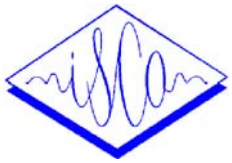

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# CURRICULA AND COURSEWARE IN SPOKEN LANGUAGE ENGINEERING IN EUROPE: A CRITICAL APPRAISAL

*Michael F. McTear*
University of Ulster
(+on behalf of the Spoken Language Engineering Working Group
of the Socrates Thematic Network on Speech Communication Science)
mf.mctear@ulst.ac.uk

## ABSTRACT

This paper summarises work by the Spoken Language Engineering (SLE) Working Group of the Socrates Thematic Network in Speech Communication Sciences. The thematic network has shown that computer-based teaching aids are vital to the future development of SLE education. This follows from the multidisciplinary and technical nature of SLE, which requires novel ways of presenting unfamiliar material. The paper analyses software resources available in relation to curricular requirements and educational criteria and makes recommendations for modules in an SLE curriculum. In addition, we identify areas for which high-quality courseware is, to our knowledge, unavailable and identify actions to fill these remaining gaps.

10.21437/Eurospeech.1999-163

## 1. INTRODUCTION

The Spoken Language Engineering (SLE) Working Group of the Socrates Thematic Network in Speech Communication Sciences, now in its third funding year, has surveyed SLE course provision in Europe [1] and has made proposals for SLE curriculum development at both undergraduate and postgraduate levels [2]. The thematic network has shown that computer-based teaching aids (on-line tutorials, demonstration packages and so on) are vital to the future development of SLE education. This follows from the multidisciplinary and technical nature of SLE, which requires novel ways of presenting unfamiliar material. In recent years, such software has begun to appear, partly as a result of initiatives taken within the network, within associated projects and independently. The increasing interest in SLE courseware was demonstrated at the recent MATISSE workshop, from which much of the following review material is taken [3]. This paper is a short version of a review of these materials which will be available in book form at Eurospeech [4].

## 2. MODULES FOR SPOKEN LANGUAGE ENGINEERING

### 2.1 Phonetics and Speech Production

Most speech technology courses start with an introduction to time and frequency domain representations, and relate these to basic phonetics. All this is natural material for multi-media presentation and a number of tools are available, varying from simple editors and annotation tools to large-scale packages which cover complete areas in phonetics.

A particularly useful feature is vocal tract animation coupled to speech synthesis. Figure 1 shows BALDI, an animated conversational agent provided with the CSLU toolkit. BALDI presents visual speech through facial animation synchronised with synthesised or recorded speech [5].

*Figure 1. BALDI with visible articulators*

One of the best tutorials in this area is the Sensimetric CDROM on speech production and perception [6]. This CDROM contains units on spectrograms, vowel and consonant acoustics, speech and vowel perception. Its interactivity provides considerable pedagogical advantage when compared to a printed textbook. For example, in the speech acoustic unit self- and teacher guided instruction promotes experience-based interactive learning that enables students to gain intuitive understanding of relationships between the place and manner of articulation with time and frequency characterization of the signal. Other interactive demonstrations include adjustable filtering of synthetic sources, demonstrations of the vowel spaces of adult and child speakers, identification and discrimination experiments with various stimuli, creating and analyzing conventional or 3D spectrograms, and examining animated vocal tracts synchronized with audio playback and spectrogram displays.

Most Web based courses are not interactive and multimedial, as the web medium is too limited in capacity for attractive animations involving video and sound. Some courses have short sound or speech samples, a good example being a Web based course in German on acoustic phonetics [7].

Speech input is a serious problem because standard HTML software including Java has not had this facility as yet. A tutorial developed at University of Cottbus, Germany and implemented using Java extension shareware explains human speech production on the basis of an LPC speech model [8]. The user can manipulate his (or a stored) voice by changing the pitch sequence and the number of prediction coefficients.

## 2.2. Perception

Sheffield's MAD (Matlab Auditory Demonstrations) software [9] is a growing resource for interactive student learning in speech and hearing. A number of these demos address topics in hearing and speech perception, for instance basilar membrane motion, sine-wave speech and auditory scene analysis. Strong features of this software are its uniform look-and-feel and documentation style. One theme is to make psycho-acoustic experiments 'come alive' for the student. For instance, in the demo of two-tone streaming illustrated in Figure 2, the student clicks on any point in the plane and hears the appropriate stimulus. When she has populated the space, the classical results can be revealed for comparison.

