A COMPARISON BETWEEN TWO APPROACHES TO AUTOMATIC SPEAKER RECOGNITION

L. Fasolo, G.A. Mian

Istituto di Elettrotecnica e di Elettronica
Università di Padova, 35100 Padova, Italy

ABSTRACT

Over the past years many different methods for automatic speaker recognition have been successfully proposed and tested. The aim of the paper is to present the results of a comparison carried out on two such approaches, viz. Atal /1/ and Sambur /2/, both using, but in a different manner, the parametric representation of speech derived from the linear prediction model. The experiment was carried out on 500 phrase length utterances of 10 speakers recorded over a three-month period. The results obtained in an identification experiment showed the superiority of the Atal approach, as for recognition rate and sensitivity to intra-speaker variability.

1. INTRODUCTION

The results obtained by many researchers over the past years suggest that the speech signal contains so much information about speaker identity that, despite the complex form in which it is encoded, speaker recognition can be successfully effected in a variety of ways using different feature extraction techniques and pattern recognition algorithms /3,4,5/.

Due to the lack of a general theory of speech-speaker interaction, the only way of comparing different methods is on the basis of their recognition performances, storage size, computation time, etc. For this purpose, an experiment was set up to assess the relative merits of two different approaches on the same data-base. Both methods are based on the parametric representation derived from the linear prediction model of speech production and use only the spectral features related to the smooth characteristics of the spectrum, as reflected by one of the parameters sets derived from LPC. The main difference is only in the manner in which they are used. Sambur /2/, on suggestion derived from synthesis experiments, found that very good speaker related features are contained in the averages and cross correlations of the time evolution of such parameters along a phrase length utterance. In Atal /1/, instead, the time information is preserved, even if the time alignement of the sentences is obtained in a rough way, and the decision relies on multiple observations carried out on single "similar" events.

We here describe how the data-base was set up and results obtained with both methods.

2. SPEECH DATA AND PROCESSING

Speech samples from an homogeneous population of 10 male speakers, whose ages ranged from 22 to 35, were recorded in a quiet room. All speakers had the same regional accent and presented the same speech particularities. They were told of the nature of the experiment and were asked to speak as normally as possible. The recordings were made on three different days with an interval of one month between each recording. Three and four recordings sessions, at distance of about an hour between each, were made on the first two days and on the third day, respectively. At each of the ten recording sessions each speaker read, in a different order, a set of fifteen sentences of which only five were used in the speaker identification experiments. The others were included to avoid laying any overdue stress on the selected sentences.

The sentences, which were devised to provide a variety of speech events, are as below:

1) /lwɔ̀mo é ún animále ratsjonàle/
   (Man is a rational animal)
2) /il názo dèlla nɔ́nna nón é normále/
   (Granny's nose is not normal)
3) /kwésto é un ezémpio délla mía vótʃe/
   (This is an example of my voice)
4) /kwésta é una prɔ́va di riconoʃiménto del parlatóre/ (This a speaker recognition experiment)
5) /káza/,/pétto/,/kúpo/,/páste/,/fílo/, /kúttʃa/,/píno/,/méla/,/ferɔ́tʃe/
   (House,chest,glomy,cakes,wire,dog's basket,pine,apple,fierce).

The first four are both complete sentences and utterances, the fifth consist of nine words spoken separately. The average duration of the sentences was 1.8, 2.,1.7, 2.2, 3.8 s respectively and the overall duration of the data-base (without pauses and silent portions) about 20 min.

Following the recording, the speech sam

CH1285-6/78/0000-0275$00.75@1978IEEE

ples were quantized uniformly to 12 bits at a 10 kHz sampling rate, after sharp low-pass filtering at 4.3 kHz (-50 dB at 4.6 kHz). Each utterance was then subjected to a linear prediction analysis (autocorrelation method): a Hamming window of 200 samples was applied to the preemphasized signal and the window advanced in steps of 200 samples. For each non-silent frame the pitch period, the gain and the first 12 reflection coefficients were computed and stored. The first two parameters were not considered in the following experiments. They were icluded in order to allow for the possibility of testing performances of the analysis process by synthesis and to enable us to carry out future experiments, which will also take into account such parameters. To save memory, the other linear prediction parameters used in the experiments have been obtained from the k-coefficients when necessary. As a result of the analysis, a vector $\underline{u}(i)$, having as its components the parameters associated with the frame, was assigned to each frame i and the utterance was represented by the time evolution of $\underline{u}(i)$, i=1,...,I, where I is the number of frames in the utterance.

