IDEM: A SOFTWARE TOOL TO STUDY VOWEL FORMANTS IN SPEAKER IDENTIFICATION

M. Falcone, A. Paoloni, N. De Sario
Fondazione Ugo Bordoni, Roma, Italy

ABSTRACT

We introduce the new version of the IDEM system, that is a software package, running under Windows, for speaker identification. The recognition algorithm is, in summary, based on the comparison of a set of parameters, e.g. the pitch and the first three formants of the five vowels /a/, /e/, /i/, /o/, /u/ estimate in the stable portion of speech. In particular we describe the SPREAD module, that is the decision module that performs the identification task.

INTRODUCTION

The IDEM system [1] is a set of software tools to perform speech analysis and speaker identification tests on a personal computer, under the Windows graphical environment. It originally utilised a special high cost professional audio board, but now it works also with the standard audio boards compatible with the MPC definition like the Sound Blaster, it only required that the board support the 16bit audio. It was designed to help operators to carry on a speaker identification test in forensic, and special attention was paid to realise an efficient and simple interface as expert and non-expert people should use this package as well. Several revisions of the product have been released. The present one is the V1.8, but new features and powerful characteristics will be added in the future. An update 32bit version for the NT platform or Chicago software platform is also under evaluation.

OVERVIEW OF THE SYSTEM

The system consists of eight applications plus the acquisition one, that is usually bounded with the audio board or with the operating system of the computer, when MPC workstation is used. In the previous versions of the system the acquisition module was part of the package as it manage, under Windows, a specific audio board, but now it is not part if the IDEM package anymore, as the audio is under the direct control of the operating system. Figure 1 shows all the modules of the latest version of IDEM.

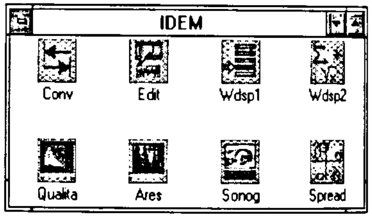


Figure 1. ARES and SPREAD are two applications of the IDEM package

"Conv" is the unique new module; it converts the speech and data file from the old to the new format and vice versa, this make possible to move from the old to the new version without any shocks. For a description of all the modules see [1], here we only describe ARES and SPREAD.

ARES

This module, written in MSVC 1.5, is dedicated to the spectral analysis of fixed length signal window. On the top of the main window a 2.5 second waveform of the audio signal is represented. The cursor is a tick line, wide as the selected zone (you may select any power of two cursor from 128 to 4096 points, default is 512). In the bottom left you have the zoom of the selected window. In the bottom right the power spectrum of the selected window, optionally in this window you can plot the LPC and the CEPSTRUM smoothed power spectrum. Fine tuning is possible by clicking the two buttons on the zoomed signal window. According to the defined number of formant 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. The position (in Hertz) of the line you are moving is monitored on the left side of the window. Just down the waveform you have plotted two scalar quantities (default are pitch and energy).

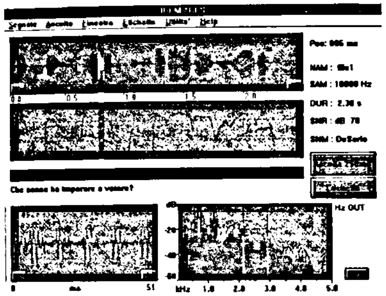
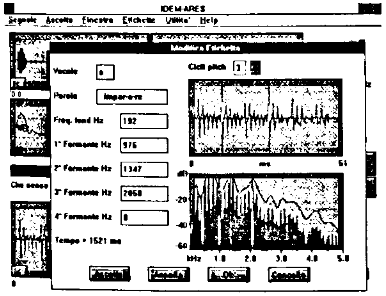


Figure 2. Main window of the ARES module.

Once you have 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 (i.e. save in a file) the information that includes the pitch value, the formants values, the vowel, the context of the word.



Vocale: e
Parola: imparare
Freq. fond Hz: 192
1° Formante Hz: 576
2° Formante Hz: 1347
3° Formante Hz: 2058
4° Formante Hz: 0
Tempo = 1521 ms
Cicli pitch: 3

Figure 3. Expanded (modify parameters) window of the ARES module.

The symbol of the labelled vowel will now appear aligned to the audio wave on the screen. If you want to modify, cancel or control any data, just double click on this symbol and a new window where you may read/modify all these parameters appears on the screen.

