08/15/01 — 10/31/02: RESEARCH AND EDUCATIONAL ACTIVITIES

In the fourth year of this project, we focused our efforts on two activities:

- **Production System Release**: completed a release of the system that contains all necessary core speech recognition functionality. We developed new approaches to cross-word decoding and parameter estimation that simplify the complexity of the system. We also developed several user interface tools to aid in configuration of the system.
- **On-line Documentation**: overhauled our on-line tutorial to reflect our experiences in supporting this software. We now provide step-by-step instructions starting with configuration and installation that should allow most users to learn about the technology in a self-paced manner.

We also continued activities in the areas central to the overall research program:

- **Foundation classes**: added many algorithms at the higher levels of the class hierarchies to support a modular speech recognition system. This makes our recognition system perhaps the most flexible engine publicly available.
- **Workshops**: hosted a software design review in January 2002, and a one-week training workshop in May'02. Our on-line resources, including lab experiments, were significantly enhanced.
- **Software engineering**: Streamlined our support activities and improved the ease of use of our distribution and verification procedures. We also upgraded our support of Linux and now perform much more detailed regression tests before a release. All such testing has been completely automated.

Overall, this year we saw a significant shift from academic support to industrial support. Because our software is now stable and mature, and widely known for being an important baseline system, industry has become much more interested in its use. We have been involved in several trials with leading speech technology companies, and have successfully convinced them that our system achieves state of the art performance when comparing similar configurations. We have also educated them on how to run leading industry-standard benchmarks.

A. Retrospective

It is hard to underestimate the impact this research program has had on our laboratory, the Institute for Signal and Information Processing (ISIP). In many ways, I feel the success of this lab is a good justification for programs such as CARE. The original goal of the CARE program was to build lasting infrastructure to support research with strong experimental needs. Today, ISIP is well-known throughout the world for providing state-of-the-art speech recognition technology and educational materials. At the beginning of this program, we were virtually unknown outside the inner circle of the speech research community.

It is often said that imitation is the most sincere form of flattery. Another extremely important way to benchmark the success of our program is to examine what has happened in the open source movement since this program was started. For many years, the speech recognition community was dominated by a small number of research groups with proprietary systems. Research progress was slow and expensive. Even when we originally proposed the open source concept in the late 1980's, and then again in the mid-1990's, we were repeatedly refuted by arguments that for-profit software was adequate and was not limiting access to the field.

Since this project was started in 1998, several major speech recognition sites have followed with open source releases. These include well-established sites such as Cambridge University (HTK),

Carnegie-Mellon (SPHINX), MIT (Galaxy), University of Colorado Boulder (PHOENIX/CU-Communicator), and SRI (Decipher/SRILM). Whether these releases would have happened naturally, or were prompted in part by the well-publicized launch of this program, will remain a mystery. Nevertheless, it is clear that more speech recognition technology exists in the public domain than ever before and that the technology is more accessible to smaller organizations.

Our system, however, still remains the only unencumbered system. It is the only system which can be exploited commercially without licensing fees. It does not come with any licensing restrictions. It does not even carry any form of GPL [1] license. We do include a licensing statement in the distribution that states:

```
=-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==
Institute for Signal and Information Processing
Mississippi State University

LICENSE AGREEMENT

This software is distributed without any restrictions.
You may copy, distribute, or modify this software as you see fit.

This package is provided "as is" and without any express or
implied warranties, including, without limitation, the implied
warranties of merchantability and fitness for a particular purpose.
=-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==-==
```

This was requested by several corporate entities whose lawyers felt uncomfortable downloading software without some from of a license statement.

It is not surprising, therefore, that this year we have enjoyed more industrial interest than ever before. We had extensive interactions with several significant speech recognition companies (e.g., Nuance Communications and Intervoice). Both are conducting internal evaluations of our technology, and used our technology to bring up new applications within their research groups. Another company, Lexia, sent an engineer to our lab for one week to develop a prototype of a child reading aid that uses speech recognition, and to assess the feasibility of the project given the current state of the technology. Other laboratories that received support attention included MITRE and DoD. We even have had European partners, such as the Fraunhofer IMK Research Institute, develop extensive systems based on our technology.

In addition to these types of activities, an important milestone was reached in 2002 when the Aurora Large Vocabulary evaluations [2-5] were conducted using our software. This was an excellent opportunity to increase our visibility on an international scale and to demonstrate the relevance of this technology to state of the art problems in this field. In the Aurora evaluation, several sites competed for the rites to a next generation standard for front end processing for cellular telephone applications. The Aurora Working Group is a sub-group under the auspices of the European Telecommunications Standards Institute (ETSI). The work performed in this collaboration will be presented in a special session at the European speech technology conference (EUROSPEECH) in September 2003.

In addition to these milestones, a significant body of work has been developed to support the annual workshops funded by this research project. These workshops were originally planned to introduce a small number of people in a very focused manner to the technology. Due to strong

demand for these workshops, we decided to double the size of the workshops by leveraging a new training facility that became available at Mississippi State. Over the three years we held the workshops, over 100 people were trained on our software. (In fact, in January 2003 we recently held a remote version of this workshop in Dallas, Texas, which represented our first attempt at hosting the workshop off-site.) All lecture notes, lab exercises, etc. are available on-line. This material has become an invaluable body of training and support literature.

Another unique feature about our technology is the availability of on-line resources such as our speech recognition tutorial [6]. This tutorial starts with the basics of downloading and installing the software, and then proceeds step-by-step to teach users about recognition, training and language modeling. It concludes with a chapter on how to build state of the art systems. What is most unique about this tutorial is that all data, configuration files, etc. are available on-line. Hence, users are shown what commands to execute, and then are provided with the expected outcomes from these experiments, making it easy to debug problems along the way. No recognition technology available on the Internet includes such a tutorial. It is a good example of how the CARE program has provided a unique opportunity to disseminate such resources.

Finally, a good measure of the success of a public domain software effort is the extent to which others have contributed to the core software base. We have had several such contributions in recent years, including code (often unacknowledged by the providers request) from private and government labs such as MITRE and DoD, and a port of the system to Windows contributed from Ralph Edrich at the University of Applied Sciences in Germany. We have also incorporated numerous low-level code modules contributed from various sources into the system. In these cases, as we expected, we served as system integrators and refactored this code according to our software engineering standards. We have several users within DoD and academia working with our extensive C++ libraries and underlying code structure to create new applications and capabilities. Hence, our software infrastructure and web site are fulfilling the original mission of the CARE program to develop advanced experimental research resources. We now describe some of these capabilities in detail.

B. Speech Recognition Toolkit

The general design philosophy for our speech recognition toolkit is discussed in prior project reports and publications [7]. In this section, we focus on a few key features of this system that differentiate it from other publicly available systems. We also discuss the rationale behind this design. We begin with a discussion of the core decoder.

Time-synchronous Viterbi beam search [8] has been the dominant search strategy for continuous speech recognition systems for the past 20 years. Search [9] is a good example of an algorithm that is conceptually simple, but extremely hard to implement in a general way in practice. Worse yet, the slightest inefficiencies in search can result in a system that cannot solve large-scale problems. Most search implementations are restricted to one approach (e.g., time-synchronous Viterbi search) and cannot be separated from the statistical modeling approach (e.g., HMM). Further, the search structure is commonly limited to three levels (e.g., word, phone, and state). The goals for the design of our generalized decoder were:

- an arbitrary number of independent levels;
- long-span context-dependent models at any level;
- lexical tree expansion at any level;
- posterior symbol probabilities at any level.

We also designed our system to provide alternate algorithm choices for statistical modeling at each state, acoustic modeling via a collection of states, and language modeling via finite state machines. In fact, our decoder is designed to support recognition for applications beyond speech and audio streams. and currently supports multiple search algorithms using a common programming interface.

A generalized hierarchical search space [10] that must be implemented in a typical speech recognition system is shown in Figure 1. In such a search space, each level can be conceptually considered as the same. Phrases, words or phones are simply symbols at different levels. Each level contains a list of symbols S_{ij} and a list of graphs G_{ik}., where, i is the index of the level, j is the index of the symbol and k is the index of the graph. Each graph has at least two dummy vertices, the start vertex and the terminal vertex both indicating the start and the end points of the graph in a search space.

The lower level graph G_{in} (such as G_{12} with **RUN** and **WALK**) is the expansion of the symbol $S_{i-1,n}$ (such as S_{02} with **VP**) at the level above. This relationship connects all the levels together, realizing the entire search space. There are two exceptions for such a symbol-graph expansion. First, at the highest level, only one graph can exist. This graph is called the *master grammar*. It is a map for the decoder to iterate through the entire search space. Second, at the lowest level, each symbol represents an underlying statistical model. These symbols can not be expanded into a sub-graph. Observations probabilities will be evaluated when the search process reaches these symbols.

The search begins from the start vertex of the *master grammar*, the starting point of the search process. The traces, indicating search paths from one vertex to another, first propagate from the top level to the lowest level, where the observation probabilities of the each frame of data are evaluated. Then the traces propagate forward from the lowest level. Whenever a trace reaches the terminal vertex of a graph, it propagates up. The decoder drives this up and down process for every frame until the end of an utterance. The hypotheses will then be retrieved by traversing backwards through the traces.