3. EXPERIMENTS

Two models were used to represent the speech-speaker interaction at utterance level, a "global" and a "local" one, and experiments were performed on speaker identification. Here, by speaker identification we mean the task of identifying to which of S registered speakers a speech sample belongs, without rejection.

3.1 Speaker Identification Based on Time Averages of an Entire Utterance

In the first experiment it was supposed that the relationship of a sentence represented by the time-series $\underline{u}_j(i)$ to a speaker j could be summarized by the average $\underline{U}_j = E[\underline{u}_j(i)]$ and by the covariance matrix $\Sigma_j = E[(\underline{u}_j(i)-\underline{U}_j)(\underline{u}_j(i)-\underline{U}_j)']$, which characterize both the speaker and the sentence. Hence, given the realization $\underline{x}(i)$ of a sentence uttered by one of the speakers of the sample population, the sentence is associated with the speaker from which its distance

$$d_j(\underline{X}) = (\underline{X}-\underline{U}_j)'\Sigma_j^{-1}(\underline{X}-\underline{U}_j) \qquad (1)$$

where $\underline{X} = E[x(i)]$, is minimal. Distance (1) corresponds to Bayes test for minimal risk with symmetrical cost functions, assuming multivariate normal distribution and equal scatter coefficients /6,7/. Alongside (1) we have also used the distance

$$D_j(\underline{X}) = (\underline{X}-\underline{U}_j)'\Sigma^{-1}(\underline{X}-\underline{U}_j) \qquad (2)$$

where $\Sigma = E[\Sigma_j]$. This allows storage economies and a more stable estimate of the covariance matrix from a finite size sample, though obviously at the cost of an information loss with regard the speakers.

As is well known, (1) and (2) are invariant under any non singular linear transformation of the measurement space, in particular under the one that leads to the principal components /6/, i.e. to the orthogonal linear prediction parameters used in /2/. Keeping this in mind, it was decided to carry out the identification experiment initially within the original space, leaving the principal components analysis to a follow-up stage. Such a method would allow us to select features that parsimoniously characterize the speakers, i.e. to go from a 12 dimensions to a circa 7 dimensions space.

In order to obtain an estimate of the performance of the classifiers, for each speaker and each of the 5 utterances in the data-base, the samples were divided into a test set and a design set using the "leave-one-out" method /6/.

The average error-rate obtained using the distance measure (1) is given in Tab.I for the five utterances and the reflection (k), filter (a), cepstral (c) and log-area (g) coefficients.

UTTERANCE	1	2	3	4	5	Av.
PARAMETERS						
k	14.	15.	16.	12.	13.	14.
a	14.	18.	18.	9.	21.	16.
c	18.	17.	24.	18.	30.	21.4
g	20.	19.	14.	12.	15.	16.

Tab.I: Error-rate (%) for the distance (1)

Similar results were obtained with (2) and also applying the multidimensional U-test /7/.

The values are significantly greater (about 10-12%) than those obtained from /2/ using an equivalent metric. A possible explanation of such deviance might be proposed in the following terms. Results obtained using (1) and (2) were not significantly different for the majority of the sample speakers (95% significance). This might entail that, on average, intra-speaker variability was of the same order as infra-speaker variability. Hence, the results were grouped according to the recording session of the test sentence. In this way we could observe, for all the sentences and speakers, a drastic reduction in error-rate, in order from the first to the last recording (with exception of the cepstral coefficients). Tab.II gives error-rate values for distance (1) averaged out over all the sentences and grouped according to the day the recording was made. Results significantly different were obtained for the third group, using (2).

DAY	I	II	III
PARAMETERS			
k	23.3	13.3	6.
a	26.	16.6	6.5
c	34.6	16.6	15.
g	30.	14.5	6.

