# SPEECH ENHANCEMENT FOR A CAR ENVIRONMENT SUPPORT BY A FIRST-ORDER DIFFERENTIAL MICROPHONE

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

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

## ABSTRACT

Speech recognition is one of the key technologies to produce voice-control devices. However, the presence of different sources of degradation is an ubiquitous problem. This paper presents a robust microphone array processing technique to enhance speech under the influence of noise and reverberation in an automobile environment. The proposed structure combines a simple two-microphone *First Order Differential Null Beamformer* with *Spectral Subtraction* techniques. The paper also includes an evaluation of the performance of the algorithm in a real car environment. The results show a noticeable reduction in word error rates when the enhancement front-end is applied to a standard speech recognizer.

## 1. INTRODUCTION

Speech produced in a running car is perturbed by noise and reverberation. In those scenarios, a main objective consisting on increasing the contribution of the direct component relative to reverberant components of signals is pursued. As far as speech enhancement is concerned, microphone arrays and beamforming techniques have been widely applied, as they are able to perform dereverberation and noise suppression at the same time [1]. Usually, array beamforming is combined with other techniques as *Independent Component Analysis* [2][3], *Spectral Subtraction* [4] or *Linear Prediction Analysis* [5]. In the car environment, those structures may also benefit from the fact that desired speakers, fundamentally the drivers, are placed in a constrained region [6].

Through this paper, a speech enhancement system based on the use of a *First-Order Differential Microphone (FODM)* for reverberation and noise estimation purposes, and its application to *Spectral Subtraction* (*SS*) techniques is presented [7][8][9]. Differential microphone arrays provide high directional gain requiring a lower number of processing elements, as compared with common approaches [10]. Essentially, the method proposed in this paper applies a *FODM* operating in the time domain. The main aim of this procedure is to determine the contribution of desired speech signals in a specified constrained region against any other sources. Once the amount of noise and reverberation is estimated, the residual interference is then eliminated a by smoothing filtering in the frequency domain.

This method is intended to be a pre-processing stage of a *Robust Speech Recognizer* in automobile scenarios, as the one presented in Figure 1.

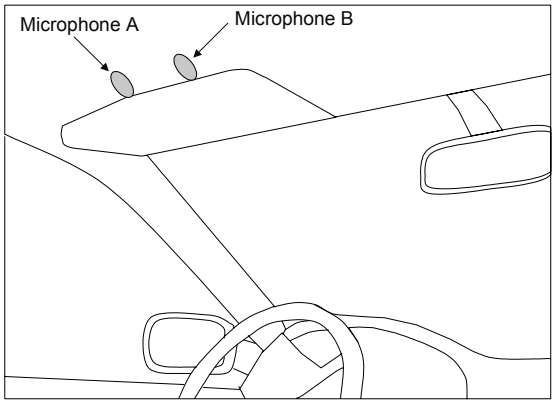


Figure 1. General framework of a two-microphone system devoted to robust speech recognition.

## 2. FIRST-ORDER DIFFERENTIAL MICROPHONE

### 2.1. Overview of the FODM

A differential microphone array consists of two omnidirectional sensors separated a distance *d*. Figure 2 shows the structure of the basic cell proposed in this work. As it may be seen that element operates in the time domain and resembles a delay-and-sum beamformer, where delays of both branches have a fixed value $\tau$. The final output is obtained subtracting the result of one individual branch from the other. Finally, channel mixing is controlled by a steering parameter $\beta$, where possible values of $\beta$ are restricted to the range *[0.0, 1.0]*.

