INTERSPEECH 2009 BRIGHTON

Application of Differential Microphone Array for IS-127 EVRC Rate Determination Algorithm

Henry Widjaja, Suryoadhi Wibowo

Department of Electrical Engineering, Institut Teknologi Telkom, Bandung, Indonesia

hry@ittelkom.ac.id, suryoadhi.wibowo@gmail.com

Abstract

Differential microphone array is known to have low sensitivity to distant sound sources. Such characteristics may be advantageous in voice activity detection where it can be assumed that the target speaker is close and background noise sources are distant. In this paper we develop a simple modification to the EVRC rate determination algorithm (EVRC RDA) to exploit the noise-canceling property of differential microphone array to improve its performance in highly dynamic noise environment. Comprehensive computer simulations show that the modified algorithm outperforms the original EVRC RDA in all tested noise conditions.

Index Terms: voice activity detection, first-order differential microphone array, EVRC rate determination algorithm

1. Introduction

Voice activity detection is a non-trivial problem when high-level noise is present, even more problematic when the noise is non-stationary or dynamic. The IS-127 EVRC rate determination algorithm compares subband-energies with a set of thresholds that are adapted according to the statistics of the background noise [1]. As with other energy-based VAD algorithms that rely on the stationarity assumption of the background noise, its performance is limited in highly dynamic noise conditions. In this study we consider a modification of the EVRC rate-determination algorithm for use with input signal acquired by a first-order differential (FOD) microphone array. Differential microphone array is known for its noise-canceling property due to the existence of pressure-differential between its elements [2]. First we shall characterize this noise-canceling property and then use it as a basis for modifying the EVRC rate-determination algorithm to take advantage of such property.

2. Method and approach

2.1. Differential microphone array

A differential microphone array consists of microphone elements that are spaced very closely compared to the acoustical wavelength. The elements' outputs are combined in an alternating fashion to achieve certain directional sensitivity patterns [4]. Figure 1 illustrates the case where a *plane* wavefront originating from a sound source in the far-field impinges on a *first-order differential* (FOD) array in *front-back* configuration. The pressure wave arriving at the front front mic can be expressed (in frequency domain) as:

$$P_f(\omega) = S(\omega)e^{-j\omega x/c} = S(\omega)e^{-jkx} \quad (1)$$

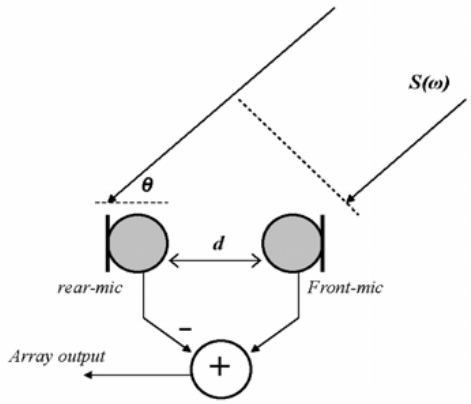

Figure 1: *First-order differential array composed of two zero-order (omnidirectional) microphone elements.*

where x is the distance from sound source to the array, c is the sound speed in air (344 m/s), $k = \omega/c$ is the angular wave-number, d is the inter-element spacing (m), θ is the angle of incidence with respect to on-axis of the array (rad), ω is the angular frequency (rad/sec), and $S(\omega)$ is a quantity proportional to the source signal. Pressure wave arriving at the rear mic can be written as:

$$P_r(\omega) = S(\omega)e^{-jk(x+d\cos\theta)} \quad (2)$$

The output signal is the difference of the elements' outputs:

$$\begin{aligned} Y(\omega) &= P_f(\omega) - P_r(\omega) \\ &= S(\omega)e^{-jkx}\left[1 - e^{-jkd\cos\theta}\right] \\ &= P_f(\omega)\left[1 - \left\{\cos(kd\cos\theta) - j\sin(kd\cos\theta\right\}\right] \\ &= j\omega\frac{d}{c}\cos\theta \cdot P_f(\omega) \end{aligned} \quad (3)$$

