Voice Activity Detection G729B Improvement Technique Using K-Nearest Neighbor Method

Suryo Adhi Wibowo [1], *Koredianto Usman* [2]
Faculty of Electrical and Communication Engineering, Institut Teknologi Telkom
Bandung, Indonesia
(Tel: +62-22-756-4500, Fax: +62-22-756-2721)
Emails: [1] suryoadhi.wibowo@gmail.com [2] kru@ittelkom.ac.id

Abstract

ITU-T G.729B describes about Discontinuous Transmission (DTX) in the system. Discontinuous Transmission (DTX) is known as a transmission method that transmits a few bits of voice under influenced of background noise or un-voiced signal. This characteristic gives advantageous in voice activity detection to classify signal. In this paper we improve the performance of voice activity detection G729B using K-Nearest Neighbor method for voice or unvoiced classification. Computer simulations showed that this method has performance better than the original VAD G729B in all tested noise conditions.

Index Terms: VAD G729B, K-Nearest Neighbor, Discontinuous Transmission, Classification

1. Introduction

Discontinuous Transmission (DTX) is one of method to improve bandwidth efficiency on audio transmission. To have a high efficiency we need accurate performance for voice activity detection. The original VAD G729B computes decisions using four parameters which *Full Band Energy difference, Low Band Energy difference, Spectral Distortion and Zero Crossing difference.* We use K-Nearest Neighbor to improve performance of the original VAD G729B, the parameters from the original VAD G729B used as input of K-Nearest Neighbor. Decision from the original 'pure' VAD G729B is used as target in K-Nearest Neighbor.

2. Method and approach

2.1 Audio transmission

As we know that communication system transmits data such as video, voice / audio, data, etc from one to the other. In this paper, we focus on audio transmission field. Audio transmission usually consists of noise such as *babble noise, street noise, machinery noise, car interior noise, music noise etc*. This noise will influence of audio transmission especially when we converse over mobile communication in noisy environment. Figure 1 illustrates clean speech waveform in audio transmission.

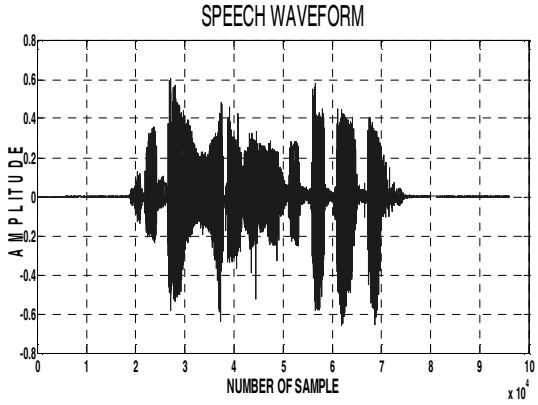
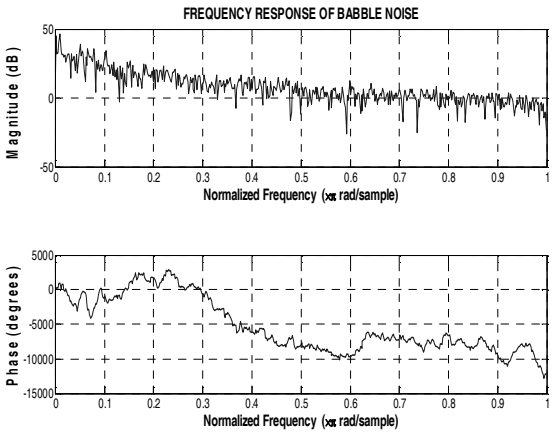
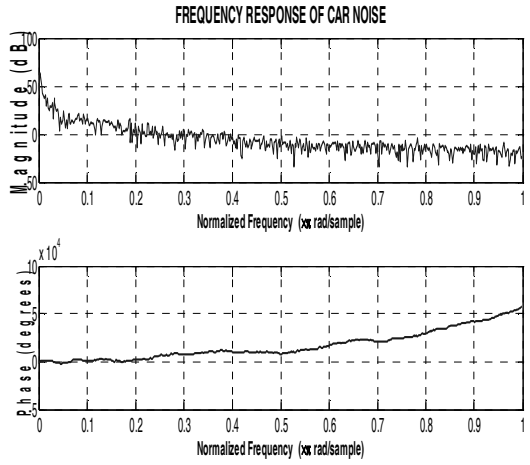


