System for visual and experimental introduction to basic speech recognition algorithms

Jan Nouza
SpeechLab, Department of Electronics and Signal Processing, Technical University of Liberec
Halkova 6, 463 11 Liberec 30, Czechia
jan.nouza@vslib.cz

Abstract

This contribution describes an educational system for visual presentation of basic speech recognition algorithms. It is aimed at introducing topics, like speech feature extraction, dynamic programming (DP) or the application of the hidden Markov model (HMM) theory. The system allows students to set up various types of recognition experiments with either pre-recorded or directly spoken words and run these experiments in the way that all the steps in the classification procedures are visualised and animated on a PC screen. The tool has been designed for the use in MSc/PhD courses on speech, language and phonetics.

1 Introduction

For long time, two closely related research domains, speech processing and natural language processing, have been developing separately, using different methods and having different goals. However, the recent progress on both the fields opens new opportunities for collaboration of speech and language experts in such tasks, like large vocabulary continuous speech recognition, spoken dialogue systems, etc. People from speech community should be acquainted to advanced NLP techniques and NLP expert may take advantage of knowing at least the basic algorithms used for speech processing and recognition. There is no doubt that learning generally applicable methods like the hidden Markov (HMM) technique, dynamic time warping (DTW) or homomorphic speech signal parameterization may be of great use even for students of NLP courses.

For educational purposes we have developed a system named VISPER (VIsual SPEech pRocessing). It has been designed namely for students taking introductory lessons in speech processing but it has found use in other language-related courses, like NLP, phonetics, artificial intelligence. The system is aimed at explaining basic speech recognition algorithms and procedures, like signal processing, feature extraction, word distance measure, hidden Markov modelling. Learning and understanding these topics is supported by VISPER's graphic design that allows for detailed visualisation and animation of the essential procedures. A large choice of options and settings makes the system a helpful educational workbench enabling students to learn on their own experiments.

2 VISPER - brief description

The idea of the VISPER matured several years. Our initial goal was to help students in understanding what is actually *hidden* in hidden Markov models. We did it by creating a tool named Visual Markov [Hajek 1996]. Later, similar teaching aids were designed for exploring the DTW algorithm and for demonstration of cepstral analysis of speech data. The next step was to integrate all these tools into a single environment that would allow students to prepare and run even complex experiments in an easy way without a need for any other programs or scripts. This was accomplished by the design of the VISPER system [Nouza 1997].

The VISPER runs on the Win95/98/NT platform. It consists of four main components: Signal Profiler, DTW Explorer, Visual Markov and Organizer.

The Organizer controls all the activities of the system by means of a simple dialogue. To start a work with the system, the user chooses from one of the ten operating modes. The modes cover both single actions like, data recording and signal analysis, DTW matching, HMM training and matching, as well as actions that utilize and combine functions of more than one component. For example, in the Speak & Recognize (HMM) mode, the system will automatically detect a spoken word, display it in the Signal Profiler's window, recognize it with the aid of the previously trained HMMs and gets ready for demonstrating the course of the Viterbi match in the Visual Markov's window.

The main task of the Signal Profiler is to handle speech data and process them to get features applicable for recognition. The recorded signal can be displayed, replayed and observed in time and frequency domains, both at detailed and global level.

