

VISPER II - Enhanced Version of the Educational Software for Speech Processing Courses

Miroslav Holada, Jan Nouza

SpeechLab, Department of Electronics and Signal Processing
Technical University of Liberec, Czech Republic
miroslav.holada@vslib.cz, jan.nouza@vslib.cz

Abstract

In the paper we describe a new version of the software tool developed for education and experimental works in speech processing domain. Since 1997, when the original VISPER was released, we have added several new modules and options that give a student a deeper look at the basic principles, methods and algorithms used namely in speech recognition. Newly included modules allow for visualization of the Viterbi search algorithm implemented either in sequential or parallel way, they introduce the idea of the beam search with pruning and guide a student towards understanding the principle of word string recognition. The VISPER concept of a single graphic environment with mutually linked modules remains untouched. The VISPER II is compatible with all recent versions of the MS Windows OS and it is freely available.

1. Introduction

Recently, we can see a constantly growing interest in education in the field of speech science and technology. It has been stimulating by a great demand from industry searching for new experts in speech and language processing. This was clearly demonstrated in the large survey [1] conducted in 1996 - 1999 within the European Socrates Thematic Network on Speech Communications Sciences. As a response to the Network proposals and recommendations, a special ISCA interest group on education (EduSIG) was established. It has launched a new line of events, such as Educational Arena and special tracks at Eurospeech conferences, or the MATTISE workshop, etc.

At academic institutions these actions helped to increase the motivation towards developing and freely sharing education oriented tools and programs. At present, education software covers such topics, like speech production and perception [2], speech signal analysis, synthesis and annotation [3], speech coding [4], speech recognition [5], speaker recognition [6] as well as dialogue management [7].

Our contribution to this domain has been the software for visualizing and explaining the basic algorithms used in speech recognition. The software, known under the name VISPER has been recently completely revised and complemented by several new tools and options. Some of these innovations followed the comments coming from the user community.

2. The VISPER's brief history

The first idea on the visual presentation of the speech recognition basics dates back to 1995 when we searched for an appropriate way of introducing the Hidden Markov Model (HMM) concepts to our students. The product named Visual Markov [8] was followed a by similar tool explaining the Dynamic Time Warping (DTW) method in 1996. At Eurospeech in 1997 we presented a unified environment that enabled a student to run off-line as well as on-line experiments based on DTW and HMM classification methods. The unique feature of the VISPER platform consisted in the fact that no programming nor scripting was needed in order to prepare and run any experiment. Hence the software could be used not only by students of engineering studies but also within speech science courses at non-engineering faculties. Another feature appreciated by the users was the animated presentation of the algorithms, which showed the training and classification procedures step after step.

Since 1997 when the VISPER was released, other research projects have not allowed us to continue in further development. We just fixed several bugs and tried to keep the software compatible with the latest versions of the MS Windows systems. A recent wave of the increased interest in speech courses at our university, however, made us realize some of the ideas that gathered during the last five years. This resulted in the new version of the VISPER that is available to the community since 2003.

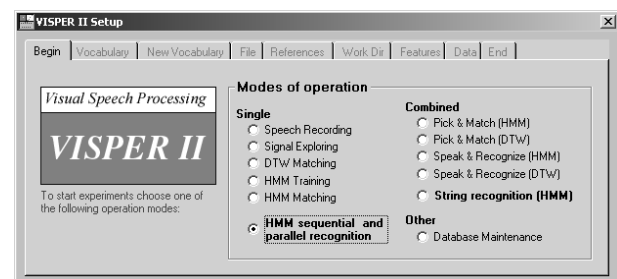


Figure 1: The VISPER starting window that allows a user to set up and launch an experiment. The new modules for sequential, parallel and string recognition are highlighted

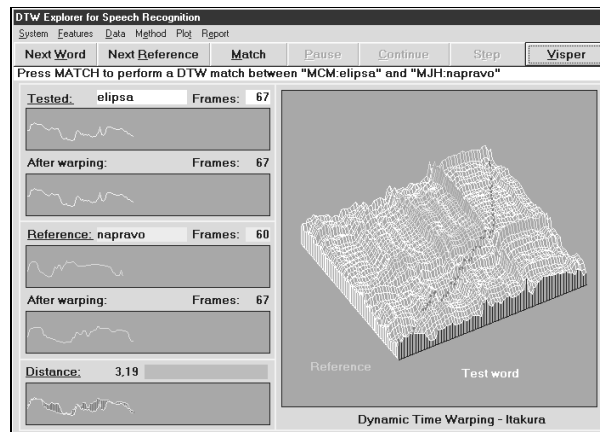
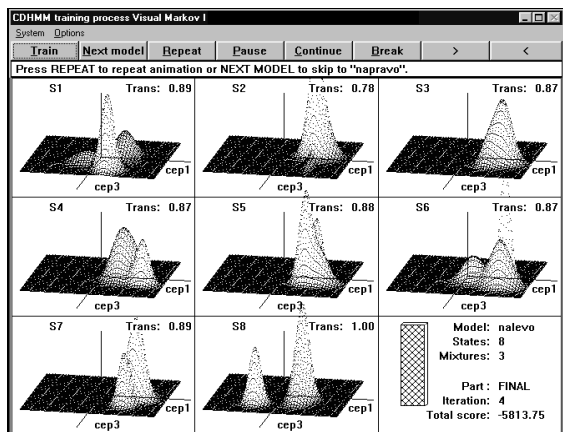