SPREAD

This application has been written in Visual Basic 3.0. SPREAD (SPeaker Recognition by Automatic Decision) reads the file created with ARES, or with any other manual or automated method that produces compatible files, and let you set up an "*experiment*" where the parameters (in our case the formants of the vowels) associated to each single speaker may be analysed and compared. You can reach the final decision following a simple step procedure: first step, load the data files; second step, check data consistence and eventually data filtering; third step, run statistical decision test and create reports and documentation. SPREAD also contains several tools that can be utilised along the "*experiment*" execution, to obtain a deeper insight of data, i.e. of the formants distribution.

Double clicking on the icon runs the application, and the main window is opened. Only two menus are active at this stage: "Utility" and "Experiment".

Utility

Under the "Utility" you have the help, that follow the Window standard, and the program configuration. It is possible to select the word processor (e.g. Notepad, Word, Write, etc.) to be utilised for document creation and manipulation, as well the symbols that will be associate to the different speakers in the graphical reports. Once you have made your choice, the configuration is automatically saved, the Utility menu will not be utilised furthermore, unless for the 'exit' command that close the session.

Experiment

It must be clear that SPREAD is based on the "experiment" object.

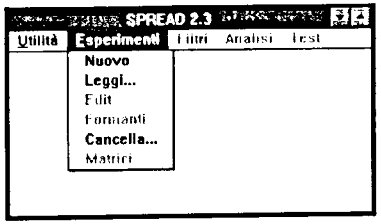
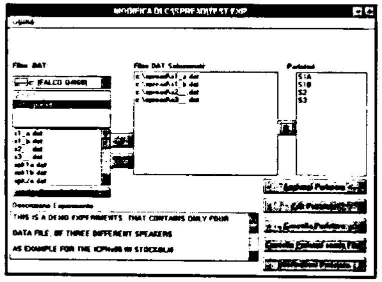


Figure 4. The main window of SPREAD, when run the application

You can work on, modify or delete previous experiments, or create new one. Inside the active experiment you must define: which formants (at least two) you want to use for each vowel; which data files you want to load (you can load and download files as you like it); the reference matrices that model the population [2]. Once you made your choice and loaded the files, you are ready to the intermediate step 2.



THIS IS A DEMO EXPERIMENTS THAT CONTAINS ONLY FOUR DATA FILE, OF THREE DIFFERENT SPEAKERS AS EXAMPLE FOR THE ICPhS95 IN STOCKHOLM

Figure 5. The edit window where the data files are added or deleted inside the experiment.

Filters and Analysis

If you believe that your data are not *clean*, you maybe want to run some filtering. There are five possible filters: one of these is mandatory and it is automatically executed before the test. This filter check data consistence, and looks for missing values or singular matrix, i.e. it checks all the possible causes that make the mathematical procedure senseless. The other filters check for "out of range" values, or performs decimation according to the standard deviation or to the population reference or internal matrix.

Once your data have been validated you may want to look at them. The "analysis" menu has three choices:

- create a report file containing all the information, for each speaker, or for each phoneme (including all data values, mean, standard deviation, occurrence, covariance matrix, etc.);
- plot the data on a Cartesian axis, you may plot the data of any numbers of speakers, for any combination of the used vowels. The variables to plot may be chosen by a selection menu, usually the standard F1 versus F2 plot is used as shown in figure 6. The graph is automatically updated when you click on the menu, so that it works as an interactive graphical environment. It is easy in that way to compare different speakers.

Many other options, as zoom, colour selection, title and legenda insertion, etc. are also available.

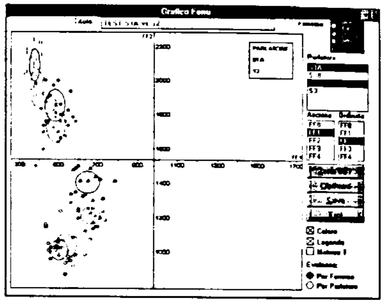


Figure 6. The F1 vs. F2 plot of the four vowels (a,e,i,o) for two speakers

Plots as the previous one only give an immediate overview of the distance among different speaker in a two dimensional space (e.g. F1, F2, or F2, F3, etc.) and their intrinsic limitation must be clear. In fact they may give a wrong indication due to the limitation that only two dimensions are shown.

