Res. Lett. Inf. Math. Sci., 2005, Vol. **7**, pp 147-155
Available online at http://iims.massey.ac.nz/research/letters/

Automotive three-microphone voice activity detector and noise-canceller

Z. QI[1] and T.J.MOIR[2]

[1] *Department of Electrotechnology, Unitec New Zealand, Auckland, New Zealand* [†]
[2]*Institute of Information and Mathematic Science, Massey University at Albany, Auckland, New Zealand*

This paper addresses issues in improving hands-free speech recognition performance in car environments. A three-microphone array has been used to form a beamformer with least-mean squares (LMS) to improve Signal to Noise Ratio (SNR). A three-microphone array has been paralleled to a Voice Activity Detection (VAD). The VAD uses time-delay estimation together with magnitude-squared coherence (MSC).

1. Introduction

One of the most challenging and important problems in Intelligent Transport Systems (ITS) is to keep the driver's eyes on the road and his hands on the wheel. Speech recognition offers one such solution to this problem. Speech control in car is a safe solution e.g. to enter a street name in a Global Positioning System (GPS) navigation system by speech is better than to do it by hand. However, speech recognition in a car has the inherent problem of acquiring speech signals in a noisy environment. There are two types of additive noises in a car cabin: stationary and non-stationary. Stationary noise in car is from the engine (though it varies with speed), road, wind, air-conditioner etc. Non-stationary noise is from the car stereo, navigation guide, traffic information guide, bumps, wipers, indicators, conversational noise and noise when passing a car running in the opposite direction (Shozakai, Nakamura, & Shikano, 1998). Therefore noise reduction methods for speech enhancement in a car have been investigated for various applications. The Griffiths-Jim acoustic beamformer is a main technology in reducing stationary or non-stationary noise in car cabin(Cho & Ko, 2004). In our approach here, three microphones are used to detect the desired and undesired periods of speech by defining a geometrical 'active zone'. With three microphones this word boundary detector can retrieve the desired speech embedded with noise from varieties of noisy backgrounds. Some simulation experiments have been shown that the algorithm is an effective speech detecting method that exceeds to an average 80% of success rate(Chen & Moir, 1999).

This paper uses a three-microphone VAD and focuses on a real environment of car. There are two parts in this three-microphone VAD system:

- Part 1: A three-microphone beamformer with least-mean squared (LMS).

[†] Email addresses: tqi@unitec.ac.nz ; t.j.moir@massey.ac.nz

- Part 2: A three-microphone Voice Activity Detection (VAD) algorithm.

The VAD acts as a switch on a double-acting Griffiths-Jim adaptive beamformer. Van Compernolle (Van Compernolle, 1990) introduced this switching adaptive filter with a 4 microphone array in a highly reverberant room with both music and fan type noise as jammers. SNR improvements of l0 dB were typical with no audible distortion.

2.VAD Algorithm

2.1 System configuration

In Figure 1 three microphones are located as shown and there is 50 cm distance between these microphones. A desired speech source is located 50 cm away from Microphone 1 and Microphone 3. The distance between the speaker and Microphone 2 is 70.7 cm.

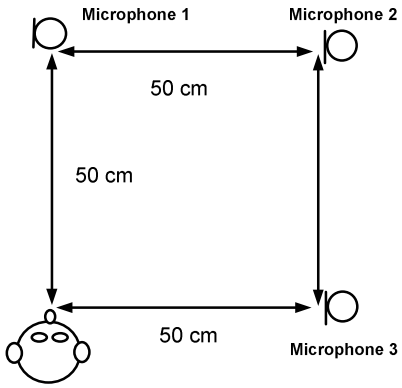

Figure 1 Automobile environment layout

Therefore, when speech travels to microphone 2 it has 20.7 cm more distance from to microphone 1 and also has 20.7 cm more than from microphone 3.
The sample rate of Microphone 1, 2 and 3 is 11025 Hz, and the speed of sound in air is 34600cm/second. Therefore during every sample the speech travels 3.1 cm so that the wave-front of speech arrives at microphone 2 delayed by 7 sample intervals with respect to the other two microphones.

2.2 Three-microphone VAD controlled three-microphone adaptive digital filter

A block diagram of the three-microphone VAD-controlled three-microphone noise canceller shown in Figure 2. The noise canceller (three-microphone adaptive digital filter) is detailed in Figure 3. The VAD switches various LMS filters on or off depending if the desired speech is presented. Moreover, the VAD allows signal output only when desired speech presented i.e. it mutes the output when there is noise present outside the desired zone but only if simultaneously there is no desired speech.

