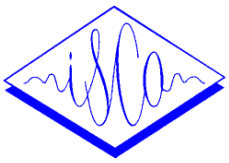

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

PARAMETRIC SPEAKER RECOGNITION OVER LARGE POPULATION OF TELEPHONIC VOICES

A. Federico , ENEA
Via Anguillarese, 301 - 00060 S. Maria di Galeria, Rome
A. Paoloni, Fondazione Ugo Bordoni
Via Baldassarre Castiglione, 59 - 00142 Rome
e-mail: pao@fub.it

ABSTRACT

The voice parametric features extraction followed by some statistical identification test is the widespread approach to the speaker recognition. The problem reaches its maximal complexity when a voice sample must be attributed to a set of speakers being non-zero the a-priori probability that the sample does not belong to the given set. Usually these open set tests are related to the highest level of responsibility like in the forensic applications.
This paper is addressed to present a balanced solution to the speaker recognition problem and to give the right statistical foundation to the decision task. All the related issues are restated, the modelization method is reconsidered for sparse experimental matrices and the algorithms for a suitable bayesan approach to the decision are derived following a more consistent theory.

1. INTRODUCTION

One important area of speaker recognition concerns forensic applications. Studies have been carried out to investigate the performance that under different degraded condition, can be reached by the recognition techniques. It is clear that if for other application a wrong decision has only material consequences, in forensic applications much more serious aspects are involved that requires highly reliable decisions to be taken. Various methodologies for approaching this problem have been proposed; they may substantially be classified into the following categories, according to the criteria adopted for the analysis of the speech signal: the auditory method, the spectrographic method and the automatic method [1], [2], [3].
The need of an objective and reliable decision suggests the adoption of a full automated method; unfortunately, due to the adverse recording conditions, to non-cooperative speakers, to very similar voices, etc., is proved that a large amount of operator's interventions should be performed to achieve adequate reliability. It seems very useful to have an user-friendly system dedicated to this scope, i.e. an environment able to manage speech acquisition and restitution, time and frequency signal representations, extraction of speech features, and finally the execution of the statistical tests.
The IDEM system, that is explained in the section 2, is a tentative to fulfil these needs, and it is (as far as we know) the only system designed ad hoc that runs on a small computer [4].
To evaluate the IDEM system the right way is to build a database to test the system. Such database would be extremely precious for several purposes. One would be to evaluate forensic speech experts and clarify if they can really do much better than an naive listener. It would also clarify to the forensic professionals what are the limitations of human and automated techniques of speaker recognition for these applications. This corpus is used, in our speaker identification objective method, to forecast the false identification errors as with reference to each particular identification case.
This paper restates some foundations of our automated method. In particular in Section 2 we describe the speaker recognition system named IDEM. Sec. 3 describes the population model and statistical model evaluation. Sec. 4 contains some guideline to the research activity to validate and complete our method.

2. THE IDEM SYSTEM

The IDEM system is the result of many years of experience in the field of speech analysis and speaker identification for forensic purposes [5]. In 1990 we start an internal project which main goal was to implement either the already existing software packages, or some new tools in a "low-cost" and "user-friendly" hardware/software environment. We decide to use a PC under the Microsoft WINDOWS operating system [6].
The IDEM method is based on the comparison of a set of parameters (in our case is the F0 and the first three formants of the Italian vowels /a/, /e/, /i/, /o/) estimated in *well stable* portion of speech. Given a set of these parameters for each speaker a reference matrix is computed. The comparison of the resulting data will give the response to the identification test. The IDEM system is based, as shown in fig. 1, on five main module.

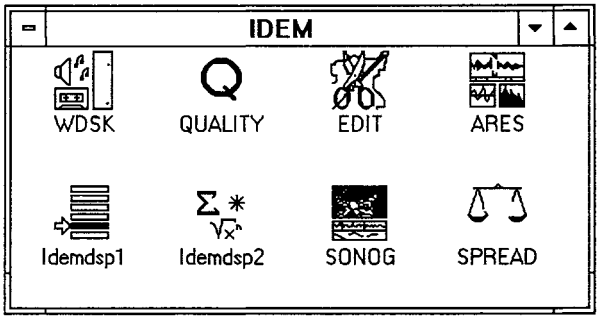

Fig. 1. Modules of the IDEM package.