Tab.II: Error-rate (%) deviance in terms of days on which the test-sentences were recorded.

Note how results in the third column are comparable whith those in /2/. Considerably better results were obtained when the design-set was based on the last recordings.

The more obvious interpretation of such a result is that the method can work well with trained speakers, but is potentially sensitive to intra-speaker variability. Such a conclusion has been confirmed by carefully listening to the original tapes: this showed that the speakers with higher error-rates were those who had varied their realizations more during the course of the recordings

3.2 Speaker Identification Based on Single Segments and on Multiple Observations.

In the second experiment the phoneme sequence constituting an utterance is represented by a time series described by an average and a covariance matrix, time-varying in strict dependence on both the particular phoneme and speaker. In such a case, given a sentence uttered by one of the sample speakers, greater reliability can be had in identification, by using the cumulative evidence collected from the observations carried out on the single components of the sentence.

To be fully exploited, such a procedure has to carry out a time registration of the utterances. In the present case, to obtain an approximate temporal alignement we used a simple linear stretching. For this purpose the time series $\underline{u}_j(i)$ of each utterance, was divided into segments with an average duration of about 80 ms. (i.e. four frames) and each segment characterized by the average value of the reflection coefficients in that segment.

The recognition experiment was carried out on single segments, using for decision purposes the distance (2), and on the whole utterance, using in this case $\overline{D}_j(\underline{x}) = \sum_1^i D_j[\underline{x}(k)]/i$ as distance.

The results obtained for the 5 sentences based on a single segment show a reduction of identification accuracy with respect to the results given in Tab. I, based on a whole sentence, but not so drastic as expected. As an example, for sentence n. 4 the error rate was 36.4, 35.2, 27.6, 36.6 % for the k,a,c,g coefficients, respectively.

Moreover, as in [1], the cepstral coeffi

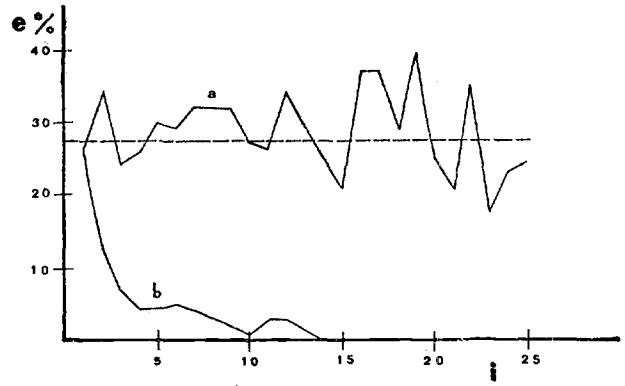
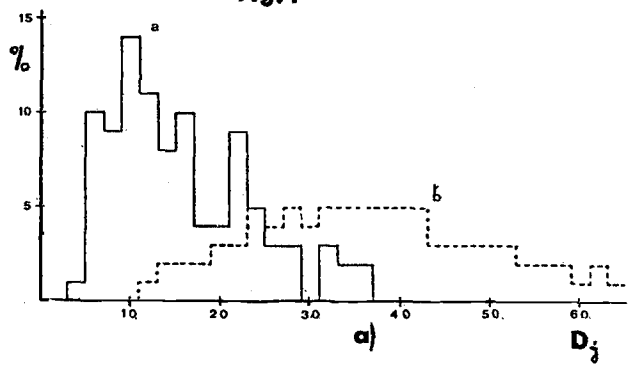


