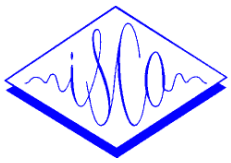

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ROBUST SIGNAL PREPROCESSING FOR HMM SPEECH RECOGNITION IN ADVERSE CONDITIONS.

Jean-Baptiste PUEL, Regine ANDRE-OBRECHT.

IRIT - URA CNRS 1399 - UNIVERSITE PAUL SABATIER
118, route de Narbonne - 31062 Toulouse Cedex
FRANCE

ABSTRACT

The detection of speech endpoints is a strategic process for speech recognition systems in adverse conditions, but it remains a rather delicate problem. We introduce two signal processing methods that offer a good robustness without requiring high level informations about the signal. The first approach uses temporal parameters, the other frequential ones. We discuss and compare their performances using the ARS ESPRIT database (isolated words pronounced in a car). We show that these methods coupled with a statistical segmentation offer very good discrimination between noisy segments and speech segments, and a better precision for locating the speech boundaries.
The preprocessing is introduced in a HMM speech recognition system.

I. INTRODUCTION

The automatic speech recognition in adverse conditions presents some difficulties that deteriorate significantly the performance of systems designed in silent conditions.
When the conditions are noisy, the noise involves an additive distortion of the signal and a modification of the sound production : the Lombard effect [Lombard 11].

Very often, endpoint detection algorithms are used to improve the preprocessing of noisy speech.
Most of these methods are based on parameters like the signal energy, the residual energy, sometimes limited to specific frequency bands; threshold automatons take the decisions. Generally, the boundaries are systematically released in order to avoid word truncations [Lamel 81], [Junqua 92].

So, in order to improve the performance of an automatic speech recognition system, we propose some signal preprocessing designed to remove noise of the signal and localize the speech endpoints. This preprocessing consists of three principal parts we detail later :

- a statistical segmentation of the signal using an AR model,
- an accurate detection of the noise / speech frontiers, we propose two methods
  - a temporal analysis for very low SNR (about 0 dB)
  - a frequential analysis for medium SNR (15 to 30 dB).
- noise spectral substraction [Lockwood 92].

This preprocessing coupled with a spectral analysis provides the input data of an HMM speech recognition system.

The organisation of the paper is as follows.
In section 2, the automatic segmentation algorithm is reminded ; in section 3, the two speech / non speech detection algorithms are described and the section 4 deals with the implementation of this pre-processing in a speech recognition system based on HMM. In the last section, an experimental evaluation and some conclusions are presented.

II. THE SEGMENTATION ALGORITHM

The signal is assumed to be described by a string of homogeneous units, each of which is characterized by an AR model :

$$y_n = \sum_{i=1}^{p} a_i\, y_{n-1} + e_n$$

$$\text{var } e_n = \sigma^2$$

The model is parametrized by the vector $\lambda = (a_1,..., a_p, \sigma)$.
This method consists in performing on line a detection of changes in the model parameters. The test is based on the monitoring of a suitable distance between the two models λ_0 and λ_1, located as indicated in figure 1.

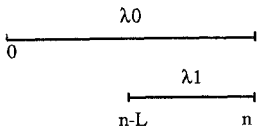

Figure 1 : localisation of the two models

This distance measure is derived from the cross entropy between the conditional distribution of these models.
The experimentations show that the performance of this segmentation algorithm are corpus, speaker independent and condition independent. No parameter learning are necessary to process new data.
The segments which are obtained can be classified into three categories :
- stationary segments which correspond to the steady part of sound,

• transient segments during which a formantic structure exists and its behavior remains monotonous,
• short segments (about 10ms) which are rapid articulatory change, as a plosive burst.

Some systematic omissions are caused by the asymmetric behavior of the statistics, so too long segments are processed again, in the backward sense, by the same test. For more details, see [Andre-Obrecht 88].
When we experiment this segmentation algorithm on noisy database, we observe that the speech endpoints are always detected, but they are not identified as such.

III. THE ENDPOINT DETECTION ALGORITHM

3.1 The temporal analysis :

We exploit the fact that the "curvilinear abcissa" must grow faster in voiced speech zones than in noise zones.
The temporal analysis is done in two steps, we first decide for each segment if it is speech or noise. Then we use a dynamic coordination to validate the label sequence.

The static labelling :

In order to find the first label, we compute s(t), the curvilinear abcissa of the speech signal y(t), where t is the sample index.

We introduce two functions :

• $S(n) = s(nL) - s((n-1)L)$

• $DS(n) = S(n) - S(n-1)$

where L a fixed number of samples .
S(n) represents an average value of the "length of the curve" per time and DS(n) its derivative.

Supposing that the surrounding noise is stationary in each segment, the S function must vary very slowly in noise zones, rapidly grow in the beginning of speech zones and decrease at the end of the speech zones.