2.1 Acquisition module

Because usually speech materials are not given in digital format, the first step is speech acquisition. In our very specific case the speech material is always originally recorded on tape from telephonic or ambient interceptions. The acquisition module has three main function: to play an

audio (speech) file, or part of it; to record a new file; to calibrate the acquisition procedure.

2.2 Quality module

After you have the speech files available in a digital format on your PC, it may be necessary to evaluate in a objective way the quality of the speech material. In fact it is a trivial fact that worster is the speech quality, harder is the task of the speaker identification decision. An estimation of the signal to noise ratio, as well as the long term spectrum of both "signal" and "noise", may give to the operator an immediate sight of the speech quality, so that he can decide if the speech signal is good enough to be processed and analysed, or it should be discarded.

2.3 Editing module

The editing module has been designed to control and extract from long signal file the selected speech material needed to estimate, in an accurate way, a given speaker (see fig. 2). This module has two important requirements: it must be simple and fast to use. To satisfy these need in designing this module we considered that in telephonic conversations (that is the most common case) we only have two speakers. The main function that an operator can do with this module is to assign selected speech zones to one speaker (say speaker A), or to another speaker (say speaker B).

2.4 Preprocessing module

Once you have selected a signal file, from a main menu where the possible parameters are listed you have to select the parameters to be computed. The main features in the menu are: energy, pitch, formant tracking, sonogram, vowel localisation and formant estimation. Each parameter is computed by a separate program, so that it is easy to add or change any of them if the user needs a different speech parametrisation.

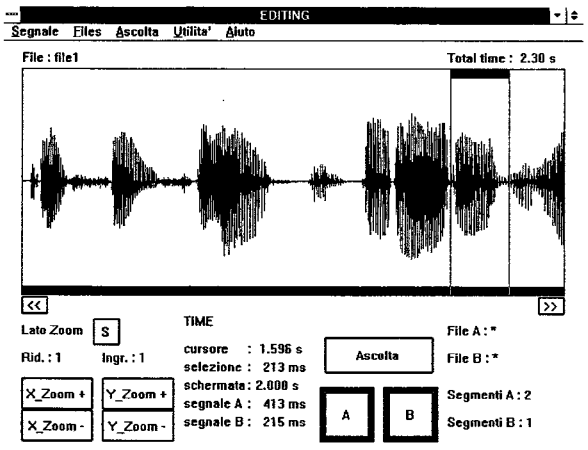

Fig. 2. Main window of the EDIT module.

2.5 Formant estimation module

This module is dedicated to the spectral analysis of a fixed length signal frame. As shown in fig. 3 on the top of the main window a 2.5 second waveform of the audio signal is represented. The cursor is a thick line, wide as the selected zone (you may select any power of two, from 128 to 4096 points, default is 512). In the bottom left you have the zoom of the selected frame. In the bottom right the power spectrum of the selected frame, optionally in this window you can plot the LPC and the CEPSTRUM smoothed power spectrum.

According to the defined number of formants that you want to estimate (from one to four, default is three), in the power spectrum window you have some vertical bars, that you may move using the mouse. Once you find the signal portion from which you want to estimate the formant value, you had to move the vertical lines on the supposed formant frequencies. Now you may fix the information that include: the pitch value, the formants values, the vowel (or phoneme), the context of the word. The symbol of the labelled vowel will now appear aligned to the audio wave on the screen.

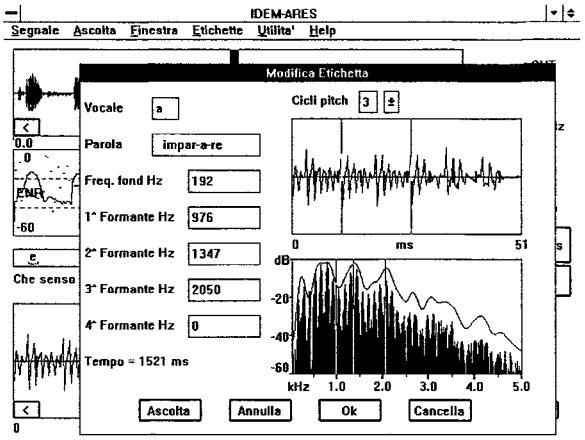

Fig. 3. "Modify parameters" window of the ARES module.