Lexical trees, also referred to as prefix trees, are required so that we can model transitions between words (or any symbol at any level) more accurately. The particular acoustic unit chosen at a boundary can depend on the first sound in the next word. The process of searching for an optimal unit is exponentially complex because we must consider all possible next symbols.

The basic idea of lexical tree expansion is to collapse pronunciation models of different words by sharing the same beginning phonemes. The use of a lexical tree significantly reduces the search space and search effort [8]. Based on the idea of sharing the common prefix, we extend the
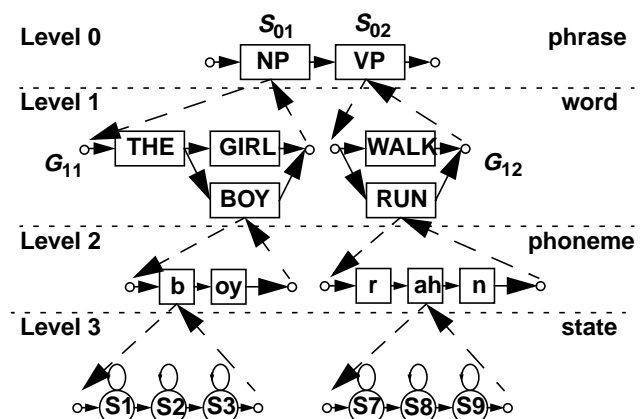

Figure 1. A representation of a generalized hierarchical search space for a recognition system consisting of four levels of knowledge: phrases (sentences), words (dictionary), phonemes (pronunciations) and states (acoustics).

prefix tree representation to any level in the search hierarchy. If users set the decoder to expand a lexical tree at level *i*, the symbols of level *i* will be expand to their corresponding sub-graphs at level *i+1* and the common prefix of these symbols will be shared, resulting in a tree-structured search.

Context-dependent phone models, defined as a model which depends on preceding and following sounds, are generally more accurate than context-independent phone models, since the former can capture coarticulatory effects. In our decoder, the concept of a context-dependent model can be used at any level. If a symbol S_{ij} is context-dependent then the underlying model is determined dynamically via it's neighboring symbols. This generalized implementation imposes no restrictions on the length of context and the number of levels using context. Similarly, N-gram models are extended to N-symbol models. The symbol can be a phrase, a word or a phone, depending on the level at which it is used. The decoder doesn't restrict the order of the N-symbol model. Therefore, users can apply arbitrarily long time-span language or acoustic models to meet the needs of their applications.

The virtue of this approach is that speech recognition research can be performed by simply manipulating a hierarchy of these networks. We have developed a tool, called Network Builder [11], that allows users to create and manipulate these networks using a graphical user interface. A detailed tutorial on this tool is available from our web site.

Another feature of our system is that we have developed several useful generalizations of hidden Markov model (HMM) training. Our acoustic trainer is a supervised learning machine that estimates the parameters of the acoustic models given the speech data and transcriptions. We support both Baum-Welch and Viterbi reestimation of model parameters for mixtures of arbitrary statistical models. In addition to standard features of HMM trainers such as state-tying and mixture splitting, additional requirements for our software design included:

- arbitrarily long context-dependent models at any level;
- parallel processing approaches to training;
- network training for enhanced pronunciation modeling.

The HMM trainer we implemented is based on a hierarchical search space described previously. Most STT systems contain three levels of information: the language model (word graph or N-gram), the pronunciation model (lexicon) and the state level (HMM topology for each phone). The motivation to develop a trainer with no restrictions on the number of levels was to further support research into nature language processing, which often involves more than three levels of representation, and uses many diverse knowledge sources.

In most HMM-based systems, a phonetic transcription is required as input to the parameter reestimation process. This transcription is derived at automatically and iterated upon as the recognition system improves. Silence must also be hypothesized between each word model, since we don't know where in the utterance silence actually occurred. In essence, the trainer decides on the optimal alignment of the hypothesized transcription as part of the training process.

Our network trainer, which is depicted in Figure 2, alleviates the need for a phonetic transcription. Instead, it is capable of hypothesizing all possible expansions of a symbol list, and performing parameter reestimation across this expanded network. The benefits are twofold: we remove the dependency on an intermediate transcription, and we can inherently train statistical pronunciation models. This data-driven approach often results in better performance.

Level 0   **transcription**
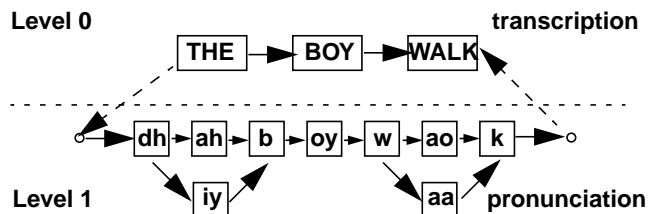

Figure 2. A graph that depicts a representation of multiple pronunciations in network training. Transition probabilities between phones (Level 1) are automatically reestimated using Baum-Welch training.

Many existing systems are limited to simple left to right topologies for their acoustic models, thereby restricting the type of research that can be performed with such a system. Our trainer imposes no constraints on the topology of any model in the hierarchical structure. In the training process, models at the top of the hierarchy are dynamically expanded to its constituent models at the lower level, forming a single network. The reestimation algorithm (e.g., Baum-Welch) operates at all levels and updates parameters related to the transitions and the states simultaneously. Hence, language modeling training is implicit.

Finally, our toolkit uses database formats for all file input and output. Though cumbersome at first, this is an extremely important step in eliminating support problems and aborted recognition experiments. Utterance transcriptions, which are required for training, are represented using an AnnotationGraph format [12] that has recently become popular in our research community. This graph format allows a hierarchical representation of information in the transcription. One transcription can contain many types of information, including time-alignment information. Users access this information in a variety of ways, including numeric indices for levels and text labels for specific transcription information.

Audio files are also referenced via an audio file database, implemented in a class named *AudioDatabase*. This allows an experiment to be conducted by simply supplying the recognizer with a list of utterance ids. Standard experiments can be easily replicated from simply a list of utterance ids. The order of the list is not an issue, whereas in many other systems the input must be a synchronized list of transcriptions and audio files. Use of these databases is a good example of how our system has been designed to minimize support effort and to be self-diagnostic. Our goal is that a bad experimental setup can be self-diagnosed before the experiment has accumulated appreciable run-time. Linking transcriptions and audio files via a database, and using structured data as input (e.g., AnnotationGraph), represent major steps towards achieving this goal.

C. Educational and Tutorial Resources

A significant effort was made in the last year of this project to enhance our on-line educational resources. In Section E, we will describe the workshop component of this activity, which also serves as our Outreach activity. Here, we will focus on on-line documentation, which comes in two forms: software manual pages [13] and tutorials [14].

Our software includes on-line html-based web pages. This was a bold experiment started early in this four-year project in an effort to make the software documentation portable and available to web users simultaneously with the actual software installation. Implementation required a rather novel approach using cookies so that users and software developers could easily transition between their private enlistment (stored on a local disk) and the released version of the software (stored in a network-based resource and used by those requiring a stable working environment). In our development environment, a developer can view man pages through their private enlistment, while users of the software view the released versions of the web pages. All this is done without

having to make any changes to the URLs loaded into the html pages via the use of a cookie (that facilitates the path resolution for the web page). It would be a disaster if the developer had to change URLs on a web page before release.

Our web pages have a uniform structure discussed in previous reports. They are organized into an ontology that begins with differentiating between utilities (such as the recognition program) and classes (our foundation classes), and continues to branch out from their in a hierarchical fashion that mirrors the organization of the software on disk. Two examples are provided below that show a typical introductory page [15], shown in Figure 3, that contains an overview of a library (e.g., Algorithms), and an actual class manual page [16], shown in Figure 4, which describes many details about a specific class (e.g., Energy).

Figure 3. An example of a man page representing the introduction to a C++ library (e.g., Algorithms).

Manual pages are an excellent example of what differentiates the results of this CARE project from most other speech research related software developments (in the sense that we can develop resources not normally justifiable in a typical research or spare time project). Our manual pages contain a description of the class (often including relevant theory), a quick start example, a description of all data and methods included in the class, and a couple of detailed examples that show how to use the class in a standalone fashion.

More importantly, all code is linked into the man page. When you click on a method, it takes you to the code for that method. Our code is written in a well-documented easy to read format. An emphasis has been placed on keeping the code simple and easy to understand — we refer to this

Figure 4. An example of a man page for a class (e.g., Energy). All documentation is html-based.

as self-documenting code. Hence, perhaps the most valuable part of a man page are the links it contains to the software. We developed tools to support this, including a simple perl script that parses the code and returns only the requested function from a file. The result is that the code is displayed automatically on a web page, providing a seamless integration between documentation and code. As part of a release, we include documentation that can be viewed on a user's local machine with or without the software distribution.

In addition to the on-line documentation, a major effort was spent this year expanding our on-line recognition tutorial [6]. This is a very detailed step-by-step tutorial that teaches users about all aspects of speech recognition. A copy of the front page from this tutorial is shown in Figure 5. There are seven sections: installation, data preparation, feature extraction, recognition, acoustic modeling, language modeling, and comprehensive examples. An appendix and glossary are also provided. The information presented in this tutorial reflects our four years of on-line support. We have tried to focus on examples that reflect the common problems our customer support. For example, the installation and data preparation sections both were created from our experience that 90% of the work in running a speech recognition system involves getting the software and data in place to run experiments.

