SOME FACTORS INFLUENCING THE PERFORMANCES OF A SPEAKER RECOGNITION SYSTEM BASED ON LPC

G. A. MIAN

Istituto di Elettrotecnica e di Elettronica - Università di Padova
35100 Padova (Italy)

ABSTRACT

Linear prediction parameters are critically dependent upon the short-term spectrum of speech and therefore to noises and distorsions introduced by transmission and recording systems. The aim of the present work was to evaluate the performances of an automatic speaker recognition system, based on LPC, on the same speech material recorded in three different conditions: on a quiet room, from dialled up telephones lines via direct hookup and via a suction cup tap. Each of ten speakers spoke an 8s long sentence four times over a two-months period. Sentences were manually segmented and performance evaluation was conducted on phonemes, on breath groups and on the whole sentence using a minimum weighted distance classifier.

INTRODUCTION

The methods used for speaker-recognition follow, broadly, two main approaches (1,2). In the first, statistical parameters independent from the phonetic sequence are extracted by long-time averaging time dependent parameters as spectrum, cepstrum, outputs of bandpass filters or parameters derived from linear prediction of speech waveform. If a stable extraction of the statistical parameters is obtained, this approach leads to text-independent speaker-recognition (3). In the second approach a preselected utterance is used and the temporal information preserved. However, as any parameter extracted from the speech signal is indicative of the speaker identity and of what is said, to exploit the speaker dependent information it is necessary to compare the same linguistic events. To this purpose, the test utterance has to be aligned to the reference, using a nontrivial and nonerror free time registration procedure. As for time dependent parameters extraction, most of the work has been done with parameters extracted at regularly spaced intervals, resulting in a large set of data with a high redundancy, but work has also been directed to selective measurements of parameters at appropriately chosen locations in selected utterances (4-7).

The present work uses the second approach and presents results obtained on an 8s long sentence, spoken by ten speakers. The features used were the smooth characteristics of the spectrum, as reflected by one of the parameters sets derived from the parametric representation of the speech process given by linear prediction. To avoid the possible errors of any time registration scheme, the sentence has been manually segmented in phonemes: this allows to compare the phonemes used in terms of information carried for speaker recognition and to estimate variations in performance due to the context. Moreover, as linear prediction parameters are critically dependent upon the short-term spectrum of speech and therefore to noises and distortions introduced by transducers and transmission systems, it seemed interesting to test the performance reduction when the utterances were recorded off a standard telephone line and to test the deconditioning effect of subtracting from the cepstral coeffi cients their time averages (1). In the following we describe how the data-base was set up and the results obtained.

DATA BASE AND ANALYSIS

Ten male speakers in age from 25 to 40 years, with the same regional accent and without noticeable speech defects were chosen for the recordings. They were aware of the nature of the experiment. In order to take into account the important effects of changes in speaker's voice over time (9), the recordings were made on two different days with an interval of about two months. On each day two separate recording sessions were made and on each the speakers read a set of five sentences, only one of which was used in the present experiment. This was the sentence: /'a li'ʃano in 'luʎo 'dʒa alle 'undit∫i la 'sabbia 'skɔtta per il 'sole e ba'ʃarsi nel 'mare di'venta una net∫essi'ta/ composed with 78 phonemes (without pauses) most of them voiced and with an average duration of about 7.8s (~6.s without pauses). The sentences were spoken in a quiet environment into

0-1379-7779/000-0781$00.75 1979 IEEE

a normal telephone handset and the thelephone calls were carefully recorded at the receiver (at the Centro di Studio per le Ricerche di Fonetica del C.N.R. in Padova) simultaneously via direct hookup (to which we will refer as type 2 recordings) and via a suction cup tap (type 3 recordings). The transmission system was through public switched lines. In order to estimate the effects of transducers and transmission conditions the sentences were also recorded through a good quality microphone placed near the telephone set (type 1 recordings).

The speech samples so obtained were passed through a 3.3 kHz low-pass filter with 60 dB/octave roll-off, digitized to 12 bits at a sampling frequency of 8 kHz and stored in the computer disk. Using an interactive system the speech files were scanned and, through a combination of careful listening and visual examination of the waveforms, divided manually into nonoverlapping segments, neighboring segments corresponding to distinct phonemes. The lengths varied from about 15. to 130. ms. The segmentation has been carried out only on type 1 rec.and the data obtained used for type 1 and 2 after manually locating the beginning of the corresponding sentences. When the above procedures failed to isolate a segment having a quality within the range "usually" associated to his phonetic label or when the phoneme was "absent" (because of elision or merging phenomena) no forced decision was made; the phoneme was then associated with an empty slot. From each segment the middle part was singled out (duration 0.8 of the nominal duration) and Hamming-windowed, with the effect of further tapering off the transitions to/from other segments. The linear prediction parameters associated with the segment were obtained by a tenth-order linear predictive analysis using the autocorrelation method. To save memory only the reflection coefficients were stored (on 16 bits). The other parameters (prediction and cepstral coefficients) used in the experiments have been obtained from them, when necessary.