2.6 Decision module

The last operation will be the decision problem, that can be taken by statistical tests on these parameters. Because of the very specific target of this application (forensic), particular attention has been devoted to the development of the decision module, to the characterisation of the speaker, and the estimation of false rejection, and false identification probabilities. The theoretical background the identification test will be discussed in the Section 3.

The SPREAD (SPeaker REcognition by Automatic Decision) is an interactive system that allow you to load in a working area as many "data-file" (in our case data are pitch and formants) as you like. To each data file a speaker name is associated, so that you may select from the speaker list a set of speakers to compare. Before running the test you may remove the outliers data using a lot of filtering utilities: based on predefined intervals, or on the standard deviation of model data, or on the standard deviation of the speaker data (this is a recursive filter). After the test execution a YES/NO matrix is displayed, and you may read the detailed result clicking on the specific box.

The decision method is based on a bayesian approach that requires the operator (subjective) definitions of what the cost functions are. The SPREAD module then compares the risk values selecting the lower risk decision.

3. POPULATION MODEL BUILDING

The decision about the identification of two parametric voice samples is taken, following the bayesian approach,

assuming as "true" or "nature state" of the unknown sample the minimum risk option. The alternatives are that the unknown voice is either a release of the known speaker voice, or of a different speaker, no matter, in the second case, who the speaker may be.
To compute the risk of the decision, the cost factors must be assigned. Their definition may be an ethic before that a technical or scientific question. Depending on the peculiar aspects of any single case the cost factors may change.
Their choice belongs to the person that is finally responsible for the decision itself. Being E the event, D the decision in our case represented by the sample collections of the population parameters belonging to a known and to an unknown voice, being the risk or decision function has the form:

$$R(D)\text{-}Cost\ Factor\ (D) \times Probability\ (D/E) \qquad (1)$$

Taking as zero the cost of correct decision, the ratio between the two wrong decisions functions allows the outcome:

$$Samples\ identification\ (E)\ if \qquad (2)$$
$$Probability\ of\ False\ Identification\ (P_{f.id.})/Probability\ of\ False\ Rejection\ (P_{f.rj.})<$$
$$<Cost\ of\ False\ Rejection/Cost\ of\ False\ Identification$$

Because in the forensic applications of voice identification the cost ratio, besides the claimed responsibility of the decisor, could be of the order $10^{-3} \div 10^{-4}$; because the P.f.rj. may be assumed equal to α, even in the bayesan approach, under restricted hypotheses the identification of two voice samples requires that:

$$P_{f.id.} \approx 10^{-5} \div 10^{-6} \qquad (3)$$

The $P_{f.id.}$ is what we call the posterior probability of false identification that takes into account either the probabilistic models or the data likelihoods or the event E Priors [7]. The $P_{f.id.}$ is therefore the true contribution to the decision that comes from the experimental data and from the case knowledge. In the embedded likelihoods it is contained the only information increment carried by the measured parameters. It may appear surprising that in most of the current literature, possibly due to the algorithmic difficulties, the calculation of the $P_{f.id.}$ is lacking. The error rates are mainly given as quality test indicators but these are not allowed to enforce any decision because they are related to the particular voices set of the experiment. Furthermore, being the error rates some frequency scores, they will not result in a decision function applicable to real but different conditions.
The IDEM SPREAD system computes a model based expected value of the $P_{f.id.}$. It needs either the speaker model or the population model. Both are hypothesised as Multivariate Normal models with parameters estimated from large Italian population databases [8].
The identification test is governed by the models of the voices average parameters (centroids) whose variability decreases as the sample size square root increases. Either in the known covariance hypothesis ($\chi 2$ test) or in the unknown but equal covariance hypothesis (Hotelling test) the m-dimensional identification region around the known speaker centroid can be shown to be an m-dimensional ellipsoid. The shape is the same of the internal speaker variability region, given by the so called W-MODEL developed by the authors but its size is reduced by the ratio $(n_1+n_2)/n_1n_2$, being n the vector sample cardinality. *How many population individuals fall into the identification region?* Their percentage equals the expected value of the Probability of False Identification. Answering to this question requires the knowledge of the population model (T-MODEL).
The elliptic integrals that compute the expected fraction of misclassified individuals in the test are computed, following the statistical models, by a Montecarlo calculation that is graphically documented in realtime by the IDEM-SPREAD system module. Being it time consuming, under the restrictive hypotheses that the W and T Models are equioriented, that is to say that they have the same correlation matrix, a closed form expression was developed [7] that readily gives a well approximated expression of the $P_{f.id.}$
Good identification test means therefore good W and T Models. Elsewhere [9] we showed the derivation of W, here we present the first reliable results in the development of the T Model that gives the variability of the speakers centroids over the entire Italian population. To gather many replication of few speakers to estimate W is quite simple; to get a representative number of speakers, each with at least 10 vectors, to compute T, is very awkward. Putting together twenty years of laboratory activity we were able to collect a sample of the order of 10^2 (10^3 vectors) that passed the qualification procedures. The provisional target is at last one order of magnitude over.

The database building criteria were:

- Each sample comes from real forensic cases;
- Each sample parametric vector was derived from recorded signals with controlled measuring methods;
- An uniform coverage of the Italian regions was targeted;
- The monodimensional normality tests were performed without any data filtering;
- Bidimensional (scatter) plot were graphically controlled for normality;
- The cross identification tests were performed without any a posteriori filtering (resulting experimental error rate=0);
- The ergodicity of the sample was controlled.

3.1 Corpus description

The database implementation was done with Italian telephonic voices sampled in real case of forensic interest. No laboratory data were entered. This restriction results in a corpus of particular value, of course at the cost of a reduced speed in database growth due to the obvious limits of availability of true forensic trials.
In a country like Italy the regional influence over a voice corpus is suspected to be impressive due to the historical development of the dialects, sometimes similar to true autonomous languages but, in any case, strongly different from each other. The parameter choice and the accurate segmentation of the vowel signals, is expected to reduce the regionalization effect on the corpus.
There is no fear of eventual tendencies to regional clustering of parameters. Nevertheless this effect will reduce the power of the corpus and many more samples could be necessary to validate the statistical models. Data were sampled from Piemonte, Veneto, Sardegna, Lazio, Puglie and Sicilia, out of doubts covering the sociolinguistic communities mainly involved in forensic cases. As shown in fig. 4 the scatter diagram F1-F2 of the vowel "a" the regionalization effect were found of minor importance. The regional scatter and the related model are significatively the same as the whole population. This result does not exclude the opportunity to

specialise the effort toward some regional corpus that may be of large interest in applications.
It must be pointed out that, whichever is the T (population) model adopted, the decision parameters (error probabilities) may be slightly different. All the decision functions must be specified with reference to well defined and documented W and T models.
In a typical forensic trial one or more voice reference samples are released from suspected people. One or more telephone calls are available as test samples. When the case is solved and the results are issued we can have either a unique sample, built up by reference and identified test samples, or two or more groups that gathers the voice parameters that the system defined as non identifiable with each other. For T MODEL building we managed some 200 items with at least 2000 vectors equivalent to 32000 measures. After the forensic trials the different voice items reduced to 100. The averages were computed and a 16 dimensional model was computed as population representative by the Multivariate Normal variance-covariance parameters maximal likelihood estimation.
The result is now available as a Corpus 95 release defined by a population overall Centroid, giving the mean values for all the Italian vowels formant frequencies, and a variance covariance matrix built up by four 4X4 submatrices each representing variability and correlations among the vowel parameters.
A general test was launched to evaluate the experimental error rate in the identification. The expected error score was minimum because the expected P.f.id. estimated from the model is of the average order of 10-5. The actual error rate was zero.

