

An Interactive Curriculum in Human Language Technology for Undergraduate and Graduate Education Research

1. Summary: Year One Goals and Accomplishments

The objectives of the proposed work, as stated in the CRCRD proposal, are:

- To develop a comprehensive, multidisciplinary curriculum in HLT incorporating the latest research advances in the field that prepares advanced undergraduate and graduate students for careers in industry and academia;
- To stimulate and sustain interest, continued study, and careers in areas of HLT, with special attention to recruiting more women, by providing an exciting and enriching learning experience through hands-on experience developing and evaluating language technologies and systems; and
- To facilitate transfer to and implementation of the curriculum at other institutions and to widely disseminate project results.

The set of courses developed and taught in year one are intended to form a multidisciplinary program in human language technology of laboratory courses in Computer Science, Electrical Engineering, Linguistics, Psychology, Speech, Hearing and Language Sciences, and Communication. In each of these courses, laboratory modules will be developed to provide students with hands-on experience using state-of-the-art tools and technologies in signal processing, speech recognition, speaker identification, natural language understanding, speech synthesis, facial animation, and dialogue modeling. In addition, these technologies will be investigated in the context of working systems designed by the students. Taken together, these courses provide the theoretical and foundation for students to design conversational systems incorporating animated agents.

Excellent progress was made toward achieving the objective of creating a multidisciplinary curriculum in human language technology. Six courses were developed and taught during the first year of the award, in five different universities (University of Colorado, Carnegie Mellon, Oregon Graduate Institute, Stanford, and UC Santa Cruz). In all courses, modules within the CSLU Toolkit (a comprehensive set of tools and technologies for researching and developing language technologies and systems) were used to provide students with hands-on experience using and developing speech and technologies (e.g., speech recognition, speech synthesis); designing and running experiments, and creating useful spoken dialogue systems.

Progress was made in integrating the courses into a multidisciplinary curriculum. As described in the next section, the University of Colorado approved a certificate program in Human Language Technology with participation by four departments. Two of these departments contain a high proportion of female students; **their** exposure to HLT has created an awareness of possible career options among these women. We will continue to develop strategies to attract and retain women into our certificate program by applying an

understanding of what appeals to and what repels women from science and technical careers based on ongoing research at Carnegie Mellon University as well as new research at the University of Colorado.

Finally, due to the efforts of Cliff Nass at Stanford, over eighty representatives of thirty different companies participated in a workshop in which students who took a course on the design of speech interfaces, supported by the CRCD grant, presented the results of their projects.

2. Creating an HLT Program at University of Colorado, Boulder

In order to facilitate creation of a multidisciplinary curriculum in Human Language Technology, faculty in four departments worked together to design a multidisciplinary certificate program in human language technology. We then submitted a proposal to the University of Colorado to establish an Interdisciplinary Certificate Program in Human Language Technology. The proposal was approved.

The curricular plan includes 5 core courses, consisting of a computer programming foundation plus four courses in various aspects of Human Language Technologies. Students are required to major in one of the four Human Language Technology disciplines (Computer Science, Electrical and Computer Engineering, Linguistics, Speech, Language and Hearing Sciences). The program is available to Masters or Ph.D. students, including students in the 5-year bachelors/masters program. The home page for the certificate program can be found at <http://www.Colorado.EDU/ling/jurafsky/curr.html>. The proposal submitted to the University of Colorado can be found at <http://www.Colorado.EDU/ling/jurafsky/curr.pdf>.

3. Courses Developed and Taught

SLHS-2010 Science of Human Communication, John Hansen, UC Boulder

Course Description:

This course represents an introduction to the fundamentals of human speech communication, introducing aspects of physical speech production, auditory processing and psychoacoustics, and language. This interdisciplinary course covers how human thought is transmitted from the brain of a speaker to the brain of a listener. The complex interaction of acoustics, speech physiology, anatomy, neurobiology, hearing, and psychology all play integral role in human communication. The course covers the fundamentals of the physics and biology of spoken language as well as how the communication process is observed and measured.

The course begins with a discussion of fundamental issues in speech and language, including basic introductory concepts in linguistics, acoustics, and the physics of sound. Next, the course shifts to physiology of human speech production. Concepts on the articulatory movement for individual phoneme production are considered. This is followed by and

discussion of the physiology and function of the human auditory system. Finally, a brief discussion of the "brain" processing needed to both form and perceive speech is considered. A number of examples of psychoacoustic experiments in sound and speech perception are considered.

Finally, the last portion of the course considers how technology plays a role in the speech chain, including voice communication systems, assistive technology for individuals with communication disabilities, and emerging commercial speech technologies for human-machine interface.

This represents an Arts and Sciences Natural Science Core course, and therefore fulfills part of the natural science core requirement of the College of Arts and Sciences. The course is taught once per term with 55-65 students enrolled. While this is an undergraduate course, I am in the process of establishing a graduate equivalent to coincide (SLHS-5010 Science of Human Communication). SLHS-5010 will meet at the same time and location as SLHS-2010, but will have a course project requirement and more extensive home work assignments as well as separate exam grading requirements. SLHS-5010 will be a required entry-level course for the HLT Certificate.

Progress:

One of my primary goals was to establish a web page and a series of interactive lectures for this course. The following web location includes (i) a course summary; (ii) class schedule with a complete set of slides for each lecture; (iii) copies of all homework assignments as well as solutions; (iv) a summary of web links of class interest; and (v) [most importantly] a class Feedback and Discussion page.

<http://cslr.colorado.edu/classes/SLHS2010/>

During the fall term, I developed a complete software based course version of SLHS-2010, which consisted of power-point presentations for each lecture, and interactive demonstrations (including audio and video examples). All lectures, homework sets, solutions, handouts were integrated into this student accessible web page that includes an anonymous feedback page for students to post comments or questions regarding lectures. Because the University Administration moved so slowly to approve financial resources for renovations for the computer lab I was establishing in SLHS, it was not possible to include computer-based labs in SLHS-2010. However, I was able to provide interactive audio demonstrations for 2/3's of the lectures. Also, I spent between 7-9 hours preparing each lecture, which resulted in a fully interactive web based course which provided students unique access to lecture and course materials. Teaching this course proved to be a challenge, since all the courses I have taught over the past 18 years have been for electrical engineering students. SLHS-2010 is a core course with undergraduates from all areas of A&S. As such, not all students were prepared to handle some of the scientific aspects considered in the course. Additional problem sessions and optional homework sets that I ran helped many students. However, in the future, I believe it will be beneficial to the students to convey early in the term that this is a core

"science" course, and to identify weaknesses in student backgrounds. I ran help sessions early in the term for those students which had no experience using computers.