In our case, the *FODM*, despite of its simplicity, recalls the behavior of a null beamformer controlled by the $\beta$ parameter. This mechanism shows a transfer function in the frequency domain, which may be formulated as

$$Y(\alpha,\delta) = 2e^{-j(\pi-\delta)/2}\left((1-2\beta)\cos\alpha \sin\delta/2 - \sin\alpha\cos\delta/2\right) \quad (1)$$

where

$$\alpha = \omega\zeta = 2\pi f = 2\pi f\zeta = \frac{2\pi fD}{c}\sin\varphi \quad (2)$$

$$\delta = \omega T = 2\pi fT = 2\pi fk\tau = 2\pi k\frac{f}{f_s} \quad (3)$$

$\varphi$ being the angle of arrival, $\zeta$ half the array travel time, $f$ the frequency of the signal, $k$ the delay order, $d=2D$ the microphone distance, and $f_s$ the sampling frequency.

This function shows for a frequency $f<f_s/2$ a sharp notch at an angle given by

$$\varphi_n = arcsin\{\frac{c}{2\pi fD} arctan[(1-2\beta)\ tan(\pi k f/f_s)]\} \qquad (4)$$

$c$ being the sound propagation speed.

The last expression establishes the relation between useful bandwidth and maximum-steering angle. As it may be noticed, the relation depends of three factors: sampling frequency, microphone separation and the fixed channel delay.

Figure 3 shows the results of evaluating *(1)*, when plotted against $f$ and $\varphi$ for one value of the steering parameter $\beta$. As it may be noticed, a fixed beta value not always implies the cancellation of source frequencies originated from a single angular direction of arrival. In fact, that is only the case for only three beta values: *0.0*, *0.5*, and *1.0*. Therefore, the value of $\beta$, which produces the highest degree of cancellation, depends not only on the value of the incoming angle $\varphi$, but also on the signal frequency $f$. This property of the *FODM* implies that for broadband signals like speech, the spectra of interest should be divided into different frequency bands, throughout the use of bandpass filters, and then replicating the beamformer cell previously presented.

More exactly, the relation among source angle of arrival $\varphi$, signal frequency $f$ and values of $\beta$, is given by the following expression

$$\beta = 1/2\left(1 - \frac{tan\left[\frac{2\pi fD}{c} sin\varphi_n\right]}{tan(\delta/2)}\right) \qquad (5)$$

### 2.2. Noise and reverberation estimation *FODM* based

The estimation of the amount of noise and reverberation presented in a speech signal captured by the array will be calculated as shown in Figure 4.

The null-beamforming task is executed only for the three directions that do not require a frequency-dependent solution. Those positions correspond to broadside ($\beta= 0.5$) and to the maximum steering angles ($\beta= 0.0$ and $\beta= 1.0$). In this case, broadside represents an average location for the car driver whereas the other two 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, the minimum of outputs, corresponding to boundary angles, constitutes a valid reference of the background disturbance level.

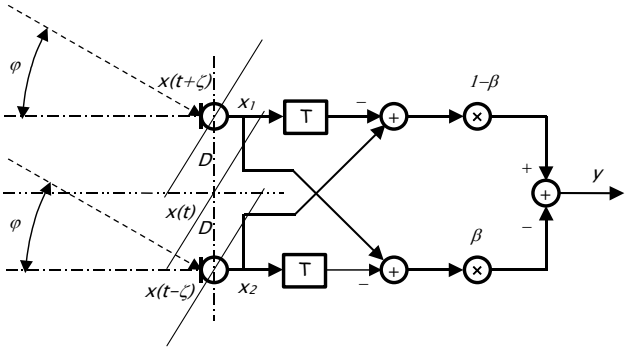


Figure 2. Structure of the two-microphone null beamformer. Microphones are separated a distance $d=2D$, being $\varphi$ the angle of arrival for an incoming sound source. The angular tracking factor is modeled through parameter $\beta$. This processing element introduces a delay interval $T=k\tau$, being $\tau$ the time delay unit.

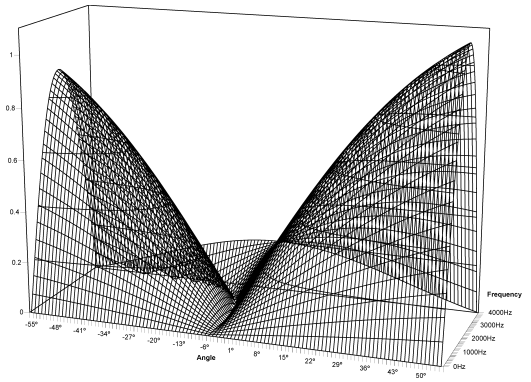


Figure 3. Module of the FODM transfer function for $d=5\ cm$, $k=1$, $f_s=8000Hz$, and $\beta=0.55$.