As a result of the analysis the sequence of phonemes constituting the utterance is represented by the time evolution of a vector $\underline{x}(n)$ $(n=1,78)$ and to each speaker four such time-series are associated with each of the 3 recording conditions.

EXPERIMENTS

Speaker identification (without rejection) was based on a minimum weighted Euclidean distance. With this distance the reference data for speaker j and phoneme n is represented by a mean vector $U_j(n)$ and a covariance matrix $\Sigma_j(n)$. In effect, to avoid that, due to the limited sample size, the difference of covariance matrices for different speakers may be less than their estimation error, the pooled covariance matrix $\Sigma(n) = E[\Sigma_j(n)]$ was used. In order to make full use of the available data, the "leave one out" method was utilized to estimate the performances of the classifier. Each of the 4 repetitions was used in turn as the test set, while the remaining three formed the design set. So for each phoneme 40 tests were made.

The error-rate averaged over all the phonemes of the utterance and the corresponding sample standard deviation are given in tab. I for the three recording conditions and for the prediction (a), reflection (k) and cepstral (c) coefficients. From tab. I one can observe that:

I) for type 1 rec. the quoted values are comparable with the ones obtained in speaker-recognition experiments which used linear time registration of the utterances (1,11). Actually, they are higher, but the effect of unvoiced phonemes has to be taken into account (see tab. II);

II) in all conditions c-coefficients are significantly the most effective LP parameters. It is worthnoting that there is some experimental evidence that this property is lost (10,11), if the cepstral parameters are time averaged;

III) the combined effect of the carbon microphone and of transmission conditions over dialled-up lines is quite drastic (at least in this experiment) and it increases the error-rate by about 12% for all parameteres;

IV) the additional distortions and noise introduced by the suction cup tap increase the error-rate by about another 2%, which is perhaps a lower bound and can easily increase, if less care is exercized in positioning the transducer.

In tab. II the average error-rate for c-parameters is referred to unvoiced (UV), voiced (V) phonemes and to stressed vowels (SV) (the results are similar for the a- and k- parameters). As expected, unvoiced phonemes seem to convey significantly less spectral information about the speaker's identity than voiced ones. Moreover, among voiced phonemes most of the speaker related features are obtained from stressed vowels, generally longer and better articulated.

In fig. 1a) the solid curve gives the time evolution of the error along the utterance for type 1 rec. and in fig. 1b) for type 2 rec. (the curves for type 3 rec. are similar). Both figures refer to c-coefficients. The most favoured vowels for speaker-identification seem to be the ones surrounded by a sonorant environment, other things being equal (i.e. vowel involved, degree of stress within the sentence etc.). Particularly intersting appear the minima in correspondence of the clusters /pan/, /par/, /mar/, which resulted also robust to system degradation. This fact combines the remarks put forward by (6,12) about N- and -r environment. As for the voiced consonants it results from the data obtained that the voiced stops /b/, /d/ and the sonorant /l/ are the worse and the nasals the better. However the scores here obtained are very different

from the ones obtained in (7) (with sounds spoken in isolation) and are strongly affected by the context.

Fig. 1 shows also some trend in the error along the utterance, with the error increasing from the beginnings /a.../ and /e baʃarsi.../ to the corresponding ends /..'sole/ and /..netʃessi'ta/. This can be due to coarticulation or word boundary effects, which would be better taken into account considering also features arising from the dynamic properties of the speech production.