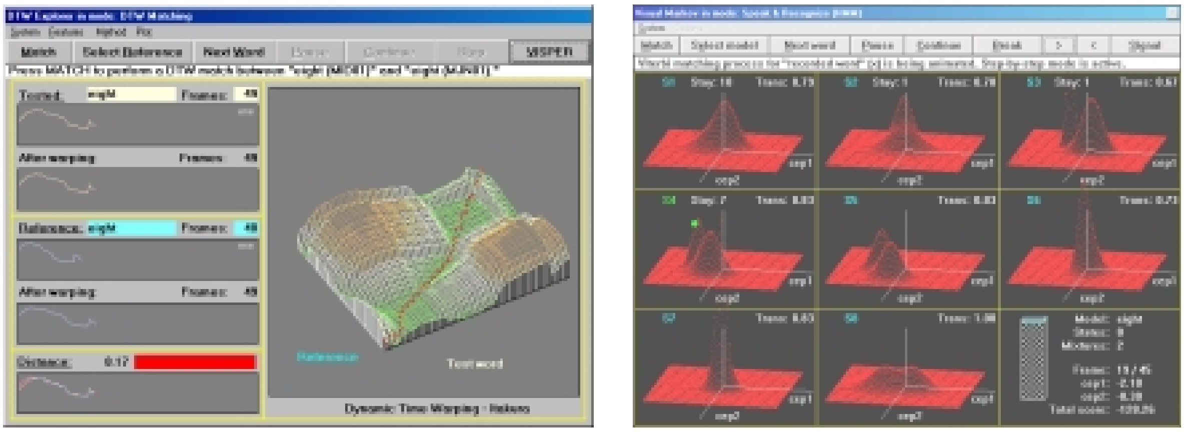

Fig.1. *Two of the VISPER tools: DTW Explorer (left) and Visual Markov (right)*

The DTW Explorer performs and presents several variants of the Dynamic Time Warping method, which is the basic technique used for measuring the similarity between two speech tokens. Animated 2D or 3D pictures provide a closer look at the principles of this classification procedure. Using the 3D graph (Fig.1), the search for the optimal alignment between the two tokens may be interpreted as a task to find in the mountains the path with the lowest total elevation. The student can compare the paths and their scores for different reference templates using various modifications of the DTW algorithm.

The Visual Markov tool has been developed to explain how words can be modelled by continuous density multi-gaussian HMMs. The graphic design of the program allows to observe either all or several selected states of a Markov model. Each state has its own window where the output probability density function (pdf) is displayed as a 3D function of two optionally selected features. The pdf is supposed to be 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 (Fig. 1.)

The Visual Markov works in two operation modes. In the first one, it performs training based on the well-known Baum-Welch reestimation formulae. 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. In the second one, the principle of the Viterbi search algorithm is visualised and explained.

3 VISPER - use in education

The system can be used by a teacher during his/her lectures on speech processing. In such case it may serve for illustrating the basic principles and their application. However, its main use is in practical seminars. The students can prepare and run experiments with their own speech data, their own parameter, model and algorithm settings. In this way they achieve a higher level of understanding of the topics. Instead of asking a teacher they try to find answers on their questions simply by running and observing an experiment. A detailed curriculum of a 5-lecturer and 5-seminar introductory course in speech recognition can be found in [Nouza 1999].

4 Conclusions a future work

Since its completion in 1997 the VISPER has been registered and used at some 60 universities and research institutes world wide. More information about it is available at URL specified below.

In very near future the VISPER's capabilities will be enhanced by allowing also the presentation of basic algorithms used in continuous-speech recognition.

Acknowledgements

The work reported in this paper was supported by the Grant Agency of the Czech Republic (grant no.102/96/K087) and by the Czech Ministry of Education (research goal project MSM 242200001).

References

Daniel Hajek, Jan Nouza (1996): *Unhiding Hidden Markov Models by their Visualization*. In „Virtual Environments and Scientific Visualization '96“. Gobel M., David. J., Slavik P. and van Wijk J. (eds.) Springer-Verlag, Wien - New York, 1996, pp.277-285

Jan Nouza, Miroslav Holada, Daniel Hajek (1997): *An Educational and Experimental Workbench for Visual Processing of Speech Data*. Proc. of EUROSPEECH'97, Rhodes, Greece, pp.661-664.

Jan Nouza (1999): *Teaching and Learning through Visual Speech Processing Experiments.* Proc. of MATISSE Workshop, London, UK, pp.121-124

VISPER at http://itakura.kes.vslib.cz/kes/visper.html