Cepstral Speech/Pause Detectors

Petr POLLÁK, Pavel SOVKA, Jan UHLÍŘ
Czech Technical University, Faculty of Electrical Engineering, K331
Technická 2, 166 27 Prague 6, CZECH REPUBLIC
E-mail: pollak@feld.cvut.cz Fax: (+42 2) 2431 0784

ABSTRACT — **This paper describes two algorithms for speech/pause cepstral detectors. Integral cepstral algorithm and differential algorithm based on differenced cepstrum. Both algorithms use smoothing procedure based on median filtering. New criteria have been used for detectors comparisons. Many experiments confirmed detectors reliability and their ability to detect speech in real car noise with high probability.**

1 Introduction

Many systems for noisy speech processing usually require reliable speech/pause detector. While energy based detectors often fail, cepstral ones give promising results. Our basic requirements for detector are:

- Detector should be implemented in the frequency domain.
- The detector should enable a real-time implementation (only a removable non-causality is allowed). Information should be got from a small number of actual signal segments only.

Two cepstral speech/pause detectors - *integral* and *differential* will be now described.

2 Integral algorithms

One step algorithm. The simplest version originates from [2], [3], [6], [5], [4]. The distance between a current cepstral vector $\mathbf{c}[n]$ and a short-time average of background cepstral vector $\overline{\mathbf{c}}[n]$ is computed. The used cepstral distance is (time index $[n]$ is omitted for simplicity)

$$\Delta c = 4.3429\sqrt{(c_0 - \bar{c}_0)^2 + 2\sum_{k=1}^{p}(c_k - \bar{c}_k)^2} \qquad (1)$$

with corresponding spectral distance [7]

$$\mathrm{Ds} = \frac{4.3429}{2\pi}\int_{-\pi}^{\pi} \ln^2 \frac{|S(e^{j\omega})|^2}{|\overline{S(e^{j\omega})}|^2}\, d\omega\,. \qquad (2)$$

The cepstral vector is $\overline{\mathbf{c}}[n]$ updated in pauses only. The decision about speech presence is based on simple statistics. Long time mean value $\overline{\Delta c_N}[n]$ and standard deviation $\mathrm{dv}\,(\Delta c_N)$ of $\Delta c[n]$ (see (3)) are computed in pauses only. (The number of terms of the sequence $\Delta c[n]$ should be greater than 30.) The speech is detected if a current value $\Delta c[n]$ is greater than the threshold $\Delta c_{th}[n]$ defined by Eq. (4). The parameter $z_{\alpha/2}$ controls the number of false speech detections. The whole detection algorithm is as follows

$$\Delta c[n] = \mathrm{dist}\,(\mathbf{c}[n], \overline{\mathbf{c}}[n])\,, \qquad (3)$$

$$\Delta c_{th}[n] = \overline{\Delta c_N}[n] + z_{\alpha/2}\,\mathrm{dv}(\Delta c_N[n])\,, \qquad (4)$$

$$\mathrm{OUT} = \begin{cases} 0, & \text{if } \Delta c[n] < \Delta c_{th}[n]\,, \\ 1, & \text{if } \Delta c[n] \geq \Delta c_{th}[n]\,, \end{cases} \qquad (5)$$

$$\overline{\mathbf{c}}[n+1] = \begin{cases} (1-\lambda)\overline{\mathbf{c}}[n] + \lambda\mathbf{c}[n], & \text{if OUT} = 0\,, \\ \overline{\mathbf{c}_{ST}}[n], & \text{if OUT} = 1\,. \end{cases} \qquad (6)$$

The parameter λ specifies the time constant of short-time exponential averaging. This detector is shown as the first block of detector on fig. 1. The detector needs some initial phase for the adaptation to background noisy characteristics.

Two step algorithm. Because the output of one step algorithm usually contains bad decisions because of fluctuations of background noise characteristics the two step algorithm was designed. If the sequence $\Delta c[n]$ is properly smoothed undesirable fluctuations are removed. Because signal edges should be preserve median filtering was used

$$\Delta c_m[n] = \mathrm{med}\,(\Delta c[n], m)\,. \qquad (7)$$

The decision about speech presence is made by the same rule as in the first step: according the Eqs. (4) to (6) applied to the smoothed cepstral distance sequence $\Delta c_m[n]$.