Figure 1: *Clean speech waveform*

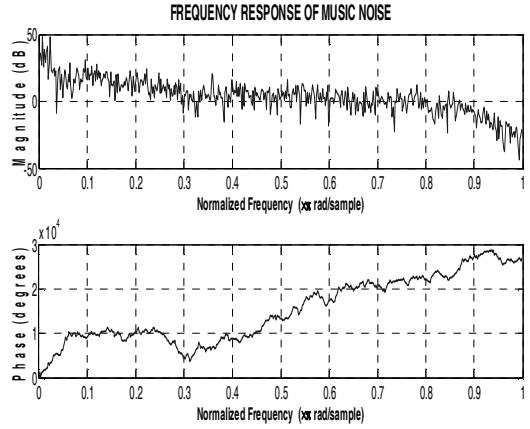
Figure 2 shows the characteristic spectrum of *babble noise, car interior noise, music noise and street noise*.



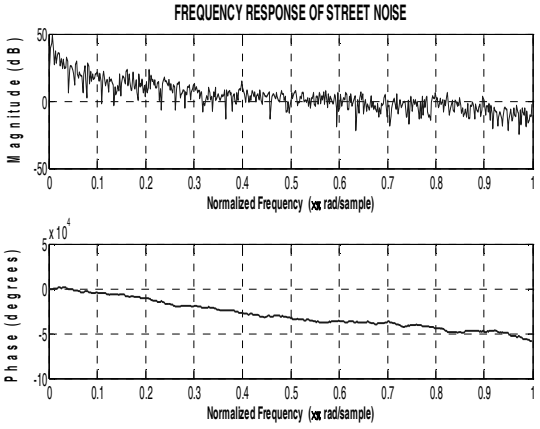
(a)



(b)



(c)



(d)

Figure 2: *Magnitude and phase of (a) Babble Noise, (b) Car Interior Noise, (c) Music Noise, (d) Street Noise.*

2.2 Voice activity detection G729B

Voice activity detection based on G729B is input signal processed frame by frames. A standard frame size from this algorithm is 240 samples.

In this simulation, we use a 3 second signals with sampling frequency of 32000 samples in each second. Therefore we have 96000 samples in total. As the frame size is 240 samples and frame overlap is 160, we obtained number of frames to be 1200 frames. Each frame of input signal is passed to high pass filter that serves as a precaution against undesired low-frequency components [2]. A second order pole/zero filters with a cut-off frequency of 140 Hz have been used as HPF. Both the scaling and high-pass filtering are combined by dividing the coefficients at the numerator of this filter by 2 [2]. The resulting filter transfer function is given by:

$$H(z)=\frac{0.4636718-0.92724705\cdot z^{-1}+0.46363718\cdot z^{-2}}{1-1.9059465\cdot z^{-1}+0.9114024\cdot z^{-2}} \quad (1)$$

After filtering, the next step is calculates the autocorrelations. Windowing is used to calculate four parameters (*full band energy, low band energy, zero crossing rates, and line spectral frequency*). We follow ITU-T G729B use hamming cosine window to smoothing the signal. The hamming cosine window is given in the following equation:

$$W=\begin{cases}0.54-0.46\cdot\cos(2\cdot\pi\cdot n), n=0,...,199\\ \cos(\frac{2\cdot\pi\cdot(n-200)}{159}), n=200,...,239\end{cases} \quad (2)$$

The next step is to compute the four parameters which will be used as input.