where in the last row we have assumed that the dimension of the array is much smaller than the signal's wavelength. $(kd \ll 1)$. Two important properties are evident from the result in equation (3). First, the cosine term which is a function of angle of incidence gives rise to the bidirectional sensitivity pattern (other directionality patterns can be achieved using delay elements in one of the mic signal path). Second, the first-order differentiator term $j\omega$ indicates that the frequency response magnitude increases linearly with frequency. In practice the frequency response does not increase without bound with frequency, instead, it levels off above a certain design frequency; in this respect, a FOD array can be said to exhibit *bass roll-off*.

2.1.1. Proximity Effect

Another interesting property of the FOD array is the increase of low frequency sensitivity when the sound source is close to the microphone, in microphone literature this is known as the *proximity effect* [3]. This occurs because of the large pressure gradient that exists in the near-field of the source. In the near-field (within half a wavelength) of the source, the pressure wave arriving at the front mic can be expressed as:

$$P_f(\omega) = \frac{S(\omega)}{x} e^{-jkx} \tag{4}$$

Similar derivation as with the far-field case yields the expression of the array's output:

$$\begin{aligned} Y(\omega) &= P_f(\omega) jk \left[1 + \frac{1}{jkx} \right] d \cos\theta \\ &= j\omega \frac{d}{c} \cos\theta \cdot P_f(\omega) + \frac{d}{x} \cos\theta P_f(\omega) \end{aligned} \tag{5}$$

The second term of equation (5) is dependent on the source distance and vanishes when x is large (far-field sources), what remains is the first term which is equal to the result given in equation (3). For small x (near-field sources) the second term can be interpreted as the gain of a lowpass filter [3]. Hence, for near-field sources the array's response is a combination of highpass and lowpass transfer function, in which the later contributes to the increase in the low frequency response.

Figure 2 shows the frequency response of a FOD array for various source distances. For near-field sources the low frequency boost cancels the bass roll-off below a certain frequency and the closer the sound source is to the microphone the flatter the frequency response becomes. This effect is enhanced as the differential order of the array increases [4].

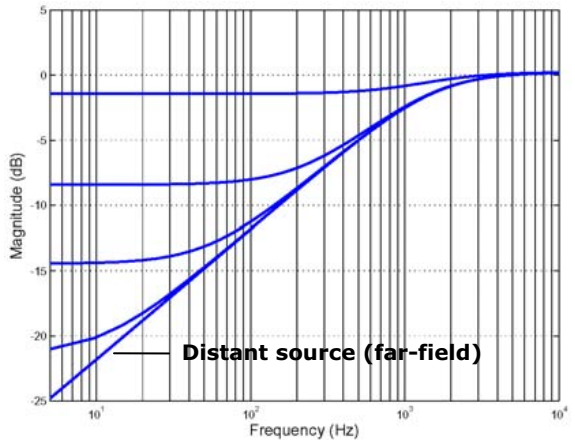

Figure 2: *Frequency response of a FOD array for various source distances.*

2.2. Modified-EVRC Rate Determination Algorithm

From the result of the previous section, it is clear that if we assume the target speaker is in the near-field of the microphone and background noise is in the far-field, then a differential microphone array could introduce substantial SNR improvement (over a single omnidirectional microphone) in the low-frequency band. On this basis, we develop a modification of the EVRC RDA to exploit such SNR boost. In this paper we will refer the modified algorithm as the *m-EVRC RDA*.

