$Th_n = \mu_n + \alpha_n\cdot\sigma_n, \quad (8)$$

where Th_s and Th_n indicate the speech threshold and noise threshold, respectively. μ_n and σ_n represent the mean and variance of the values of the combined MDSSACF parameters, $Comb(m)$, respectively. Similarly, α_s and α_n are thresholding coefficients for speech threshold and noise threshold, respectively

The VAD decision rule is defined as follows:

$$\begin{aligned}&\text{if } (Comb(m) > Th_s) \quad VAD(m)=1\\ &\text{else if } (Comb(m) < Th_n) \quad VAD(m)=0;\\ &\text{else } VAD(m)=VAD(m-1).\end{aligned} \quad (9)$$

If the detection result shows a noise period, the mean and variance of the values of the combined feature parameters are updated by as follows:

$$\mu_n(m) = \varepsilon\cdot\mu_n(m-1) + (1-\varepsilon)\cdot Comb(m), \quad (10)$$

$$\sigma_n(m) = \sqrt{[Comb_{buffer}^2(m)]_{mean} - [\mu_n(m)]^2}, \quad (11)$$

$$\begin{aligned}[Comb_{buffer}^2(m)]_{mean} = \varepsilon\cdot[Comb_{buffer}^2(m-1)]_{mean}\\ + (1-\varepsilon)\cdot Comb^2(m).\end{aligned} \quad (12)$$

Here, $[Comb_{buffer}^2(m-1)]_{mean}$ is a mean of the buffer of the value of combined feature parameter during a noise-only frame. We then update the thresholds by using the updated mean and variance of the values of the parameters.

3. Simulation results

In order to evaluate the performance of the proposed VAD, the recordings of each sentence were spoken by 40 native speakers

(20 males and 20 females) and were sampled at the rate of 8 KHz with 16-bit resolution. The noise signals were all taken from the noise database NOISEX-92 [22]. The speech analysis window was the Hamming window and window size, L_{frm}, equal to 32ms. We average speech/non-speech hit rate (HR1/HR0) for each type of noise over -5 to 30 dB for comparison. The speech/non-speech hit rate (HR1/HR0) is calculated as:

$$HR0 = \frac{\text{number of non-speech frames correctly classified}}{\text{number of real non-speech frames}} * 100\% \quad (13)$$

$$HR1 = \frac{\text{number of speech frames correctly classified}}{\text{number of real speech frames}} * 100\% \quad (14)$$

So, the hit rates as a function of the false alarm rates are shown as:

$$FAR0 = 100 - HR1; \quad (15)$$

$$FAR1 = 100 - HR0. \quad (16)$$

In Table 1, the proposed VAD obtains the best behavior in detecting speech with 93.18% average value for SNR levels from 30 to -5 dB under various types of noise (The parameters used for the proposed VAD were: $\alpha_s = 40$, $\alpha_n = 10$, $\beta = 0.7$, $\gamma = 0.5$, $\varepsilon = 0.6$ and $L_{frm} = 256$), while the G.729, AMR-2, LTSE and wavelet sub-band based VADs yield 85.81 %, 85.72 %, 88.22% and 91.82 %, respectively. On the other hand, the proposed VAD provides the best the non-speech hit rate with 78.98% average value over other VADs. Table 2 indicates the proposed VAD has lowest false alarm rates for SNR levels from 30 to -5 dB. It can be noted from Table 1 to Table 2 that the average speech/non-speech hit rates and false alarm rate are superior to other VADs..

4. Conclusions

The paper shows a novel VAD scheme using weighted sub-band ACF with sub-band feature combination. The proposed VAD is developed on a strategy of sub-band feature combination that incorporates sub-band weighting method depended on sub-band SNR estimation the fact that the degree of the noise corruption on each sub-band is different from each other. Our experimental results show that the MDSSACF depends only on the amount of variation of the normalized ACF, not on the energy level of the signal. In addition, the combined feature parameter derived from the combination of sub-band weighted MDSSACF is superior to other standards VADs. We prove that the proposed combined parameter can be successfully used in the real noisy environments.

5. Acknowledgements

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 99-2221-E-158 -006.

6. References

[1] Rabiner L. and Juang B. H., Fundamentals of Speech Recognition, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] Freeman D. K., Cosier G., Southcott C. B., and Boyd I., "The voice activity detector for the pan European digital cellular mobile telephone service," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, 369-372, 1989.