The technology used included MS PowerPoint presentations. Integrated audio examples of speech production, auditory processing, and psychoacoustic experiments on perception of sound. I also included demonstrations of the CSLU Toolkit including BaldiSync, to show speech articulation. Finally, four lectures were dedicated to speech technology based on the CU Communicator System that was developed at CSLR. This included examples of speech recognition, text-to-speech synthesis, and natural language generation via a speech based telephone dialup airline travel system that was developed for DARPA. One homework assignment required all students to complete several interactive travel scenarios, and then to respond to a formal set of evaluation questions for user acceptance of the speech technology.

Limitations/Plans for Next Year:

Clearly, the largest shortcoming of the class was the lack of computer labs for students to learn about speech, hearing, and language science. While several simple experiments could have been assigned, the computer laboratory was not operational during the fall 1999 term [the CU administration required 7 months of discussion before approving the funds to renovate lab space while the computers sat in their boxes]. I am happy to say that the lab will be functional when this course is taught this fall. Presently, we are working on a series of computer exercises for students in this class, which use the CLSU Toolkit, as well as other speech based analysis tools. The Feedback and Discussion Page will also be extended to provide more directly evaluation of the use of technology for students who are typically non-engineering/computer science majors for HLT.

SLHS-5674 Signals & Systems in Speech & Hearing Sciences, John Hansen, UC Boulder

Course Description:

This course provides an in-depth study of signals and system concepts for use by audiologists and speech science students in signal generation and modification, signal measurement, analog and digital signal representation, filtering, amplification. Presents an overview of the functional application of instrumentation used in the field of speech and hearing science.

Progress:

The goals here were to establish a lecture series which include interactive speech and hearing analysis. I established the following web page which includes a course description and slides for all lectures.

<http://cslr.colorado.edu/classes/SLHS5674/>

Specifically, the following page summarizes concepts/goals presented in each lecture for the term.

<http://cslr.colorado.edu/classes/SLHS5674/sched2.html>

This course focuses more on the signal processing of speech for students interested in understanding frequency structure, filtering concepts, and systems for hearing impaired (hearing assist devices, digital hearing aids, etc.). I included a number of examples of speech analysis using computer-based tools. Students were also exposed to basic concepts in system transfer characteristics, time characterization of systems, digital signal processing in both time vs. frequency domains, speech spectrograms, and applications to hearing aids and interactive systems. One lecture was devoted to construction of a dialog system using the CSLU Toolkit (RAD: rapid application developer), where the class participated in a hands-on overview given by a CSLR staff person. This exercise was considered in order to expose SLHS MA graduate students to the speech technology tools available in formulating interactive systems that could be used to assess speech and hearing motor function.

Limitations/Plans for Next Year:

Again, the students in this class were able to get hands-on lab experience. However, this was only possible by holding class in the CSLR computer classroom on CU East Campus. The computer classroom being established in SLHS was not available, so follow-up exercises could not be considered. This summer, I have been supervising a Electrical Engineering PhD. student to develop several laboratory exercises which will be used for this class. Further work is also needed for the web page for this class [the class was much smaller than SLHS-2010, so students generally did not depend on web information to the same level]. For example, while students regularly posted comments on the feedback page in SLHS-2010, students were not likely to use this route to give feedback. This clearly was due to the differences in class size (7 vs. 60 students in the two classes). In the future, we will consider alternative ways to obtain better and more real-time feedback of lectures and lab exercises.

ECEN-5022 Speech Processing & Recognition; John Hansen, UC Boulder

Course Description:

Speech Processing and Recognition is an interdisciplinary course that provides an introduction to the analysis, modeling, and recognition of speech signals. Topics covered include: fundamentals of speech science, acoustic-phonetics, discrete-time linear predictive (LPC) models, short-term analysis of speech signals, voice coding methods (CELP, GSM), speech recognition principles (hidden Markov models), and special topics in human-computer interfaces. A speech processing term project is required for this course.

Progress:

The goals here were to establish interactive lectures for students in electrical engineering who have interests in speech processing and algorithm design. I established the following web page that includes a course description and slides for all lectures.

<http://cslr.colorado.edu/classes/ECEN5022/>

The following page summarizes concepts/goals presented in each lecture for the term: <http://cslr.colorado.edu/classes/ECEN5022/sched2.html>

Again, significant effort was spent in establishing PowerPoint slides for each lecture, and to construct classroom discussion to encourage student participation regarding speech-processing concepts. This portion was a major success. Many lectures ran 20-30 minutes longer than regular class time because students were so engaged in lecture concepts presented. Many integrated audio examples were included during the term lectures. Also, examples of interactive speech systems were demonstrated using the CSLU Toolkit and the CU Communicator system.

Since an individual class project is required of all students (I have taught this course now for 12 years, and require a class project in speech processing/recognition), computing support was needed. Since the INTEL computer lab I was setting up in the Dept. of SLHS was not available, we decided to use a Sun computer lab in the Dept. of Computer Science. I made this decision, because of a previous software license (Entropic Xwaves/HTK toolkit) which purchased for classroom use on Sun Unix machines. The decision to use this Computer Science Sun workstation lab turned out to be a major problem as the term progressed. Since CS maintains their own computing infrastructure that is isolated from the University of Colorado Boulder computing network, it was impossible to have reasonable collaboration with students. The firewall security CS enforces meant that students could not log in from a remote machine, and only after countless days of effort on my part, were we able to get the Entropic Xwaves/HTK toolkit software running for the last two weeks of class. In spite of this hardship, students were able to make great progress on their class projects while license issues were being worked out for this analysis software. I organized several computer laboratories that used this software to illustrate how to build a speech recognition system. Students then had homework exercises that required that they modify and run the recognition system on new data. This provided necessary computer experience for class term projects.

