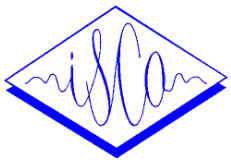

ESCA Tutorial and Research Workshop on
Speech Processing in Adverse Conditions
Cannes-Mandelieu, France
November 10-13, 1992

Improving recognition rate in adverse conditions by detection and noise suppression

Dominique PASTOR , Christian GULLI

Sextant Avionique, rue Toussaint-Catros, B.P.91,
33160 SAINT MEDARD EN JALLES, FRANCE.

This paper describes the signal processing performed in order to improve the recognition rates of a DTW algorithm in flight noisy environments, and the results obtained during flight tests including G-load .

1. INTRODUCTION

Sextant Avionique Research and Development programs include the design of a Speech Recognition System which is to be used in combat aircraft cockpits, therefore in noisy environments and under G-load.

The Speech Recognition algorithm is based on the Dynamic Time Warping algorithm (D.T.W.).

So, the goal is to improve the recognition rate in noisy environments and under G-load by signal preprocessings. This approach is complementary to the work described in [8], where the G-load effect alone (without noise) has been studied.

Speech Detection

A major cause of errors in speech recognition in noisy environments lies in the inaccurate detection of the endpoints of test and in the difficulty to separate speech-segments from non-speech.

A word boundary detector has been designed, according to the recommended characteristics described in [1]: reliability, robustness, accuracy, adaptation, real-time processing and no a priori knowledge of the noise.

Noise cancellation

In order to improve the speech recognizer performances, noise reduction algorithms based on spectral substraction and Wiener filtering have been performed and tested, such techniques being consistent with the basically spectral analysis used, (cf section 4). These algorithms require estimations of the average noise spectrum magnitude. Consequently, a robust discrimination between speech-segments and non-speech is required too, making this point common with the previous one.

Detecting noise frames allows an estimate of noise parameters.

2. SPEECH-NOISE DISCRIMINATION

2.1 General description

As in [1], [2], [3], the main parameters used are : pitch extraction, energy thresholding, heuristic information.

From a general point of view, the sentence to process is decomposed as following :

- firstly, we point out the set of all the frames between the first voiced frame and the last voiced frame,
- secondly, on each side of this set of frames, may still remain unvoiced speech frames as in "STOP", or not, as in "AUTO".

Therefore, the first step intends to determine the first and the last voiced frames, and the second one to determine the unvoiced speech frames (if they exist) next to these voiced frames.

2.2 PITCH EXTRACTION

Two main techniques have been pointed out and tested.

The first one uses the correlation function defined as it follows. Considering N signal samples x(0),..., x(N-1), the correlation function is :

$$\Gamma_x(k) = (1/N)\Sigma_{0 \le n \le N-1}\ x(n)x(n-k)$$

The second method uses the Average Magnitude Difference Function computation. Considering N signal samples x(0),..., x(N-1), the AMDF function is defined as :

$$AMDF(k) = \Sigma_{0 \le n \le N-1}\ |\ x(n) - x(n-k)\ |$$

AMDF(k) is tied to the correlation function by the

relation : $AMDF(k) \leq \lambda\ 2[\Gamma_X(0)-\Gamma_X(k)]^{1/2}$ (cf [4])
In each case, we compute a function r(k) ($r(k) = \Gamma_X(k)$ ou $r(k) = AMDF(k)$) and it can be shown that the maximum of r(k) is always obtained when $k = 0$. The pitch value is Fe / k_0, where $k = k_0$ is the second maximum of r(k) . In order to determine k_0, the function r(k) has to be thresholded. The choice of the threshold still remains an heuristic one. The correlation function remains more robust to adverse conditions than the AMDF. This can be explained by mathematics: the AMDF is only a non-euclidean distance between the signal and its delayed version, even though the correlation function is a scalar product, an orthogonal projection which eliminates white noise.

The pitch extraction algorithm briefly described above will only indicate if a frame is voiced or not. Since the goal is to determine the first and the last voiced frames of the utterance, an expertise based approach, taking into account heuristic informations about the voiced speech structure, is required and has been performed.