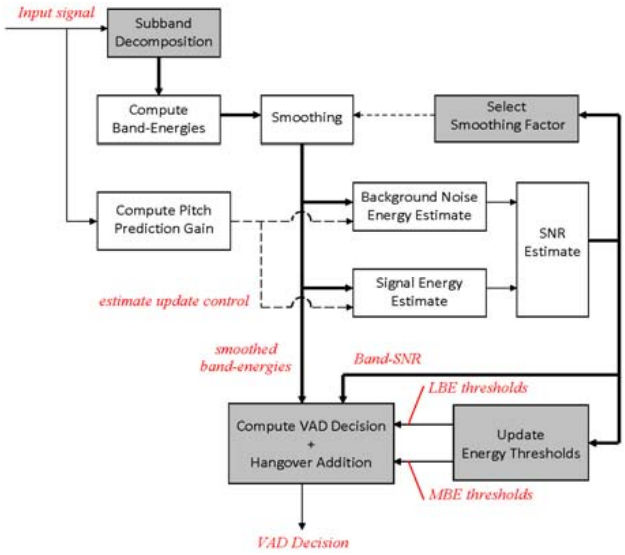

Figure 3: *Block Diagram of m-EVRC RDA.*

Block diagram of m-EVRC RDA is shown in Figure 3. The basic structure of EVRC rate determination algorithm is evident, plus a number of modifications which shall be discussed next.

2.2.1. Sub-band decomposition

The input signal is decomposed into two bands: the low-band (0 to 400 Hz) and the mid-band (300 Hz to 2000 Hz), from which the band-energies are computed. The low-band frequency range is determined by considering the extent of SNR improvement provided by the FOD array as follows. Taking the ratio of magnitude of equation (3) and (5) we obtain :

$$\frac{|Y(\omega)|_{near-field}}{|Y(\omega)|_{far-field}} = \sqrt{1 + \frac{1}{k^2 x^2}} \tag{6}$$

If target signal is in the near-field and noise is in the far-field, then the SNR improvement is given by :

$$\begin{aligned} SNR \text{ improvement} &= 20 \cdot \log_{10} \left(\frac{|Y(\omega)|_{near-field}}{|Y(\omega)|_{far-field}} \right) \\ &= 10 \cdot \log_{10} \left(1 + \frac{c^2}{\omega^2 x^2} \right) \end{aligned} \tag{7}$$

which shows how the SNR varies as a function of frequency, as depicted in Figure 4.

From this analysis we chose 0 to 400 Hz to be the low-frequency band in which the low-band energy (LBE) is computed. Significance of this low-frequency band SNR enhancement is illustrated in Figure 5. The top graph shows the waveform of a clean speech signal, the middle graph shows the waveform of that speech signal after a car's engine noise has been added to it at -5dB SNR. The bottom graph shows the log low-band energy computed from the output of a FOD array and a single omnidirectional microphone. The car's engine noise was recorded with the microphone array positioned 4 meters away from the car, and the clean speech is

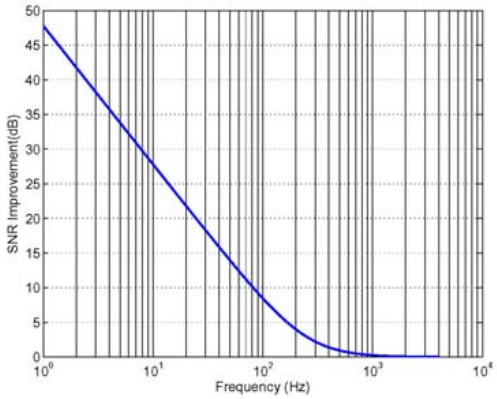

Figure 4: *Theoretical SNR enhancement over a single omnidirectional microphone offered by a FOD array as a function of frequency*.

recorded with the microphone positioned 5 cm away from the speaker's lips. The speech portion of the signal is no longer visible from the waveform of the noisy signal mixture; however it is quite apparent from the LBE computed from the FOD array output. It is evident from the bottom graph that when we switch from a single omnidirectional mic to a FOD array the low-band SNR is boosted by approximately 10 dB.

The mid-band energy (MBE) is much less affected by the SNR enhancement and therefore is of less importance in our algorithm. We used the 300 Hz to 2000 Hz frequency band which was defined as high-frequency band in the original EVRC RDA as the mid-band range, and used the MBE as a complementary metric to refine the VAD decision, more detail on this in section 2.2.3.