[3] Telecommunications Industry Association, Enhanced variable rate codec, speech ser-vice option 3 for wideband spread spectrum digital systems, TIA doc. PN-3292, 1996.

[4] Rabiner L. R. and Sambur M. R., "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in Proc. Int. Conf. Acoustics, Speech, Signal Processing: 323-326, 1977.

[5] Haigh J. A. and Mason J. S., "Robust voice activity detection using cepstral features," in IEEE TEN-CON.: 321-324, 1993.

[6] Hoyt J. D. and Wechsler H., "Detection of human speech in structured noise," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, 237-240, 1994.

[7] Tucker R., "Voice activity detection using a periodicity measure," in Proc. Inst. Elect. Eng.: 377-380, 1992.

[8] Wu G. D. and Lin C. T., "Word boundary detection with Mel-scale frequency bank in noisy environment," IEEE Transactions on Speech and Audio Processing, 8(5): 541-553, 2000.

[9] B. F. Wu and K. C. Wang, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," IEEE Transactions on Speech and Audio Processing, 13(5): 762-775, 2005.

[10] Marzinzik M. and Kollmeier B., "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Process., 10(2): 109-118, Feb. 2002.

[11] Ramirez J., Segura J. C., Benitez C., Torre A. D. L., and Rubio A., "Efficient voice activity detection algorithms using long-term speech information," Speech Commun., 42(3): 271-287, Apr. 2004.

[12] Chen S. H., Wu H. T., Chang Y. and Truong T. K., "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator," Pattern Recognition Letters, 28(11): 1327-1332, August 2007.

[13] Juang C. F., Cheng C. N. and Chen T. M., "Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network," Expert Systems with Applications, 36: 321-332, 2009.

[14] Chen S. H. and Wang J. F., "A wavelet-based voice activity detection algorithm in noisy environments," International Conference on Electronics, Circuits and Systems, vol.3, pp. 995- 998, 2002.

[15] Stegmann, J. and Schroder, G., "Robust voice-activity detection based on the wavelet transform," IEEE Workshop on Speech Coding for Telecommunications Proceeding: 99-100, 1997.

[16] Benyassine A., Shlomot E., and Su H., "ITU-T recommendation G.729, annex B, a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data supplications," IEEE Commun. Mag.: 64-72, 1997,

[17] Rabiner L. R., "On the use of autocorrelation analysis for pitch detection," IEEE Transactions on Acoustics, Speech, and Signal Processing, 25(1): 24-33, February 1977.

[18] Ouzounov A., "A Robust Feature for Speech Detection," Cybernetics and Information Technologies, 4(2): 3-14, 2004.

[19] Martin R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process, 9(5): 504-512. 2001.

[20] Doblinger G., "Computationally efficient speech enhancement by spectral minima tracking in subbands," Proc. Eurospeech 2: 1513-1516, 1995.

[21] Gerven S. V. and Xie F., "A comparative study of speech detection methods," In Proceedings of Eurospeech, 3: 1095-1098, 1997.

[22] Varga and Steeneken H. J. M., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., 12: 247-251, 1993.

[23] ETSI EN 301 708, Digital cellular telecommunications systems (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description (GSM 06.94 version 7.1.1 Release 1998), V7.1.1, 1999.

[24] Wu B. F. and Wang K. C., "Speech Activity Detection Based on Auto-Correlation Function Using Wavelet Transform and Teager Energy Operator," Computational Linguistics and Chinese Language Processing, 11(1): 87-100, March 2006.

[25] GSM 06.94. (1999, Feb.) Digital cellular telecommunication system (Phase 2+); voice activity detector VAD for adaptive multi rate (AMR) speech traffic channels; general description. ETSI, Tech. Rep. V.7.0.0.

Table 1. Average speech/non-speech hit rates under SNR levels ranging from 30 to -5 dB