Each section in this tutorial begins with an introduction describing the salient theoretical background for the technology discussed, and then continues with examples. For each example, specific command lines are given to execute a task. All files required to set up the experiment are provided, along with links to the expected output. Hence, users can diagnose their own problems, and know what to expect for an outcome. We also often point out common mistakes and address

Figure 5. A screenshot from the front page of our on-line tutorial. Detailed instructions are supplied for everything from installation to complete LVCSR application development.

anticipated questions, again all a result of our extensive experience with customer support. We hope that this tutorial will eventually become part of a speech recognition textbook, and will be used in graduate level coursework. We plan to use this tutorial in our graduate-level speech recognition course scheduled for Spring'04.

D. Java Applets

One of the original unique features of our proposed research program in 1998 at the project's inception was the demonstration of some advanced educational capabilities in Java [17]. Over the course of this project, we have developed two novel Java-based capabilities: a remote job submission facility [18] that allows users to submit speech recognition jobs over the Internet to our servers, and applets [19,20] that demonstrate fundamental machine learning concepts. Our applets are accessible from a single page on our speech web site [21].

The page shown in Figure 6 consolidates all of our on-line demonstrations. These are organized into four categories: Java applets which are described below, the Job Submission demo which has been described extensively in previous reports, audio demos which are designed to familiarize readers with some fundamental concepts about speech recognition and perception, and interactive tools which allow users to step through the recognition process in a debugging-type mode.

As documented previously, the Java programming language continues to be a support nightmare since there are significant changes in the language from one version to the next. Our applets have recently been upgraded to support Java 1.4, and this appears to be a much more stable and better supported version of the language. Nevertheless, we still have problems with robustness and unpredictable behavior forcing applications to spontaneously crash or hang — requiring the application to be restarted. Our job submission demo makes use of a new web server technology (Apache's Tomcat server) which facilitates remote connections to a server, but is still suffering

Figure 6. A screenshot from our web page that consolidates all of our on-line demonstrations.

from robustness problems.

Throughout the course of this project, we have developed and enhanced six Java applets that demonstrate concepts ranging from convolution to pattern recognition. These applets can be split into two categories: digital signal processing (DSP) and machine learning. The DSP applets include convolution, digital filter design, pole/zero analysis and spectrum analysis. The machine learning applets include dynamic time-warping and pattern recognition. A screenshot from the Java Applets page is shown in Figure 7.

The Pattern Recognition applet is a good example of what has been accomplished with these applets. This applet allows users to define custom data sets, or to choose from an inventory of classic data sets, and then to apply different pattern recognition techniques to the data. Users can choose from conventional techniques such as Principle Components Analysis and K-MEANS clustering, or more advanced algorithms Linear Discriminant Analysis. In Spring'03, we are adding two new techniques, Support Vector Machines and Relevance Vector Machines, which are considered on the leading edge of machine learning technology.

This applet allows users to explore how these algorithms work. Step by step instructions showing intermediate calculations are shown in a dialog panel on the applet. Users can step through the algorithms, or run them to completion. Decision surfaces and classification results are presented as well. Users can also apply different algorithms to the same data sets to see how the results vary as a function of the algorithm. This applet has been highly successful and extremely stable. We plan to continue evolving this applet to support new research into machine learning techniques. We are in the process of releasing a new version of the applet that includes a completely reengineered control structure to make it simpler to add new algorithms.
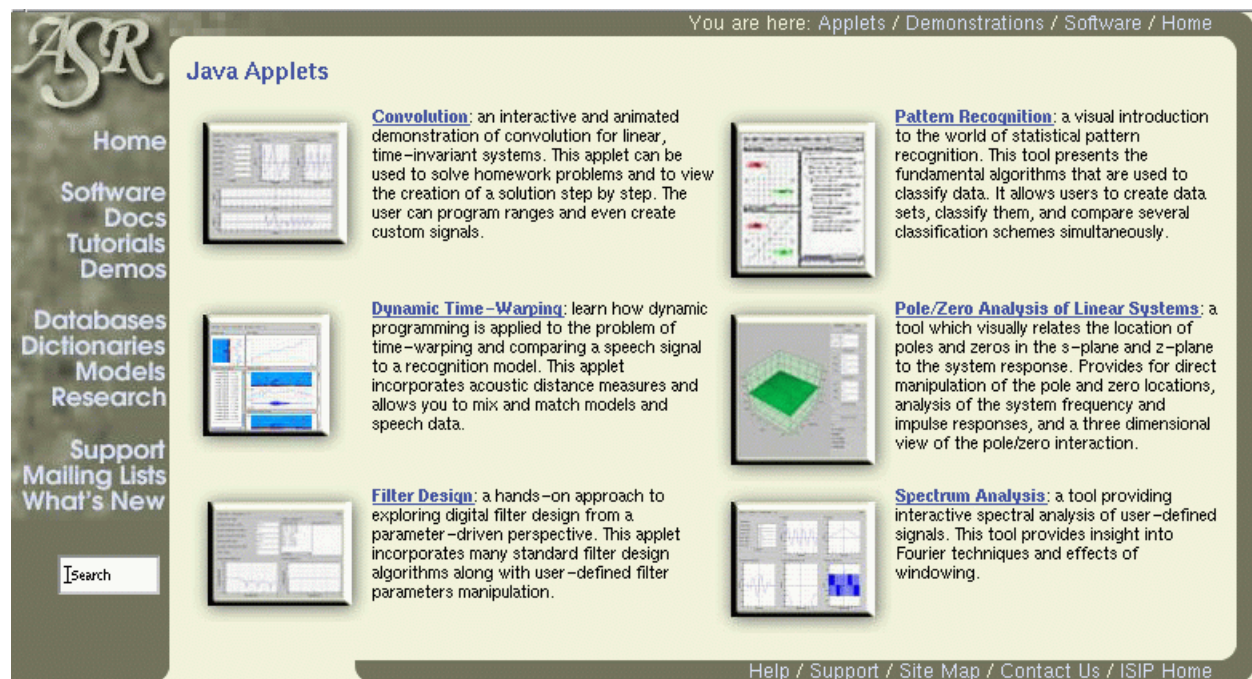
Figure 7. A screenshot from our Java Applets page that consolidates all of our applets. Our flagship applet is the pattern recognition applet, which demonstrates advanced concepts such as linear discriminant analysis and support vector machines.

E. Workshops

The last major category of effort which we describe here is our workshops. For the past three years of the research project, we have hosted two annual workshops. In January, we host a two-day software design review focused on industrial partners and technologists. In May, we host a one-week training workshop focused on graduate students. These workshops have been described in previous years of this report. Here we focus on a summary of some key issues related to this type of activity, as well as some unique capabilities introduced in the last year of this project.

Participation in these workshops has been consistent. We originally planned for a dozen participants, but ended up hosting about twice that number. Our single largest group of participants, as expected, has been entry-level graduate students. Most have been from small research groups wanting to get started in this area. Hence, we have fulfilled our goal of increasing access for research groups not coupled into mainstream speech recognition research. Several of these students have maintained a long-term collaboration, representing MIT, Georgia Tech, and University of California at San Diego. The latter has been the most serious level of collaboration, resulting in a proposed joint research project with the student, who is now a new faculty member starting a career in academia at Federal University of Para in Brazil. The plan is for his group to continue development of new features for recognition system, including maximum mutual information estimation (MMIE) training [22], under funding from a joint program between the Brazilian government and NSF.

Our interactions with MIT have included participation of one of their graduate students, and transfer of knowledge of how to develop conversational speech systems. Their interest in this case was more focused on procedures and data rather than direct use of our software (since they have their own technology). Our interactions with Georgia Tech have been with a small research group attempting to look at more fundamental signal processing and pattern recognition research. We were able to help them reach a point where their research could be easily and efficiently added to our system, and evaluated on a state of the art benchmark. We have also developed relationships with a number of other US institutions and European institutions that resulted in some level of collaborative research and/or proposal writing.

Over the course of the three years we have hosted these workshops, we have been pleasantly surprised by the level of industrial involvement. In the third year, due largely to the bad economy and hi-tech downturn, participation of industry was down a bit. Nevertheless, we have seen approximately 50% of our participants come from the industrial and government laboratory sectors. In fact, in past years we had to limit the attendance of industrial partners to reserve enough space for graduate students. Industrial participants are typically extremely motivated to transfer this technology into their laboratories.

We also established a goal from the inception of the program to solicit underrepresented institutions. Each year, we have had one participant from such an institution, which was well below our target. In the first year of the program, we solicited over 100 underrepresented institutions by email. In subsequent years, we sent personalized letters and followed up by email. We also aggressively contacted underrepresented universities within our state because we have on-going relationships with some faculty at these institutions. Unfortunately, the success rate of this approach was not impressive. We attribute this to a lack of qualified graduate students at these institutions interested in speech research, and, as one participant explained to us, an intimidation factor ("would they be competitive at a place doing significant research"). In retrospect, I think

increasing underrepresented participation requires a more active research collaboration first, so that interested students can be gradually introduced to the topic area. Our experience with North Carolina AT&T (formerly an HBCU), which is involved in a DARPA speech research program, seems to support this conclusion though our sample size is obviously small. I suspect face to face meetings at these universities prior to the workshops would help also.