The data considered until now refer to experiments of speaker-identification on the base of one phoneme. However, as known, if the classification decision can be based on the cumulative evidence collected from many observations, the probability of correct recognition can be usually increased over that obtained with decisions based on a single observation. This is important in automatic speaker-recognition and seems to parallel the human speaker recognition process. In fact the cumulative confusion matrices for type 1, 2 and 3 rec. revealed that the mode of the classification count for each speaker was quite peacked on the principal diagonal. So a second experiment has been carried out. After the distances of each test-phoneme are obtained, the distances are accumulated and used to make the decision. The utterance was divided into breath groups (as they appear from the measured lengths of pauses) and the accumulation procedure was started at the beginning of each breath group. Unvoiced phonemes and phonemes for which not all four realizations were found (as the /i/ in the last word), were discarded. To reduce computation time, with no practical effect on the recognition rate, after calculating the distances for five consecutive phonemes the half of the references giving the higher distances were rejected. The results are given by the dashed curves of fig. 1 a) and b) for type 1 and 2 rec., respectively. The curves evidence the error-rate reduction as the number i of phonemes entered into the distances computation is increased. Furthermore the starting point affects the error-rate reduction, so that it may be supposed that better results can be obtained on a sequence of words separately spoken. Pooling together the results obtained for each starting point, it resulted that a relationship $E(i)\cdot i^{\alpha}=k$, with $\alpha \sim 1.4, 1.2, 1.$ respectively for type 1,2 and 3 rec., gives a first approximation to the data, at least for $E(i) \geq 5\%$.

In addition, an examination of istograms of the intra- and infra-speaker distances accumulated on the whole sentence showed an overlap strongly increasing from type 1 to type 3 rec. This suggests that the error rate reduction as shown in fig. 1 gives perhaps an optimistic picture.

As for type 1 it is interesting to note that the error reduction is faster than in (1,11) as a function of i (probably for the greater independence of distances), but is on the average the same if expressed in terms of the duration of speech observed.

Finally, in (1) it was suggested the removal of the time averages of cepstral coefficients as a mean to eliminate time-invariant frequency distortions (at the risk of removing also long-term speaker-related information). To test the practical effectiveness of such a procedure, the time-averages were computed for each repetition of the utterance (using only voiced phonemes and weighting with their duration) and subtracted from the original coefficients.

The result was an average error-rate (referred to voiced phonemes) of 39.7,48.9,52.5% for type 1,2,3 rec., respectively, i.e. an increase of $\sim 7\%$ with respect to tab.II, practically independent of the type. In fact, the determinant of $\Sigma(n)$ decreased of about one order of magnitude and the average intra- and infra- speaker distances resulted nearer, so that the overall effect of the operation was a contraction of the points of each speaker around the mean, more than compensated by the contraction of the speakers means.

Instead, when the new distances were summed, a faster error reduction was observed so that about the same final results given in fig. 1 a) and b) were obtained. However, the overlap between istograms of intra- and infra- speaker distances was consistently greater than the one previously obtained.

In conclusion, it seems that subtracting the time-averages from cepstral coefficients results in more speaker-information loss than what is gained from eliminating the transmission apparatus influence.

ACKNOWLEDGMENT

The author wishes to acknowledge dott. F.E. Ferrero for supplying the speech material and prof. A.M. Mioni for useful discussions and suggestions.

REFERENCES

(1) B.S.Atal, "Automatic Rec. of Speakers from their voices",Proc. IEEE, 64, 460-475, 1976.

(2) P.Jesorsky, "Principles of automatic speaker-recognition", in Speech Communication with Computers, Carl Hanser Verlag, 1978.

(3) J.D.Markel et al.,"Long-term feature averaging for speaker recognition", IEEE Trans.ASSP,25, 330-337, 1977.

(4) J.J. Wolf, "Efficient Acoustic Param.for Speaker Recognition", J.A.S.A., 51, 2044-2056, 1972.

(5) M.R. Sambur, "Selection of acoustic features for Speaker ident.",IEEE Trans.ASSP,23,176-182,1975.