The biggest success was clearly the student class projects. As a rule, I have 1-2 exams in this class with no final exam. Instead, students are required to write-up a 4-page conference style paper on their term project and present their work at the end of the term in place of a final exam. A summary of the class projects (with abstracts) is included at the following class web site:

http://cslr.colorado.edu/classes/ECEN5022/Spg00_projs.htm

Again, in order to assess student presentations, I had all students in class write up comments on their fellow student presentations (these evaluations were collected separately, and passed back to me only after grades were submitted; so students were free to respond without impacting the grades of fellow students. The class projects, online lectures and interactive speech demonstrations in class were areas that students identified as being a real strength in the course.

Limitations/Plans for Next Year:

The area where students were most frustrated with was computer support for lab exercises and the term project. Some students were more PC oriented, while others prefer to use Unix workstations. In the future, we plan to shift to analysis tools which will run on Window based

PCs, as well as Linux running on PCs. In this manner, we can avoid many of the cross-platform issues that came up during the term.

A better mechanism was needed to set up and support lab exercises for students in the class. Some students had limited background in real-time signal processing of speech signals (about a third of the students were more use to traditional textbook, math oriented, lectures with no computer lab use at all). In the future, better support to level the background before class labs will be initiated early in the term.

Text-to-speech Synthesis, Alan W. Black, Carnegie Mellon University

Six male and four female first year graduate students at CMU majoring in language technologies, computer science or robotics attended this graduate course. Several of these students are intended to pursue careers in speech and language technology, and a few will continue in speech synthesis research itself.

The course was designed to cover all aspects of speech synthesis from both a theoretical and practical point of view. Students were given the opportunity to learn about new research in the areas of text processing, prosodic modeling and waveform synthesis as well practical experience in using existing synthesis technologies. The course consisted of the following sub-parts:

- History and general use of speech synthesis;
- Text analysis: text conditioning, markup languages, homograph disambiguation
- Lexicons and letter to sound rules;
- Prosodic modeling: phrasing, duration and intonation;
- Waveform synthesis: diphones, and unit selection;
- Building new voices in new languages, and
- Limited domain synthesis for practical applications.

The course was based heavily around the Festival Speech Synthesis System. Although the CSLU toolkit itself was not used, Festival is an integral part of the toolkit, so the course can be taught using the toolkit under Windows. Moreover, all voices, techniques, models, etc developed within this course can be used directly in toolkit applications.

As Festival offers an environment for building new synthetic voices as well as an end user delivery vehicle for black box text-to-speech, it offers an ideal platform for teaching students what can be done with today's speech output technologies. Each week simple exercises were assigned involving different aspects of the system so the students could learn from practical experience how the technology worked. The system is designed such that no low-level C++ programming was required, thus opening the course to a much wider audience. In all cases, existing simple rules and functions used in Festival were presented to students for modification using the Scheme scripting language; this enabled students to learn without having to delve too deeply into the complexities of the system.

In addition to synthesis techniques, the students were led into the field of building new synthetic voices in new and currently supported languages based on the released documentation and scripts that are part of the CMU FestVox project (<http://festvox.org>). These scripts and tools sit on top of Festival (and the Edinburgh Speech Tools) and offer a complete environment for developing new synthetic voices.

In addition to the weekly exercises, a larger project was set towards the end of the course. Quite ambitious projects were attempted (two of which have led to publications). These included: cross language limited domain synthesis (a talking clock in Chinese), Thai letter to sound rules, a talking Eliza program, and complete new female US English voices.

The complete course notes and slides have been made available at <http://festvox.org/festtut/> along with the course schedule. This site will also be updated with some of the example projects that were done and model answers to the exercises (some are already on the general section of the FestVox site). We are continuing to update these notes and there will be new releases making it easier for both individual students to follow the course notes and institutions to use these notes to teach their own course.

Psychology 218 Speech Perception, Dom Massaro, University of California, Santa Cruz

This graduate seminar offered in the spring quarter by the Psychology department at UCSC covered all aspects of human speech perception, focusing on topics of primary interest to the students taking the seminar. The goal of the course was to enable students to design their own experiments to investigate aspects of speech perception and production. The plan was to have students pursue a specific focus throughout the quarter to develop expertise in a particular domain.

Students designed their experiments using modules of the CSLU toolkit, developed a prototype demonstration or experiment and made a class presentation about their project. Topics included speech production, articulatory characteristics of speech, acoustic characteristics of speech, speech perception, speech synthesis, development of speech production and perception, speech perception by nonhumans (machines and other animals), speech in human/machine interfaces, speech translation, and substitutes for speech.

Under the guidance of a teaching assistant and other skilled toolkit users, students had the opportunity to learn to use the CSLU Toolkit to design experiments (see <http://mambo.ucsc.edu/psl/pslfan.html> and <http://cslu.cse.ogi.edu>). Tutorials on various aspects of the toolkit are given at <http://mambo.ucsc.edu/psl/tools/tutorial.html> and <http://cslu.cse.ogi.edu/toolkit/docs/2.0/apps/rad/tutorials/index.html>. The user-friendly nature of the toolkit modules makes speech science accessible and engaging to students while simultaneously allowing them to master quite difficult material.

The Perceptual Science Laboratory (PSL) modules of the CSLU Toolkit allow the user to control and manipulate auditory speech and the speech movements of an animated face. The Rapid Application Developer (RAD) enables rapid prototyping and development of spoken language dialogs between an animated talking head and the language user. Students used

these modules in a hands-on computer lab to design, implement, carry out, and analyze experiments in face-to-face language processing.

The course was in general successful and met most of the expectations of the instructor and the students. Although the existing tutorials on using RAD and PSL were very helpful, a large amount of instructor and teaching assistant help were essential to a successful experience. One reason was that each project usually required several extensions of the current capabilities of the toolkit, which required new programming in Tcl or C. For future classes, it would be helpful to organize and catalog all embellishments and extensions of the toolkit so that students can build on these for their new applications and therefore lessen the amount of new programming that is required. Finally, it would be worthwhile to attract students with different types of expertise (linguistics, psychology, computer science) to work together in groups and therefore enhance individual accomplishments.