Our workshops have enabled us to develop an impressive body of speech recognition educational material on the web. All lecture notes and laboratory exercises are available on-line at [23,24]. The one-week training workshop is structured according to the schedule shown below in Figure 8. Morning lectures focus on theory and theoretical aspects of implementation issues. Afternoon sessions focus on hands-on self-paced laboratories.

It is interesting to note how our experience with these workshops has evolved our teaching style. In the beginning, we implemented a "leave no one behind" strategy in which we tried to keep the entire class on the same page in a lock-step fashion. However, since the participants have a diverse set of skills and interests, this quickly became unmanageable since some users were unfamiliar with basic Unix operating system issues. Within a short time, people were skipping an assignment or not finishing things. The range of users attending the workshop was surprising. In the latter years of the workshop, we designed the labs to be much more self-paced, and provided solutions and intermediate results so failure on one part of a lab would not impede progress on the next part. Our feedback collected at the end of the workshop verified our intuitions that this was a more successful approach even though it meant some of the participants would not work all laboratories to their completion.

	08:30 – 10:00	4.1: Signal Processing	Measurements Statistical Modeling Typical Implementations	Joe Picone
	10:00 – 10:30	**BREAK**		
	10:30 – 12:00	5.1: Transformations	Algorithms and Recipes Front-End Overview A Graphical User Interface	Hualin Gao
TUESDAY MAY 14	12:00 – 13:30	**LUNCH (SELF-PAY)**		
	13:30 – 15:00	5.2: Evaluations	Generating Features Running Evaluations Adding New Algorithms	Hualin Gao
	15:00 – 15:30	**BREAK**		
	15:30 – 17:00	5.3: TIDIGITS	Building Digit Recognizers Evaluating New Front-Ends Picone's $20 Challenge	Hualin Gao
	18:00 – 21:00	**SPORTS NIGHT**		

Figure 8. An overview of the agenda for a typical day (signal processing) in our one-week summer workshop. Morning lectures are followed by afternoon laboratories.

F. Summary and Future Work

The original goal of the CARE program was to build lasting infrastructure to support research with strong experimental needs. We have presented over the past four years a significant amount of evidence that ISIP has achieved this goal in the area of speech recognition. Our web site and commitment to education and resource development is unparalleled in the community. We now have a highly flexible and modular speech recognition system in release, and are constantly adding features to it.

It is important to reflect upon our successes and failures over the course of this project. For example, I believe we have done a good job of software development and customer support. On the other hand, our vision for Java-based tools has been a struggle to implement. Our programmers have grappled with Java language problems from day one, and have not been productive in that programming environment. The development of high quality documentation, which is essential to any such research program, has also been a major struggle, and much harder to achieve than we expected. It requires staff who are fluent in the technology and interested in the software engineering side of the research. Such people are hard to find. Reliance on students for technical documentation has been a problem.

Over the next year, we plan to continue making substantial improvements to our software. Two important features that need to be added are maximum likelihood linear regression [25] and maximum mutual information estimation [22]. A first pass of an MLLR implementation has been completed but not released yet. An MMIE implementation is planned using a collaboration with Federal University of Para in Brazil. We are also continuing work to make the system more efficient, and will release a new version of the system that is 5x faster shortly. Finally, we are exploring the integration of speech and natural language processing in two new research projects, and hope to begin adding capabilities to support dialog systems to our software.

G. REFERENCES

[1]   "GNU General Public License," Free Software Foundation, Inc., 59 Temple Place — Suite 330, Boston, MA 02111, USA (see *http://www.gnu.org/copyleft/gpl.html*).

[2]   N. Parihar, J. Picone, D. Pearce, and H.G. Hirsch, "An Analysis of the Aurora Large Vocabulary Evaluation," submitted to the European Conference on Speech Communication and Technology, Geneva, Switzerland, September 2003.

[3]   "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm," *ETSI*, April 2000.

[4]   D. Pearce, "Overview of Evaluation Criteria for Advanced Distributed Speech Recognition," ETSI STQ-Aurora DSR Working Group, October 16, 2001.

[5]   N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation," AU/384/02, Aurora Working Group, Dec. 2002 (*http://www.isip.msstate.edu/projects/aurora*).

[6]   J. Picone, "Fundamentals of Speech Recognition: A Tutorial Based on a Public Domain C++ Toolkit," Institute for Signal and Information Processing, Mississippi State University,

Mississippi State, Mississippi, USA, June 2001 (see *http://www.isip.msstate.edu/projects/ speech/software/tutorials/production/fundamentals/current/*.

[7]   J. Picone, "Internet Accessible Speech Technology," National Science Foundation, May 25, 2001 (see *http://www.isip.msstate.edu/publications/reports/nsf_care/2001/*).

[8]   N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 84-107, September 1999.

[9]   K. Huang and J. Picone, "Internet-Accessible Speech Recognition Technology," presented at the IEEE Midwest Symposium on Circuits and Systems, Tulsa, Oklahoma, USA, August 2002.

[10]   B. Jelinek, F. Zheng, N. Parihar, J. Hamaker, and J. Picone, "Generalized Hierarchical Search in the ISIP ASR System," *Proceedings of the Thirty-Fifth Asilomar Conference on Signals, Systems, and Computers*, vol. 2, pp. 1553-1556, Pacific Grove, California, USA, November 2001.

[11]   K. Huang and J. Picone, "Network Builder — An LVCSR Configuration Tool," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip.msstate.edu/projects/speech/software/ tutorials/monthly/2003_03/*).

[12]   I. Alphonso and J. Picone, "Annotation Graphs," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip.msstate.edu/projects/speech/software/tutorials/monthly/2002_09/*).

[13]   "Documentation," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http:// www.isip.msstate.edu/projects/speech/software/documentation/*).

[14]   "Software: On-line Tutorials and Courses," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http:// www.isip.msstate.edu/projects/speech/software/tutorials/*).

[15]   "Introduction to Algorithms," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http:// www.isip.msstate.edu/projects/speech/software/documentation/class/algo/*).

[16]   "Algorithm: Energy," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http:// www.isip.msstate.edu/projects/speech/software/documentation/class/algo/Energy/*).

[17]   "Convolution," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip.msstate.*

edu/projects/speech/software/demonstrations/applets/util/convolution/current/).

[18] "Remote Job Submission," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip. msstate.edu:8080/isip/jsa/index.jsp*).

[19] J. Shaffer, J. Hamaker and J. Picone, "Visualization of Signal Processing Concepts," *Proceedings of the IEEE International Conference on   Acoustics, Speech and Signal Processing*, pp. 1853-1856, Seattle, Washington, USA, May 1998. ICASSP paper no. 2

[20] "Java Applets," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip.msstate. edu/projects/speech/software/demonstrations/applets/*).

[21] "Java Applets: Pattern Recognition," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http:// www.isip.msstate.edu/projects/speech/software/demonstrations/applets/util/ pattern_recognition/current/*).

[22] P.C. Woodland and D. Povey, "Large Scale MMIE Training for Conversational Telephone Speech Recognition," *Proceedings of the Speech Transcription Workshop*, College Park, Maryland, USA, May 2000.

[23] "Speech Recognition System Design Review," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, January 2002 (see *http://www.isip.msstate.edu/conferences/srsdr/*).

[24] "Speech Recognition System Training Workshop," Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, March 2003 (see *http://www.isip.msstate.edu/conferences/srstw*).

[25] C. J. Leggetter, and P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," *Proceedings of the ARPA Spoken Language Technology Workshop*, Barton Creek, Texas, USA, February 1995.

08/15/00 — 08/14/01: RESEARCH AND EDUCATIONAL ACTIVITIES

In the third year of this project, we focused our efforts in two core areas:

- **Java Applets**: overhauled our Java interfaces to use servlet technology, and launched two new applications: feature extraction and recognition.
- **Production System Release**: completed two alpha releases of a new version of the production system that greatly enhance its flexibility and functionality. Extended the interface to support more diverse file formats (both input and output). Integrated our front end application building software.

We also continued activities in the areas central to the overall research program:

- **Foundation classes**: added many algorithms at both the math and signal processing layers of the system. Introduced classes to handle general statistical pattern recognition. Revamped many of the underlying classes to make better use of templates and templatized functions.
- **Workshops**: hosted a software design review in January 2001, and two one-week training workshops held in May ('00 and '01). An impressive collection of on-line resources related to these workshops is publicly available.
- **Software engineering**: upgraded our software development process to use a new problem-tracking tool that was written specifically to deal with bug life-cycle issues. Streamlined our support activities and improved the ease of use of our distribution and verification procedures.

The workshops continue to be extremely successful as demand has far surpassed our original estimates for enrollment (and taxed our facilities). We have seen a dramatic increase in the number of commercial users of the recognition system. In fact, some of our most active users are now commercial users. On-line support has improved, but still remains a challenge given the wide range of experience levels from the users.

A. Java Applets

A key component of our CARE project is the development and operation of an Internet-based job submission facility [1]. This web site allows users to experience speech technology through an easy-to-use web interface that does not require installation of local software. The main reason for this is that the infrastructure required to run a speech recognition experiment can be considerable both in terms of computing resources and intellectual resources (language models, dictionaries, etc.). One important goal for the third year of the project was to bring on-line a large number of PC-based compute servers to serve as the computational engines for this applet, and to make these machines available through this job submission facility.