Figure 2: Visualization of the HMM and DTW procedures on real speech data

3. New modules and features in VISPER

The original version had these basic working modes:

Speech recording and parameterization, signal and parameter exploring, DTW matching, continuous HMM training and matching, on-line recognition with the DTW and HMM algorithms, and simple tools for speech and model database maintenance. Two of the working mode are displayed in Fig.2

The new version further includes a module for visual presentation of sequential and parallel recognition and a module for word string recognition.

3.1 Sequential HMM recognition

In this mode the user can learn the classic implementation of the Viterbi algorithm, where an unknown word (of length F frames) is matched sequentially to HMMs of all words in the vocabulary.

The matching procedure is displayed in lattice $f \times s$ (f are the frames and s stands for model states - see Fig.3). For each model the evaluation is made and animated in columns, representing frame f . The best path leading to any point (f,s) is indicated so that the local decisions of the Viterbi algorithm can be seen. After processing the last column (i.e. the last frame), the log likelihood in the point (F,S) is displayed and the optimal alignment between the word frames and the model states is revealed by backtracking the local pointers. The same is done and visualized sequentially for all the models. After processing the last model the best score is shown and the word is classified.

The whole procedure can be animated on several levels. On the highest one only the resulting scores and the revealed paths are shown and the recognition result is available in one instant. If desired, the procedure can be animated frame after frame, word after word, which allows the student to understand not only the algorithm itself but also its typical implementation.

3.2 Parallel HMM recognition

This mode utilizes the same environment as in the previous one. However, here the matching is done frame by frame but in parallel across all the models. In this arrangement the student can learn the principle of time synchronous processing. Such implementation allows for comparing the best local scores for individual models and make a decision about the least promising hypothesis. The student can choose a threshold for pruning and immediately see how it will influence the recognition. The evaluation of the models whose best paths do not fit with the threshold is canceled and the number of models and paths under consideration gradually decreases. In this way the student can learn the principle of the Viterbi beam search, which is the basic concept of large vocabulary recognition. The animation options are same as in the previously described mode.

3.3 Word string recognition

Our experience says that after learning the principle of the parallel evaluation, understanding the most difficult concept of word string recognition becomes much easier. It is also because the VISPER demonstrates it in same environment as in the previous modes. The student just notices a new entry state added to each model at its start. This state serves for transferring the scores from predecessor words.

The visualization of the string recognition procedure is shown in Fig. 4. In principle it is very similar to the parallel classification mode. The animated evaluation is done again frame after frame. When one or more paths reach the final states of any of the models, the routine finds the model with the best score in the final state and the score value as well as the information about the winner is passed to the entry states of all the models. The tool visually indicates all the information used for the decision at the internal and word-end level.