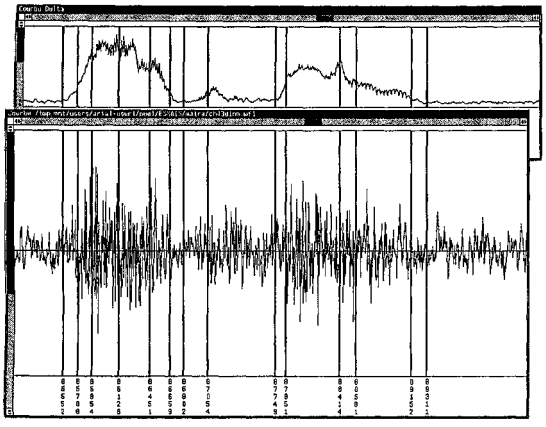

Figure 2 : Pronouciation of the name "Verlet" and S function

In other words, the derivative of this function, DS, may indicate appearance and disappearing of the speech in the signal. In order to detect those variations, we introduce two thresholds λ_1 and λ_2 and the following tests :

• if $DS(n) < \lambda_1$, the frame n corresponds to noise,
• if $DS(n) > \lambda_2$, the frame n corresponds to noisy speech.

The results of this detector are compared with the segmentation results
for every segment $[T_o, T_f]$:

• if the average value of DS(n) where $T_o < nL < T_f$ is less than λ_1, the segment is classified "noise".
• if this average value is greater than λ_2, the segment is classified "speech".
• between λ_1 and λ_2, the segment is classified according to the decision taken for the contiguous segments, and taking the duration of the segment into account.

The threshold λ_1 is automatically determined with the average value of DS computed on the first signal frames, supposed to be noise only.
The threshold λ_2 is proportional to λ_1.
This calculation is independent of the noise, but requires very noisy conditions.

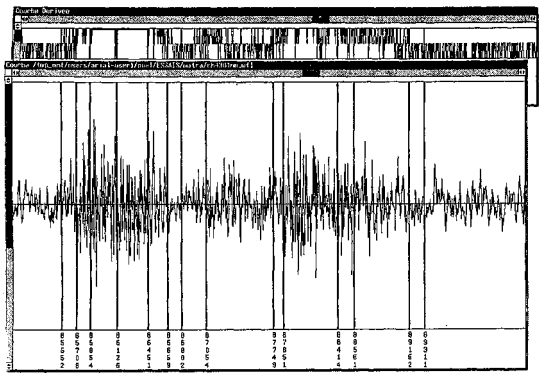

Figure 3 : Pronouciation of the name "Verlet" and derivative DS.

The dynamic coordination :

The maximal duration parameter, that determines if a doubtful segment is joined to the adjacent word is a strategic setting of the system : an important value will avoid truncatures, but may cause word omissions, when two words are separated by a short noisy segment.
A low value of this parameter will bring inverse advantage and disadvantage : few risks of concatenation, but risk of losing the first or last segment of the pronounced word.
This solution may also eventually cut a word pronounced slowly.
In conclusion, we give this parameter a rather big value (about 300ms), that offers a good robustness during experimentation, for our application (see section IV).

3.2 The frequential analysis :

When the noise spectrum can be considered as a monotonous curve, with a very low derivative, the noisy speech presents a high dynamic.
The idea of the frequential analysis using the spectral derivative is to compute, at each time, the sum of the energy variations between adjacent frequency bands.
The maxima of this curve fit with vocalic nucleus of the speech, and minima with noisy parts as previously.
The detector includes three principal parts :
- localisation of primary frontiers by detecting maxima
- segmentation of the speech signal
- decision module

Primary frontier localisation :

The signal is sampled at 8 kHz and analysed every 16ms on a 32ms window (256 samples).
After pre-accentuation and Hamming windowing, a Fourier transform is performed, and energy is calculated with 24 triangular filters distributed on the Mel frequential axis.
The spectral derivative is given for each signal frame by

$$Dspect_n = \sum_{i=1}^{23} \left(En_{i+1}^n - En_i^n\right)^2$$

Local maxima of this curve, called lobes, are searched by threshold crossing : the beginning (resp the end) of a lobe is detected when the spectral derivative crosses a threshold and stays monotonous during a minimum duration.
The primary speech endpoints lobes are gathered using criterions like sound duration, or instability of the spectral derivative : short noise zones can be detected inside a sentence, (for exemple plosive silence).
Those frontiers delimit the outlines of the speech part between two extreme vocalic nucleus.

The decision module :

We refine these primary frontiers because of possible non voiced sounds at the beginning or at the end of the speech.
We modify their position using the statistical segmentation results.
The automatic segmentation algorithm is used before and after the detected speech zone, in a 200ms range.
We keep as definitive frontiers of the speech the boundaries detected by the segmentation algorithm the most distant of the primary frontiers in a 200ms range.

