TABLE I
Results of the test on the method proposed in this paper

<table>
<thead>
<tr>
<th>Speakers</th>
<th>V→V</th>
<th>V→UV</th>
<th>% UV→V</th>
<th>UV→V</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>383</td>
<td>1</td>
<td>0.3</td>
<td>91</td>
</tr>
<tr>
<td>F2</td>
<td>363</td>
<td>4</td>
<td>1.1</td>
<td>106</td>
</tr>
<tr>
<td>M1</td>
<td>413</td>
<td>0</td>
<td>0</td>
<td>59</td>
</tr>
</tbody>
</table>

TABLE II
Results of the test on the method of Atal - Rabiner

<table>
<thead>
<tr>
<th>Speakers</th>
<th>V→V</th>
<th>V→UV</th>
<th>% UV→V</th>
<th>UV→V</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>372</td>
<td>12</td>
<td>3.1</td>
<td>94</td>
</tr>
<tr>
<td>F2</td>
<td>361</td>
<td>6</td>
<td>1.6</td>
<td>106</td>
</tr>
<tr>
<td>M1</td>
<td>410</td>
<td>3</td>
<td>0.7</td>
<td>62</td>
</tr>
</tbody>
</table>

V→V denotes the number of Voiced segments classified as Voiced;
V→UV denotes the number of Voiced segments classified as Unvoiced;
and so on. % denotes the misclassification rate.

On the whole the method here proposed appears offering attractive performance for further evaluation; computer performances are indeed comparable with those obtained by using the method of Atal-Rabiner. Then, in particular, the examination of the sequence of the classifier decisions has shown the uselessness of a smoothing algorithm after the decision, useful in other methods (e.g. [1]).

4. CONCLUSIONS

In this paper we have presented a method of Voiced/Unvoiced classification based on the maximum a posteriori probability criterion, which has given good experimental results. This method is a simplified version of more sophisticated Voiced/Unvoiced/Silence classification criterion, in which either the a priori probabilities of the classes either the p.d.f. of the measurements vector xk are updated. We are actually testing this more complex method of classification.

REFERENCES


1. INTRODUCTION

In the area of analysis and synthesis of the speech signal and, in particular, in that of continuous speech recognition, a relevant factor consists in discriminating, among the sounds produced by the vocal tract, those which are voiced from those which are unvoiced. For this purpose, the signal is subdivided into segments, typically of 10 to 20 msec duration. On each segment 4 types of measurements are made, on the basis of which the Voiced/Unvoiced classification is chosen. The methods proposed by previous authors are of two types: probabilistic (see e.g. [1]) and deterministic (see e.g. [2]).

The method here presented (sec.2) is probabilistic. In particular, for each segment, on the basis of the measurements, the a posteriori probabilities of the two classes are determined and the class for which the above probability is maximum is chosen. In addition, using the a posteriori probabilities that served for choosing the segment class and assuming that the sequence of segment classes constitutes a first order time-dependent Markov chain, the a priori probabilities of the classes, which are necessary for the classification of the successive segment, are updated. Finally, it is assumed that each vector composed by the 4 different measurements has a multidimensional gaussian p.d.f., with parameters which depend upon the class to which the segment is supposed to belong.

In this paper we give results obtained by preliminary experimentalization of the method proposed (sec.3), carried out by using 4 different measurements, appropriately chosen among those which are commonly employed [1], [2]. The training has been made by employing two female speakers, while the testing has been made on two female speakers, one of which was used also in the training, and one male speaker.

