AN EDUCATIONAL AND EXPERIMENTAL WORKBENCH FOR VISUAL PROCESSING OF SPEECH DATA

Jan Nouza, Miroslav Holada, Daniel Hajek

SpeechLab, Dept. of Electronics and Signal Processing
Technical University of Liberec, Halkova 5, 461 17 Liberec, Czech Republic
Tel. +420-48-254 41 / 208, FAX: +420-48-510 71 26, E-mail: jan.nouza@vslib.cz

ABSTRACT

In the article the focus is put on educational aspects of the speech processing science. A set of tools that have been developed with the aim at presenting, visualizing and explaining basic topics of speech recognition is described. The set consists of programs, like a signal analysis unit, a dynamic time warping algorithm (DTW) explorer and hidden Markov model (HMM) investigation tools, that are integrated into a single environment and allow for easy and highly illustrative learning through experiments with real speech data.

1. INTRODUCTION

The recent progress in the speech and natural language processing (SNLP) domain has been reflected not only by a large interest of the commercial sphere but also by a growing number of educational activities. Many universities open undergraduate and postgraduate courses on SNLP. In Europe, a scheme of collaboration in education has been created under the Socrates program [1]. Also an increasing number of textbooks on speech processing has appeared recently (e.g. [2], [3]) together with latest versions of supporting software like, for example, that of Entropic Research [4].

Yet, there seems to be a lack of tools that could help students and other interested people in getting an appropriate starting knowledge and experience in speech topics. Some of the existing software allow to learn on the experimental base, however, it is often just a run-test-see-result approach. We believe that a good education program should offer more. It should demonstrate and explain the methods and their principles rather than only emitting numbers, scores, etc., which is an essential demand, particularly, in such complex areas like, for example, the hidden Markov models (HMM).

Our initial attempts on the educational software field date to 1995 when we presented the first version of the Visual Markov package [5]. It has been offered for free use at universities and research institutes and found a warm response in the speech community. Later, some other teaching and visualization tools followed [6]. The next, quite natural, step was to integrate all the tools into one environment that would allow for easy use. Moreover, the environment should provide a necessary support to all kinds of experiment work, such as data recording, speech database maintenance, communication between the individual programs and data transfer.

As a result of our effort, a workbench-like system named *VISPER* (*Vi*sual *Spe*ech P*r*ocessing) was created in 1997. It is a graphic (PC based) system consisting, at present, of five autonomous units: the Signal Profiler, the DTW Explorer, the Visual Markov training and testing tools and the VISPER Organizer. The whole system is oriented on explaining major aspects of the isolated-word recognition task. Such a task seems to be an appropriate one for introduction into the speech processing area.

2. SPEECH DATA VISUALIZATION

There exist a lot of speech analysis tools that provide most of the standard signal processing functions together with a visual presentation of the input and output data. Among them, the *Waves+* package [4] is probably the best-known one. However, not many of such tools have been designed for the PC platform, which makes their broader use in education rather expensive. Also none of the tools, at least to the authors' knowledge, allows direct communication and data transfer to other tools that perform tasks associated with speech recognition.

When designing the *Signal Profiler* program (see Fig.1) we aimed at creating an easy-to-use tool that would accomplish most of the signal processing tasks. It was to provide a user by speech recording options with automatic detection of spoken utterances, by signal visualization and audition facilities, by computation and plotting of widely used speech features and by displaying spectral characteristics of the signal. Moreover, the program was to ease recording of speech databases and provide a simple access to other speech recognition tools.

The largest panel in the Signal Profiler is devoted to the speech signal corresponding to a single word (or single utterance). The displayed signal is partitioned to 3 zones; a main zone containing the (automatically detected) speech signal and two margin zones with pieces of the signal preceding and following the utterance. Any frame of the signal can be selected and displayed in detail in the frame panel. Moving a mouse pointer along the signal waveform results in a synchronized show of frames. Optionally, the frame panel can present some other plots relative to the displayed frame, e.g., the „hamminged“ signal or short time spectrum plots.