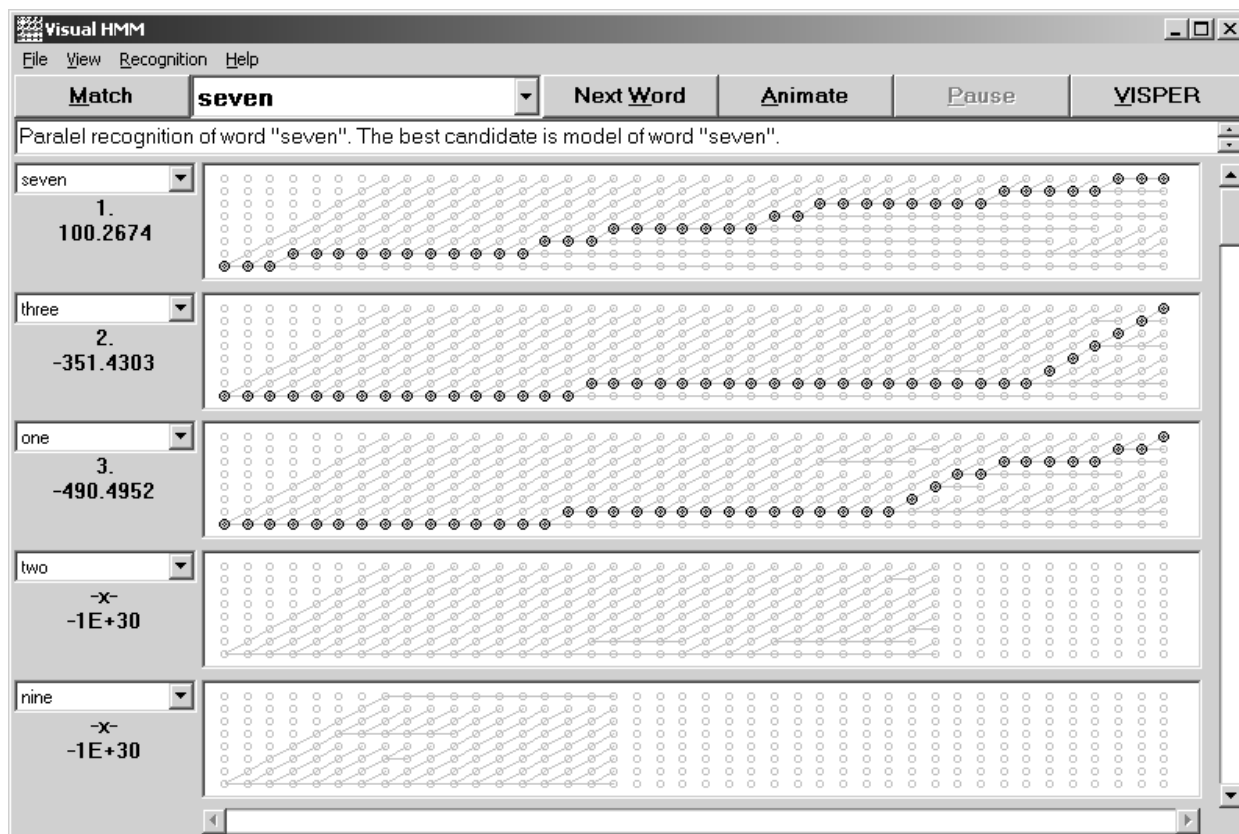


Figure 3: Visualization of the parallel HMM classification procedure with pruning.

(For each of the displayed models, the local path decisions are shown, for the unpruned models also the globally optimal path is backtracked at the end. The complete procedure is animated frame by frame. The evaluation of the least promising models is canceled on the way, which was the case of the models „two“ and „nine“.

The student learns that the backtracking procedure is of the essential importance here, because it is the only means to unveil the most probable sequence of the spoken words.

The word string recognition must be preceded by the definition of the allowed word combinations. In a special window the user can allow either all word pairs or he/she can select those combinations that will be allowed/disallowed. The values of the bigrams are calculated and displayed in order to introduce at least the basic concept of the probabilistic language model.

3.4 Additional new features and options

In all the three modes the classification routines can run with up to 50 word models, although for understanding the principles a vocabulary of 10 words (like e.g. digits) is sufficient enough. The two scroll bars in the animation window allow for displaying any part of the model and time space. Moreover, the order and the position of the displayed models can be arbitrarily chosen, so that the user can focus on those models that are the most relevant ones in the particular experiment. The user also has a large choice of options that make the visual

presentation even more transparent. It is possible to modify the color and size of the lattice points and paths, the models can have same or different length, the position of the models can be automatically reorganized according to the scores, in the string recognition the word-end decisions can be displayed for each frame, the control buttons can have a touch sensitive help, etc.

The organization of all the experiments within the VISPER II has been further simplified in the way that the configuration of the experiments can be stored and later retrieved. This also helps the teacher to prepare a line of experiments that guides the student from very simple to the most complex ones.

3.5 Compatibility with the other modules

All the new modules are fully compatible with the original ones. Hence, within the single environment the student can record his/her own speech database. Using the database he/she can run initial DTW experiments or train word models for HMM recognition. The HMM models can be 3D visualized and at the same time they are ready for sequential, parallel or string recognition - without any programming or scripting work.