Fig.1



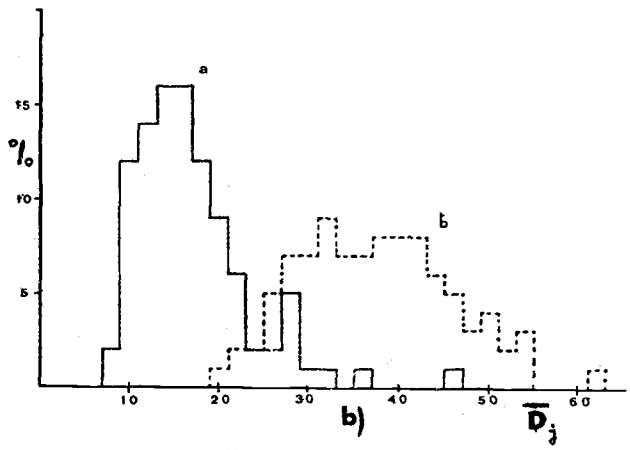


Fig.2

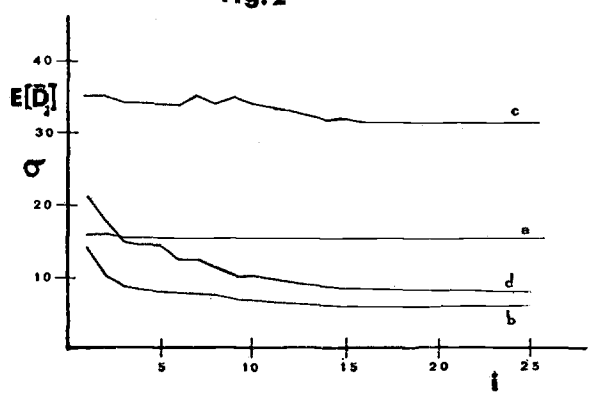


Fig. 3

cients were found significantly better than the other parameters considered. However, when more segments are used for decision purposes, all parameter sets hold good, the difference being mainly in the speed with which the error-rate is reduced.

Fig. 1 shows, for the sentence n. 4 and cepstral coefficients, typical dependency of the error-rate on the particular segment (curve a) and the improvement (curve b) in the identification accuracy, when the number of segments is increased starting from the beginning of the utterance (similar behaviour is also exhibited when the segments are connected randomly). The error - rate reduction is obviously a consequence of reducing the overlap between the speakers distributions. This is reflected in fig. 2 where istograms of $D_j(\underline{x})$ for the first segment (fig. 2a) and of $\overline{D_j(x)}$ for the whole sentence (fig. 2b) are given for $\underline{x} \in j$ (curves a) and $\underline{x} \notin j$ (curves b).

It is interesting to note that, due to intrinsic correlation of the distances computed in the different segments of test-utterances, the variances of the distributions of both the "intra" and "infra-speaker" distances (curves b and d of fig. 3, respectively) attain a non-zero limiting value (quite quickly with the number i of segments used) not very dependent from the parameter set and from the sentence. The same holds for the average values of "intra" and "infra-speakers" distances, given by curves a and c of fig. 3, respectively.

4. CONCLUSIONS

A comparison of two methods for speaker identification has been carried out on the same data-base.

From the results obtained it seems possible to conclude that the method based on multiple observations of single segments of an utterance /1/, despite of the crude time alignement technique used, is quite superior the one suggested in /2/. Averages on a phrase length utterance lead to better accuracy than averages on a short segment, but are very less accurate with respect to the decision made on the basis on more segments. This seems confirm, also if indirectly, the results given in /8/, in which it was shown that quite long time-averages are necessary to attain a good (text-independent) characterization of a speaker.

REFERENCES

/1/ B. Atal: "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J.A.S.A., Vol. 55, N. 6, pp. 1304 ÷ 1312, June 1974.

/2/ M. Sambur: "Speaker recognition using orthogonal linear prediction", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, pp. 1632 ÷ 1645, Aug. 1976.

/3/ V. Sarma, B. Yegnanarayana: "A critical survey of automatic recognition systems", Journ. Comp. Soc. of India, Vol. 6, N. 1, pp. 9 ÷ 19, Dec. 1975.

/4/ B. Atal: "Automatic recognition of speakers from their voices", Proc. IEEE, Vol. 64, N. 4, pp. 460 ÷ 475, April 1976.

/5/ A. Rosenberg: "Automatic Speaker Verification: a review", Proc. IEEE, Vol. 64, N. 4, pp. 475 ÷ 487, April 1976.

/6/ K. Fukunaga: "Introduction to statistical pattern recognition", Academic Press, 1972.

/7/ H. Cramér: "Mathematical methods of statistics", Priceton Univ. Press., 1963.

/8/ J. Markel, B. Oshika, A. Gray: "Long term feature averaging for speaker recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, pp. 330 ÷ 337, Aug. 1977.