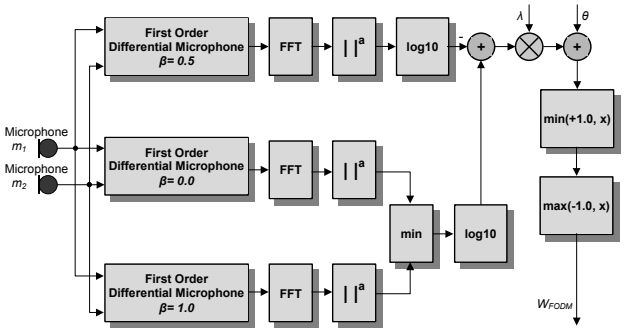


Figure 4. General framework of the structure based on the FODM to obtain subtraction weights $W_{FODM}$

This procedure takes place for every bin in the frequency domain, so that, *FODB* 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

$$W_{FODM}(m) = \lambda \log_{10}\left(\frac{\min\left\{\|S_{0.0}(m)\|^a, \|S_{1.0}(m)\|^a\right\}}{\|S_{0.5}(m)\|^a}\right) + \theta \qquad (6)$$

being $M$ the window size, $0 \le m \le M/2-1$ the frequency bins, $\lambda$ a gain factor, and $\theta$ a bias parameter.

As a final step, weights $W_{FODM}$ are fitted within the range *[0.0, 1.0]*.

## 3. SPECTRAL SUBTRACTION

To implement the filtering in the spectral domain, weights $W_{FODM}(m)$ will be considered relevant estimators. The procedure we proposed may be seen in Figure 5.

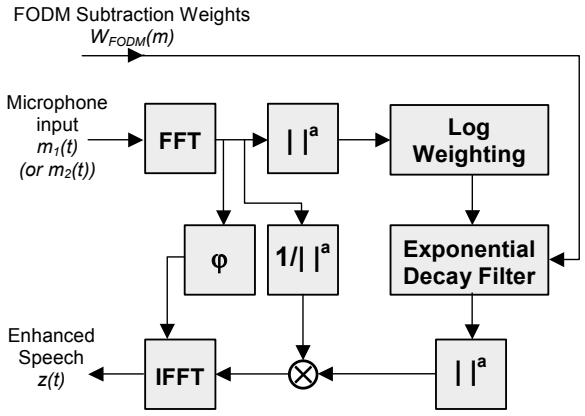


Figure 5. Structure of the spectral subtraction module that exploits $W_{FODM}$ estimators.

First of all, one of the microphone input signals $s(t)$, corresponding to $m_1(t)$ or $m_2(t)$, is segmented in overlapped windows and transformed into the frequency domain applying the same configuration introduced in the last section. After that, these values are passed to a filter with exponential decay given by

$$g_S(m) = \alpha \log_{10}\left(\|S(m)\|^a\right) + (1-\alpha) g_{n-1}(m) \qquad (7)$$

where $\alpha$ is a coefficient that controls the log-power rate update.

Once we have adapted the incoming signal energy, the calculation of the subtracting-signal at frame index $n$ and frequency index $m$ or $g_n(m)$, is accomplished by a new exponential decay filter controlled by the set of weights $W_{FODM}(m)$, previously studied

$$g_n(m) = (1 - W_{FODM}(m)) g_S(m) + W_{FODM}(m) g_{n-1}(m) \qquad (8)$$

As it may be noticed, the above expression implies that a weight equal to 1.0 prevents from updating the estimation of $g_n(m)$ at all. On the other hand, a weight close to 0.0 produces a fast adaptation.