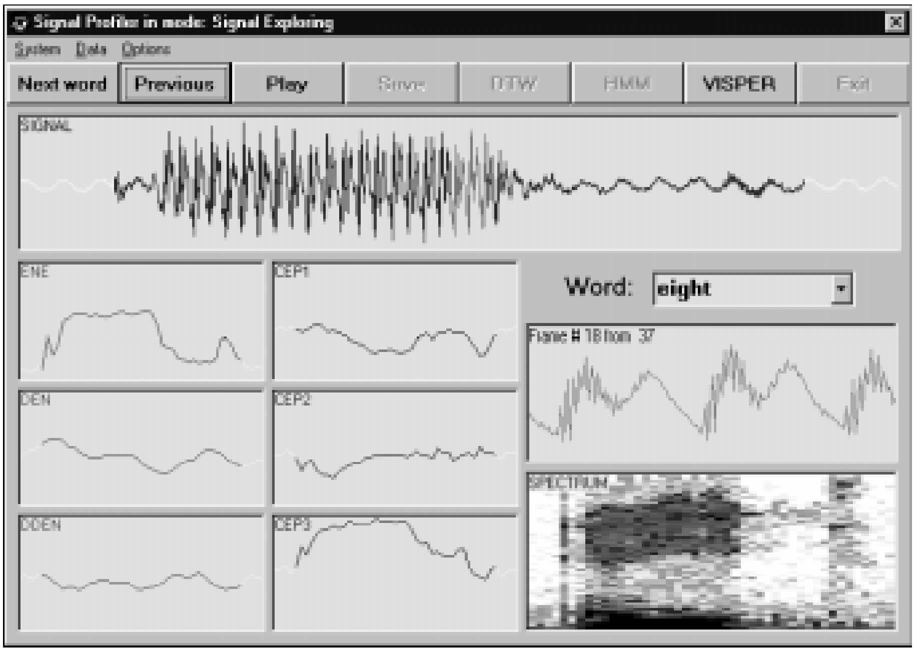


Fig.1. *The Signal Profiler unit operating in the data observation mode*

A group of six feature panels serves for displaying parameters from a given feature set. In our case the set consists of 20 components: 8 cepstral coefficients together with their deltas, energy and its 1st and 2nd derivatives and a spectral variation function (SVF). The latter parameter, defined in [7], was chosen because it may help a user in identifying and realizing boundaries between stationary speech segments. The last panel displays a rough estimate of the signal spectrogram with 2 options available; either a color 2D plot or a 3D view.

The Signal Profiler operates in 2 basic modes: a speech recording mode and data observation mode. In both the modes the program can communicate with the other speech processing tools described in the next sections.

3. VISUALIZATION OF RECOGNITION PROCEDURES

What makes the VISPER system unique is that it focuses not only on classic visualization of *data* but also on visual demonstration of *procedures* and *algorithms* applied in speech recognition. The two techniques chosen for the demonstration are the Dynamic Time Warping (DTW) algorithm and the Hidden Markov Models in their continuous version (CDHMM).

3.1 Dynamic Time Warping

The DTW technique was the first of the methods that made speech recognition feasible under some limited conditions. Though it has been later overcome by the introduction of HMMs, its knowledge is still essential for understanding the principles of modern speech recognition systems. That is why a unit named *DTW Explorer* has been added to the VISPER package.

The DTW Explorer's screen is shown in Fig.2. We may notice two regions in the screen layout. In the left one, there is a pair of speech signals to be matched. The upper is a tested word, the lower is a reference. Both of them are represented by a selected feature (e.g. energy). The plots, though simplified by a one-dimensional projection of the original feature space, give the user at least an approximate view on the signal contours and duration. The latter is explicitly shown as the number of (10 ms long) frames.