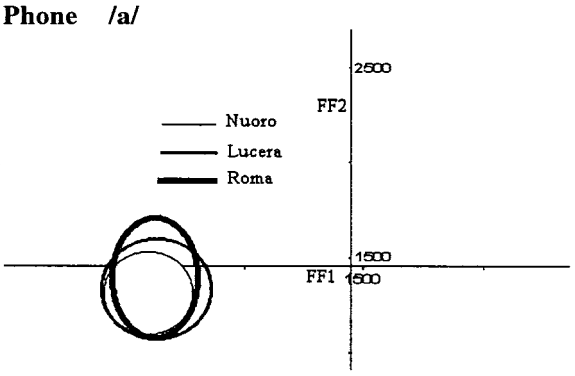

Fig. 4. Scatter diagram F1-F2 of the vowel /a/

3.2 Experimental results

The main expectation was on the sample quality. Remembering that the normality of the formants parameters in the single speaker variability universe was already tested by the authors a normality test was performed on the speaker population means.
The normality is a first approach hypothesis. A more refined approach to regionalities or other clustering factors could have been the multiple normal model.
All two by two parameters scatter diagrams were plotted to evaluate a qualitative behaviour of the two variate normality of the sample. No evident violation was found. The Multivariate Normal model was therefore accepted for the Italian vowel population.

The relation of the expected identification error rate to the test environment are mainly determined by the models, by the reference and test samples dimensions and by the position of the reference sample mean within respect to the general population centroid. Only varying the last item an exponential increase of the identification test quality (the resolution power) is expected. A logarithmic plot of this quality index against the Mahalanobis distance of the reference voice centroid from the population centroid confirms the exponential hypothesis. This result demonstrates that the robustness of a forensic voice identification may increase strongly if the reference voice is uncommon, that is to say a low probability Italian voice.

4. CONCLUSION

With the publication of this paper the first release of a database built up with forensic true data is documented while the statistical tests define the data available as having the right behaviour, it is clear that to increase the database size is mandatory.
The T model describing the Italian population of talkers is here quoted as a multivariate normal model in 16-dimension. The next step will be the extension to 20 dimensions, inserting the vowel "u" and the data fitting with a more flexible multivariate multinormal model that does not question the algorithms to calculate the false identification probabilities here presented. Two further modelization will take into account with more sensitivity factors as regionalities and other data aggregating features.
To partially compensate the slow growth speed of the corpus, due to the limited number of real forensic cases, an accurate control of the measuring procedure, i.e. the formant extraction, will be carried out and a continuous monitoring of the database current state will be performed.

5. REFERENCES

[1] G.R. Doddington. *Speaker recognition - Identifying people by their voices*. Proc. of the IEEE , vol. 73 , n. 11 , 1985.

[2] G. Ibba, A. Paoloni, A. Di Carlo, A. Federico. *Proposta per uno schema di decisione nella identificazione del parlatore mediante prove d'ascolto*. Atti del Convegno AIA '86, Sorrento, 1986.

[3] K. N. Stevens et al. "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material". *J. Acoust. Soc. Am.*, vol. 44, 1968, pp. 1596-1607.

[4] M. Falcone, A. Paoloni, N. De Sario. *IDEM: a software tool to study vowel formant in speaker identification*. Submitted at XIIIth International Congress of Phonetic Sciences '95, Stoccolma, 1995.

[5] A. Federico, G. Ibba, A. Paoloni. *A new automated method for reliable speaker identification and verification over telephone channels*. Proc. of ICASSP, 1987.

[6] M. Falcone, N. De Sario. *A PC based speaker identification system for forensic use: IDEM*. Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, 5-7 aprile 1994, pp. 169-172.

[7] A. Federico. *Reliable statistical Models, for decision making in speaker identification*. FUB Report 5B07694.

[8] A. Federico. *Algoritmi di decisione nei problemi di verifica del parlante*. FUB Report 5B09193.

[9] A. Federico, A. Paoloni. *Bayesian decision in the speaker recognition by acoustic parametrisation of voice samples over the telephone line*. proc. EUROSPEECH'93, Berlin, 1993, pp.2307.