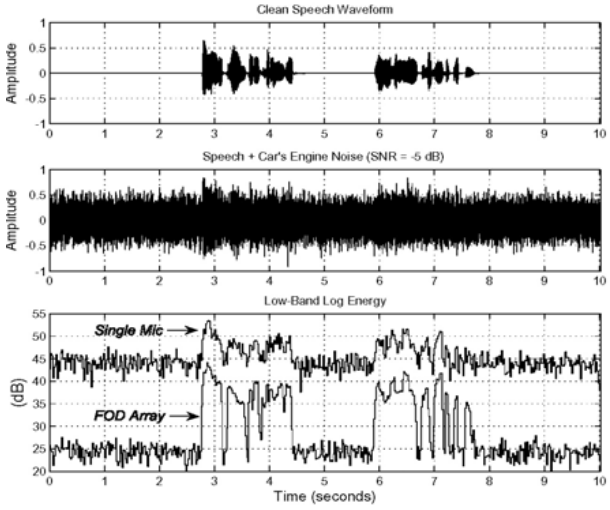

Figure 5: *Log low-band energy of speech + car's engine noise signal (SNR = -5 dB).*

2.2.2. Band energies smoothing

Having computed the band-energies, EVRC RDA compares the band-energies with a set of adaptive thresholds to yield the VAD decision. However, our algorithm primarily relies only on the low-band energy which may be very low for unvoiced speech segments, thus it would be unreliable to perform thresholding directly on this quantity. We solved this problem by smoothing the band-energies prior to the decision process. We further improved the scheme by making the smoothing factor adaptive according to values of the low-band SNR estimates. The smoothing factor is adapted as follows :

$$Smoothing\ Factor = \begin{cases} 0.45 & ; \quad low\text{-}band\ SNR \geq 6 \\ 0.5 & ; 6 > low\text{-}band\ SNR > 1 \\ 0.75 & ; 1 \geq low\text{-}band\ SNR \end{cases} \quad (8)$$

where the (normalized) low-band SNR takes integer values in the range of 0 to 7. At very high SNR (clean speech) lower smoothing factor value is selected to avoid prolonged speech detection at the offset, where as at lower SNR higher smoothing factor reduces the chance of miss-detecting low energy unvoiced speech segments.

2.2.3. VAD decision

We modified the EVRC RDA decision process by introducing different ways of combining the band-decisions depending on the smoothed low-band energy ($LBE_{smoothed}$) and SNR estimates. This reflects our increased confidence in the low-band energy parameter. The VAD decision is computed for every frame and the decision logic is given as follows :

$$Threshold_f = \begin{cases} T_{lower} & ; SNR_f > 5 \\ T_{upper} & ; \text{else} \end{cases} \quad (9)$$

$$Decision_f = \begin{cases} 1 & ; Band\text{-}Energy_{smoothed} > Threshold_f \\ 0 & ; \text{else} \end{cases} \quad (10)$$

$$\begin{aligned} &if\ ((SNR_{f1} > 5)\ \&\ (SNR_{f2} > 5)\ |\ (LBE_{smoothed} > 80dB)) \\ &\quad Decision = Decision_{f1}\ \&\ Decision_{f2} \\ &else, \\ &\quad Decision = Decision_{f1} \\ &\quad if\ (T_{upper} > LBE_{smoothed} > T_{lower}) \\ &\quad\quad Decision = Decision_{f1}\ |\ Decision_{f2} \end{aligned} \quad (11)$$

where subscript *f1* denotes low-band, *f2* denotes medium-band, and '&, |' denote the logical operator 'and, or' respectively.