We have created this facility, and have a large number of servers available for users [2], as shown in Figure 1. These machines have dual-Pentium processors (typically, 600 MHz or faster), large memories (typically 1 GByte), and run the Sun Solaris operating system. Jobs are distributed amongst them by some simple load balancing software that attempts to maintain a reasonable job load on each machine. Our web server handles the initial interactions and job distribution.

As part of our goal to enhance this applet, we embraced the Java servlet technology [3] — a new extension of the Java language designed to make it much easier to pass data to/from a user's browser. Given our needs to do fairly sophisticated client/server communications, and our concerns about security, we needed a more powerful interface than that provided by languages such as html or cgi. Servlet technology, though quite new, has been embraced by web server technology providers such as Apache, and will become an important industry-wide standard
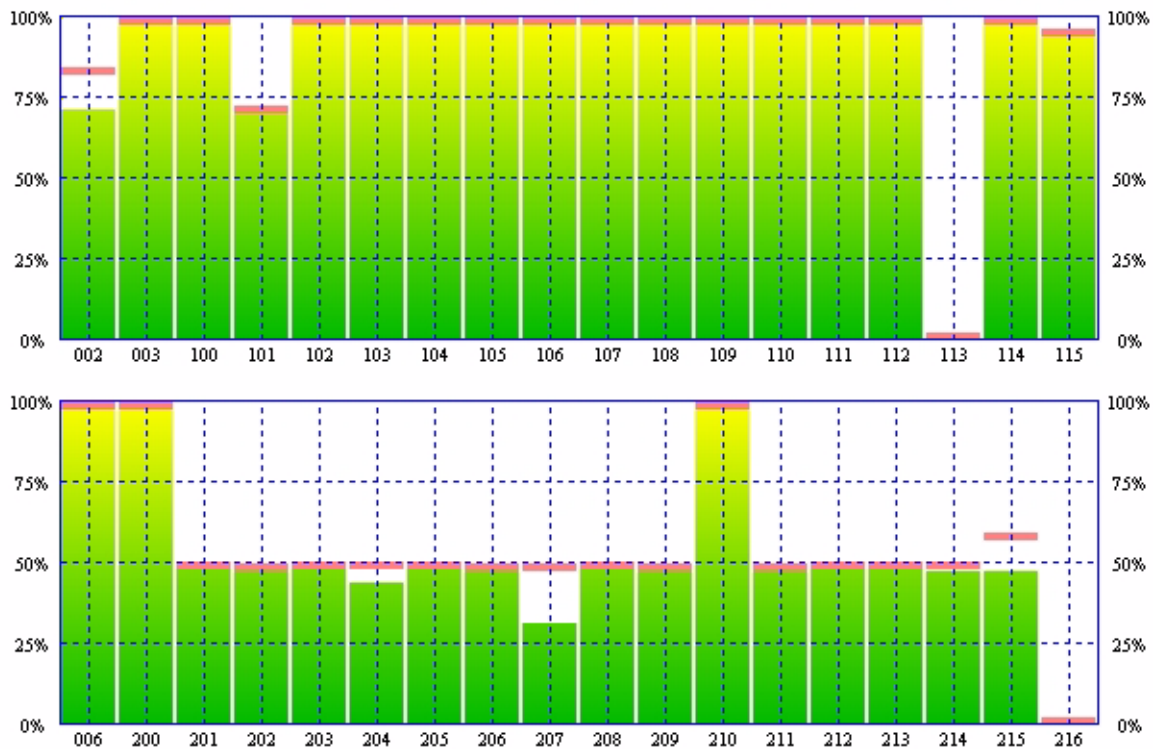
Figure 1. A view of the ISIP remote job processing facility. The machines shown in the bottom row, numbered 200 — 216 are available for remote jobs submitted through the Java applet. These machines represent various generations of dual Pentium processors ranging in speed from 333 MHz to 800 MHz, with a maximum of 1G of memory. The Sun Solaris operating system is used on all machines.

component of a web server. We spent time this year learning this programming paradigm, and reshaping our existing job submission applet to make use of its features. Previously, we used Java's Remote Method Invocation (RMI) protocol, and found it wasn't sufficiently powerful. Servlet technology is now being used in several places on our web site.

The new job submission applet [1] has a dramatically improved interface, as shown in Figure 2. In this version, users select the type of experiment desired from the front page, and then traverse subsequent pages to configure and run the experiment. Users can edit parameters, listen to data, configure the particular data sets to be analyzed, or even upload their own files. Servlet technology has been indispensable to making all this work in a seamless interface. The only downside is that the current version of the servlet compliant server (an interim release by Sun) is a bit slow and lacks robustness. However, an official release of a servlet compatible server by Apache is expected to resolve most of these problems.

The new version of the applet supports two important applications: feature extraction and speech recognition. Feature extraction was introduced as a dedicated application because this task has been one of our most common support problems. Inexperienced users often have trouble settling details such as byte-ordering, sampled data encodings, compression, etc. Hence, we provide users an ability to process standard files, such as the DARPA Switchboard Corpus, or to upload their own files. The latter feature allows users to compare their local installations to our reference implementations on their own data. It also allows us to efficiently interact with them on their specific problems since we can both view (and modify) the same experiment. The interface for the feature extraction mode of the applet is shown in Figure 3.
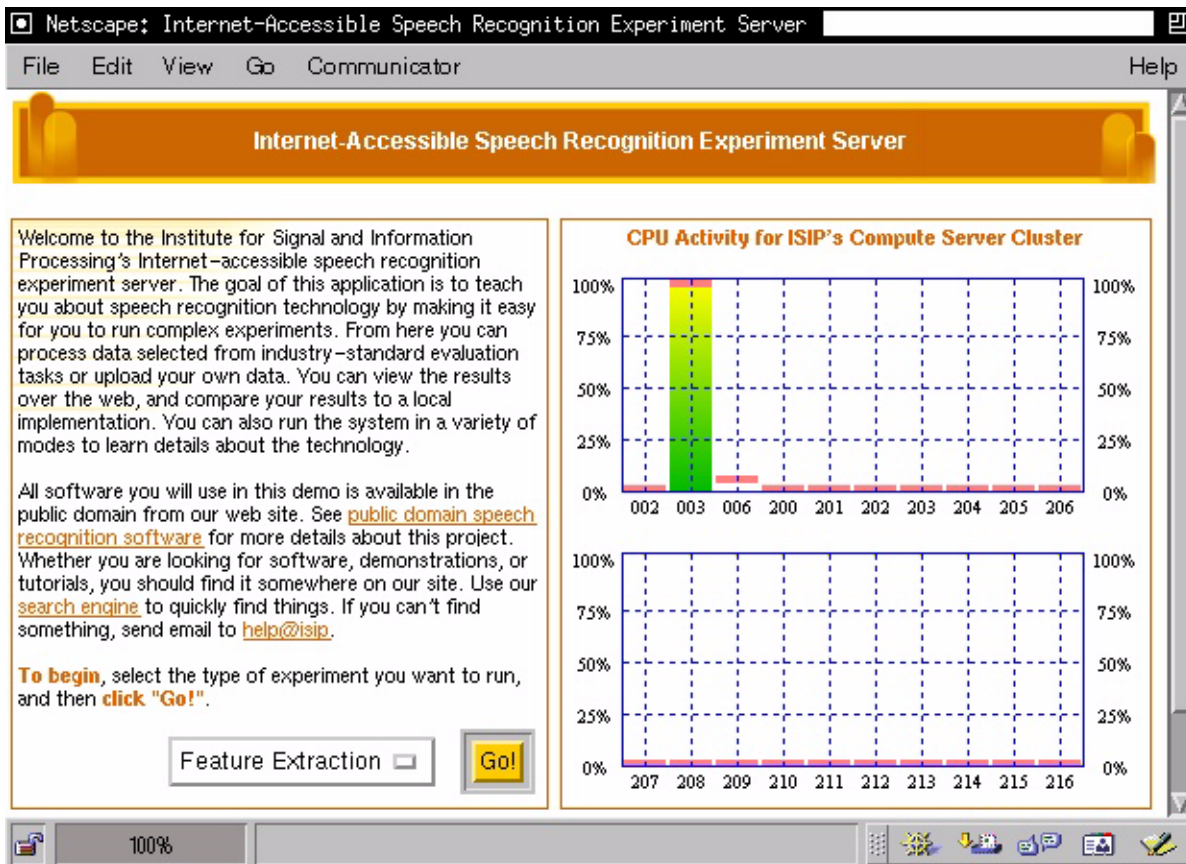
Figure 2. The front page of the remote job submission applet is shown. On the right, users can view the CPU activity of the servers that are accessible. On the left panel, users select the type of experiment. Subsequent menus allow configuration of this experiment and monitoring of the job. All results and intermediate calculations are accessible from the web. Users are emailed when the job completes.

The second application we have introduced is an upgraded version of our recognition decoding applet. This application again allows users to perform recognition on a standard set of prestored utterances, or to upload their own utterances. The main page of this portion of the applet is shown in Figure 4. At this stage, users essentially are selecting a set of models to be used for recognition. Parameters for the recognizer can be modified in subsequent menus. Results are returned in the form of a URL that allows the user to access an experiment directory on our servers. The recognition applet was introduced for the first time at SRSTW'01, and was well-received.