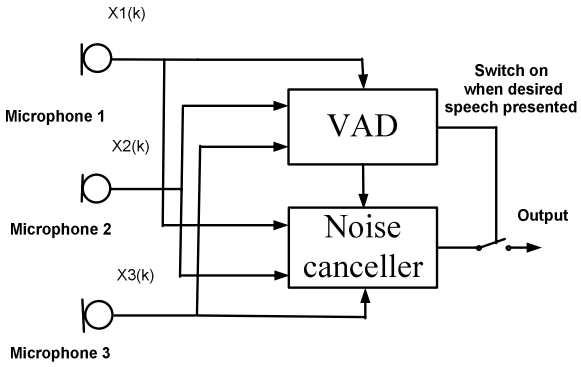

Figure 2 Overview of three-microphone VAD controlled three-microphone noise canceller

2.3 Three-microphone adaptive digital filter

A three-microphone noise canceller based on Van Compernolle's work is showed as Figure 3. There are four LMS units in a three-microphone noise canceller. The top path of the beamformer has a summation term which forms the primary input whilst both of the bottom paths have a difference term which forms the reference input. The three microphone signals contain speech as well as noise. The left section of the system serves at improving the noise reference by eliminating speech so that the VAD switches this part on when speech energy is dominant. The right section consists of LMS 2 and LMS 4, which are only switched on to adapt during the absence of speech (i.e. during noise periods). For these experiments the number of weights used in W1 and W3 were 100 and in W2 and W4, 450.

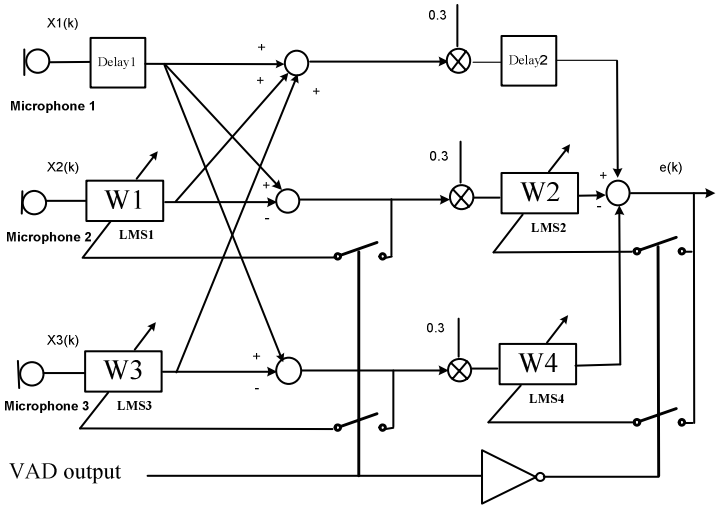

Figure 3 Three-microphone noise canceller block diagram

2.4 A three-microphone VAD

Carter et al. (Carter, Knapp, & Nuttall, 1973) describe a method for estimating the magnitude-squared coherence (MSC) function for two zero-mean wide-sense-stationary random processes. The estimation technique utilizes the weighted overlapped segmentation fast Fourier transform (FFT). Analytical and empirical results for statistics of the estimator are presented. The analytical expressions are limited to the non-overlapped case. Empirical results show a decrease in bias and variance of the estimator with increasing overlap and suggest a 50-percent overlap as being highly desirable when cosine (Hanning) weighting is used. Once the MSC is found the Generalized Cross-Correlation (GCC) method is used to give a robust estimate of time-delay. The technique can be summarized as follows for three microphones and two estimated time-delays.

At each FFT frame index $i = 1,2,3,...$ assign the three vectors

$$x_1 = [n_0, n_1, ...n_{N-1}]^T \tag{4}$$

$$x_2 = [m_0, m_1, ...m_{N-1}]^T \tag{5}$$

$$x_3 = [l_0, l_1, ...l_{N-1}]^T \tag{6}$$

which are composed of N samples of the three microphone inputs and have been suitably windowed with their corresponding frequency vectors corresponding to X_1, X_2 and X_3 respectively.