Full Band Energy:

$$E_f=10\cdot\log_{10}[\frac{R(0)}{N}] \quad (3)$$

After computing Full Band Energy, we compute the Full Band Energy difference. Equation to compute Full Band Energy Difference is

$$\Delta Ef=\overline{Ef}-Ef \quad (4)$$

Low Band Energy:

$$E_l=10\cdot\log_{10}[\frac{h^T Rh}{N}] \quad (5)$$

Low Band Energy Difference will also compute uses this equation:

$$\Delta El=\overline{El}-El \quad (6)$$

Zero Crossing Rate

$$Zc = \frac{M}{2} \cdot \sum_{i=0}^{M-1} [|\operatorname{sgn}[x(i)] - \operatorname{sgn}[x(i-1)]|] \qquad (7)$$

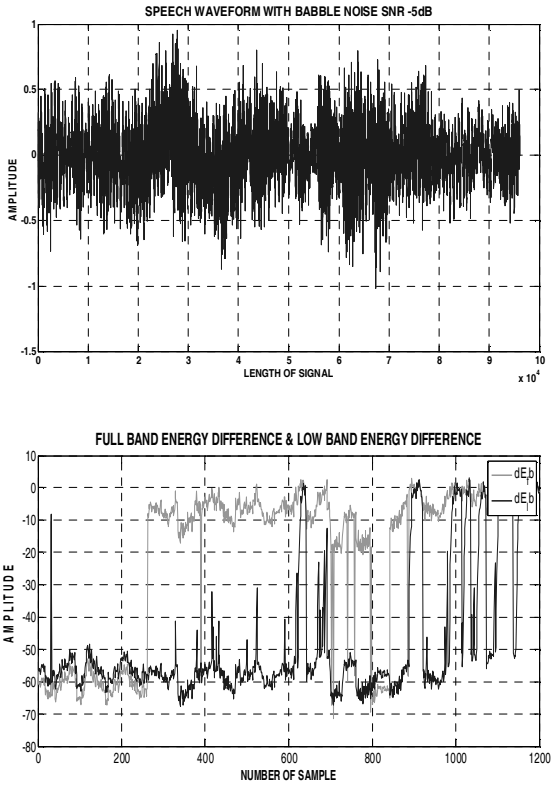
And the equation to compute Zero Crossing Difference is

$$\Delta Zc = \overline{Zc} - Zc \qquad (8)$$

Spectral Distortion

$$\Delta S = \sum_{i=1}^{p} (LSF - \overline{LSF})^2 \qquad (9)$$

After four parameters has been calculated (*full band energy difference, low band energy difference, zero crossing rate difference and spectral distortion*), it used as input to K-Nearest Neighbor for classifying. For more clear about our explanation, we illustrate in figure 3.



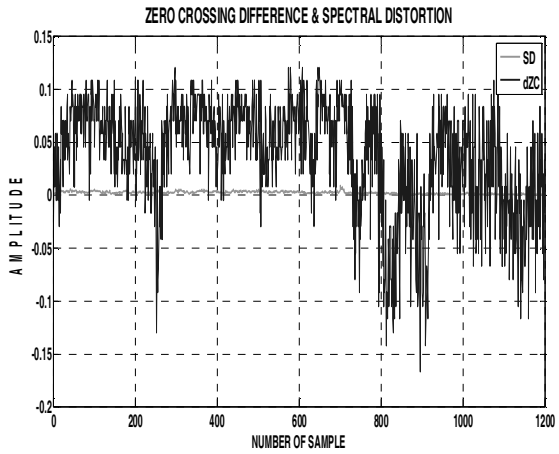


Figure 3: *Speech with SNR -5dB and the spectral of VAD G729B parameter.*

2.3 K-Nearest Neighbor

K-Nearest Neighbor algorithm has a strong point which it can give good decision to classify from noisy training data and more effective if training data has big size.

The step of K-Nearest Neighbor algorithm:

1. Define 'K' parameter.
2. Compute distance between input and all training sample. This computation use Euclidean distance method.

$$\sqrt{\sum_{i=1}^{n} (Pi - Qi)^2} \qquad (10)$$

   The descriptions from the equation 10 are:
   n = number of sample data
   P = Input data-i of testing
   Q=Input data-i of training
3. Make distance group and arrange nearest neighbor based on minimum distance 'K'.
4. Make group from value of nearest neighbor as Y.
5. Select the most frequent value from nearest neighbor as a prediction value to next data.

2.4 The VAD based on K-Nearest Neighbor

K-Nearest Neighbor algorithm is used in this VAD, which the original four VAD G729B parameters as input. Figure 4 illustrate the flowchart of K-Nearest Neighbor method.