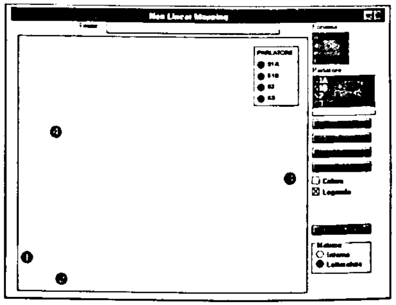
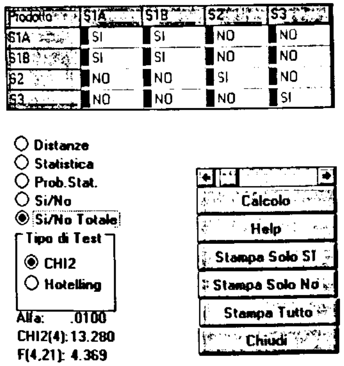


Figure 7. Result of non linear mapping procedure. File 1 and 2, belonging to the same speaker are quite near.

An alternative measure that take into account all the variables [3], may be utilised. In this case you have a mapping of an n-dimensional distance in a two dimensional space. The non linear mapping is an iterative procedure based on a randomised initial disposition of the points. When convergence is reached, speakers with the 'same' voice looks as neighbourhood in a plane space, as shown in figure 7. This representation is useful only for diagnostic purpose, it is not a real test as it is the result of a iterative approximation methodology, in other words it 'may' happens that two speakers that have short distance in this space are not identified, while others with greater distance are identified.

Test and report creation

The last step is the execution of the identification test. Under this windows is shown a table containing on both axis the different speakers name. Each cell contains the value/result between the indexed speakers. You may have numbers in these cells (e.g. the distance between speaker for a given vowels), or the YES/NO identification result of a CHI-2 or a Hotelling test. In this case you have the default value of p=0.01 but it it possible to change this value by a simple click. In figure 8 the results for the four files we have used in this paper are shown. As the non linear mapping indicated, the file S1 and S2 belong to the same speaker, this according a CHI-2 test with p=0.01, i.e. with a probability of 99% of correct identification.



Prodotto	S1A	S1B	S2	S3
S1A	SI	SI	NO	NO
S1B	SI	SI	NO	NO
S2	NO	NO	SI	NO
S3	NO	NO	NO	SI

○ Distanze
○ Statistica
○ Prob.Stat.
○ Si/No
● Si/No Totale

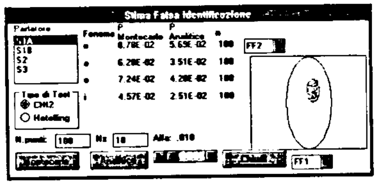
Tipo di Test
● CHI2
○ Hotelling

Alfa: .0100
CHI2(4): 13.280
F(4,21): 4.369

Figure 8. The result of the identification test, is easy to understand

It is also possible to estimate the false identification error, i.e. the probability that a unknown speaker will be identified with some of the speakers used in the experiment. This measure is possible with both analytical and simulated methods, as shown in figure 9.

A set of ASCII reports, with different degree of information, may be automatically created after the execution of the tests. It is also possible to create report for single speaker, or in relation of the test result itself. For example you may have a report that describes only the data that give a positive (or negative) identification score. Many other combinations are possible, but we have no space to describe them here.



Parlatore: S1A, S1B, S2, S3

Fonema	P Montecarlo	P Analitica	#
[illegible]	8.78E-02	5.65E-02	100
[illegible]	6.20E-02	3.51E-02	100
[illegible]	7.24E-02	4.20E-02	100
i	4.57E-02	2.51E-02	100

Tipo di Test
● CHI2
○ Hotelling

N.punti: 100 Ns: 10 Alfa: .010

Figure 10. The False Identification Error is computed with analytic method, and with Montecarlo simulation

CONCLUSION

We introduced the last version of the IDEM system. In particular we describe two modules: ARES that let you create data files containing the value of the vowel formants, it has a easy to use interface with audio and visual feedback; and SPREAD that starting from these files let you analyse and filter the data and perform identification tests. This package is now available only for Italian language and it is used, mainly, for forensic purpose. We are currently evaluating a new and powerful (English) version for the Windows NT platform.

ACKNOWLEDGEMENT

We would like thank all that colleagues that contribute to the practical and theoretical development of IDEM, in particular Berardo Saverione and Antonio Federico, as well all the users that gave us useful hints and feedback.

REFERENCES

[1] Falcone, M., De Sario, N. (1994), "*A PC Speaker Identification System for Forensic Use:IDEM*", ESCA Workshop, Martigny, April 1994, pp.173-176

[2] Federico, A., Paoloni, A. (1993) "*Bayesian Decision in the Speaker Recognition by Acoustic Parametrisation of Voice Samples Over the Telephone Line*", Proceedings EUROSPEECH'93, Berlin, September 1993, pp.2307-2310

[3] Sammon, I.W. (1969) "*A Nonlinear Mapping for Data Structure Analysis*", IEEE Trans. on Computer, Vol.18, N.5