Finally, the exact amount to be subtracted is generated and the subtraction itself is performed, producing an enhanced signal in the time domain $z(t)$

$$G(m) = 10^{g_n(m)} \qquad (9)$$

$$\|Z(m)\|^a = \|S_{en}(m)\|^a - G(m) \qquad (10)$$

## 4. RESULTS AND DISCUSSION

In order to examine the validity of the method proposed, several speech recognition systems were built and tested, using the framework presented in Figure 6. The method proposed through the paper provides a robust preprocessing stage to the shared speech recognition engine. A pair of examples, related to an original speech signal and a processed or enhanced one may be seen in Figure 7 and Figure 8.

A subset of the Aurora3-SpeechDat Car Finnish database is used for testing purposes. The corpus, which contains realizations of connected digits uttered in a realistic automobile environment, is divided in two different groups: train and test. Each group has three different categories related with the amount of distortion contained in the recordings: quiet, low, and high. In our experiments, we use the recordings associated to channels *ch2* (microphone placed at the ceiling of the car in front of the speaker behind the sunvisor) and *ch3* (microphone installed at the ceiling of the car over the mid-console and near the rear mirror). As it may be noticed, that configuration is exactly the one previously introduced in Figure 1.

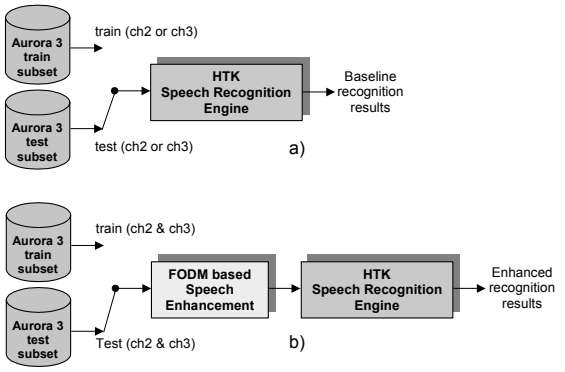


Figure 6. a) Baseline HTK based speech recognizer. b) Enhanced system incorporating the method proposed in this paper to the same speech recognition engine.

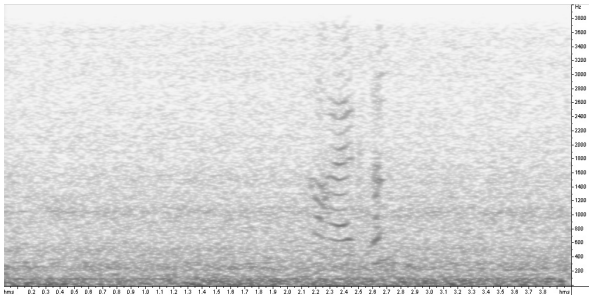

Figure 7. Power spectrum of an utterance contained in the Aurora 3 database and produced by a male speaker.

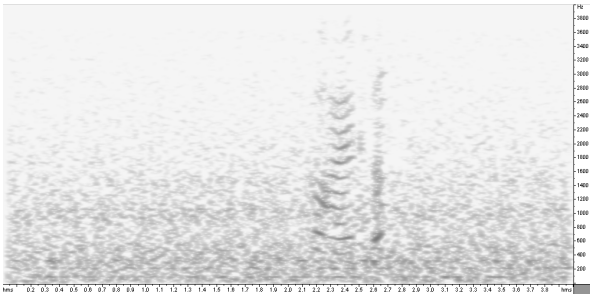

Figure 8. Power spectrum of the enhanced signal associated to signal in Figure 7, being $\alpha=0.33$, $\lambda=5.0$ and $\theta=0.5$.

It is important to remark regarding SpeechDat Car databases that although several microphones inputs are simultaneously captured, those compilations are not array-processing oriented. In particular, two aspects are relevant related to microphones $m_2$ and $m_3$: sensors have different characteristics as they are models produced by different manufacturers and, distance between them is not defined

precisely. However, the spread availability of the corpus makes it suitable for robust speech recognition evaluations.