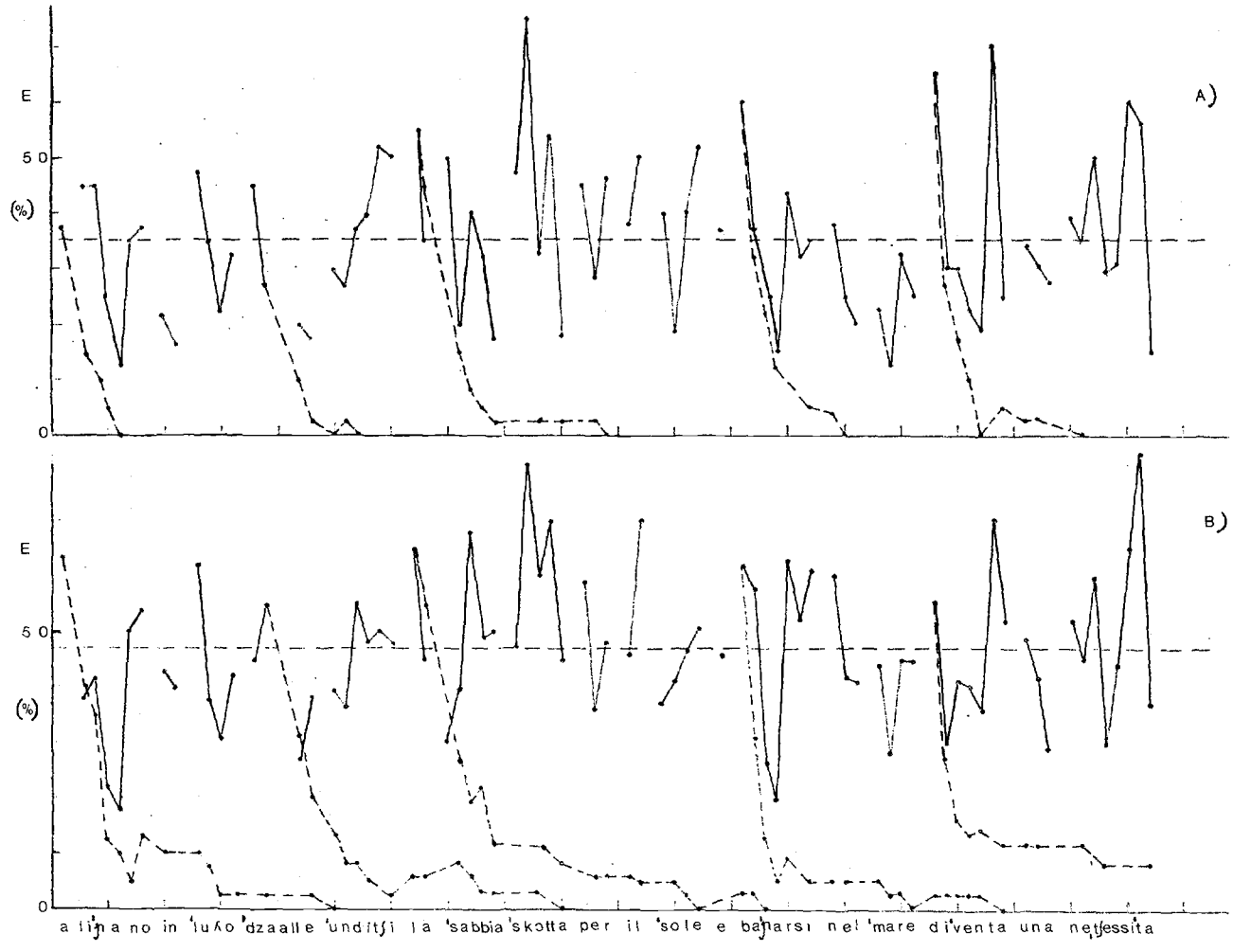

FIG. 1

(6) U. Goldstein, "Speaking-identif.feat. based on formant tracks", J.A.S.A., 59,176-182, 1976.

(7) U.Hoefker, "Phoneme-Ordering for Speaker recognition", Proc.9 I.C.A., Madrid, July 1977.

(8) R.L. Kashyap, "Speaker rec. from an unknown utterance ..",IEEE Trans ASSP, 24,481-488, 1976.

(9) A.Rosenberg, "Evaluation of an Automatic Speak. Verif..", B.S.T.J., 55, 723-744, 1976.

(10) R.S. Cheung, B.A.Eisenstein, "Feature selection via ...", IEEE Trans ASSP, 26, 397-403, 1978.

(11) L.Fasolo, G.A.Mian, "A comparison between two approaches...",Proc. 1978 IEEE ICASSP,273-276.

(12) K.Lo-Soun, P.Li, K.Fu,"Identification of speakers by use of nasal coarticulation", J.A.S.A., 56, 1876-1882, 1974.

Param. / Type	k	a	c
1	45.1 (1.5)	40.6 (1.5)	35.4 (1.6)
2	55.8 (1.4)	53.1 (1.4)	47. (1.7)
3	58.2 (1.6)	54.6 (1.5)	49.7 (1.4)

Tab.I Average error-rate (%) and sample stand.dev. for single-phoneme speaker identification.

Type	UV	V	SV
1	52.8	30.2	21.
2	60.2	42.5	34.7
3	65.	44.9	36.8

Tab.II Average error-rate for c-coefficients referred to unvoiced(UV),voiced(V) phonemes and to stressed vowels (SV).