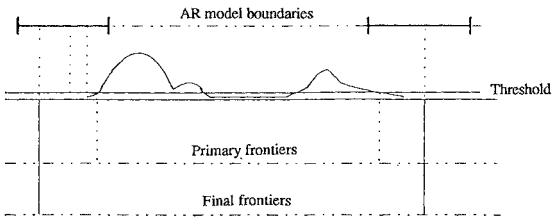

Figure 4 : Exemple of frontiers calculated by frequential analysis.

IV. THE SPEECH RECOGNITION SYSTEM

This system is dedicated to automatic speech recognition for an hand-free car telephone : isolated word recognition in an address book.
In this context, data are inevitably different between learning and recognition use, because of security reasons : learning is done engine stopped, but recognition is processed driving the car.
The corpus are provided by Matra Communication and recorded for the ESPRIT ARS project.
Four speakers have been recorded pronouncing four times 43 words (names of the address book and system commands).
Recording has been made with three level of noise :
- no additive noise, car stopped,
- car driven at 90 km/h (SNR about 0 dB),
- car driven at 130 km/h (SNR about 0 dB).

To realise the recognition system, we use an HMM compiler developped in the laboratory.
We modelized the application with a general network for the 43 words, each word represented by a subnetwork of pseudo-diphones.
The recognition system uses the results of the preprocessing module as illustrated in this figure.

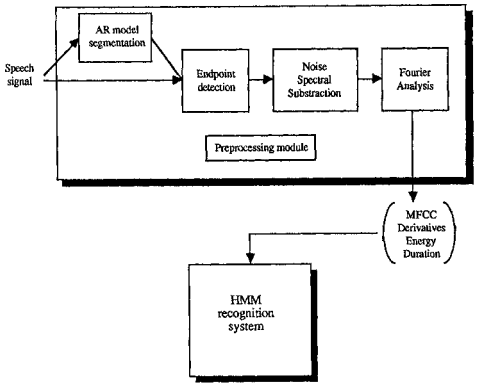

Figure 5 : Diagram of the preprocessing

The output data of the preprocessing phase are the results of Fourier analysis, and some information more :

- 8 MFCC
- energy
- 9 derivatives
- segment durations

These data are used as input for the HMM recognition system.

V. EXPERIMENTAL EVALUATION AND CONCLUSION

As we said when presenting the methods, the two preprocessing algorithms offers best results in specific noise conditions.
The temporal analysis performs better with very low SNR (0 dB).
The frequential analysis obtains its best precision with medium SNR (15 to 30 dB).
The computed endpoints, compared with a manual segmentation of the corpus presents an average difference of 100ms in the worse conditions.
The word beginnings are easier to detect as for an automatic system and a manual labelling : the average difference is about 60ms.
The endings of the words are always more difficult to found in noisy conditions, the automatic system presents an average difference of 140ms.

	130 km/h	90 km/h	Average
Speaker 1			
Beginning	64	60	62
End	156	110	133
Speaker 2			
Beginning	49	52	51
End	175	127	151
Speaker 3			
Beginning	63	56	59
End	188	103	145
Speaker 4			
Beginning	68	64	66
End	169	113	141

Figure 6 : Difference between manual and automatic segmentation (ms).

Experiments are currently performed to assess the entire system .
We compare the recognition of noisy speech after Noise Spectral Substraction and the recognition of the same words pronounced in silent conditions.
We compare the accuracy of the endpoints in all these conditions.

We try to evaluate the respective importance of each module of the preprocessing, and measure the performance improvment the system brings.

In the future, the preprocessing will allow us to separate the modification of speech due to additive noise and the modification due to sound production.
Our objective is to try to model Lombard effect and dedicate a module of the preprocessing to it.

BIBLIOGRAPHY

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Segmentation of Continuous Speech Signals", IEEE Trans. on ASSP, vol. 36 pp 26-40, January 1988.

[Junqua 92] B.Mak, J.C. Junqua, B. Reaves, "A robust speech/non-speech detection algorithm using time and frequency-based features", ICASSP 1992.

[Lamel 81] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon "An Improved Endpoint Detector for Isolated Word Recognition", IEEE Trans. on ASSP, vol. 29, pp 777-785, August 1981.

[Lockwood 92] P. Lockwood, J. Boudy, M. Blanchet, "Non-Linear Spectral Subtraction (NSS) and Hidden Markov Models for Robust Speech Recognition in Car Noise Environment", ICASSP 1992.

[Lombard 11] Lombard, "Le signe de l'élévation de la voix", ANN. Maladies oreille, larynx, nez, pharynx, 37, pp 101-119, 1911.