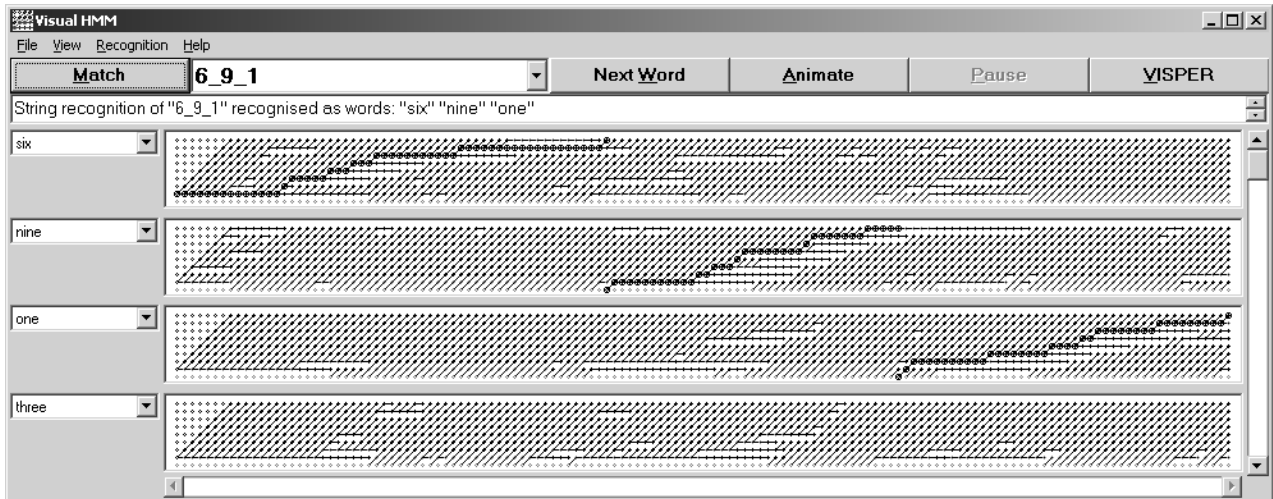


Figure 4: Visualization of time synchronous Viterbi search in the word string recognition task

4. Application in speech related courses

The VISPER tool has been used in our MSc courses in Pattern and speech recognition for more than 5 years. Our experience show that a tool of this type not only makes the course more attractive for the students but it also helps the teacher in preparing the lectures. While previously we used VISPER screen-shots drawn on slides, recently we run the software on the data projector during the lecture and provide the students by an animated and more comprehensive explanation of the algorithms. At the end of the one-semester module at least some of the students are able to write their own programs for DTW and HMM recognition. The VISPER also can help them in developing and debugging their programs, because they can test them on the same data that are shown in the visual environment.

The previous version has been registered on some 80 institutions world-wide. According to our knowledge the software was used not only at technical departments but also in speech science courses taught at non-engineering faculties (e.g. at the Dept. of phonetics at the Charles University in Prague.) The software is freely available for any non-commercial education purposes - for details see out WWW pages [9].

Acknowledgements

This work has been supported by the Grant Agency of the Czech Republic (grant no. 102/02/0124) and through research goal project MSM 242200001.

References

[1] Bloothoof G. et al: The Landscape of Future Education in Speech Communication Sciences. Vol-

umes 1, 2 and 3. OTS Publications. Utrecht, 1997, 1998, 1999.

- [2] Fellbaum K., Richter J.: Human Speech Production Based on a Linear Predictive Vocoder - An Interactive Tutorial. Proc. of MATTISE conference. London, April 1999, pp.57-60.
- [3] Speech Filing System (SFS) by UCL. Available at <http://www.phon.ucl.ac.uk/resource/sfs/>
- [4] Uhler J.: The Set of Exercises in Digital Speech Processing. Proc. of MATTISE conference. London, April 1999, pp.53-56.
- [5] Nouza J., Holada M., Hájek D.: An Educational and Experimental Workbench for Visual Processing of Speech Data. Proc. of EUROSPEECH'97, Rhodes, Greece, September 1997, pp.661-664.
- [6] Drygajlo A., Molina D.G.: JavaSpeakerRecognition - Interactive Workbench for Visualizing Speaker Recognition Concepts on the WWW. Proc. of Eurospeech2001. Aalborg, Sept.2001, pp. 2787-2790
- [7] Sutton S. et al.: *Universal Speech Tools: The CSLU Toolkit*. Proc. of ICSLP'98, Sydney, 1998, pp.3221-3224.
- [8] Hájek D., Nouza J.: MARKOV - System for Graphic Presentation and Investigation of Continuous Hidden Markov Models Used in Speech Processing. Proc. of 5th Czech-German Workshop "Speech Processing", Prague, September 1995, pp.49-51.
- [9] SpeechLab WWW pages at URL: <http://itakura.kes.vslib.cz/kes/indexe.html>
VISPER page: <http://itakura.kes.vslib.cz/kes/visperi.html>