The recognition experiments are established by selecting different material 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. The front-end extracts energy plus 36 MFCCs (12 cepstrum, 12 delta cepstrum and 12 delta-delta coefficients). The HMMs are built with 16-state whole word models for each digit in addition to a begin-end model and a word-separation one. Finally, models have 3 diagonal Gaussian mixture components in each state.

Table 1 contains *Word Error Rate* (*WER*) results for the two baseline systems. Baseline means no enhancement is applied to signals linked to microphone ch2 or microphone ch3. As it may be clearly seen, there is a high disparity between recognizer performances when channel ch2 is selected as system input as opposed to when ch3 is chosen.

| ch2 | Deletions | Substitutions | Insertions | WER |
|---|---|---|---|---|
| Train set A | 104 | 491 | 916 | 48.34% |
| Train set B | 62 | 71 | 234 | 11.74% |
| Train set C | 62 | 53 | 196 | 9.95% |
| **ch3** | **Deletions** | **Substitutions** | **Insertions** | **WER** |
| Train set A | 271 | 204 | 275 | 23.99% |
| Train set B | 80 | 47 | 172 | 9.56% |
| Train set C | 70 | 51 | 208 | 10.52% |

Table 1. Baseline recognizer results for microphones ch2 and ch3.

Table 2 presents the results when the method proposed in this paper is applied as a pre-processing stage to the same front-end. The improvement is significant for both channels and the three training sets except for channel *ch3* and train *set A* (high mismatch between training and testing material). Finally, the relative improvement is summarized in Table 3.

| FODM | Deletions | Substitutions | Insertions | WER |
|---|---|---|---|---|
| Train set A | 94 | 401 | 498 | 31.77% |
| Train set B | 58 | 84 | 121 | 8.41% |
| Train set C | 48 | 65 | 162 | 8.80% |

Table 2. Recognition results for the enhanced system.

| ch2 | Deletions | Substitutions | Insertions | WER |
|---|---|---|---|---|
| Train set A | 9.62% | 18.33% | 45.63% | 34.28% |
| Train set B | 6.45% | -18.31% | 48.29% | 28.34% |
| Train set C | 22.58% | -22.64% | 17.35% | 11.58% |
| **ch3** | **Deletions** | **Substitutions** | **Insertions** | **WER** |
| Train set A | 65.31% | -96.57% | -81.09% | -32.40% |
| Train set B | 27.50% | -78.72% | 29.65% | 12.04% |
| Train set C | 31.43% | -27.45% | 22.12% | 16.41% |

Table 3. WER reduction for the enhanced system compared to baseline recognizers.

## 5. CONCLUSIONS

The combination of a *First-Order Differential Microphone* structure and *Spectral Subtraction* techniques constitutes an efficient approach to the speech enhancement problem in noisy and reverberant environments. The proposed method requires neither a *Voice Activity Detecto*r nor *a priori* knowledge of the working framework. Speech recognition experiments carried out with real data taken form the Aurora 3 database show a noticeable reduction in word error rates, especially if strong sensor response mismatches have to be assumed.

The simplicity of the audio acquisition equipment and a moderate computational complexity of the solution allow building end-user products at a reasonable cost.

## 6. ACKNOWLEDGEMENTS

This research is being carried out under grants TIC99-0960, TIC2002-02273 and TIC2003-08756 from the *Programa Nacional de las Tecnologías de la Información y las Comunicaciones (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] Barros, A. K., Itakura, F., Rutkowski, T., Mansour, A.; Ohnishi, N., "Estimation of speech embedded in a reverberant environment with multiple sources of noise" Proc. of ICASSP´01, May 7-11 2001, Vol. 1, pp: 629- 632.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", IEEE Trans. on ASSP, Vol. 27, 1979, pp. 113-117.

[10] Buck M., Rößler M., "First Order Differential Microphone Arrays for Automotive Applications", *Proc. of the 7th International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, 10-13 September 2001.

[11] 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.