The order of the median filter m has the strong influence upon detector results. False decisions

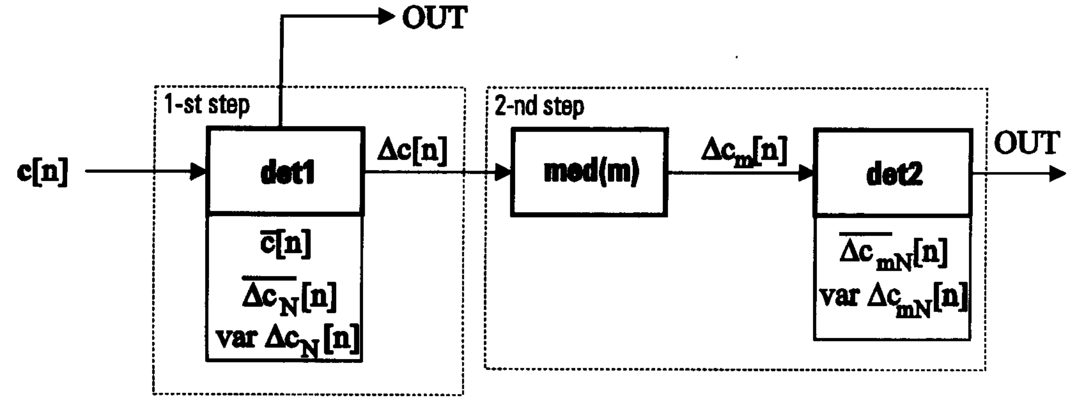

Figure 1: Statistically based integral speech detector

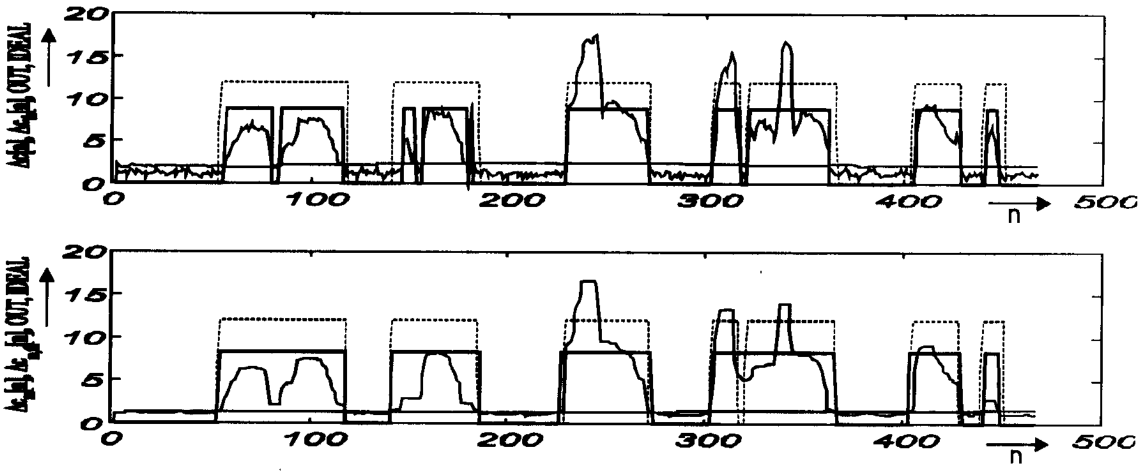

Figure 2: Illustrative example of smoothing and detection process.

are removed better for higher values of m but, on the other hand, bad determinations beginnings and ends of speech sequences appear. The block scheme of the whole algorithm with two steps is on Fig. 1. Time responses of main detector parameters given by Eqs. (4) to (6) are on Fig. 2.

3 Differentially based algorithms

Eq. (1) can be rewritten in the form

$$D[n] = \sqrt{\delta_0^2[n] + 2\sum_{k=1}^{p} \delta_k^2[n]}, \qquad (8)$$

where $\delta_k[n]$ stands for a time derivative of cepstral coefficients $c_k[n]$ (the polynomial approximation to the derivative of $c_k[n]$ [1]), which is called the differenced cepstrum. Differenced cepstrum written in a vector form is given by

$$\boldsymbol{\delta}[n] = \left(\sum_{i=-K}^{K} i \cdot \mathbf{c}[i+n]\right) \Big/ \sum_{i=-K}^{K} i^2. \qquad (9)$$

This equation is labelled as the block "diff" on Fig. 3. Eqs. (8) and (9) lead to other possible modifications of cepstral detectors.