Each student produced a well-motivated experiment. Although actual experiments were not conducted due to time constraints, the projects produced several positive outcomes. Students studied the scientific literature, learned key issues that could be addressed through research, and designed experiments to test hypotheses. Descriptions of the projects can be found at <http://mambo.ucsc.edu/psl/pslint/psldocs/psych218> (login baldie, password pingvin).

Projects:

Audible Cues to Sarcasm

This project involved the audible cues to sarcasm. Based on previous literature, Greg hypothesized that rate of speaking and the pitch range are cues to sarcastic utterances. He chose the standard Festival text-to-speech system parameters as “non-sarcastic,” and created additional utterances with successively faster rates of speaking and wider pitch ranges. He programmed the toolkit to vary these two variables independently of one another, in a 5 by 5 factorial design, with participants making a rating on a scale from one to seven. His tutorial was an informative description of his study. He introduced the existing literature and generated a testable hypothesis based on this research. It would be valuable to carry out the experiment, which would advance the field in this area and provide a new test of integration models in language processing.

Audible and visible cues in Co-articulation

Qiang’s project investigated perception of auditory and visual cues to perception of vowels in consonant vowel syllables. Specifically, he examined the effects of audible and visible cues in the consonant to perception of the following vowel. He programmed a speech identification experiment in which he varied the audible frication information and the visible speech in a computer-animated talking head independently of one another, in a 5 by 5 expanded factorial design. Participants identified the vowel when given only the frication portion of the consonant-vowel syllable. His tutorial was an informative description of his study. He provided a valuable description of the articulatory and acoustic properties of fricatives in the context of rounded and unrounded vowels. The pilot results established some parameters for a systematic study, and it would be valuable to carry out the experiment, which would advance the field in this area and provide a new test of integration models and perceptual units in speech perception.

Spoken language influences on written word recognition

Alexandra's project addressed whether spoken language would influence written word recognition. She programmed the toolkit to vary a priming stimulus presented just before the onset of two written words. The prime was either a tone or a word. The test display was the written word and a pseudoword that would be pronounced in the same way as the word. For example, an experimental trial might be the spoken word "true" followed by the written words "true" and "trew". The question was whether the prime would influence the accuracy of choosing the correctly spelled letter string. Her tutorial was an informative description of the study. She presented a thorough review of the current literature, which generated the impetus for the proposed experiment. She gave a detailed description of the additional code needed to carry out experiments on written word recognition in the toolkit platform.

Facial and vocal cues to memory

JoAnn's project involved the role of auditory information from the voice and the face in memory. She programmed the toolkit to present a continuous list of words that were spoken by a computer-animated talking head. The participant had to indicate whether the word was presented previously. Previous studies had shown that having the word presented in the same voice facilitated the memory relative to the word being presented in a different voice. JoAnn asked if having the face occur with a different voice would improve the memory for the word even though the voice had changed. JoAnn's tutorial was an informative description of his study. After a detailed description of the influences of talker variability on memory, she created an innovative test of the influence of visual information. She uncovered several limitations in the toolkit functions and worked to extend the toolkit to carry out memory experiments. An extension and modification of her proposed experiment would make a valuable contribution to the memory literature.

Situational Cues in Sentence Processing

Thomas's project involved the role of visible speech in situations when the situational context is highly related to the spoken utterance. Based on previous literature, he hypothesized that the visible speech would not contribute to performance when the situational context is highly related to the spoken utterance. He programmed the toolkit to vary the contextual situational information and the presence of visible speech independently of one another. His tutorial was an informative description of his study. He reviewed the literature of context effects in perception and their relevance to everyday processing, and generated an innovative hypothesis to provide a test of the fuzzy logical model of perception. The toolkit was used to present both pictures and spoken sentences in order to test the hypothesis. An extension and modification of her proposed experiment would make a valuable contribution to the sentence processing literature.

Multimodal speech perception by infants

Nicole's project addressed the question of how accurately infants can determine the relationship between auditory speech and facial articulatory movements. Previous literature had shown that at 4.5 months of age infants can determine which face is saying /i/ when the alternative face is saying /a/. She asked if infants would be able to distinguish more subtle distinctions such as Spanish /ba/ versus English /ba/ and the words error versus mirror. She

programmed the toolkit to present two talking heads on the screen and mouthing different segments of speech. One of these segments agreed with the audible speech. The test would involve whether the infant would look longer at the face that was articulating the sound being presented. Nicole was open to suggestions for improvements of the experiment. The toolkit was extended to present two Baldi's on the screen simultaneously mouthing two different segments. The auditory speech agreed with one of the articulations. This experimental setup offers a technological advancement in infant research, and a modification of her proposed experiment would make a valuable contribution to the speech development literature.

Audible and visible cues to emotion

Ravi's project involved the audible and visible cues to emotion. Based on previous literature, he hypothesized that both the face and the voice would contribute to the perception of emotion. He chose the standard Festival parameters as representing a neutral emotion, and based on descriptions in the literature created additional utterances with acoustic parameters that supposedly indicated fear, anger, and sadness. He used the built in parameters for showing these different emotions on the talking head. He programmed the toolkit to vary the information from the face and the voice independently of one another, in a 5 by 5 expanded factorial design. His tutorial was an informative description of his study. This is the type of experiment most easily carried out within the toolkit, and its analogies with studies of speech perception should provide valuable advances in processing multimodal inputs.

Computer Speech Recognition; John Paul Hosom, Oregon Graduate Institute

This course provided an introduction to Hidden Markov Models (HMMs) as applied to the task of speech recognition. HMMs are currently the most commonly used technique for automatic speech recognition. The course covered several variants of HMMs, including discrete, semi-continuous, continuous, and ANN-hybrid approaches, as well as the strengths and weaknesses of the HMM approach in comparison to other ASR techniques such as template matching and segment-based systems.