B. Production System Release

The prototype system [4], which is a fully functional speech recognition system, has been in release for over two years as a proof of concept of basic algorithm ideas. It has been used in several formal speech recognition evaluations [5-7], and delivers state of the art performance. In the third year of this project, we have focused on delivering a new version of this system, which we refer to as the production system, that is built on top of our powerful foundation classes. We have made two releases of this version of the system and solicited user feedback on desired functionality. We have also begun running experiments with this system and benchmarking its performance relative to the prototype system.

Figure 3. The feature extraction page allows users to select files from existing databases, or upload their own files. The parameters for the front end can be configured by selecting Edit button in the upper right of the menu. Users can browse their local filesystems when selecting files to upload.

Figure 4. An excerpt from the interface to the speech recognition evaluation applet. This page allows users to select data and run a recognition experiment. Most parameters for the recognizer can be configured from subsequent pages.

The production system is built from a hierarchy of software modules implemented in C++ that provide a wide variety of generic mathematical and data structure-oriented operations. Collectively, these modules are know as the ISIP Foundation Classes (IFCs) [8]. The hierarchy is shown in Figure 5. At the highest level, the production system employs a class called Speech Recognizer, that accepts signal data as input and produces output such as a transcription. This class in turn invokes a pattern recognition class, HiddenMarkovModel, to perform a particular type of speech recognition. These two levels of the system are highlighted in Figure 6.

The Hidden Markov Model class is built upon a powerful search library that implements a dynamic programming-based hierarchical search [9]. This search algorithm has been under development for several years, and provides an ability to emulate a wide variety of approaches to speech recognition through a user-defined hierarchy of state machines. It also mixes network decoding and N-gram decoding in a novel way that will support an interesting mixture of knowledge sources in a single probabilistic format. The alpha release currently only supports context-independent models and network decoding.

The production decoder reads the decoding process configuration from a parameter file specified as a command line option. This parameter file is in Signal Object File (Sof) format. The parameter file contains a full front-end and hidden Markov model (HMM) specification. The HMM specification contains an algorithm and implementation field, the number of levels in the search hierarchy, the model file location and a desired output description. Beam width values can be also specified. A typical parameter file is provided in Figure 7.

Speech Recognition (/asr):	Speech Recognizer, Transcription
Pattern Recognition (/pr):	Hidden Markov Model, Support Vector Machines
Search Algorithms (/search):	Hierarchical Search, History, Search Level, Search Node, Trace ...
Signal Processing (/sp):	Audio Front End, Features, Front End, Featrure Buffer, Recipe ...
Statistics (/stat)	Calculus, Correlation, Covariance, Filter, Prediction, Spectrum ...
Algorithms (/algo):	Histogram, Gaussian Model, Mixture Model, Uniform Model ...
Multimedia (/mmedia):	AudioFile
Shell Interactions (/shell)	Command Line, Debug Level, Filename, Signal Data Base
Data Structures (/dstr):	Circular Buffers, Hash Tables, Graphs, Linked Lists, Set, Stack, Vector ...
Numeric (/numeric):	Bark, LinearAlgebra, Mel ...
Math(/scalar/vector/matrx):	Boolean, Byte, Char, Complex, Float, Double, Short, String, Templates
Input/Output/Files (/io):	Signal Object Files, Lists, Parsers, Name Map
System-Level (/system)	Checksum, Console, Error, File, Integral, Memory Manager, SysChar ...

Figure 5. The ISIP Foundation Classes (IFCs) represent a hierarchy of software modules designed to make implementation and modification of complex systems straightforward. New to this release is the numeric library, which implements some standard mathematical functions.

	LEVEL 0: /asr	LEVEL 1: /pr
Classes	Speech Recognizer [Image Recognizer, Music Recognizer]	Hidden Markov Model [Neural Network, Mixture of Experts]
Algorithms	HMM, NN, NLP	VITERBI, BAUM-WELCH
Features	• opens log file • loads models (algorithm-specific) • loops over files • executes a compute method for a file (algorithm-specific)	• determines the processing mode (decode or train) • creates supervision grammar for training • creates language model for decoding • handles model initialization • performs core frame-based statistical modeling computations • loops over a file and processes each frame

Figure 6. An expanded view of the functionality provided in the top two levels of the production system. It is here that the system branches on fundamental pattern recognition approaches. Such flexibility is one thing that makes the ISIP very unique amongst existing speech recognition systems.

The decoder output configuration can be modified from the command line. We follow a rule that a command line option always overrides any option read from the parameter file. An input file list can be specified from the command line. The three output modes that can be specified from the command line are: single file (FILE), output file list (LIST), and input list transformation (TRANSFORM). The format given below specifies the output mode from the command line:

- *-output_mode* <FILE, LIST, TRANSFORM>: output mode specification can either be a file, a list or a transformed input list

Examples of typical command lines are shown below:

(1) isip_recognize file1.raw file2.raw -p param.sof -output_mode FILE -output_file hypothesis.out
(2) isip_recognize file1.raw file2.raw -p param.sof -output_mode LIST -output_list out_list.sof
(3) isip_recognize file1.raw -p param.sof -output_mode TRANSFORM -output_dir output -dir_pres 3

In (1), the FILE mode outputs all hypotheses to the single output file specified by the *-output_file* command line option. The output hypotheses are written to the file *hypothesis.out.* In (2), LIST mode generates one hypothesis per output file. The list of these output files is specified by the *-output_list* option. The file *out_list.sof* contains the locations of output files.

In (3), the TRANSFORM mode outputs hypotheses to the *N*th subdirectory inside the existing output directory specified by the *-output_dir. N* is specified by the *-dir_pres_level* command line option. The *N* subdirectories are formed as same as the corresponding *N* input subdirectories. For example, if the input file is in */isip/data/examples/**test/boy/b**g/bg_119oo39a.raw*, the output will be placed in the file */output/**test/boy/bg**/bg_119oo39a.sof*. In this case *N* is 3. Here the sub-directory structure **/test/boy/bg** is preserved.

```
@ Sof v1.0 @
# This is the parameters file for the production system

@ FrontEnd 0 @
name = "AudioFrontEnd";
audio = {
  byte_order = BIG_ENDIAN;
  num_channels = 1;
};
frame_duration = 0.01;
target = "BASE&D&A";
output_mode = "SOF_FEATURES_TEXT";

@ Algorithm 1 @
 name = "WINDOW";
algorithm = RECTANGULAR;
 duration = 0.025; alignment = LEFT;

@ Algorithm 2 @
name = "ENERGY";
implementation = LOG;
.
.
@ Algorithm 17 @
 name = CoefficientLabel;
 variable = "BASE&D&A";

@ DiGraph<Long> 0 @
weighted = true;
vertices =
      {0, {0}},
      {1, {1}},
      {2, {2}},
.
arcs =
      {0, 1, 0},
      {0, 2, 0},
      {1, 3, 0},
.
@ HiddenMarkovModel 0 @
algorithm = "DECODE";
implementation = "VITERBI";
num_levels = 3;
model_file = "$ISIP_DEVEL/doc/examples/data/models/tidigit_preview.sof";
output_mode = "FILE";
output_file = "hypotheses_302010.out";

# beam width at the 2-nd level
@ beam_2 0 @
value = 300;
# beam width at the 1-st level
@ beam_1 0 @
value = 200;
```

Figure 7. An example of the Sof version of the recognizer parameter file. This format is temporary and comprises a representation of the core information. More flexible formats supported by generalized parsers will be available soon.

Additional facilities accessible by the command line specification are:

- *-help*: displays a help message,

- *-verbose* <NONE, BRIEF, DETAILED, ALL>:   specifies the different verbosity levels about the decoding process,

- *-verify*: verification mode flag - no decoding is done in this mode. This mode verifies the availability and format of all required input files. It also generates output files.

- *-log_file* report.text: system generates the log file instead of printing information about the decoding process to console,

- *-debug_level* <NONE, BRIEF, DETAILED, ALL>: debugging information level - specifies the amount of debugging information generated by the system.

These modes have been provided to facilitate debugging and to minimize the number of user configuration errors. Verify mode is particularly important because time-consuming runs can be tested before a large amount of time is invested in an experiment. This mode is modeled after the Unix "make" facility.

Performance of the production system is comparable to the prototype system, which is extremely encouraging given the increased overhead of the production system (due to its flexibility). It is hard to directly compare these results to the prototype system since there are significant differences between the two. For example, the production system does feature extraction internally while the prototype system does feature extraction with a separate utility. Nevertheless, we can note that the production system is only about 50% slower than the prototype system on decoding of TIDigits, which is very respectable given the flexibility of the system. Further, the production system is much more efficient in terms of memory use than the prototype, because the prototype system in network decoding is not able to effectively prune state-level histories. The prototype system requires at least an order of magnitude more memory for most conditions.

C. Foundation Classes

An overview of the structure of the foundation classes is shown in Figure 5. This year we have focused on higher-level libraries such as Search and Pattern Recognition, which are integral parts of the new version of the recognition system. In the process of developing these, we made significant modifications to our Algorithm classes, which are the core the signal processing components of the system. These modifications, in turn, allowed us to significantly simplify some of our supporting tools, such as the TransformBuilder show in Figure 8.