2. CLASSIFICATION METHOD

Let C1 and C2 indicate the classes of the voiced and unvoiced sounds respectively. Let S, S ∈ {C1, C2}, represent the unknown class to which
the kth segment to be classified belongs and $S_k$ the corresponding measurements vector. By denoting with $f(x/C_j)$ the p.d.f. of $x_k$ conditioned to $C_j$, it is assumed for sake of simplicity, but with no loss of generality, that:

$$f(x_k/C_j) = (2\pi)^{-1/2} |B_j|^{-1/2} \exp \left(-1/2 \left(x_k - x_k^j\right)^T B_j^{-1} \left(x_k - x_k^j\right) \right) = \frac{1}{B_j} f(x_k/C_j), \quad j = 1, 2, \forall k$$  

(1)

The classification of the kth segment and the procedure for the preliminary updating of the successive classification are made as follows:

a) On the basis of the a priori probabilities $P_k(C_j)$, $j=1,2$, and of the observation $S_k$, the a posteriori probabilities $P_k(C_j/S_k)$, $j=1,2$, are computed;

b) The decision is made for the class for which the a posteriori probability is maximum:

c) Using the a posteriori probabilities $P_k(C_j/S_k)$, $j=1,2$, and the transition probabilities $P_j(C_j) = P(C_j/C_k)$, $j=1, 2$, the a priori probabilities $P_k(C_j)$, $j=1,2$, which are necessary for the classification of the (k+1)th segment are computed. The a priori probabilities $P_j(C_j)$, $j=1, 2$, are estimated on the basis of the training data.

A schematic diagram of the algorithm is illustrated in fig.1.

\[ \begin{array}{c}
\begin{array}{c}
\text{P}_k(C_j) \quad \text{A posteriori probabilities} \quad \text{computation} \\
\text{P}_k(C_j/S_k) \quad \text{P}_k(C_j/S_k) \quad \text{Classifier} \\
\text{Unit delay} \quad P_{k+1}(C_j) \quad \text{A priori probabilities} \\
\text{updating} \quad P(C_j/C_k) \\
\end{array}
\end{array} \]

fig.1

Let us now consider the various steps of the algorithm:

a) Since $P_k(C_j) = P(S_k/C_j/S_k)$ and $P_k(C_j/S_k) = P(S_k/C_j/S_k/S_k)$, from Bayes formula we obtain:

$$P_k(C_j/S_k) = \frac{P_k(C_j) f(x_k/S_k, C_j, S_k)}{\sum_j P_k(C_j) f(x_k/S_k, C_j, S_k), \quad j = 1, 2, \forall k}$$  

(2)

In addition, supposing that, $\forall k$, it holds:

$$f(x_k/S_k, C_j, S_k) = f(x_k/S_k, C_j), \quad j = 1, 2$$  

(3)

from eqs (1), (2) and (3) we have:

$$P_k(C_j/S_k) = \frac{P(C_j) f(x_k/S_k, C_j, S_k)}{\sum_{j} P(C_j) f(x_k/S_k, C_j, S_k), \quad j = 1, 2, \forall k}$$  

(4)

b) The classification is made by putting $\hat{C}_k = C_1$ when $P_k(C_1/S_k) > 1/2$ , $\hat{C}_k = C_2$ when $P_k(C_1/S_k) < 1/2$ , and choosing at random if $P_k(C_1/S_k) = 1/2$.

c) For the purpose of classifying the (k+1)th segment we observe at first that: $P_k(C_j/S_k) = P(S_k/C_j/S_k, C_j)$ . Moreover, since it is reasonable to believe that the probabilistic structure of the language is independent of the speech production mechanism, we can assume that:

$$P(S_k/C_j/S_k, C_j) = P(S_k/C_j/S_k, C_j) \quad j=1,2, \forall k$$  

(5)

In addition, by introducing the hypothesis that the sequence of the classes constitutes a 1st order time homogeneous Markov chain, we have:

$$P(S_k/C_j/S_k, C_j) = P(S_k/C_j/S_k) \quad j=1,2, \forall k$$  

(6)

Then, from eqs (5) and (6) it follows:

$$P_k(C_j) = \frac{1}{\sum_j P(C_j) P_k(C_j/S_k)} \quad j=1,2, \forall k$$  

(7)

3. EXPERIMENTAL RESULTS

In this paragraph preliminary results concerning the application of the method previously illustrated are given. The method of classification has been experimented at the "Centre Scientifique de Paris, IBM France", in the framework of the "Deaf Children" project. The laboratory in which the test of the algorithm has been carried out disposed of an IBM Series/1 computer. A dynamic microphone has been used. The speech signal has been filtered at 4.8 kHz by means of a Butterworth low-pass and 8 pole filter, which permits the sampling of the signal at a frequency of 10 kHz. The signal so sampled has been stocked on a disk and then analyzed, by subdividing it into segments of 12.8 msec duration, corresponding to 128 samples of the signal. From each segment we have extracted, with standard procedures, the following features: 1) Energy of the signal, 2) Zero-crossings, 3) Autocorrelation coefficient at unit sample delay, 4) First predictor coefficient.