The course began with an introduction to speech, and illustrates why automatic speech recognition is a difficult problem, even at the acoustic level. An overview of theories of human speech recognition was given, as well as an overview of several ASR techniques. Then, Markov Models and Hidden Markov Models are introduced in a systematic way using several examples. The elements of an HMM were defined, and HMM structures commonly used in ASR were described. Next, the Viterbi search algorithm was presented in detail, with both a mathematical framework and a significant number of examples. The students used the CSLU Toolkit to construct their own HMM/ANN based recognizer, learning the process of partitioning a corpus into training, development, and test sections, generating features, training the acoustic models using ANNs, and evaluating the resulting system. (The tutorial for constructing this recognizer is on-line at http://cslu.cse.ogi.edu/tutordemos/nnet_training/tutorial.html). In addition, there was a class project to implement the Viterbi search in C code, and test it on some sample data.

Once an existing HMM, including structure, probability estimation, and search techniques, was understood, the class proceeded to details of training HMMs. The course covered flat-

start, forced-alignment, *k*-means segmentation, and forward-backward training. A significant amount of time was spent on the forward-backward technique, and another class project was assigned in which students learned to implement the algorithm in C code.

Having mastered the basics of HMM initialization and training, the course then covered implementation-related aspects of HMMs, such as state tying and cloning, pause models, and improved duration modeling. Features used in ASR, such as PLP and MFCC, were covered in some detail, as well as noise-removal techniques and a derivation of the LPC equations. Then, connected-word recognition techniques were introduced, including the 2-level, level-building, and one-pass algorithms. Finally, *N*-gram language modeling was covered. At this point, the students used the CSLU Toolkit to train their own HMM system for digit recognition using Gaussian mixture models, *k*-means segmentation and the forward-backward algorithm. (This tutorial is on-line at <http://cslu.cse.ogi.edu/tutordemos/csluhmm/doc/csluhmm.ps>).

A comparison was made of HMMs to other ASR techniques such as template matching, segment-based recognition (in the SUMMIT and EAR systems), and techniques that emphasize temporal change, such as TRAPS and diphone-based systems. ASR evaluation (including accuracy, correctness, and significance testing) and state-of-the-art performance on various tasks was presented. Finally, several applications of ASR systems were demonstrated using the CSLU Toolkit. Interspersed throughout the course were introductions to techniques such as vector quantization (VQ), dynamic time warping (DTW), data-driven clustering, and the A* search.

Summary:

This course provided a fairly detailed introduction to HMM technology. Concepts were made concrete to students through use of two tutorials in the CSLU Toolkit (for training HMM/ANN and standard HMM recognizers) and through class projects implement the Viterbi search and forward-backward algorithms. The lecture notes were all made using PowerPoint, which took significant effort but should make future versions of this course easier to prepare. An effort was made to encourage class discussion and participation, which sometimes took the class away from the main topic but provided a dynamic and relaxed learning environment. The students were, on the whole, quite advanced; several had already used the HTK software package to build speech recognizers, and others were knowledgeable about Markov models but had little background in speech. This made it challenging to keep the class together as a group, but made class discussion (and after-class discussion) interesting. One student from industry was, at the end of the course, planning to do further research on HMM training techniques and, if successful, submit a paper to a conference, all on his own time. The tutorials using the CSLU Toolkit went smoothly, overall, with the largest problems being in setting up the environment.

Plans for the Future:

For future courses, there are several things I would like to do differently. First, I plan to break the two programming projects into three or four projects, and change some of the homework to small programming assignments. Second, the lecture notes should be brushed up and put on the web. Third, I want to encourage the students to not only complete the

tutorials “as is”, but try different training parameters and features. I would also like to integrate the Toolkit into the programming assignments, but that may require a re-write and more modularization of the recognition component of the Toolkit.

Experiments in Speech User Interfaces, Cliff Nass, Stanford

Executive Summary

Experiments in Speech User Interfaces was a mixed undergraduate/graduate seminar (approximately 60% undergraduate) in which 35 students participated. In an eight-week period, small groups of students (3-4 students per group) were required to design, run, and analyze data from an original (the study had not been done previously), unique (each group pursued different research questions and experiment), practical (the study had to have direct implications for design and should have an externally-valid context), publication-quality experiment in speech user interfaces. Eight of the ten groups used the CSLU Toolkit. Nine of the ten groups produced publishable studies (Powerpoint presentations of the studies can be found at www.stanford.edu/~nass/comm369temp).

At the end of the eight-week period, I convened a one-day industry conference consisting of student presentations. 80 participants from 35 countries (including Europe and the East Coast) participated, including several of the major players in speech recognition, voice interfaces and voice portals. The conference was enormously successful in achieving industry transfer and industry involvement. For example, approximately half of the participants indicated that they would be delighted to be group mentors next year, contributing time and money to work on future student research projects. We are now developing a model for pursuing these relationships.

Structure of the Course

In the Winter of the 1999-2000 school year, I circulated an announcement of a new course in Speech User Interface to undergraduates, graduate students, and faculty in the departments of communication, computer science, industrial engineering/product design, psychology, sociology, and symbolic systems (cognitive science). I also circulated the announcement to students who had taken my lecture course on interface design (approximately 150 students).

Approximately 40 students applied to the course, writing statements of interest. I was sufficiently impressed with the quality and diversity of the applicants that I admitted all of the students. Approximately 35 students eventually took the course.

At the initial meeting of the course, I provided a list of fourteen research questions that the students could pursue. Each student then submitted a form providing information about their: major; gender, knowledge/experience with experimental design and social science methodology; knowledge/experience with computer programming; knowledge/experience with design; rank order of interest in each of the research questions. Examination of the responses suggested that the course yielded students drawn from majors across the university. All of these students were interested in the psychology of technology, although some were more oriented toward technology, others experimental research, and yet others design.

The most important purpose of the questionnaires was to assign students to the project groups. Research performed by Professor Larry Leifer (Product Design, Stanford U.) has

suggested that teams work best when they are: a) diverse on as many dimensions as possible, b) three or four students, and c) assigned rather than self-selected. Informed by this research and given the breadth and high quality of the students, I was able to ensure that each group had

- 3-4 students
- At least one undergraduate and one graduate student
- At least one male and one female
- At least one person with experience in EACH of social science, computer programming, and design
- More than 80% of students were assigned to their first or second choice project

Outline of the Experiments

Each of the experiments was designed to address one or more of the basic research questions outlined at the start of the course. The Appendix immediately following this summary provides brief descriptions of each of the group projects. The web site www.stanford.edu/~nass/comm369 provides PowerPoint descriptions of all aspects of each of the studies.