In the process of advancing the higher levels of the system, we have also been able to simplify lower levels of the system using some newer features of C++. In an effort to consolidate the features of C++ we are dependent upon, we developed an on-line tutorial of aspects of C++ that we find extremely important and useful [10]. This tutorial leads users through the syntax, provides examples linked to actual ISIP code, and references sections from a definitive reference textbook in C++ [11] for further reading. This tutorial has been a very useful instructional aide for our undergraduate student programmers, and was actually authored by one of our undergraduates.

Through a combination of newer features of the language that are now supported by the compilers we use (GNU/EGCS gcc), and some novel code generation techniques implemented in Make files, we have been able to dramatically reduce the code size of our lower level classes by relying
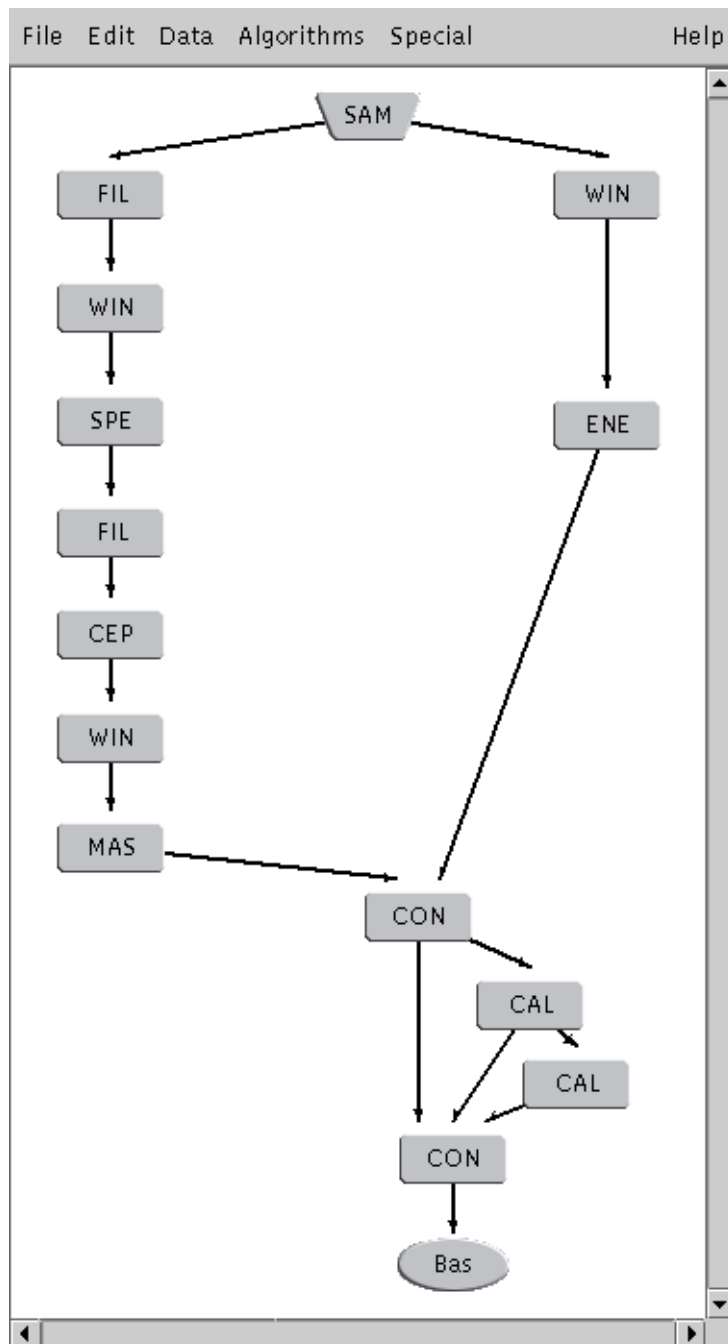
Figure 8. An example of a front end created using our interface building tool. Users can flow chart algorithms and generate the required code automatically.

on the make process to generate data-specific versions of the class. This has made maintenance and support of these classes much easier, and made the entire Math class package much more elegant.

We had also been struggling recently with the extensibility of the Algorithm classes in terms of issues such as operation of these classes on abstract data types. For example, some classes, such as those that perform linear prediction analysis, can operate on several types of data (e.g., signals, autocorrelation coefficients, and prediction coefficients), can produce different types of outputs (filters, temporal sequences, frequency domain representations, etc.), and can use different algorithms to perform such an analysis. Our most recent attempt at supporting this type of functionality uses explicitly defined input and output data types [12], and appears to be sufficiently flexible. We are now able to prototype a wide range of front ends using the tool shown to the left in Figure 8, and to implement them using our generic signal processing tool *isip_transform*.

These components are now integrated into the speech recognition system as well. The recognition system is capable of performing recognition on a variety of input streams, including sampled data files, industry-standard audio formats, and feature vectors. The latter format is extremely important because it allows non-speech recognition users to use this technology for other pattern recognition applications. The integrated front end also allows users to generate derived feature vectors, such as derivatives of features, inside the recognizer, rather than storing these in external files. Considerations such as these result in significant disk space savings, and improve ease of use of the system.

D. Workshops

In January 2001, we hosted our second annual design review workshop [15]. Attendance for this workshop tends to be light, and included 14 participants representing 6 universities and 5 industrial sites. The primary functions of this workshop are to review our software design and to solicit the community for input on new features. The workshop is a two-day event that includes a half-day of training. A complete archive of the workshop, including presentation slides, is available on-line at the workshop web site [15].

Feedback from this workshop was generally positive. Some examples are shown below in Figure 9. Several good suggestions flowed from this workshop, including a need to place a greater emphasis on reproducible benchmarks. The signal processing software and foundation classes continue to be an attractive component of the system, which is probably more an indication of the type of people attending the workshop (engineers vs. computer scientists) than the relative success of a particular component of the system.

Our second workshop, which is a major focus of this research program, is a one-week training workshop geared towards entry-level graduate students pursuing research into speech. Last year, when I submitted this report, it was prior to holding our first summer workshop. This year, I delayed the submission of this report until after we hosted our second workshop. We now have had two very successful summer workshops, and are in a position to comment on the effectiveness of this forum. Of course, all materials related to these workshops, including lecture notes and lab exercises are on-line at the respective web sites for these workshops [13-16]. Feedback for the two

...SW engr. was cool. but i think you should have focused less on that and more on the SW design of the Production SR. of course that's from a SW Engr. perspective...

...I don't think my current research effort would be at all feasible if I didn't have the ISIP system to build on. I'm very pleased to have it available. The workshop helped me get a better understanding of how the system is put together and the constraints under which it has been / is being built. I would have liked to have spent more time on the IFCs, though. ...

...Workshop was very informative. I was really impressed with how well your students conducted each session of the workshop. Now, relating to the SRSD, as a novice in this area, I felt well informed and gained an enormous amount of knowledge about the system...

...It is very interesting to see the evolution of the ISIP system over the past few months. A lot of work was done to improve the functionalities and the reusability of the classes...

...Congratulations for a very well organized workshop. I found it informative, helpful and enlightening. Although I have used the ISIP recognizer before, now I have a much better understanding of it overall...

...I thought the workshop was very informative! All of the speakers were very knowledgeable and extremely helpful! Looking forward to using the software at our site....

...Good workshop. Excellent effort. We all anxiously await the production system. I would like to see more about the class hierarchy and learn where to find classes/methods in the source tree...

Figure 9. Samples of feedback from SRSDR'01. The second year of this workshop was a much more efficient and effective operation.

summer workshops is also available on-line at [17,18]. This year, we even broadcast live images from the workshop using our web camera.

The format for the second summer workshop followed closely the format for the first workshop (half-day lectures and half-day labs). However, we made significant changes to the laboratories this year based on the feedback from last year's workshop. Labs this year were much more self-paced and open-ended, and required active participation from the users. From the feedback this year, it is clear the lab sessions were vastly more successful. On the other hand, we did have some problems with some mechanics related to downloads and installation of code. Based on past experiences, we felt it was a good idea to have every participant download and install code. Unfortunately, 24 simultaneous downloads from our servers taxed the network and took longer than expected. Participants felt this was unproductive time. We need to reevaluate our process for educating users about the finer points of installation and configuration.

Participation in the summer workshop this year was dominated by industry relative to last year. SRSTW'00 included 18 participants from universities and 4 participants from industry. SRSTW'01 included 9 participants from universities and 15 from industry. This skew occurred despite heavy marketing of the workshop to graduate students through various forms of electronic announcements. One possible explanation is that our system has matured to a point where it is attracting significant industry interest. There is, of course, truth to this as a number of participants this year explicitly discussed our interest in such commercial endeavors, and we know of several companies developing products around pieces of the system.

On the other hand, it is clear that we aren't quite as successful as we had hoped in meeting our goal of increasing the graduate student population in speech. To do this, we need to increase the number of graduate students attending the summer training workshop. Discussions with students who attended this workshop seem to indicate we need to do a better job advertising the workshop to graduate students. What the best forum is to reach them remains a question. This year we sent email messages to the department heads at the top-100 universities in the country, as well as 100 historically black institutions (HBCUs) with engineering to computer science programs (in addition to posting to all relevant newsgroup, web sites, and community-wide mailing lists). Apparently, these messages targeted at department heads are not getting forwarded to the students. We need a better approach to reaching students. We are considering surface mailing printed announcements.