For each of the two signals there are two similar windows available. The upper always shows the original, the lower displays the signal after the warping procedure. The time-aligned signals as well as the local, accumulated and global distances are displayed in the lowest of the windows.

The large panel in the right half of the program screen gives a detailed clarification of the time-alignment problem. This is achieved by visualizing the space where the test-reference mapping is searched. Here, the Explorer offers three visualization modes. In the simplest one, the warping path is shown - in the classic way - as a polyline in the *xy* plain. The second mode adds colors to the plain and presents the DP problem as a search in a cartographic map. Yet, the most illustrative explanation of the DTW alignment is provided by a 3D plot (shown in Fig.2) where the *z*-axis elevation corresponds to the local distances.

In order to increase the visualization effect, all the plots associated to a DTW match are synchronously animated. As the warping path is drawn (e.g. in the mountainous landscape), the warped and aligned signals in the left part are plotted simultaneously. The animation can be controlled by a set of command buttons.

A wide variety of system options allows for extensive investigations. The user can choose from various DTW algorithms as they were introduced in classic literature (e.g. [8],[9]), set up different local and global constraints, select features and distance measures, etc. It is also possible to make automatic reporting from the experiments and print out the results together with graphic illustrations.

3.2 Continuous HMMs

The technique of continuous HMMs has been widely used in speech recognition since late 1980s. Though some good textbooks on HMMs have been published (e.g. [10]), for many students it is difficult to achieve full understanding of the technique that utilizes a high level of abstraction.

Our *Visual Markov* tools have been designed for a special case of modelling whole words by left-to-right HMMs. This type of models finds a wide use in many practical isolated-word systems operating with small and medium size vocabularies.

The graphic design of the tools (Fig.3) allows to observe either all or several selected states of a Markov model. Each state has its own window where the output pdf is displayed as a 3D function of two optionally selected features. The pdf is supposed to have a form of a mixture of one or more gaussian functions. In the upper right corner of each of the windows, the probability of staying in the state is printed. The last window (info window) is used to show some relevant statistics concerning either the training or matching procedure (e.g. the number of iterations, the current likelihood score, etc.)