Experiments 1-5 used the ability of the CSLU Toolkit to answer the telephone and run a conversation with speech recognition and voice output. For these experiments, participants were given the address of a web site that gave initial instructions and provided the phone number subjects would call to interact with the system (each experiment had a different phone number; different conditions were assigned different passwords that participants spoke to gain entry). After working with the experimental interface run by the Toolkit, the participants answered a post-test questionnaire on the web site. In all of these cases, the experiments used the phone support, speech recognition, text-to-speech output, and recorded speech output features of the interface.

Experiments 6-8 were web-based studies. These studies used the CSLU Toolkit to create the experimental stimuli. Particularly important was the facility to easily and powerfully manipulate various parameters of the text-to-speech output.

Experiments 9 and 10 did not use the Toolkit.

The students worked with myself and the Head Teaching Assistant, Eva Jettmar, to focus the broad research questions and to develop an experimental design that would address the questions. We ensured that the studies met the most stringent criteria for clarity of questions, rigorous design, and dependent variables that answered the questions. All of the experiments had a minimum of two factors, each with a minimum of two levels, and all experiments used a questionnaire with multiple dependent variables. Many of the studies involved behavioral measures (e.g., bidding behavior) or recognition memory measures. As noted earlier, each of the studies was very different and each addressed one or more fundamental questions in speech interface theory and design.

Experimental participants were drawn from the 200 undergraduates in my lecture course on interface design. Each student was required to participate in five experiments. Each experiment used between 50 and 100 experimental participants.

To my knowledge, these experiments are the first ones that execute an experiment by combining a telephone-based speech recognition system with a web site for administration of the questionnaire. The critical advantages here are:

- Participants can interact with the stimuli in a natural and externally valid setting;
- Lengthy questionnaires can be administered with greater flexibility, efficiency, and accuracy. On the telephone, one cannot administer non-numeric Likert scales; it is slow and tedious to answer questionnaires via phone; and the Web provides psychological separation from the stimuli, avoiding the politeness results demonstrated by Nass, Moon, and Carney (Nass, C., Moon, Y., & Carney, P. (1999). Are respondents polite to computers? Social desirability and direct responses to computers. *Journal of Applied Social Psychology*, 29(5), 1093-1110); and
- Multiple participants can be run at one time (this is extremely difficult in the typical laboratory setting).

The results of the group projects were extraordinary! Nine of the groups obtained enough significant results to warrant publication in major journals of social science and/or computer science. It is anticipated that the studies will be summarized in a book (one study per chapter) to be co-published by University of Chicago Press and CSLI Publications (CSLI Publications has already agreed to publish the book; there are on-going negotiations with University of Chicago Press to act as a co-publisher). Each student group will be identified as authors on the chapter that describes their particular study, although I also expect that each group will produce a journal publication based on their research.

I have not yet received the formal student evaluations of the course. I have heard that many of the students expressed to other faculty and the industry visitors that this was the best educational experience of their academic careers. Virtually every student indicated that they obtained a set of skills and knowledge that they could not have obtained any other way. When the course evaluations become available, I will post the results on the course web site.

Leveraging success: Industry Involvement

When this course was initiated, I was very dubious about its success. To my knowledge, there had never been a project-based course grounded in original experimental research, and it seemed unlikely that the students could put together such ambitious projects in such a short-time frame (approximately eight weeks!). However, by the middle of April, I began to sense that the students were exceeding my most optimistic hopes, in that each group had developed very crisp questions and elegant experiments to answer those questions. Most importantly, the CSLU Toolkit made it possible for students to implement these experiments remarkably quickly. Simply put, it would have been impossible for virtually any of the groups to have done their experiments without the Toolkit.

As I described the course to my colleagues in industry, they exhibited remarkable interest and enthusiasm about the projects, and urged me to develop a method for linking the research projects and students to industry. Despite significant student reluctance about committing to presentations when the studies had not yet run, by May 1, we had agreed that there would be a small industry event on June 2 (this was the first day of finals in the Spring Quarter, and hence the last possible day in which all students would be on campus). I

warned the students that giving only four weeks notice for a one-day seminar of student presentations would likely result in a very small conference.

Invitations to people in industry, sent via email, went out between May 2 and May 16. Approximately 40 invitations were circulated. The invitees were derived from a list of industrial affiliates of the Center for the Study of Language and Information at Stanford and industry colleagues whom I had encountered in conferences and via consulting—as there was extremely little time to put together this event, I did not use a systematic method for inviting participants.

There were three stated goals for the event:

1. Disseminating state-of-the art discoveries in the design of speech interfaces to industry;
2. Exposing student researchers to industry people concerned with these issues; that is, all attendees were required to be researchers or designers of speech interfaces, and;
3. Exploring industry-university collaboration

The industry response was extraordinary! Every single company that was contacted elected to send one or more individuals, and a number of companies I had not contacted requested that they be invited. All told, there were over 80 industry participants at the conference, representing approximately 35 different companies. As can be seen from the attendee list (see Appendix), virtually every U.S.-based speech recognition company, U.S.-based voice portal company, and voice interface company was represented at the conference. Attendees came from Europe (Philips) and all over the U.S. The fact that this was a *one-day* event of *student-only presentations* organized in about *four weeks* demonstrates the tremendous industry interest in these types of events.

The conference was an enormous success. There was clearly high interest throughout the day. Indeed, many of the visitors requested that we add a second day of presentations and opportunities to meet with students, visit the labs, etc. Unfortunately, this could not be done because of final exams.

Perhaps an even stronger measure of interest was that we asked the industry guests about mechanisms for moving forward. Approximately 40 of the visitors indicated that they and their companies would be willing to contribute industry mentors to the project, i.e., having one or more of their employees serving as members of the project teams. The idea was that this would add to the diversity and strength of the groups, and even more readily facilitate industry transfer. The companies also indicated that they would be willing to provide economic and technology support for these groups.