3. Performance evaluation

3.1. Clean speech corpus

For evaluation purposes a clean speech corpus database has been developed. The speaker subjects comprised of 7 males and 7 females with various speaking styles. Recordings were made by placing the microphone array (element spacing is 1.1 cm) approximately 4 to 5 cm from the speaker's lips; this distance range is selected to emulate the practical condition where a handset or a headset equipped with microphone boom is used. Each speaker utters a predefined sentence set of 20 seconds long. Speech - silence composition is approximately 45% - 55%.

3.2. Background noise corpus

In order to evaluate the performance of the algorithm in various noise conditions, we obtained a comprehensive set of background noise recordings using the FOD array prototype. The recordings are classified into 7 background noise types: *car interior*, *machinery*, *street*, *babble*, *office*, *background-music*, and *background-speech*.

3.3. Performance measure

Performance of the algorithm is evaluated in terms of *Speech Detection Error Rate* (SDER), *Noise Detection Error Rate* (NDER), and *Overall Detection Error Rate* (OVER), defined as follows :

$$SDER = \frac{\sum miss\text{-}detected\ speech\ frames}{\sum speech\ frames} \times 100\% \qquad (12)$$

$$NDER = \frac{\sum miss\text{-}detected\ noise\ frames}{\sum noise\ frames} \times 100\% \qquad (13)$$

$$OVER = SDER\left(\frac{\sum speech\ frames}{\sum frames}\right) + NDER\left(\frac{\sum noise\ frames}{\sum frames}\right) \qquad (14)$$

where $\sum speech\ frames$ and $\sum noise\ frames$ are the reference number of frames classified as *speech* and *noise* respectively. The above performance measures are averaged over 20 noisy speech signals and 4 independent recordings for each of the 7 background noise classes.

Noisy speech signals are constructed by artificially adding background noise signals to clean speech signals. The clean speech signals are normalized to -20 dBov active speech level and the noise is scaled such that the resulting speech + noise mixture has the desired SNR level.

4. Simulation results and discussion

We compared the performance of m-EVRC RDA and the original EVRC RDA, where the later is evaluated with two different setups: the *normal setup*, where input signal to the algorithm is acquired using a single omnidirectional microphone, and the *array setup*, where the input signal is obtained from the output of the FOD array. This was done to give insight on how the distant noise-canceling characteristics of the FOD array alone might improve the performance of the un-modified EVRC algorithm. Table 1 shows the simulation results for 0 dB, 5 dB, and 15 dB SNR levels. m-EVRC RDA combined with FOD array substantially improves the original EVRC RDA performance in virtually all tested noise condition, more significantly for lower SNR and dynamic background noise such *babble*, *office*, *background-music*, and *background-speech*. As expected, FOD array does improve somewhat the performance of the un-modified EVRC RDA but the combination still does not match the performance of the modified algorithm used in conjunction with the FOD array.

5. Conclusions

This paper presented a modification of the EVRC rate determination algorithm for use with a FOD array. The modified algorithm exploits the SNR enhancement in the low-frequency band that is introduced by a FOD array to improve the algorithm's performance in highly dynamic background noise.

A thorough performance evaluation has been carried out via computer simulation for 7 classes of background noise over various SNR levels. The modified EVRC RDA combined with a FOD array has been shown to work very well and outperforms the original EVRC RDA in all noise conditions and SNR levels, more significantly for low SNR levels and highly dynamic noise environments.

Table 1. *Performance comparison*

		Car Interior			Machinery		
SNR	Algorithm	SDER	NDER	OVER	SDER	NDER	OVER
		(%)			(%)		
0 dB	EVRC	6.61	**1.22**	3.39	31.46	5.02	15.45
	EVRC + FOD	3.99	2.66	**3.12**	37.49	**0.65**	13.74
	m-EVRC + FOD	**3.13**	3.73	3.52	**4.24**	1.34	**2.85**
5 dB	EVRC	7.62	**1.02**	3.65	26.68	**0.07**	10.51
	EVRC + FOD	4.46	1.78	2.71	13.39	1.35	5.67
	m-EVRC + FOD	**3.81**	1.13	**2.22**	**4.07**	1.28	**2.59**
15 dB	EVRC	7.35	0.54	3.18	9.01	0.32	3.78
	EVRC + FOD	**1.84**	0.37	**0.88**	5.57	1.89	3.18
	m-EVRC + FOD	3.38	**0.33**	1.55	**4.72**	**0.04**	**1.87**