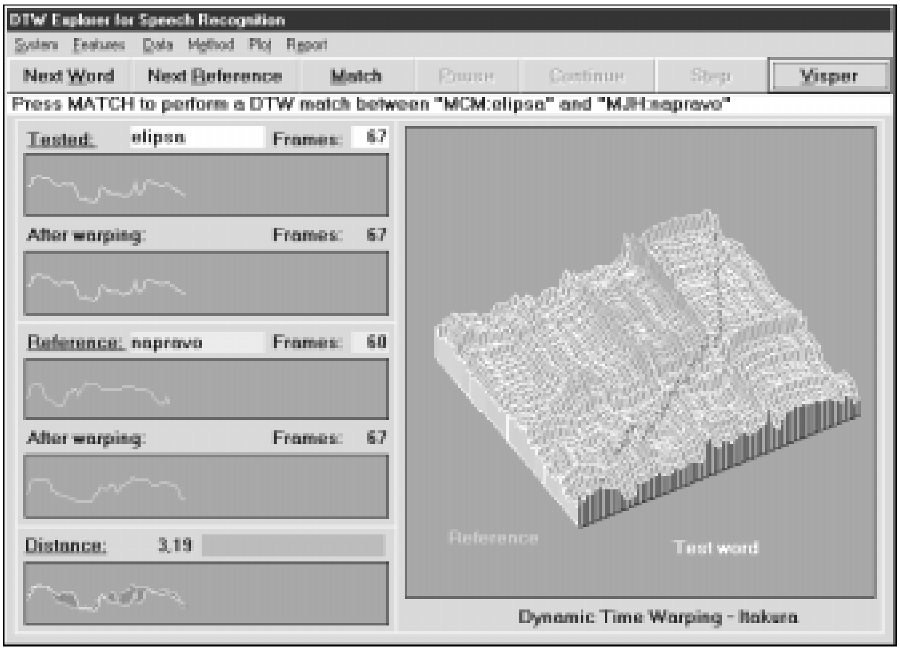
3.2.1 CDHMM training

The training is based on the standard CDHMM scheme: first a k-means initialization and then the Baum-Welch reestimation. The process of training is animated and displayed in iteration steps so that the evolution of the model states and their parameters can be observed. (The visual demonstration is interesting, particularly, in case of a multi-mixture pdf.) The info window gives a global overview of the procedure by displaying the current process status (initialization, reestimation, termination), the iteration counter and the total likelihood score.

The user can choose from a wide range of options; he or she can set up the model parameters (numbers of states, mixtures and features) as well as the system parameters (the features to be displayed, 3D plot settings, etc.)

3.2.2 CDHMM matching

The matching between a word and a model is presented in the same graphic layout. The procedure used is the Viterbi algorithm. It is demonstrated by means of a green ball traveling through the visualized model states. Each of its travel stops corresponds to one speech frame. The actual position of the ball is determined by the currently processed frame vector and by the Viterbi decoder that had estimated the most likely state sequence. The evolution of the log likelihood score can be observed in the info window and compared with the score achieved for the best model. This gives the student a nice opportunity to see how much the model fits the utterance, both on the local and global level.


DTW Explorer for Speech Recognition
System Features Data Method Plot Report
Next Word
Next Reference
Match
Pause
Continue
Step
Visper
Reference
Test word
Dynamic Time Warping - Itakura

Press MATCH to perform a DTW match between "MCM:elipsa" and "MJH:napravo"

Tested: elipsa Frames: 67

After warping: Frames: 67

Reference: napravo Frames: 60

After warping: Frames: 67

Distance: 3,19

Fig.2. *The DTW Explorer performing a visual match between two words*

As in the previous case, the user have a variety of choices and options; the selection of the word and the model, the two displayed features, the plot parameters, etc. A set of buttons allows him or her to control the animation flow.