During the training, for each of the two classes, the expected value and the covariance matrix have been estimated. For this purpose, the segments of five sentences (for a total of 1946 segments), pronounced by two female speakers, have been analyzed. In addition, on the basis of the analysis of 4936 segments, the a priori probabilities of the classes for the first segment and the transition probabilities have been evaluated. The various initial guesses have been estimated by using classical methods. The segments have been classified manually, by using standard procedures (examination of the spectrogram and waveform).

The algorithm has been tested on three sentences: the first pronounced by one of the two female speakers of the training (F1), the second by another external female speaker (F2) and the third by a male speaker (M1). On the basis of the same training data the algorithm proposed by Atal-Rabiner I has been tested (without speaking) with the same sentences and speakers. The results obtained with the two methods are illustrated in table I and Table II. As regards the two female speakers, both methods have given satisfactory results. In the case of male speaker, for both methods the results can be considered still satisfactory, though male speakers were not used in the training. On 1446 segments the misclassification rate was 1.4% for the method here proposed and 2.1% for the Atal-Rabiner method.
TABLE I
Results of the test on the method proposed in this paper

<table>
<thead>
<tr>
<th>Speakers</th>
<th>V→V</th>
<th>V→UV</th>
<th>%</th>
<th>UV→UV</th>
<th>UV→V</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>383</td>
<td>1</td>
<td>0.3</td>
<td>91</td>
<td>4</td>
<td>4.2</td>
</tr>
<tr>
<td>F2</td>
<td>363</td>
<td>4</td>
<td>1.1</td>
<td>106</td>
<td>7</td>
<td>6.2</td>
</tr>
<tr>
<td>M1</td>
<td>413</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>5</td>
<td>7.8</td>
</tr>
</tbody>
</table>

TABLE II
Results of the test on the method of Atal - Rabiner

<table>
<thead>
<tr>
<th>Speakers</th>
<th>V→V</th>
<th>V→UV</th>
<th>%</th>
<th>UV→UV</th>
<th>UV→V</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>372</td>
<td>12</td>
<td>3.1</td>
<td>94</td>
<td>1</td>
<td>1.1</td>
</tr>
<tr>
<td>F2</td>
<td>361</td>
<td>6</td>
<td>1.6</td>
<td>106</td>
<td>7</td>
<td>6.2</td>
</tr>
<tr>
<td>M1</td>
<td>410</td>
<td>3</td>
<td>0.7</td>
<td>62</td>
<td>2</td>
<td>3.1</td>
</tr>
</tbody>
</table>

V→V denotes the number of Voiced segments classified as Voiced; V→UV denotes the number of Voiced segments classified as Unvoiced; and so on. % denotes the misclassification rate.

On the whole the method here proposed appears offering attractive performances for further evaluation; computer performances are indeed comparable with those obtained by using the method of Atal-Rabiner. Then, in particular, the examination of the sequence of the classifier decisions has shown the uselessness of a smoothing algorithm after the decision, useful in other methods (e.g. [1]).

4. CONCLUSIONS

In this paper we have presented a method of Voiced/Unvoiced classification based on the maximum a posteriori probability criterion, which has given good experimental results.

This method is a simplified version of more sophisticated Voiced/Unvoiced/Silence classification criterion, in which either the a priori probabilities of the classes either the p.d.f. of the measurements vector $\mathbf{x}_k$ are updated. We are actually testing this more complex method of classification.

REFERENCES