When asked to discuss the reasons for the tremendous interest in this type of course, the responses fell into a few basic categories:

- Experimental design and research skills are much stronger in academia than in industry so courses of this kind provide a compact mechanism for transferring this knowledge and skill set;
- Deployment of and demand for voice interface systems far exceeds the understanding of user responses to these systems; courses of this kind rapidly produce specific answers to critical design questions;

- Because these courses are not grounded in the design of *products*, concerns about intellectual property are eliminated; and
- These types of courses enable companies to develop rich relationships with potential future hires.

Through the Center for the Study of Language and Information (CSLI) and the Department of Communication, I, in collaboration with Professor Byron Reeves (Director of CSLI) and Professor Kristine Samuelson (Chair of the Department of Communication) are attempting to identify the most appropriate and effective means for engaging industry in these types of courses.

Conclusions

1. Project-based courses organized around experimental research on interfaces can be extremely effective in producing important research and in training students.
2. The CSLU Toolkit was an invaluable tool in executing such a course. It would not have been possible to run this course without the Toolkit.
3. There is tremendous industry interest in these types of activities. There are significant opportunities for industrial support in terms of mentorship, training, and financial support.

Abstracts of the Projects: Experiments in Speech User Interfaces

1. Effects of Prompt Voice (TTS vs. Recorded; Gender) and Participant Gender on Disclosure

One of the key issues for a variety of voice and web portals and commerce sites is how to obtain more accurate and more plentiful information about users. Our project examines this question in the context of a survey of highly-personal information. Participants answer the questions on a web site or a telephone-based phone system with speech recognition and various voices for the prompts. Both male and female participants are presented with one of five conditions: text, female recorded speech, female TTS, male recorded speech, and male TTS. Will people be more willing to reveal personal information when it is typed as opposed to spoken? Are people more willing to answer personal question when asked by an impersonal (i.e., TTS voice)? Does the match between gender of voice and gender of person make a difference? Do these effects change when the information becomes less personal? The results can be applied to applications ranging from the acquisition of information from customers, the assessment of a population's opinions on current issues (e.g., political polling, market research), or even the decennial U.S. Census.

2. Personal vs. Impersonal Speech

There are two general ways of speaking: personal (e.g., personal pronouns such as “I” and “me”) or impersonal (e.g., passive voice, impersonal language). In the context of a telephone-based auction system, our project examines how the style of language interacts with whether the system uses recorded speech or synthesized speech. Is it odd for synthesized speech to use I/Me? Conversely, is it more natural of recorded speech to use personal language than impersonal language? Possible effects include trust, comfort, ease-

of-use, bidding behavior, etc.. The study has implications for any system that produces voice.

3. Mixing TTS and Recorded Speech

In many voice applications, one hears recorded speech followed by TTS, as in this example: (recorded) “The date is” followed by (TTS) “June 2.” Of course, there are cases when one hears only recorded speech and only text-to-speech. Is it better to use as much recorded speech as one can, or is a single modality better, even if it is all text-to-speech? The context for this study is a telephone-based system that provides three types of information: financial information, email dictation, and housing information. The results have implications for any system that provides both constant and varying information.

3. In-Group Prompts

When one listens to a voice prompt, one of the first determinations is whether the voice “sounds like me” or “doesn’t sound like me.” This project examines the question of whether prompts with a foreign accent (Swedish) affects users’ attitudes and behaviors toward the system. In the context of inquiries about socially-desirable information, male and female participants are asked questions by a voice-recognition-based computer system. Does gender affect people’s willingness to admit to failings? Will people react differently to prompts that match their accent as opposed to mismatch? We also examine effects on social consciousness, feelings of control, social presence, liking of and frustration with the system, and amount of words spoken. This research is important for any voice output system.

4. When Voice and Content Don't Match

With content on the Web being developed and changed incredibly rapidly, one element of voice-based interfaces that will be difficult to control is whether the content matches the characteristics of the voice reading the information. In a telephone-based system presenting a wide variety of contexts (e.g., news, movie descriptions, health information), our project explores the effects of the happiness or sadness of both recorded and synthesized speech when paired with happy or sad content. We determine whether matched voice and content lead to great liking, credibility, and memory for the information presented. This research has implications for any voice-based service that presents information.

5. Misrecognition: Empathy, Criticism, or Neutral Repetition

Because speech recognition systems are not perfect, voice-based systems must indicate mis-recognitions. Our project examines this issue in the context of a telephone auction system. We examine two variables: whether the misrecognition rate is low or high, and the type of feedback when an utterance is not recognized: a) blame the speech recognition system, b) blame the user, c) don’t blame either party, or d) “please repeat the message. We are interested in user perceptions of the system. This research has implications for all speech recognition systems. It also addresses cases in which error rates are unusually low (microphone with computer) or unusually high (automobile PC).

6. Fact vs. Opinion and TTS vs. Human Voice

Voice is being used to present a wide range of content. Our project examines the effects of human voices and text-to-speech (TTS) voices on how information is interpreted. In the context of Net Radio, we presented news stories that were labeled either “news” (fact)

or “editorial” (opinion). Questions answered include: 1) Does TTS make articles seem more “factual”? 2) Is there an interaction between the label and the mode of presentation? 3) How does the label influence people’s perception of the content? The research has implications for the selection of presentation modality, the effects of labels on content, and the issue of consistency between modality and content.

7. Consistency between TTS and Personality in Email

Email readers have become a very popular application for voice portal. This project investigates personality and identity in the context of a TTS system reading e-mail messages. In this context, there are three personalities involved: the personality of the person sending the e-mail message, the personality of the person receiving the message, and the personality of the TTS voice reading the message. What are the consequences of matches or mismatches between each of these three personalities? This research has implications for voice portals, computer agents, and any product that attempts to manifest brand through voice.

8. Choice of Voice

Although permitting a selection from multiple recorded voices is often practically impossible, it is very simple and computationally efficient to provide participants with a choice of multiple synthesized voices. In the context of a book-buying site, we provide extroverts and introverts with a choice of which of two synthesized voices (introvert vs. extrovert) they would like to use to hear book descriptions. Will individuals choose the voice that matches their own personality? Will the act of choice influence the interaction between the participant’s personality and the voice’s personality? We examine purchase behavior, liking, trust, and feelings of social presence. This research is important for all systems that provide voice output, particularly in the e-commerce context.