Estimate the auto-power spectra (periodograms) of the signals from each of the three microphones

$$\hat{S}_{x_1x_1}(i) = \beta \hat{S}(i-1) + (1-\beta) X_1 X_1^* \tag{7}$$

$$\hat{S}_{x_2x_2}(i) = \beta \hat{S}(i-1) + (1-\beta) X_2 X_2^* \tag{8}$$

$$\hat{S}_{x_3x_3}(i) = \beta \hat{S}(i-1) + (1-\beta) X_3 X_3^* \tag{9}$$

where (7)**,** (8) and (9) is a method of smoothly updating the spectrum recursively at each FFT frame. In the above equation * represents complex conjugate and $0 \le \beta \le 1$ is a forgetting factor. For the results used in this paper $\beta = 0.5$ was used as a compromise between fast tracking and smoothing. If chosen to be too large then the tracking ability of the GCC time-delay estimator is severely compromised. Some experimentation is required depending on the application. Two cross-spectrum (cross-periodograms) are found in a similar manner.

$$\hat{S}_{x_1x_2}(i) = \beta \hat{S}(i-1) + (1-\beta) X_1 X_2^* \tag{10}$$

$$\hat{S}_{x_2x_3}(i) = \beta \hat{S}(i-1) + (1-\beta) X_2 X_3^* \tag{11}$$

The MSC at each FFT frame is found from

$$\left|\hat{\gamma}_{x_1x_2}(i)\right|^2 = \frac{\left|\hat{S}_{x_1x_2}(i)\right|^2}{\hat{S}_{x_1x_1}(i)\,\hat{S}_{x_2x_2}(i)} \tag{12}$$

$$\left|\hat{\gamma}_{x_2x_3}(i)\right|^2 = \frac{\left|\hat{S}_{x_2x_3}(i)\right|^2}{\hat{S}_{x_2x_2}(i)\,\hat{S}_{x_3x_3}(i)} \tag{13}$$

and at each frame i, average over frequency k the MSC thus

$$\left|\bar{\gamma}_{x_1x_2}(i)\right|^2 = \sum_k \left|\hat{\gamma}_{x_1x_2}(i)\right|^2 \tag{14}$$

$$\left|\bar{\gamma}_{x_2x_3}(i)\right|^2 = \sum_k \left|\hat{\gamma}_{x_2x_3}(i)\right|^2 \tag{15}$$

Estimate the term $\psi_{g_1}(i)$ and $\psi_{g_2}(i)$ from

$$\psi_{g_1}(i) = \frac{\left|\hat{\gamma}_{x_1x_2}(i)\right|^2}{\left|\hat{S}_{x_1x_1}(i)\right|\left[1-\left|\hat{\gamma}_{x_1x_2}(i)\right|^2\right]} \tag{16}$$

$$\psi_{g_2}(i) = \frac{\left|\hat{\gamma}_{x_2x_3}(i)\right|^2}{\left|\hat{S}_{x_2x_3}(i)\right|\left[1-\left|\hat{\gamma}_{x_2x_3}(i)\right|^2\right]} \tag{17}$$

Estimate the time-delays of arrival d_1 and d_2 from the generalized cross-correlations.

$$\hat{R}^{g_1}_{x_1x_2}(d_1) = \max F^{-1}\left\{\psi(i)\,\hat{S}_{x_1x_2}(i)\right\} \tag{18}$$

$$\hat{R}^{g_2}_{x_2x_3}(d_2) = \max F^{-1}\left\{\psi(i)\,\hat{S}_{x_2x_3}(i)\right\} \tag{19}$$

That is the maximum of the inverse FFT of $\psi(i)\,\hat{S}_{x_1x_2}(i)$ and $\psi(i)\,\hat{S}_{x_2x_3}(i)$. A positive delay can be inferred if the maximum occurs in the region $0 \le d \le N/2-1$ i.e. the first half of the inverse FFT and a negative delay if the maximum occurs in the upper half of the inverse FFT.

Valid speech is then assumed when

$$\mathrm{d}_1 \le \mathrm{d}_{\max} \quad \text{and} \quad \mathrm{d}_2 \le \mathrm{d}_{\max} \tag{20a,b}$$

Also we require that both

$$\left|\hat{\gamma}_{x_1x_2}(i)\right|^2 \geq C_{\min} \text{ and } \qquad \left|\hat{\gamma}_{x_2x_3}(i)\right|^2 \geq C_{\min} \qquad (21a,b)$$

The latter two equations are necessary to prevent reverberant speech from being detected as desired speech e.g. when a reflection of a nearby undesired noise finds its way into the active zone. It is well established however that reverberant speech has a higher MSC than non-reverberant speech and this gives rise to (21a,b).
For the experiments carried out in this paper a sampling interval of 11025Hz was used so that each sample interval corresponds to $90.7\,\mu s$. Typically d_{max} was chosen to be no more than 5 samples and $C_{\min}$ was chosen as 0.5.
A three-microphone VAD block diagram is presented at Figure 4.

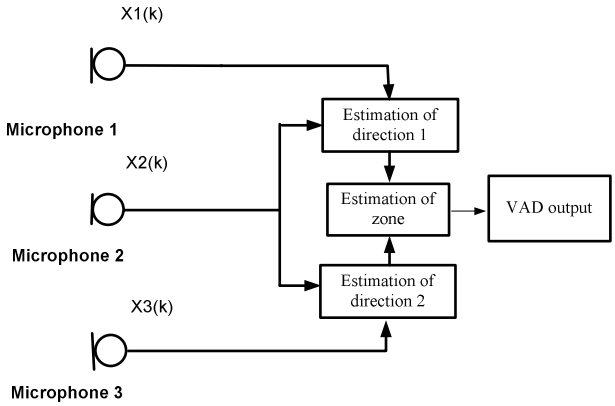

Figure 4 Three-microphone VAD Block diagram

An estimation of time delay (time-difference of arrival TDOA) defines Estimation of Direction 1 (EOD 1) located on the line adjoining Point 1 and microphone 1 as in Figure 5. This delay is estimated between microphone 1 and 2. Another estimation of TDOA between microphones 2 and 3 defines Estimation of Direction 2 (EOD 2) on the line adjoining Point 1 and Microphone 3. If the two TDOA's are zero, EOD 1 will be on the line adjoining Points 2 and 5, and EOD 2 will be on the line adjoining Points 3, 5, 6 and 7. Since EOD 1 and EOD 2 are defined, Point 1 will be the centre of the Estimation of Zone (EOZ). When the VAD is set to be within some defined number of samples e.g. 5 sample TDOA's from each microphone pair, speech is picked up from a zone around point 1. For the case of 5 sample TDOA's, the desired zone has approximately a diameter of 15 cm from point 1 as shown in Figure 5.In fact the actual zone is in three-dimensions and has the form of a two-sheet hyperboloid when two microphones are used and for this three-microphone case it will be the intersection of two such two-sheet hyperboloids. (Agaiby & Moir, 1997).

The VAD works as to switch to freeze or enable the various LMS algorithms. Also VAD switches off (mutes) the signal output when speech does not come from the desired zone.

3. Experiments

Seven testing points have been set as in Figure 5. Test point 1 is where the head of the desired speech is coming from. These tests were carried out in a stationary automobile with the engine running. While speaking at test point 1, microphone 1, 2 and 3 pick up the signal and output the enhanced signal for test point 1 by using the discussed algorithms. However, noise cancellation takes place at test points 2, 3, 4, 5, 6, 7 and 8 which are outside of the desired zone. (EOZ denotes the end of the desired zone)

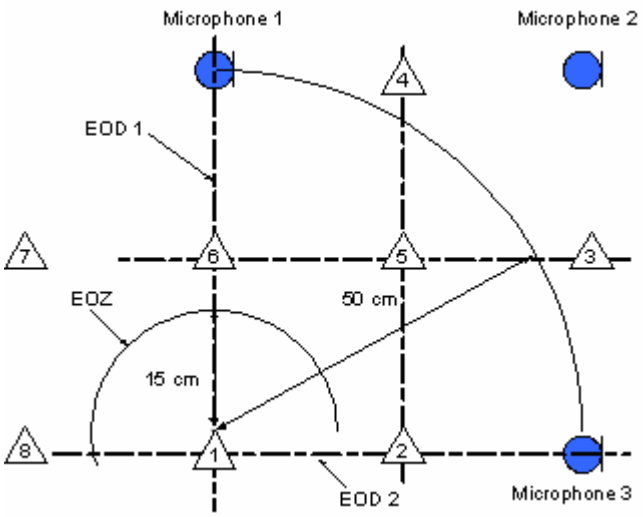

Figure 5 Seven testing points

The experiment was conducted as follows: a loud-speaker outputs a pre-recoded phrase "Open the door" once at test point 1, then repeats this for test point 2 and so on to test point 8. Therefore Microphone 1, 2 and 3 pick up the phrase "Open the door" eight times with differing strength as shown in Figure 6. Waveform "Output A" in Figure 6 shows the output at the error *e(k)* from Figure 3. It indicates that speech from point 1 is enhanced but the speech picked up from points 2-8 are attenuated. The VAD can be programmed to switch off (mute) when the speech is not from point 1 so in effect the only noise canceling that needs to be done is when speech is detected in the active zone. This is shown as "Output B" in Figure 6.

Since the waveforms in Figure 6 are the same sources at Speech 1 or 2 and so on, SNR can be compared directly from

$$SNR_i = 10\log_{10}\frac{Output\,Power}{Mic_i\ Input\,Power} \qquad \text{i=1,2,3} \tag{23}$$

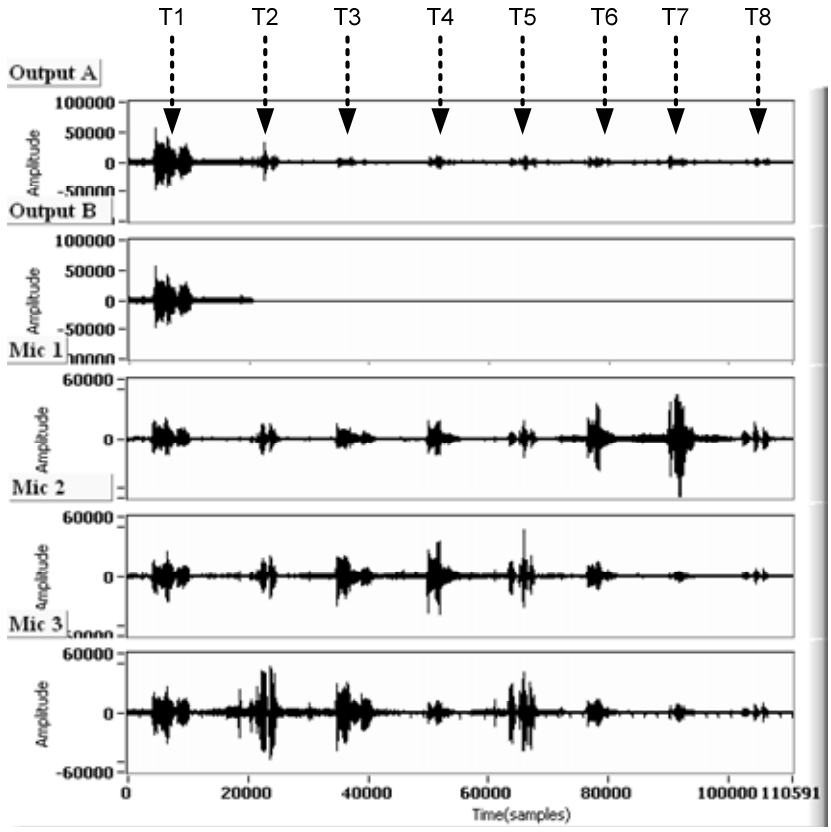

Figure 6 Speech waveforms.

The SNR results are presented at Table 1.For T1 in Table 1 the SNR should be as high as possible as this is desired speech whilst for the other test-points the SNR should be as small as possible indicating an attenuation in the speech as it appears outside the desired zone. At "Output A" in Figure 6, the un-desired speech cannot be cancelled completely. However, points 2 – 8 are very close to microphones indicating that much effort has to be done to reduce their power. Since we have a robust VAD it makes little difference whether there is in fact any residual speech after noise-cancellation since this can easily be muted as shown as Output B in Figure 6.

4.Conclusion

Experiments have been conducted in real-time on a combined three-microphone VAD and noise-canceling system. The VAD assumes that the desired speech falls within a desired geometric zone which is most appropriate for an automobile environment. The noise-canceling is only required when noise is present during desired speech as the VAD will mute any solo noise-source outside the zone. Future work will include the use of a speech-recognition engine to see the improvements in recognition hit-rate in such environments.

Table 1 SNR improvement in different test zones

	SNR_1 dB	SNR_2 dB	SNR_3 dB
T1	7.35	6.58	3.9
T2	0.93	-1.95	-10.76
T3	-1.3	-7.67	-9.04
T4	-4.96	-10.21	-4.82
T5	-7.1	-9.46	-8.76
T6	-8.48	0.58	0.65
T7	-9.62	-0.43	-2.56
T8	-10.17	-4.07	-5.64

References

Agaiby, H., & Moir, T. J. (1997). *A robust word boundary detection algorithm with application to speech recognition.* Paper presented at the Digital Signal Processing Proceedings, 1997. DSP 97., 1997 13th International Conference on.

Carter, G., Knapp, C., & Nuttall, A. (1973). Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. *Audio and Electroacoustics, IEEE Transactions on, 21*(4), 337-344.

Chen, W. N., & Moir, T. J. (1999). Adaptive noise cancellation for nonstationary real data background noise using three microphones. *Electronics Letters, 35*(23), 1991-1992.

Cho, Y., & Ko, H. (2004). *Speech enhancement for robust speech recognition in car environments using Griffiths-Jim ANC based on two-paired microphones.* Paper presented at the Consumer Electronics, 2004 IEEE International Symposium on.

Shozakai, M., Nakamura, S., & Shikano, K. (1998). *Robust speech recognition in car environments.* Paper presented at the Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on.

Van Compernolle, D. (1990). *Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings.* Paper presented at the Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on.