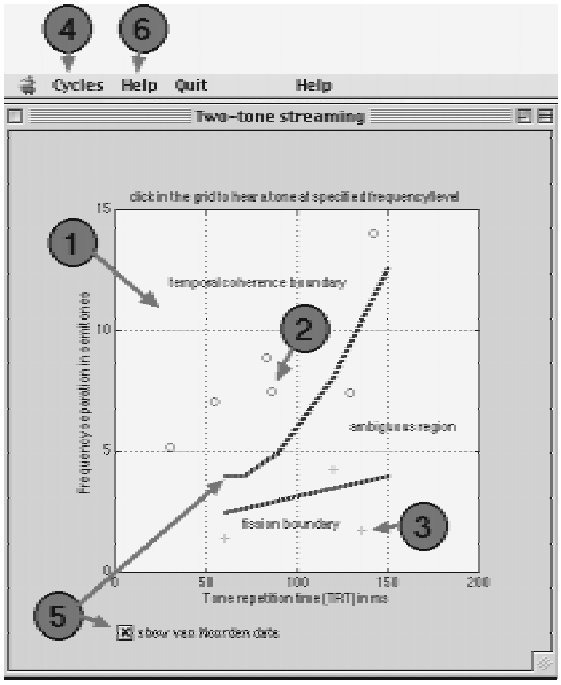


*Figure 2. MAD demonstration of two tone streaming*

*Documentation : Launch the demonstration with the command 'streamer'. Clicking anywhere in region 1 results in the delivery of a stimulus with the tone repetition time and high-low frequency separation specified by the position in the grid. The number of cycles delivered is governed by a menu option (4). Ten or more cycles may be required to hear one or other organisation. After some practice at hearing the two rhythmic possibilities, you may wish to record your responses. After each stimulus presentation, pressing 's' or 'f' will result in a red circle (2) or green cross (3) appearing at the click location. After a number of such identifications, compare your responses with those summarising van Noorden's subjects by checking the checkbox (5).*

## 2.3 Speech Analysis

The key to a successful speech analysis training is to create tools which open up the field to interactive investigation by the student. These tools should allow the student to interact with parameters in order to acquire practical skills in listening, analysis and interpretation performance as well as create new algorithms. The aim is to provide dedicated educational software (executables and sources) instead of exercises based on commonly used, only executable research tools such as Waves+ (Entropic) or MultiSpeech (Kay Elemetrics). Another goal is to create platform-independent, interactive, on-line laboratories accessible on the Internet to increase the efficiency of student self-study in the speech analysis domain.

Recently, the MATLAB programming environment has been employed by the Speech Science community to develop some educational aids for teaching and learning. Examples are the MAD programme from University of Sheffield (see Figure 2) and the set of exercises from Czech Technical University in Prague [9, 10]. MATLAB provides facilities for numerical computation, user interface creation and data visualisation. Its cheap academic edition allows students to use the demonstration tools at home. Unfortunately, MATLAB is neither completely portable, nor can it run within a Web browser.

The nearest contender to MATLAB appears to be Java. The ability of Java applets to be used within WWW pages certainly offers advantages for platform-independent distance learning. Purely Java-based speech analysis software was developed at EPFL Lausanne [11]. A second example is the Snack speech analysis module from KTH Stockholm that uses Java applets, Tcl/Tk language and C/C++ [12]. Unfortunately, Java is in its infancy and educators working with it must create many of their own classes to handle tasks for which they would be able to obtain ready-to-use libraries in other languages.

The educational speech analysis software packages available today, written in MATLAB, Java or in other languages (e.g. KHOROS [13]), are far from complete and their development can be seen as an ongoing concern to produce tools for the teaching and learning of speech analysis concepts.

## 2.4 Coding

Speech coding is an area of speech technology directly connected with telecommunications applications. Some electrical engineering background is involved,

particularly in digital signal processing, acoustics, computer programming, and information theory.

Most courses deal with the presentation and design of speech coders for different telecommunications standards. Waveform coding (time and frequency domains), parametric coding and hybrid coding (based on analysis by synthesis ideas) are discussed both as theoretical principles and applications. Since mobile communications are rapidly increasing, speech compression and encoding in GSM is receiving special attention. Also vector quantization and wavelet transform are required for low bit rate transmissions.

In order to illustrate how speech compression algorithms work and how they influence the quality of reconstructed speech, several educational courseware packages have been developed, although many of the resources on the Web were not developed with educational purposes in mind. For training and computer aided learning, speech compression software is available for: ADPCM and IMA-ADPCM encoding [14], LPC encoding [8], [9], [15], regular pulse excitation with long term prediction, vector quantization demo [14], exercices on speech coding [10]. Most of these focus on understanding the compression schemes in a graphical and interactive manner and also on evaluating the signal to noise ratio in different channel error conditions.