9. Speech Recognition-Based Search

Because of the rapid growth in the number of pages associated with a given web site, search has moved from being the exclusive province of portals to be part of every major site. In this study, we explore how voice can be used in the search context. The context of the study is a locator service for various stores and professions. Our project explores the effects of length of prompt and the number of interactions needed to achieve a successful search. We examine trade-offs in these two criteria in terms of success, speed, desirability, perceived effectiveness, etc.. The results have implications for transitioning sites from the web to the phone, as well as voice portals.

10. Lifelike Agents

Remarkably human-like agents are appearing more frequently on the web and GUI computers. This project examines how the quality of representation of the agent and the belief that it is an agent rather than a videotaped person affects attitudes and behaviors in an e-commerce context. Will people forgive flaws in agents more than videotaped people (if identical)? Will humans or agents be more persuasive and trustworthy? Do inferior representations imply inferior intelligence, reliability, and persuasiveness? This research has implications for the deployment of agents in e-commerce, advice systems, and computing

more generally, as well as cases in which bandwidth constraints require limitations in representation quality.

4. Evaluation

Dr. Lecia Barker of the University of Colorado provided independent evaluation services. The final section of this report, written by Dr. Barker, provides a summary of her initial evaluation efforts:

The primary goal for the initial six-month period (January 1, 2000 – June 1, 2000) of the CRCDD-supported HLT curriculum program was to establish the necessary infrastructure. Below is a brief description of the establishment of infrastructure followed by evaluation goals for next year. Lessons learned from the start-up period will be used to improve program offerings in the next year.

Infrastructure

The PI group successfully established an infrastructure for supporting students and faculty in the HLT program. Their efforts included course selection issues, official approval of a certificate at the University of Colorado, establishment of computer labs, and development of a program web site:

1) ***Courses – selection of courses to be offered, frequency and timing of course offerings, course sequence and/or prerequisites, and pedagogy issues (providing for students of different types of abilities, teaching model).*** Courses offered are consistent with the original proposal. The PI group decided that the courses should be non-sequential, since the curriculum is interdisciplinary and students will be coming to the program with different backgrounds. Since it takes about one year to get listings into the schedule of courses, decisions were made as to the timing and frequency of course offerings. While courses are cross-listed, the department of the lead professor took responsibility for ensuring that courses are officially offered. Some students entering the program will have non-technical backgrounds. These will be encouraged to take additional programming courses; also, technical and non-technical students will be paired/grouped, since the skills/abilities of both are needed in HLT. For more information on courses chosen and requirements for the certificate (see 2 below), see <http://www.colorado.edu/ling/jurafsky/curr.html>.

2) ***Certificate.*** The PI group gained formal approval for the Interdisciplinary Certificate in Human Language Technology on May 17th, 2000. The PI group made decisions as to who will track students in the certificate program and student advising.

3) ***Computer labs.*** Student labs enabled with the Toolkit and other appropriate software for student use have been established, though not without difficulty. Difficulty was due to the university's computing infrastructure. For example, Sun labs were unable to support Windows (the Toolkit will soon be available for the Linux operating system, eliminating this problem). Technical issues have been worked out, such that Toolkit use is possible on all participating campuses.

4) **Web site.** The URL listed above describes the certificate program and the courses. Courses offered in Spring 2000 also have web sites (see reports from individual professors). A more comprehensive web site for the program will be developed and will include information on the certificate, links to each course web site, and a variety of other materials designed to support both teaching and learning. A template for course web sites has been developed so that each site will be consistent, including lectures, assignments, etc.

Evaluation Plans

Evaluation efforts will ultimately focus on ensuring that the HLT curriculum incorporates the latest advances in HLT into the curriculum and prepares the most highly proficient students for industry and academic positions. Interim evaluation goals during the start-up period included monitoring the progress of the PI group in reaching its goals of establishing the infrastructure as well as gathering information on courses offerings. An e-mail questionnaire was sent out to professors teaching during the first semester to find out how many students (male and female, major) are enrolled in each course, any teaching or technical issues, perception of student interest in HLT careers and other courses, and experiments and activities using the Toolkit and other speech technologies. This kind of data will continue to be gathered in the future and used for formatively for improving the curriculum.

With the infrastructure in place, evaluation will focus more closely on students' learning experiences and ensuring that students are receiving the instruction needed for careers. Particular attention will be paid to three issues: recruiting and retaining women; bringing students with less technical expertise up to speed; and developing an understanding of how the courses contribute to academic and industry careers in HLT. Classroom observations, surveys and interviews with faculty, industry representatives, and others, and document reviews will provide data to inform these issues.

Faculty Evaluation Questionnaire

The following information was request via email from the course instructors.

HLT Curriculum Information Sheet

Please take the time to answer the following questions thoughtfully.

1. Have the course description, practices, or goals changed significantly from how it was described in the NSF proposal? If so, please describe.
2. How many students are enrolled in your course? Males _____ Females _____
What majors are represented?
3. What is your greatest challenge in teaching this course? (interactions with students, getting discussions underway, hardware/software problems, lab or materials availability, etc.)
4. Please speculate on student perception of the value of the course.

5. Are any of your students (that you know of) considering HLT careers? In industry or academic? Have any found jobs or been accepted into graduate programs?
6. Have any of your students inquired about availability other courses in the curriculum or courses related to yours?
7. Do you have a web site where you are posting course syllabus, lecture notes, assignments, readings, lab modules, etc.?

___ Yes – please indicate the url and password (if needed)

___ No – please let us know how we can get these materials

8. Questions about the CSLU Toolkit:

- a. In what way are you using the CSLU Toolkit? (demonstrations, lab exercises/experiments, application development, etc.)
- b. What kinds of laboratory experiments have you done with the toolkit? What components have been used?
- c. What problems have you encountered?
- d. What support have you received for managing problems? From whom?

Please rate the support:

Poor-----fair-----good

- e. How would you improve the Toolkit in terms of operations, ability to create applications, etc.?
- f. How would you change the Toolkit to facilitate teaching and learning?
- g. Additional comments about the Toolkit: