Application of a First-Order Differential Microphone for Efficient Voice Activity Detection in a Car Platform

A. Álvarez, P. Gómez, V. Nieto, R. Martínez, and V. Rodellar

Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Facultad de Informática, Universidad Politécnica de Madrid, Madrid, SPAIN
pedro@pino.datsi.fi.upm.es

10.21437/Interspeech.2005-256

Abstract

Handsfree interfaces provide a nice solution to add-on devices in car platforms. However, the amount of acoustic disturbances existing in automotive environments usually prevents satisfactory results. In most of the cases, noise reduction techniques involving a voice activity detector (VAD) are required. Through this paper, a robust microphone array processing technique for speech detection under the influence of noise and reverberation in an automobile environment is proposed. This method applies a simple two-microphone First Order Differential Microphone in order to estimate the power spectral density of the background perturbations embedded in speech signals. Afterwards, specialized order-statistics filters (OSFs) are applied in order to obtain a consistent speech/non speech decision. The paper also includes a performance evaluation of the algorithm using Aurora3 database recordings. According to our simulation results, the proposed algorithm shows a significantly better performance than standard VADs such as G.729B or ARM and, a slight advantage over other reported methods.

1. Introduction

Speech uttered inside a car is perturbed by an appreciable amount of noise and reverberation. Several solutions to overcome this problem may be emphasized, as microphone-array based techniques. Usually, Array Beamforming [1] is combined with other techniques as Independent Component Analysis [2], Spectral Subtraction [3] or Linear Prediction Analysis [4].

In a car cockpit those structures may also benefit from the fact that desired speakers, mainly the driver, are placed in a constrained region [5].

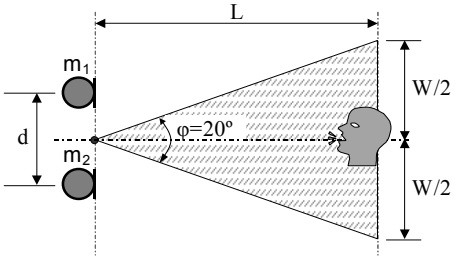

Figure 1: Framework for the working scenario indicating driver active area.

Through this paper, a VAD based on the use of a *First-Order Differential Microphone (FODM)* [6][7] for reverberation and noise estimation purposes, and its application to order statistics filters (OSFs) [8] is presented. Essentially, this approach applies a *FODM*, operating in the time domain to determine the contribution of desired speech signals in a constrained region (see Figure 1) against any other sources. In a second step, detected interferences are smoothed in the frequency domain for a robust estimation of the speech/non-speech divergence.

2. First-Order Differential Microphone

2.1. Overview of the First-Order Differential Microphone

The *First-Order Differential Microphone* (*FODM*) array proposed in this work may be seen in Figure 2. As it may be noticed, signals provided by a pair of omnidirectional microphones, separated a distance d, are combined each other applying a fixed delay $T = k\tau$. That delay is always expressed as a multiple of the sampling period τ, where $k \geq 1$ is a natural number. On the other hand, channel mixing is controlled for both branches by a steering parameter β, restricted to the range $[0.0, 1.0]$. Finally, the output is obtained subtracting the result of one individual branch from the other.

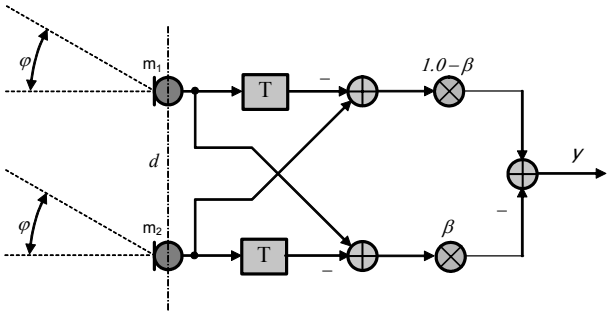

Figure 2: Structure of the two-microphone *First Order Differential Microphone* (*FODM*).

The *FODM*, despite of its simplicity, recalls the behavior of a null beamformer controlled by the β parameter. Its transfer function shows for a frequency $f < f_s/2$ a sharp notch at an angle φ given by [9]

$$\varphi = \arcsin\left\{\frac{c}{\pi f d}\arctan\left[(1-2\beta)\tan\left(\frac{\pi k f}{f_s}\right)\right]\right\} \qquad (1)$$

where f is the frequency of the signal, f_s the sampling frequency, d the microphone separation, and c the sound propagation speed.

Moreover, the effective steering range depends on three different parameters: f_s, k and d. (see Table 1).

f_s	k	d	Steering range
8000 Hz	1	5 cm	[-55.6º, +55.6º]
		10 cm	[-24.4º, +24.4º]
		20 cm	[-11.9º, +11.9º]
11025 Hz	1	5 cm	[-36.8º, +36.8º]
		10 cm	[-17.4º, +17.4º]
		20 cm	[-8.6º, +8.6º]
16000 Hz	1	2.5 cm	[-55.6º, +55.6º]
		5 cm	[-24.4º, +24.4º]
		10 cm	[-11.9º, +11.9º]
		20 cm	[-5.9º, +5.9º]

Table 1: Attainable steering ranges for different combinations of parameters f_s (sampling frequency), k (a multiple of the sampling period) and d (microphone separation).

Equation (1) establishes a useful connection between the incoming angle, corresponding to a particular source φ and the *FODM* steering parameter β. It is important to note that the above association depends also of signal frequency f. This is an undesired effect, as for broadband signals like speech, there is not a unique steering factor in order to produce the cancellation of incoming sources arriving from an angular direction φ. In fact the only exceptions to the previous rule are the following beta values: $\beta=0.5$, $\beta=0.0$ and $\beta=1.0$, that is, the broadside (0^o) and both endfire angles, as shown in Table 1.

2.2. Noise and reverberation estimation

Frequency-dependent estimations of the amount of noise and reverberation presented in the system input will be carried out comparing *FODM* outputs for the three cases previously cited.

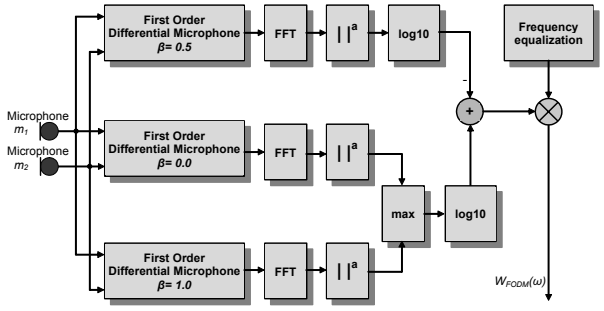

Figure 3: General framework of the structure based on the *FODM* to obtain noise-adaptation weights W_{FODM}.

The structure of the method is shown in Figure 3. In this case, the output corresponding to broadside ($\beta=0.5$) will describe signal contents regarding possible locations for the car driver, whereas the other two outputs will be linked to background interferences. It is important to notice that *FODM* outputs ideally contain signal components not arriving from the tracking angle. This means that an active source that is located in the center of the array must be eliminated in greater measurement when the beam is focused towards this position. In the same way, constrained endfire outputs ($\beta=0.0$ and $\beta=1.0$) constitute valid references of environmental degradation.

More exactly, this procedure takes place for every bin in the frequency domain, so that, *FODM* output signals ($s_{0.5}(t)$, $s_{0.0}(t)$ and $s_{1.0}(t)$) are segmented in overlapped windows and transformed using the short-time Discrete Fourier Transform. A set of subtraction weights is then given by

$$W_{FODM}(m,k)=\eta(k)log_{10}\left(\frac{max\left\{\|S_{0.0}(m,k)\|^a,\|S_{1.0}(m,k)\|^a\right\}}{\|S_{0.5}(m,k)\|^a}\right) \quad (2)$$

where m is the frame index, k is the frequency bin and, $\eta(k)$ is an equalizing function for compensating individual frequency bandwidths, inherent to the operation of the FODM structures. The role of η(k) (see Figure 4) is to compensate measures obtained for low and high frequencies. The function is calculated applying (2) for an incoming signal with φ=0º and containing white noise. Normalized coefficients are obtained by means of

$$\eta(k)=\frac{\sum_m W_{FODM}(m,0)}{\sum_m W_{FODM}(m,k)} \quad (3)$$

where $0\leq k\leq L-1$ and L is the FFT window size.

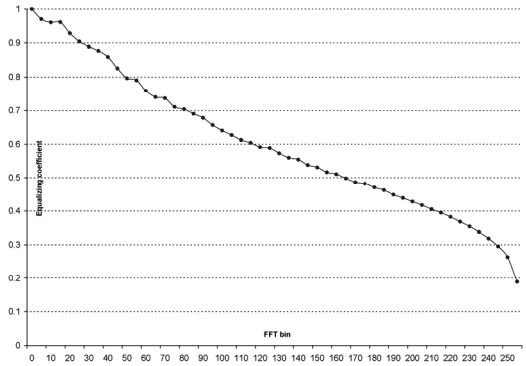

Figure 4: Representation of function $\eta(k)$.

The meaning of the maximum operation in (2) is to obtain a unified estimation of background activity since further attenuations on signal levels indicate a more accurate tracking.

3. Speech Detection Algorithm

Weights $W_{FODM}(m,k)$ may be considered a set of frequency-dependent VAD measures for frame m as they constitute the basis for the subsequent speech detection algorithm. The method uses different order-statistic filters for the estimation of the SNR. The first filter is applied to individual weights at frame m, in order to obtain a smooth estimation $W(m)$ which is given by

$$W(m)=Q_\alpha\left[W_{FODM}(m,0),\ldots,W_{FODM}(m,L-1)\right] \quad (4)$$

where α= 0.6 and $Q_x[Y]$ stands for the *x-quantile* of a serie Y.

A second OSF is used to enhance the robustness of the speech/noise decision by exploiting long-term spectral information over a N-frame neighborhood

$$W_x(m)=Q_x\left[W(m-N),\ldots,W(m),\ldots,W(m+N)\right] \quad (5)$$

where practical values of N are in the range $[5, 10]$.

The above filter is applied twice for instantaneous speech and noise measures

$$S(m) = W_{\varepsilon}(m) \tag{6}$$

$$N(m) = W_{\gamma}(m) \tag{7}$$

where ε= 0.7 and γ= 0.1 are respectively the quantiles associated to the speech and noise estimations at frame m,. In addition, SNR is computed as follows

$$SNR(m) = S(m) - N_{\lambda}(m) \tag{8}$$

where $N_{\lambda}(m)$ correspond to smoothed estimations of noise in order to track non-stationary noisy environments. Their adaptation, only during non-speech periods, is done as follows

$$N_{\lambda}(m+1) = \lambda N_{\lambda}(m) + (1-\lambda)N(m) \tag{9}$$

where λ= 0.99.

Thus, if the SNR is greater than a threshold η, the actual frame is classified as speech (H_1), otherwise it is classified as non-speech (H_0)

$$SNR(m) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{10}$$

Finally, for initialization the algorithm, the first frame is assumed to be non-speech

$$N_{\lambda}(0) = N(0) = W(0) \tag{11}$$

4. Results and Discussion

In order to evaluate the performance of the method described through this paper, several experiments were conducted. First, misclassification errors were studied at different SNR levels by means of the receiver operating characteristics (ROC) curves. Also, the influence of the proposed VAD on a speech recognition system was assessed. Standard VADs as G.729B [10], AMR [11] and AFE [12], as well as methods reported by Sohn [13] and Ramírez [8] were used for reference.

4.1. Analysis of the ROC curves

The speech material used for the analysis is part of Aurora3-SpeechDat Car Finnish database [14]. This corpus, containing realizations of connected digit and uttered in a realistic automobile environment, is divided in two sets: train and test. Both sets are divided into three different categories related to the amount of distortion existing in those recordings: quiet (SNR 12dB), low (SNR 8 dB), and high (SNR 5dB). In our experiments, channels *ch2* and *ch3* (microphones placed at the ceiling of the car in front of the speaker) were used. The speech corpus was hand-labeled on the close talking microphone (channel *ch0*) to obtain the speech/non-speech hit rates, *HR1* and *HR0* respectively.

Figure 5 shows the trade-off between speech pause hit rate and false alarm rate *FAR0* (*FAR0= 1.0-HR1*). The proposed VAD works with lower alarm rate and higher speech-pause hit rate when compared to standard methods, especially over G.729B. Although, the working point for the AFE method is up in the ROC space, that situation is far from optimal, since *HR0* is lower than 70%. The result is that an excessive number of noisy frames are hold for the rest of the recognition process. The FODM VAD also outperforms Sohn and Ramírez VADs. However, that improvement is reduced in the area corresponding to high *HR1*.

It is important to notice that although the FODM VAD exploits a two-microphone array, SpeechDat Car databases are not array-processing oriented, as far-talk sensors are not matched. In addition, the disposition and distance (approximately 20 cm) between microphones *m2* and *m3*, are far from optimal, thus representing non ideal working conditions. However, that fact also reveals an inner robustness of the method against microphone disparities.

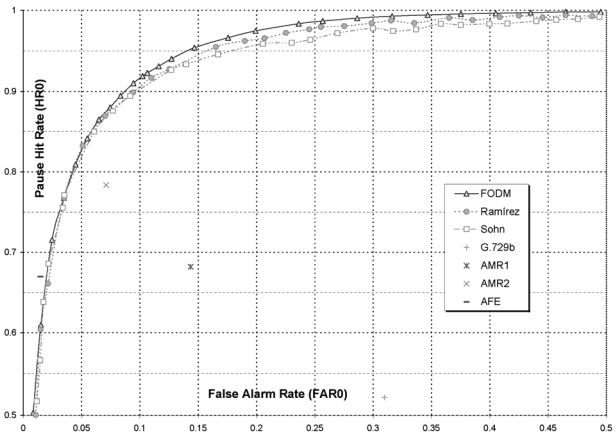

Figure 5: ROC curves for the three driving conditions as threshold varies. $W_{FODM}(m,k)$ weights were computed using *L=512* points for the FFT and $W_x(m)$ were calculated with *N=8*.

4.2. Speech Recognition Results

Several speech recognition systems were tested using the framework presented in Figure 6. As it may be seen, VAD decisions are used for frame dropping purposes in order to improve recognition accuracy. A shared recognition system is built based on HTK (Hidden Markov Toolkit) package [15].

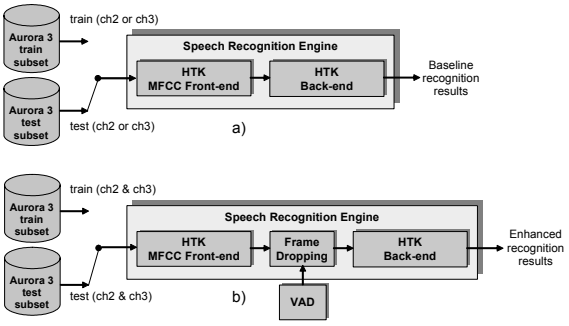

Figure 6: a) Baseline non-robust speech recognizer. b) Enhanced system incorporating a frame dropping mechanism applying VAD decisions to the same speech recognition engine.

Recognition experiments were established by selecting different materials from the training set of the database: s*et A* includes files labeled as quiet, s*et B* incorporates also files with low distortion and, finally, *set C* comprises all the training material available. The test material is the same for the threes cases and consists on 3126 words covering the full range of noise conditions.

Train subset	Word Error Rate							
	No VAD	*Standard VADs*				*Other reported VADs*		FODM
		G.729B	AMR1	AMR2	AFE	Ramírez	Sohn	
Set A	48.66%	33.88%	26.94%	25.92%	28.73%	24.29%	22.11%	**25.10%**
Set B	10.72%	16.25%	9.28%	8.39%	8.77%	7.95%	8.24%	**7.87%**
Set C	9.60%	14.40%	8.51%	7.80%	7.93%	7.69%	7.68%	**7.62%**

Table 2: Recognition results for the AURORA3 SpeechDat Car Finnish database.

Table 2 shows the word error rate (WER) for the three speech recognition systems. As it may be seen, the proposed FODM algorithm outperforms the VADs used for reference, being the improvements more relevant when compared to the standard methods. The recognition accuracy is slightly better than Sohn and Ramírez VADs except, when there is a highly mismatch between training and testing sets. However, this configuration does not correspond to a real working speech recognition system, as the recognition accuracy is excessively low.

5. Conclusions

This paper presents a robust microphone array processing technique for speech detection under the influence of noise and reverberation in an automobile environment. The proposed method uses a simple two-microphone First Order Differential Microphone to obtain a first estimation of background perturbations embedded in speech signals. Afterwards, several order-statistics filters are applied in order to produce a consistent speech/non speech decision. This work includes several evaluations carried out with real data taken form the Aurora 3 database. The proposed algorithm outperforms G.729B, AMR1, AMR2 and AFE standards and, other VADs among the best reported, in speech/non speech detection capabilities. Also, speech recognition experiments show a noticeable and consistent reduction in word error rates when the VAD is included as a component of the recognition engine.

6. Acknowledgements

This research is being carried out under grant Nos. TIC2002-2273 and TIC2003-08756 from Programa de las Tecnologías de la Información y las Comunicaciones, Ministry of Education and Science, Spain.

7. References

[1] Van Compernolle, D. and Van Gerven, S. "Beamforming with Microphone Arrays", *Applications of Digital Signal Processing to Telecommunications*, pp. 107-131, E.U. 1995. COST 229.

[2] Saruwatari, H., Kawamura, T., Sawai, K., Kaminuma, A., Sakata, M., "Blind source separation based on fast-convergence algorithm using ICA and beamforming for real convolutive mixture", *Proc. of ICASSP'02*, 13-17 May 2002, Vol. 1, pp. 921, 924.

[3] Mokbel, C. E, and Chollet F. A., "Automatic Word Recognition in Cars", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, September 1995, pp. 346-356.

[4] Grbic, N., Nordholm, S., Johansson, A., "Speech enhancement for hands-free terminals", *Proc. of 2nd International Symposium on Image and Signal Processing and Analysis (ISPA 2001)*, pp. 435- 440.

[5] Low, S. Y., Grbic, N.; Nordholm, S., "Speech enhancement using multiple soft constrained subband, beamformers and non-coherent techniques", *Proc. of ICASSP'03*, Vol. 5, pp. 489-492.

[6] Elko, G. W., "Microphone array systems for hands-free telecommunication", *Speech Communication*, Vol. 20, No. 3-4, 1996, pp. 229-240.

[7] Teutsch, H., Kellermann, W., Elko, G., "First- and Second-order Adaptive Differential Microphone Arrays", *Proc. of the 7th International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, 10-13 September 2001.

[8] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, A. Rubio, "A New Voice Activity Detector Using Subband Order-Statistics Filters for Robust Speech Recognition", *Proc. of ICASSP 2004*, Vol. I, pp. 849-852.

[9] P. Gómez, A. Álvarez, R. Martínez, V. Nieto, V. Rodellar, "Time-Domain Steering of a Differential Beamformer for Speech Enhancement and Source Separation", *Proc. of 6th International Conference on Signal Processing*, Vol. I, 2002, pp. 338-341.

[10] ITU, "A Silence Compression Scheme for G.729 optimized for terminals conforming to recommendation V.70", *ITU-T Recommendation G.729-Annex B*, 1996.

[11] ETSI, "Voice Activity Detector (VAD) for Adaptive MultiRate (AMR) speech traffic channels", *ETSI EN 301 708*, 1999.

[12] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", *ETSI ES 2002 050*, v1.1.3, November 2003.

[13] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection", *IEEE Signal Processing Letters,* vol. 16, no. 1, pp. 1-3, 1999

[14] A. Moreno, et al., "SPEECHDAT-CAR: A Large Speech Database for Automotive Environments", *Proc. of 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000, paper 373.

[15] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, 2002.