## 2.5 Speech Synthesis

Concerning web tutorials, the area of speech synthesis is not sufficiently covered. One can find some elements of it, e.g. LPC or formant synthesis, in connection with tutorials on more general aspects of speech processing (e.g. [7]).

One example, presented recently, interactively demonstrates the different processing steps of a text-to-speech (TTS) system [16]. The tutorial is arranged around a viewgraph of a TTS system. The user can type any text as input. By clicking on the different building blocks, the user receives information about the results of the different processing steps. Finally, the synthesised signal is given out. A special section of the tutorial is devoted to the crucial problem of the correct segmentation of the speech elements used for the concatenative synthesis. The user may select his own diphone segments from a given speech data base. The quality of the segments can be evaluated acoustically.

As additional material for teaching speech synthesis there are numerous examples for synthesisers throughout the web, such as the Bell-Labs system [17]. However these synthesiser demos are not primarily designed for pedagogical purposes, their main function being to demonstrate the quality of the synthesised speech.

## 2.6 Speech Recognition

We concentrate here on courseware which aims to help in the explanation of recognition methods and algorithms, rather than how this technology is deployed in ASR systems. We take as an exemplar of such software the VISPER (Visual Speech Processing) package developed at the Technical University of Liberec, Czech Republic [18].

Recognition by Dynamic Time Warping is readily illustrated by visualising the time-time plane and the best path to traverse it. Many packages include such a facility. In VISPER, the display is in 3D, and students can explore a range of variants on the DTW algorithm. The associated exercises expose the limitations of DTW and lead students naturally to the conclusion that ASR requires statistical models rather than template matching.

Statistical HMM-based recognition presents a greater challenge for the courseware developer. It is important to provide an alternative way in which this 'doubly-statistical process' can be explained in addition to a mathematical account. VISPER attempts to 'unhide the hidden Markov models in their continuous density version'. The display is based on 3D projections of state-output density functions (Figure 3). One can observe how the state distributions are aligned during decoding and how they evolve during training.

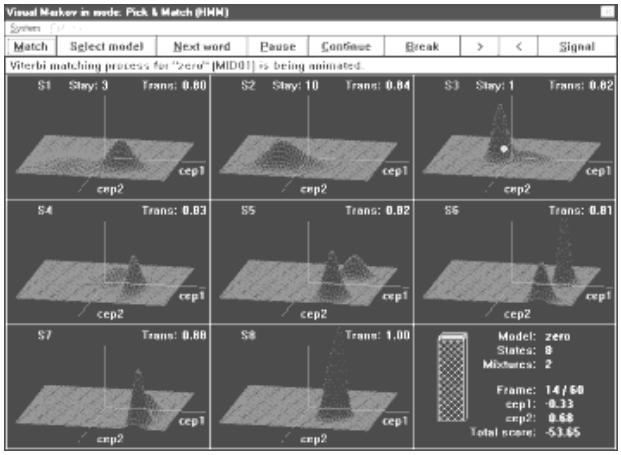


*Figure 3 Visual Markov. This module visualises procedures associated with training and testing continuous Hidden Markov Models.*

However, VISPER's visualisation of HMM recognition portrays the result of training and recognition, rather than how the algorithms work. Viterbi decoding might be illustrated in a similar way to the DTW time-time plane: it is essentially the same algorithm. It is more difficult to see how to demonstrate the Baum-Welch training algorithm, yet this is central to an understanding of modern speech technology.

Another gap in available resources is something to illustrate how this acoustic modelling is linked to

language modelling through what is sometimes called the 'fundamental equation of ASR'. This could perhaps be added to an illustration of the Viterbi algorithm: how do the recognition paths evolve with and without a language model?

More advanced ASR techniques for robust recognition - speaker adaptation, resistance to noise, environmental modelling and so on - also provide challenging topics for future courseware.

## 2.7 Spoken Dialogue Systems

Spoken dialogue systems involve the integration of all the components of spoken language technology, such as speech recognition, natural language processing, and speech synthesis, as well as a facility for modelling the dialogue flow. Various tools have been developed, some of which are available under commercial licences, such as SpeechMania™ from Philips Speech Processing and the Developers Toolkit from Nuance Communications. The Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute of Science and Technology provides the RAD toolkit free-of-charge for non-commercial usage [5]. The toolkit includes core technologies for speech recognition and text-to-speech synthesis, as well as a graphically-based authoring environment (RAD) for designing and implementing spoken dialogue systems. Building a dialogue system with RAD involves simply selecting and linking graphical dialogue objects into a finite-state dialogue model, and specifying the recognition vocabulary and actions to be taken at each state. The toolkit provides easy-to-use facilities to support courses in spoken dialogue modelling, with the main disadvantage being the potential inflexibility of the dialogues due to the finite state dialogue model provided [19], [20].

# 3. CONCLUSIONS

In this paper we have shown that a number of high-quality products are now available and that these are making a considerable impact on the teaching of SLE. Some products are available on the web, others on CDROMS. Inevitably, there are gaps in the coverage, particularly where it is more difficult to think of a way of visualising an algorithm.

# REFERENCES


[1] Green, P., C. Espain et al. (1997), Education in Spoken Language Engineering in Europe. *Proceedings of EUROSPEECH 97*, pp. 1935-1938.

[2] Espain, C. et al (1998), Spoken Language Engineering. In G. Bloothooft et al. (eds.) *The Landscape of Future Education in Speech Communication Sciences 2: Proposals*, Krips Repro, Meppel, NL.

[3] Hazan, V. and M. Holland (eds.) (1999), *Proceedings of the ESCA/SOCRATES Tutorial and Research Workshop on Method and Tool Innovations for Speech Science Education (MATISSE)*, University College London, 16-17 April 1999.

[4] K.Fellbaum et al. (1999), Spoken Language Engineering. In G. Bloothooft et al. (eds.) *The Landscape of Future Education in Speech Communication Sciences 3: Recommendations*, Krips Repro, Meppel, NL.

[5] http://cslu.cse.ogi.edu/toolkit

[6] Berkovitz, R. (1999), Design, development, and evaluation of computer-assisted learning for Speech Science education. In Hazan and Holland (eds.) pp. 9-16.

[7] http://www.phonetik.uni-muenchen.de/AP/APHome.html

[8] Fellbaum, K. and J. Richter (1999), Human Speech Production Based on Linear Predictive Vocoder - An Interactive Tutorial. In Hazan and Holland (eds.), pp. 57-60.

[9] Wrigley, S., M. Cooke and G.J. Brown (1999), Interactive Learning in Speech and Hearing. In Hazan and Holland (eds.), pp. 21-24, http://www.dcs.shef.ac.uk/~martin

[10] Uhlir, J. (1999), The Set of Exercises in Digital Speech Processing. In Hazan and Holland (eds.), pp. 53-56.

[11] Drygajlo, A. and G. Delafontaine (1999), Using Java to Develop Interactive Learning Work-Bench for Speech Analysis Basics on the World-Wide Web. In Hazan and Holland (eds.), pp. 25-28, http://scgwww.epfl.ch/JavaSpeech Lab

[12] Sjölander, K. et al. (1999), Web-based Educational Tools for Speech Technology. In Hazan and Holland (eds.), pp. 141-144, http://www.speech.kth.se/snack

[13] Kacic, Z. (1999), Laboratory Course on Speech Processing Using KHOROS Development Environment. In Hazan and Holland (eds.), pp. 117-120.

[14] Giurgiu, M. (1999), Teaching Digital Speech Processing for Telecommunications. In Hazan and Holland (eds.), pp.109-112.

[15] http://www.arl.wustl.edu/~jaf/lpc

[16] Hoffmann, R., B. Ketzmerick, U. Kordon and S. Kürbis (1999), An interactive tutorial on text-to-speech synthesis from diphones in time domain. *Proceedings of EUROSPEECH 99*, Budapest.

[17] http://www.bell-labs.com/project/tts

[18] Nouza, J. (1999), Teaching and Learning through Visualised Speech Processing Experiments. In Hazan and Holland (eds.), pp. 121-124.

[19] McTear, M. (1998), Modelling spoken dialogues with state transition diagrams. Procedings 5th International Conference on Spoken Language Processing, Dec. 1998, Sydney, Australia, pp. 1223-1226.

[20] McTear, M. (1999), Using the CSLU toolkit for practicals in spoken dialogue technology. In Hazan and Holland (eds.), pp. 117-120.



+ **Authors**: *Marian Boldea*, Politehnica University of Timisoara, *Andrzej Drygajlo*, EPFL Lausanne, *Klaus Fellbaum*, BTU Cottbus, *Mircea Giurgiu*, Technical University of Cluj-Napoca, *Phil Green*, University of Sheffield, *Ruediger Hoffman*, TU Dresden, *Michael McTear*, University of Ulster, *Bojan Petek*, University of Ljubljana, *Jan Uhlir*, Czech Technical University in Prague.