Noise type	SNR(dB)	VAD Method									
		G.729B [16]		AMR-2 [23]		LTSE [11]		wavelet-based [14]		**Proposed**	
		HR0(%)	HR1(%)	HR0(%)	HR1(%)	HR0(%)	HR1(%)	HR0(%)	HR1(%)	HR0(%)	HR1(%)
White	30	92.6	80.5	94.6	83.5	96.4	92.5	95.6	85.6	98.9	96.3
	10	82.4	45.3	85.2	50.6	90.2	74.2	93.1	51.5	95.3	78.6
	-5	64.2	30.4	66.4	38.7	74.3	56.5	82.5	43.9	85.4	70.4
Factory	30	94.1	82.9	96.6	88.9	97.7	91.5	98.6	89.3	99.6	92.5
	10	91.3	50.7	92.5	55.4	92.4	69.4	95.3	58.6	97.3	79.4
	-5	89.2	35.6	74.5	43.6	78.1	55.3	88.4	45.7	88.9	66.3
Babble	30	93.3	82.6	94.4	85.6	95.2	88.4	98.2	86.3	98.1	90.6
	10	89.4	56.4	89.5	60.3	90.4	67.5	93.9	59.8	93.6	72.5
	-5	75.8	46.3	77.8	51.5	79.3	54.4	80.6	50.7	81.6	64.3
Average (%)		85.81	56.74	85.72	62.01	88.22	72.19	91.80	63.48	**93.18**	**78.98**

HR0: non-speech hit rate, HR1: speech hit rate.

Table 2. Average speech/non-speech false alarm rates for SNR levels from 30 to -5 dB

False-alarm rates	G.729B [16]	AMR-2 [23]	LTSE [11]	wavelet-based [14]	**Proposed**
FAR0(%)	14.19	14.28	11.78	8.20	**6.82**
FAR1(%)	43.26	37.99	27.81	36.52	**21.02**

FAR0: false alarm rate for non-speech, FAR1: false alarm rate for speech (FAR0=100-HR1; FAR1=100-HR0)

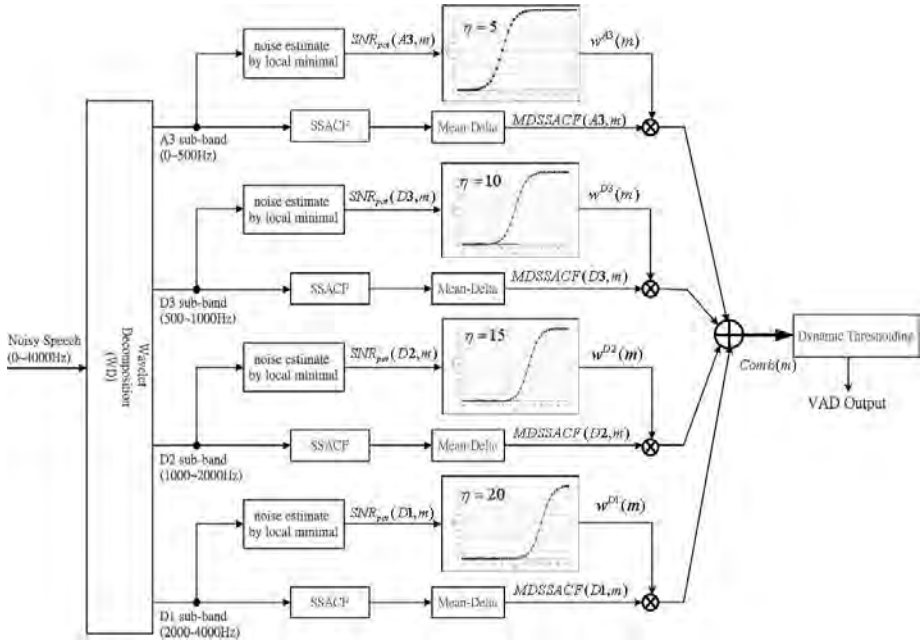

Fig. 1. The architecture of proposed VAD method

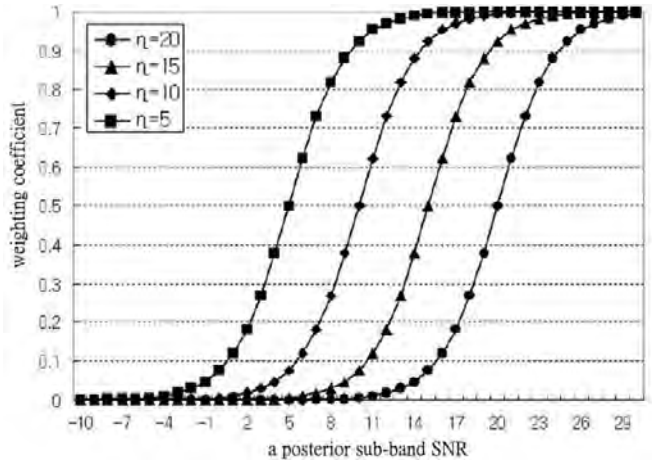

Fig. 2. The plots of weights coefficients against a posterior sub-band SNR under variable η