2.3 Unvoiced Frames Detection

Since the set of frames between the first and the last voiced frames has been determined, unvoiced frames which may exist here and there from this set have to be detected. Several algorithms based on mathematical results about the density of probability of the ratio of two energies have been studied. The best results are obtained by the following process :
- amid the frames preceding the set bounded by the first and the last voiced frames, new detection algorithms based on energy thresholding and application of the maximum likelihood principle detect the noise frames,
- these ones allow to compute average noise spectrum estimations required by the noise reduction algorithm, and used in order to "whiten" the noise,
- the noise being "whitened", it can be shown that detecting unvoiced frames by energy thresholding becomes easier, using the maximum likelihood principle. As the pitch detection algorithm, this one is improved by an expertise based approach which takes into account the special nature of the unvoiced speech frames.

3. NOISE REDUCTION

Since average noise spectrum magnitude have been computed , it becomes easier to compute spectral substraction or Wiener filtering (cf [5], [6]). The process is described here under :
- the transfert function of a spectral substraction is : $H(\nu) = 1 - \mu/|X(\nu)|$ where $X(\nu)$ is the current signal spectrum, μ the average noise spectrum magnitude. In theory, $\mu = E[|B(\nu)|]$ where $B(\nu)$ represents the noise spectrum, and μ has already been estimated by average on the detected frames of noise,
- the transfert function of Wiener Filtering is $H(\nu) = 1 - E[|B(\nu)|^2]/E[|X(\nu)|^2]$. $E[|B(\nu)|^2]$ is estimated by using the detected frames of noise, and $E[|X(\nu)|^2]$, by smoothed correlogram (cf [7]) on the current frame.

Such spectral techniques are perfectly consistent with the classical speech analysis used as illustrated by figures 1 and 2 and described in the next section.

4. IMPROVED SYSTEM SYNOPTIC

The process briefly described above has to be inserted in the final speech recognition system. Figure 1 shows the original synoptic, without any detection and noise reduction. Figure 2 presents the

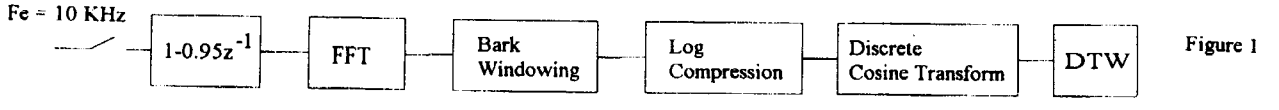

Figure 1

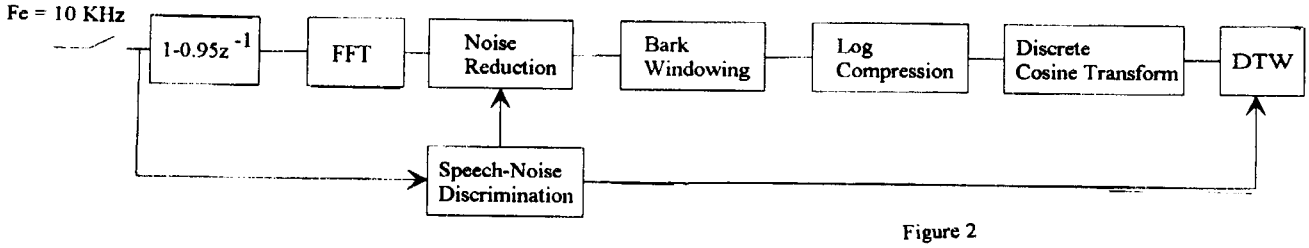

Figure 2

final synoptic. It shows the new ties existing now between detection, noise reduction and the DTW algorithm. The improved system requires more computation, and is more complex than the original one. Consequently, the complete speech recognizer including signal processing and DTW algorithm, will be implemented using a 96002 DSP (dedicated to signal processing) and a R3000 processor (dedicated to DTW algorithm).