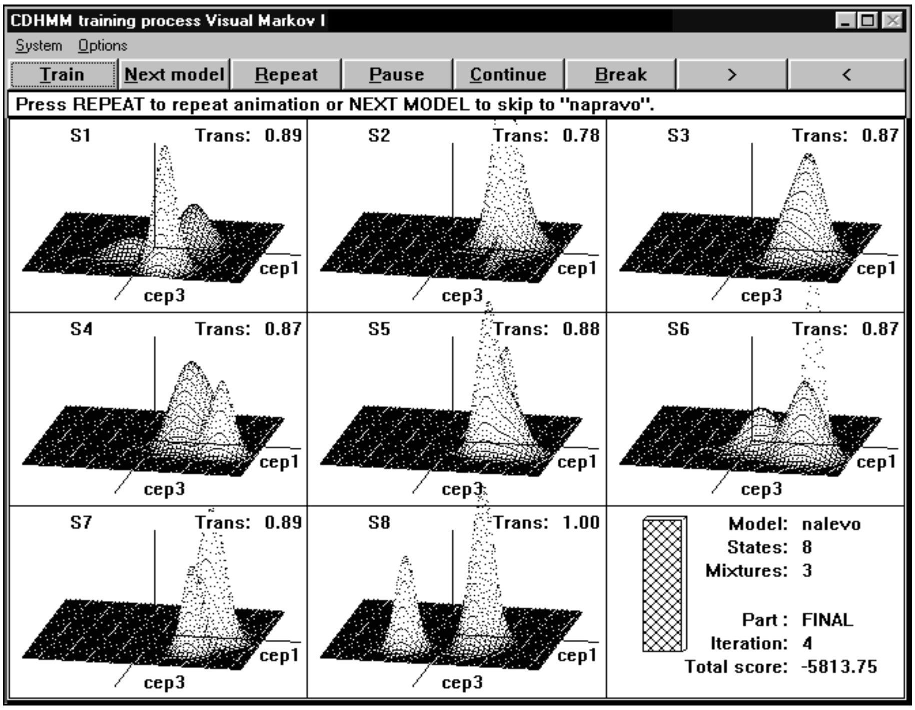
4. INTEGRATED EXPERIMENTAL ENVIRONMENT

The *VISPER Organizer* is responsible for preparing and managing the experiments. Its main tasks include the maintenance of the speech database and the disk space, creation of the vocabularies, defining the conditions for the experiments as well as preparing training, reference and test data. It also starts and controls all the programs necessary for a session and ensures their communication.

To simplify the organization and launching of the experiments, the VISPER environment offers several standard operation modes:

1. Database recording
2. Signal observation
3. DTW matching
4. HMM training
5. HMM matching
6. Signal investigation and recognition
7. Real-time speech recognition

While the first 5 option are so called single modes (only one functional unit operates at a time), in the last 2 modes all the program units are employed simultaneously. For example, in mode 7: an uttered word is automatically detected, displayed by the Signal Profiler, immediately recognized (using either of the recognition techniques) and ready for demonstrating the visual match.



Press REPEAT to repeat animation or NEXT MODEL to skip to "napravo".

Model: nalevo
States: 8
Mixtures: 3

Part : FINAL
Iteration: 4
Total score: -5813.75

Fig.3. *The Visual Markov demonstrating a training of a 8-state 3-mixture model*

The main advantage of this arrangement is that the student does not have to care about any extra and auxiliary work. Instead, he or she can focus just on the idea and the experiment. A sample session including a definition of a new vocabulary, recording reference data and testing them in real-time DTW recognition may take just a couple of minutes.

5. CONSLUSIONS

The system and the tools described in the paper have been developed with the aim to make teaching and learning topics of the speech processing science more attractive and better understandable for students. The visual tools might be appreciated also by other interested people who want to learn more about speech and the methods of its recognition.

ACKNOWLEDGMENTS

The work described in the paper have been supported by the Czech Grant Agency (GACR) under the project No.102/96//K087.

REFERENCES

[1] Information on Socrates Thematic Network „Speech Communication Science“ available at WWW address: http://tn-speech.essex.ac.uk/tn-speech/

[2] Deller, J.R., Proakis, J.G., Hansen, J.H.L.: Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, New York, 1993.

[3] Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. MacGraw Hill, New York, 1993.

[4] ESPS Waves+ and HTK Manuals, Entropic Research Laboratory, Cambridge, 1996 - 1997.

[5] Hajek D., Nouza J.: Unhiding Hidden Markov Models by Their Visualization (Application in Speech Processing). In Gobel M., David J., Slavik p., van Wijk J. (eds.): Virtual Environments and Scientific Visualization’96. Springer Verlag, Wien - New York, 1996, pp.277-285.

[6] Hajek D., Nouza J.: Visualization of Data and Procedures in Speech Processing. In Studientexte zur Sprachkommunikation, Heft 13 (Elektronische Sprachsignalverarbeitung), Berlin, 1996, pp.218-223.

[7] Nouza J: Spectral Variation Functions Applied to Acoustic-Phonetic Segmentation of Speech Signals. An article to appear in Forum Phoneticum, Goethe University, Frankfurt, 1997.

[8] Itakura F.: Minimum prediction residual principle applied to speech recognition. IEEE Trans. on Acoust., Speech and Signal Processing, vol. ASSP-23, 1975, pp.67-72

[9] Sakoe H., Chiba S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on Acoust., Speech and Signal Processing, vol. ASSP-26, 1978, pp.43-49.

[10] Huang X.D., Ariki Y., Jack M.A.: Hidden Markov models for speech recognition. Edinburgh University ress, Edinburgh, 1990.