Simple differential algorithm. The polynomial approximation to the time derivative (8) can be replaced by a simple backward difference operation, but $\delta[n]$ is rather noisy in this case. But if $\delta[n]$ is smoothed using a cumulative sum and then the cepstral distance similar to (8) is computed (second box on Fig. 3) , we can get a simple and robust algorithm for the speech/pause detection. The threshold is again determined by Eq. (4).

Combined algorithm. The sequence $D[n]$ computed using (8) is rather erratic. Every sudden change in an input signal generates a sharp peak in $D[n]$. Many peaks, however, are generated even if speech is not present. That is why a smoothing procedure must be applied to suppress false peaks and to combine and smooth correct peaks. Three possibilities of smoothing were tested: linear filters, median filters and nonlinear median hybrid filters [8]. The median hybrid filter had the best results. The improvement of a detector behaviour can be achieved by a combination of the cepstral distance (8) with cepstral distances computed from liftered and weighted cepstra. We used the raised sine lifter and the index weighting $n\mathbf{c}[n]$ (see Fig. 3). The liftering allows to suppress artifacts of the LPC analysis and the weighting enables to detect more details in unvoiced sounds. Sequences $\mathbf{c}[n]$, $n\mathbf{c}[n]$ and liftered $\mathbf{c}[n]$ give three cepstral distances D, DII and $DIII$ respectively (time index [n] is ommited). These three distances are normalised and then combined to give the resultant cepstral distance $Do = \sqrt{k_1 D^2 + k_2 DII^2 + k_3 DIII^2}$, where k_i are normalisation parameters. The final detector output is the result of using the median hybrid filter [8] on the sequence Do. The computation of the threshold and the speech detection are given by (4) and (5) respectively.

The behaviour of the whole system is dependent on the choice of the derivative $\delta_k[n]$, the value of the parameter $z_{\alpha/2}$ and the order of recursive median filter. We found that possible range for $z_{\alpha/2}$ is from 1.5 to 3.5. An appropriate interval for the median filter order is from 6 to 10. The precise determination of speech activity requires a nonlinear smoothing rather than a linear one and a lower order of the polynomial approximation (9) ($K < 6$).

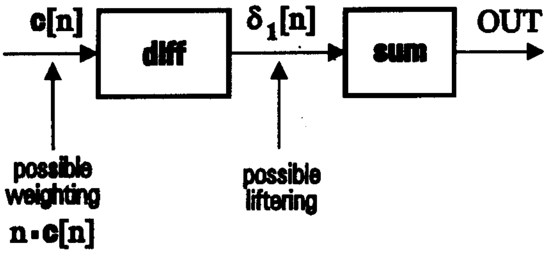

Figure 3: Block scheme of differential algorithms

4 Experiments

All algorithms were experimentally tested on the PC. All experiments were realized with the real signals recorded in running cars and collected in the database of signals with manually marked speech sequences.

Classification. We tried to determine reliable objective criteria for the comparison of different algorithms. These criteria are based on the computation of correct detection rates. *Particular criteria* were established : *correct speech detection rate* - $P(A/S)$ and *correct non-speech detection rate* - $P(A/N)$. Another possibility is the using of ***global criteria***: *correct detection rate* - $P(A)$ and *speech/non-speech resolution factor* - $P(B)$ defined as $P(A) = P(A/S)P(S) + P(A/N)P(N)$ and $P(B) = P(A/S)P(A/N)$, where $P(S)$ and $P(N)$ are rates of speech and pauses in the processed signal.

Detectors were tested under different noisy conditions. The more detailed classification was gained from histograms and distribution functions of each criterion. A rough information was got from mean values and standard deviations of each criterion. Tab. 1 contains global numerical results of all experiments. There is shown the improvement of results after the second step of the integral algorithm with median filtering there. The pause detection is worse but both global parameters are better. Differential algorithms give very good detections of pauses but worse detections of speech. Distribution functions are given on (Fig. 4 and Fig. 5).

	$P(A/S)$	$P(A/N)$	$P(A)$	$P(B)$
INT-1	0.857 0.106	0.958 0.051	0.902 0.062	0.820 0.104
INT-2	0.925 0.062	0.911 0.112	0.919 0.047	0.839 0.103
DIF-1	0.765 0.207	0.981 0.031	0.859 0.125	0.748 0.197
DIF-2	0.864 0.247	0.977 0.031	0.895 0.141	0.762 0.235

Table 1: Averaged results of experiments. Mean values and standard deviations:

INT-1 - one step integral algorithm,

INT-2 - two step integral algorithm,

DIF-1 - simple differential algorithm with the backward difference,

DIF-2 - combined differential algorithm.