5. EXEMPLE

Figures 3 represents the spectrogram (Bark's representation) and the temporal representation of an original noisy speech signal (observed during real flight under G-load). Above the spectrogram, is

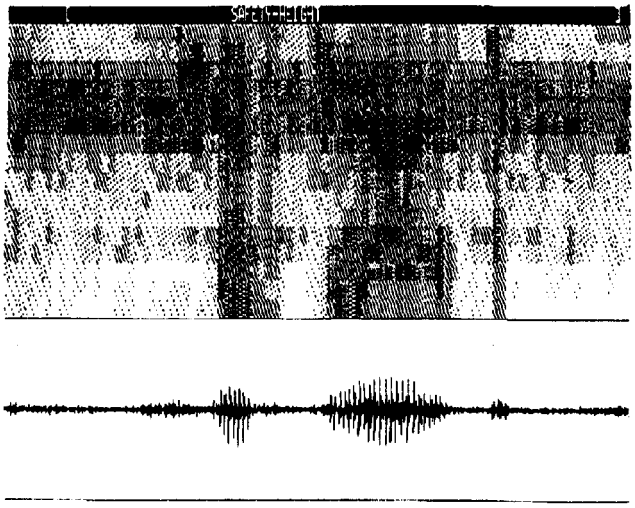

Figure 3

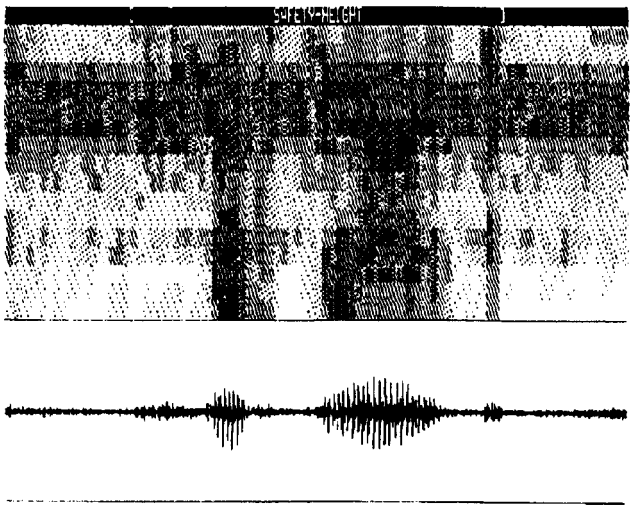

Figure 4

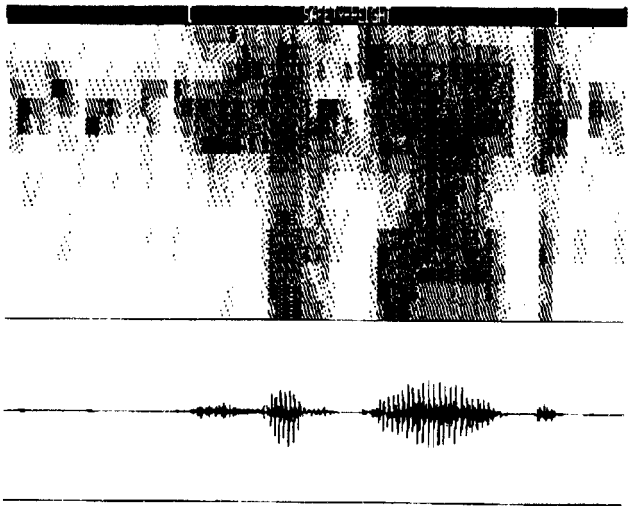

Figure 5

given the utterance to process and the brakets indicates the pilot's original push to talk (PTT).

The time interval bounded by the PTT switching on and off is too wide and doesn't allow a good speech recognition. Figure 4 represents then the result obtained after word boundary detection.

Figure 5 gives the result after speech detection and noise reduction.

6. ADVANTAGES

More than an algorithm, the technique suggested here is basically a process :

- based on signal and noise modelisations,
- using statistical theories.

For example, the statistical and mathematical hypothesis used are tied to different parameters whose effects are, physically and mathematically, well-known. Therefore, if, for any reason, the process fails, looking what happens at each step allows to modify the appropriate parameter.