		Street			Babble		
SNR	Algorithm	SDER	NDER	OVER	SDER	NDER	OVER
		(%)			(%)		
0 dB	EVRC	15.69	21.26	19.28	15.96	40.96	31.59
	EVRC + FOD	6.62	24.52	18.14	12.79	28.08	22.77
	m-EVRC + FOD	**3.73**	**7.97**	**6.39**	**3.82**	**2.56**	**3.06**
5 dB	EVRC	9.92	18.94	15.53	16.12	24.23	21.37
	EVRC + FOD	3.98	23.11	16.26	6.06	26.98	19.61
	m-EVRC + FOD	**3.44**	**5.79**	**4.87**	**3.93**	**1.50**	**2.46**
15 dB	EVRC	7.89	14.22	11.82	7.76	19.27	14.97
	EVRC + FOD	6.45	14.53	11.57	**4.36**	23.49	16.72
	m-EVRC + FOD	**4.12**	**1.19**	**2.35**	4.67	**0.28**	**2.01**

		Office			Background Music		
SNR	Algorithm	SDER	NDER	OVER	SDER	NDER	OVER
		(%)			(%)		
0 dB	EVRC	7.20	55.66	36.76	9.95	60.70	40.98
	EVRC + FOD	**3.45**	52.03	34.64	8.49	43.11	37.94
	m-EVRC + FOD	3.59	**6.77**	**5.47**	**3.37**	**18.36**	**12.60**
5 dB	EVRC	5.50	52.35	34.08	9.41	49.45	33.92
	EVRC + FOD	**3.63**	45.31	30.48	3.94	40.36	34.35
	m-EVRC + FOD	4.14	**3.92**	**3.97**	**2.56**	**15.98**	**10.85**
15 dB	EVRC	7.44	37.30	25.77	5.36	42.40	28.17
	EVRC + FOD	6.80	26.46	19.45	3.76	32.60	27.58
	m-EVRC + FOD	**4.09**	**1.66**	**2.58**	**3.68**	**6.57**	**5.54**

		Background Speech		
SNR	Algorithm	SDER	NDER	OVER
		(%)		
0 dB	EVRC	5.81	67.07	43.36
	EVRC + FOD	4.30	59.94	40.22
	m-EVRC + FOD	**2.82**	**19.21**	**12.90**
5 dB	EVRC	5.61	58.65	38.19
	EVRC + FOD	**3.11**	55.72	37.03
	m-EVRC + FOD	3.30	**18.12**	**12.37**
15 dB	EVRC	4.73	47.38	30.92
	EVRC + FOD	**2.42**	41.81	27.79
	m-EVRC + FOD	3.54	**8.55**	**6.73**

6. Acknowledgements

The authors thank Dr.Ir. I Gde Nyoman Merthayasa for granting permission to use the semi-anechoic room at the Physics Engineering Dept., Institut Teknologi Bandung.

7. References

[1] TIA/EIA, Enhanced Variable Speech Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, IS-127, 1997.

[2] Elko, G.W., West, J.E, Titus Kubli, R.A., "Adaptive Close-Talking Microphone Array", IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, page(s) 163-166, 2000.

[3] Torio, G., "Understanding the Transfer Functions of Directional Condenser Microphones in Response to Different Sound Sources", AES UK Conference, 1998.

[4] Elko, G.W., "Differential Microphone Array", in Y. Huang [Ed] , Audio Signal Processing for Next Generation Multimedia Communication Systems, 11-65, Kluwer, 2004.

[5] Stricher, R., Dooley, W., "The Bidirectional Microphones: The Forgotten Patriarch", JAES, Vol.51, No.3, 2003.