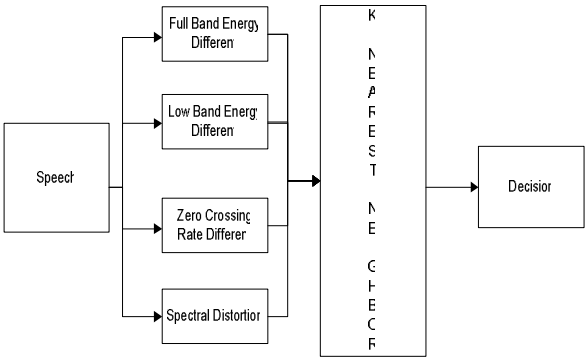


Figure 4: *VAD G729B K-Nearest Neighbor block diagram.*

Size of each the input is 1 x 1200. Because of the number of input parameter are four, so size in the input of each speech are 4 x 1200. Total of input matrix are 4 x (1200 x number of data). The number of target is 1x (1200 x number of data). We use 'K' parameter to be one.

3. Performance evaluation

3.1 Clean Speech Corpus

For evaluation purposes a clean speech corpus databases has been developed. The speaker subjects comprised of 10 males and 10 females with various speaking style. To record the speech, we only use one microphone.

3.2 Background Noise Corpus

In order to evaluate performance of voice activity detection G729B K-Nearest Neighbor method, we use background noise that it recorded from noisy area. This noise can be classified in to four background noise types: *babble noise, interior car noise, music noise and street noise.*

3.3 Performance measurement

Performance evaluation of this method are *speech detection error rate* (SDER), *noise detection error rate* (NDER) and *overall detection error rate* (OVER).

$$SDER = \frac{\sum Miss - DetectedSpeechFrames}{\sum SpeechFrames} x\ 100\% \qquad (11)$$

$$NDER = \frac{\sum Miss - DetectedNoiseFrames}{\sum NoiseFrames} x100\% \qquad (12)$$

$$OVER = SDER \frac{\sum SpeechFrames}{\sum Frames} + NDER \frac{\sum NoiseFrames}{\sum Frames} \qquad (13)$$

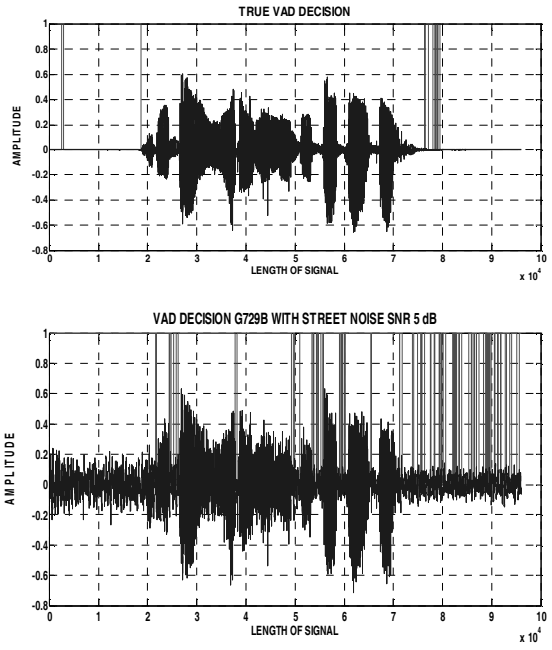
A low value of SDER, NDER and OVER indicates a good performance. In this method, noisy signal are constructed by artificially adding background noise to clean speech signal. We compute energy from clean speech signal and energy from background noise. After that, we compute level clean speech signal and assume active speech level at 20dB. The last, clean speech signal and background noise are combined in to one signal use scaling and desired level SNR.

4. Simulation results and discussion

We compare performance of voice activity detection G729B using K-Nearest Neighbor method and the original voice activity detection G729B. *Table 1: Performance of VAD G729B versus VAD G729B using K-Nearest Neighbor* shows the simulation results for -5 dB, 0 dB, 5 dB, 10 dB and 15 dB SNR level. This performance all tested noise condition.