The whole process is therefore robust to adverse conditions and accurate, while preserving flexibility during the development phase.

There is no a priori knowledge of the noise, and each detection is adapted to the current utterance. The next one will be processed independtly.

Due to these reasons, even if the statistical properties of noise change during flights, good results will be still provided by the algorithm.

7. EVALUATIONS

Table 1 gives the results obtained using tests recorded during flights on a Mirage IIIB, under G-load. The vocabulary is a restricted one, because of the difficulties to pronounce long sentences under G-load effects: the syntax involves only 36 words, allowing 8 linked words.

Two speakers (speaker 1 and speaker 2) took part significantly in these experiments.

Speaker 1 appears twice (denoted by speaker 1'), because he used two different oxygen masks.

The nomenclature used is the following one :

- PTT : results obtained, using only the pilote's original Push To Talk,
- SD : results provided by Speech Detection (without Noise Cancellation),
- SD + NC : results provided by the complete algorithm, including Speech Detection and Noise Cancellation,
- PWB : results which would be obtained if we used a Perfect Word Boundary, performed in laboratory.

In each column, the number of errors and the number of utterances are given, as well as the recognition rate: for example, 12/30 (60%) indicates 12 errors amid 30 utterances, and the recognition rate is then 60 %.

Additional evaluations can be found in [8], where the G-load effect was studied alone, using experiments in centrifugal machine. The results of [8] confirm the results presented above.

8. CONCLUSIONS

New algorithms have been described, improving speech recognition by DTW in noisy environments.

Our method provides a general way of speech preprocessing, which tends to comply with the constraints recalled in the first section (reliability, robustness, real-time processing...) .

This process could also be applied to others speech recognition systems used under adverse conditions.

We are going to realize such tests, using a word recognition based on Hidden Markov Models

REFERENCES :

[1] J.C.Junqua, B.Reaves, B.Mak, A Study of Endpoint Detection Algorithm in Adverse Conditions : Incidence on a DTW and HMM Recognizer, Eurospeech 1991

[2] R.Boite, M.Kunt, Traitement de la Parole, Presses polytechniques romandes

[3] V.Petit, F.Dumont, La Discrimination Parole-Bruit et Ses Applications, Revue technique Thomson-CSF- Vol.12, N° 4, Dec. 1980

[4] P.Wacrenier, Problème de détection des frontières de mots en présence de bruits additifs, Mémoire de D.E.A., Université Paris-Sud, Centre d'Orsay.

[5] Speech Enhancement, Edited by J.S.Lim, Prentice-Hall, Signal Processing Series,

[6] S.F.Boll, Suppression of Acoustic Noise in Speech Using Spectral Substraction, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, N°2, April 1979.

[7] P.Comon, Traitement Fréquentiel de Signaux Multivariables, Rapport CEPHAG 58/85

[8] C.Gulli, D.Pastor, A.Leger, P.B.Sandor, J.M.Clere, P.Grateau, G-Load Effects and Efficient Acoustic Parameters For Robust Speaker Recognition, AGARD 1992

Table 1 : Results on Real Flights

flights	PTT	SD	SD+NC	PWB+NC
speaker 1 - 2g	5/36 (86.1 %)	2/36 (94.4 %)	0/36 (100 %)	0/36 (100 %)
speaker 1 - 4g	4/60 (93.3 %)	5/60 (91.6 %)	3/60 (95 %)	3/60 (95 %)
speaker 1 - 5g	3/28 (89.2 %)	4/28 (95.1 %)	1/28 (96.4 %)	1/28 (96.4 %)
speaker 1' - 2g	12/30 (60 %)	6/30 (80 %)	3/30 (90 %)	2/30 (93.3 %)
speaker 2 - 2g	39/48 (18.75 %)	11/48 (77 %)	4/48 (91.6 %)	2/48 (95.8 %)
speaker 2 - 4g	53/55 (4 %)	23/55 (58 %)	13/55 (76.3 %)	8/55 (85.4 %)