5 Conclusions

A lot of experiments confirmed posibble using of algorithms for the speech detection and the preprocessing in a real car noise environment. Algorihms seem to be reliable for the precise speech/pause detection. The most critical point

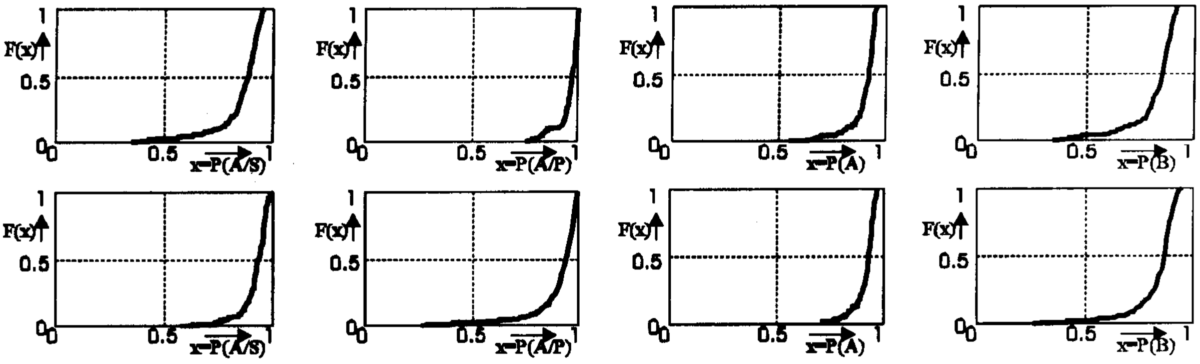

Figure 4: Distribution of results of experiments on integral detectors. Upper fig.: one step algorithm - INT-1, lower fig.: two step algorithm.

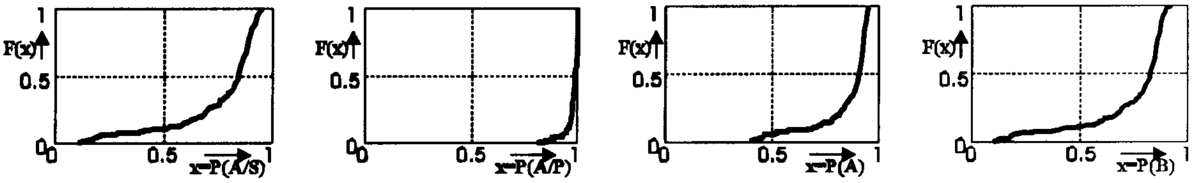

Figure 5: Distribution of results of experiments on differential detector with the first backward difference - DIF-1.

of all algorithms is the computation of thresholds. More sophisticated solution of this problem is now under investigation. Also the combination of integral and differential algorithms is studied, because the integral algorithms are better in the speech detection and the differential ones in the pauses detection.

References

[1] L. R. Rabiner and B-H Juang. *Fundamentals of Speech Recognition.* Murray Hill, New Yersey, USA, 1993.

[2] A. H. Gray, Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-24(5):380–391, October 1976.

[3] W. A. Harrison and J. S. Lim and E. Singer. A new application of adaptive noise canceller. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-34(1):21–27, February 1986.

[4] J. A. Haigh and J. S. Mason. A voice activity detector based on cepstral analysis. In *EUROSPEECH*, pages 1103–1106, Berlin, September 1993.

[5] L. R. Rabiner and M. R. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-25(4):338–343, August 1977.

[6] A. Le Floc'h and R. Salami and B. Mouy and J.-P. Adoul. Evaluation of linear and nonlinear subtraction metods for enhancing noisy speech. In *Speech Processing in Adverse Conditions*, pages 131–134, Cannes-Mandelieu (France), November 1992.

[7] I. A. Kozjuchovskaja. Ocenka tocnosti vycislenij parametrov recevovo signala dlja progozirujuscevo vokodera. *Elektrosvjaz*, 1984.

[8] Y. Neuvo R. Wichman and P. Heinonen. Firmedian hybrid filters with excelent transient response in noisy conditions . In *Digital Signal Processing '87*, pages 171–175, Italy, September 1987.