5. Conclusion

This paper describes a technique improvement performance of voice activity detection G729B using K-Nearest Neighbor method. From comprehensive computation, we can see comparisons performance voice activity detection G729B using K-Nearest Neighbor is better than the original voice activity detection G729B in all tested noise condition (*babble noise, car interior noise, music noise, and street noise*) from level SNR -5 dB until 15 dB.



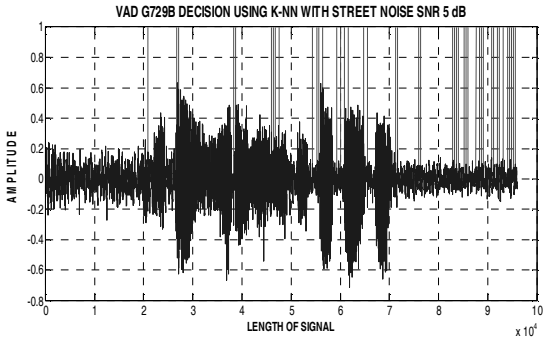


Figure 5: *Example of Decision VAD G729B versus VAD G729B Using K-Nearest Neighbor with street noise SNR 5dB*

Table 1: *Performance of VAD G729B versus VAD G729B using K-Nearest Neighbor*

SNR	Algorithm	Babble Noise (%)			Street Noise (%)		
		SDER	NDER	OVER	SDER	NDER	OVER
-5	G729B	**16.6**	74.6	48.4	21.4	69.4	47.8
	G729B+KNN	18.1	**14.8**	**16.6**	**14.6**	**16.6**	**17.3**
0	G729B	**13.4**	77	48.4	17.8	69.2	46
	G729B+KNN	19.5	**14.6**	**17.5**	**13**	**12.8**	**12.3**
5	G729B	**11.8**	74.8	46.2	**13.4**	68.8	43.6
	G729B+KNN	19.6	**12.3**	**16.3**	14.6	**10.2**	**12.2**
10	G729B	**11.8**	75	46.2	13.2	68.8	43.6
	G729B+KNN	17	**9.3**	**14.1**	**8.1**	**14.5**	**11.1**
15	G729B	**11.8**	75	46.2	13	68.6	43.4
	G729B+KNN	13.5	**8**	**11.8**	**5.8**	**16.3**	**9.6**

SNR	Algorithm	Interior Car Noise (%)			Music Noise (%)		
		SDER	NDER	OVER	SDER	NDER	OVER
-5	G729B	8.2	89.6	53.2	22.6	71.2	49.8
	G729B+KNN	**8.1**	**12**	**9**	**20.8**	**19.1**	**19.6**
0	G729B	9.4	88.4	52.8	**17.8**	69.8	47
	G729B+KNN	**7.3**	**8**	**6.1**	20.1	**13.3**	**17.1**
5	G729B	**8**	86.2	51	**15**	64.8	42.4
	G729B+KNN	8.6	**15.6**	**8.8**	17.3	**15.1**	**15.8**
10	G729B	8	85.4	50.4	14.6	64.8	42.2
	G729B+KNN	**5.8**	**12.8**	**7.5**	**9.6**	**10.5**	**10.3**
15	G729B	8	84	49.4	14.4	64.6	42
	G729B+KNN	**5.1**	**5.5**	**4.3**	**11.5**	**9.3**	**10**

6. Acknowledgements

The author thanks to Mr. Henry Widjaja, BSc. ME for giving the times to discuss about this topic and corpus databases.

7. Reference

[1] Yiteng Huan and Jacob Benesty, Audio Signal Processing for Next Generation Multimedia Communication System, Kluwer Academic Publisher, 2004

[2] ITU-T G729B, Annex B: A silence compression scheme for G729 optimized for terminals conforming to recommendation V.70, ITU-T, 1996

[3] Way C. Chu, Speech Coding Algorithms Foundation and Evolution of Standardized Coders, Wiley Inter science, 2003

[4] Peter Noll, Digital Audio for Multimedia, Proceeding Signal Processing for Multimedia Nato Advanced Audio Institute, 1999

[5] Virginie Gilg, Christophe Beaugeant et all, Methodology for The Design of a robust Voice Activity Detector for Speech Enhancement, International Workshop on Acoustic Echo and Noise Control Japan, 2003