E. Software Engineering

If there is one statement in this report that rings true, it is the following: "With every release of our system, we accumulate more experience with the challenges of supporting research software in a Unix environment, where things tend to be less standardized." Last year, we took major steps forward by introducing two key technologies into our software process: a formal report tracking system and automatic configuration during installation. Both had profound impacts on our ability to improve the quality of our software and the efficiency with which we worked. However, the report tracking software proved to have to major drawbacks: slow response time for a large database due to inefficient code, and no comprehension of the bug tracking and resolution process in the basic interface.

Hence, we embarked on an effort to develop a simple, highly-customized bug tracking tool known as Varmint [19]. Before doing so, we evaluated several industry-wide packages, and carefully

analyzed successful processes used in leading software companies (such as Microsoft, where several of our students worked as interns in Summer'00). Once we determined that none of the existing tools met our needs (and that many companies use expensive commercial packages or proprietary internal packages), we had two undergraduate programmers develop a tool. This tool is now publicly available as part of our software distribution, and has been successfully ported to at least two other groups within our university.

A screenshot of the tool is shown in Figure 10. This tool maintains bugs using a one database per project format. The primary screen, shown in the background in Figure 10, displays all current bugs and their status. Users can program their own queries using SQL-like commands ("Show me all the open bugs in the current release."). Bugs can be sorted in many modes. Perhaps the most important view of the bug list is the view of all open bugs for the current release. It is part of our
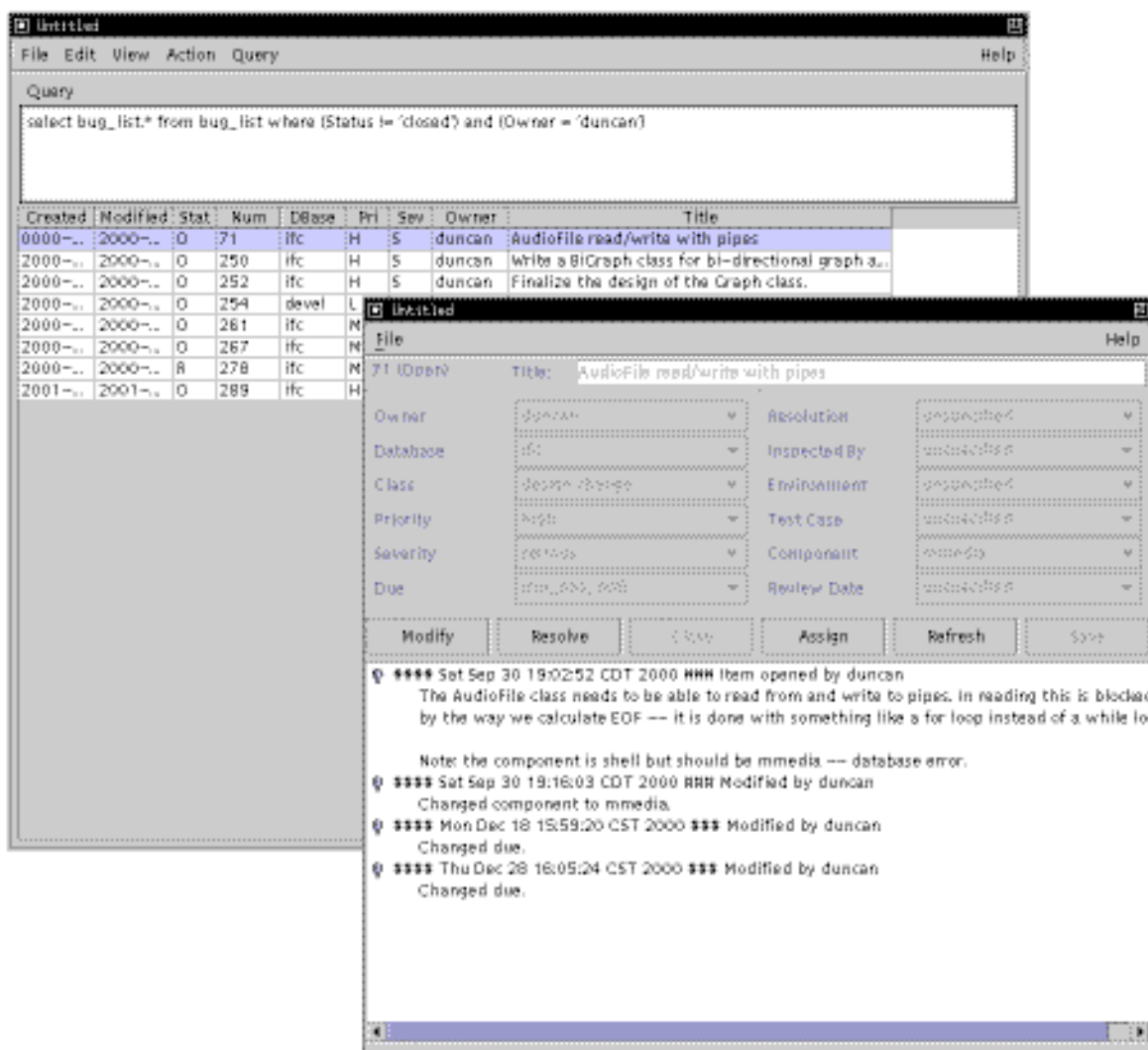
Figure 10. Screenshots from our bug tracking software tool known as Varmint. This tool is written in Tk/Tcl for portability and uses a database tool (such as Oracle or mySQL) as a back-end. Its main virtue is that it is customized to support tracking and resolution of a bug involving multiple programmers, and yet does this with a minimum of operating system infrastructure. The net result is a tool that is fast, efficient, easy to use, and equally important, easy to install.

software development process to resolve all bugs before making a release. This tool has been invaluable for managing open bugs, and for helping our student programmers to maintain a clear vision for their priorities (a time management issue).

Like all good software, the Varmint tool has found its way into many other aspects of our project management. We now organize all request management in our lab, including support requests from help@isip and web site maintenance, using this tool. The ability to track and manage this diverse set of needs from a single tool has proven to be invaluable.

F. REFERENCES

[1]   K. Huang and J. Picone, "Internet-Accessible Speech Recognition Experiment Server," *http://www.isip.msstate.edu/projects/speech/experiments*, Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[2]   K. Huang and J. Picone, "ISIP CPU Load Statistics," *http://www.isip.msstate.edu/data/statistics/cpu_stats*, Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[3]   J. Hunter, *Java Servlet Programming*, O'Reilly and Associates, Cambridge, Massachusetts, USA, 1998.

[4]   J. Hamaker, R. Duncan, N. Parihar, and J. Picone, "A Public Domain Speech Recognition System," *http://www.isip.msstate.edu/projects/speech/software/asr/download/asr,* Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[5]   R. Sundaram, A. Ganapathiraju, J. Hamaker and J. Picone, "ISIP 2000 Conversational Speech Evaluation System," Speech Transcription Workshop, College Park, Maryland, USA, May 2000.

[6]   B. George, B. Necioglu, J. Picone, G. Shuttic, and R. Sundaram, "The 2000 NRL Evaluation for Recognition of Speech in Noisy Environments," presented at the SPINE Workshop, Naval Research Laboratory, Alexandria, Virginia, USA, October, 2000. SPINE

[7]   R. Sundaram, J. Hamaker, and J. Picone, "TWISTER: The ISIP 2001 Conversational Speech Evaluation System," Proceedings of the Speech Transcription Workshop, Linthicum Heights, Maryland, USA, May 2001.

[8]   R. Duncan and J. Picone, "The ISIP Foundation Classes," *http://www.isip.msstate.edu/projects/speech/software/asr/download/ifc*, Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[9]   N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," IEEE Signal Processing Magazine, vol. 16, no. 5, pp. 84-107, September 1999.

[10] J. Langley and J. Picone, "Practical C++ ISIP-Style," *http://www.isip.msstate.edu/projects/speech/education/tutorials/c++*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, August 2000.

[11] S. Lippman and J. Lajoie, C++ Primer, Third Edition, Addison Wesley, Reading, Massachusetts, USA, 1998, ISBN: 0-201-82470-1.

[12] S. Srivastava, R. Duncan, and J. Picone, "The Algorithm Classes," *http://www.isip.msstate.edu/projects/speech/education/tutorials/isip_env/class/algo*, Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 2001.

[13] J. Picone, "Speech Recognition System Design Review", *http://www.isip.msstate.edu/conferences/srsdr00*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, January 2000.

[14] J. Picone, "Speech Recognition System Training Workshop," *http://www.isip.msstate.edu/conferences/srstw00*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, May 2000.

[15] J. Picone, "Speech Recognition System Design Review", *http://www.isip.msstate.edu/conferences/srsdr01*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, January 2001.

[16] J. Picone, "Speech Recognition System Training Workshop," *http://www.isip.msstate.edu/conferences/srstw01*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, May 2001.

[17] K. Muir and J. Picone, "SRSTW'00 Feedback", *http://www.isip.msstate.edu/conferences/srstw00/misc/feedback/feedback.html*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, May 2000.

[18] J. Langley and J. Picone, "SRSTW'01 Feedback", *http://www.isip.msstate.edu/conferences/srstw01/misc/feedback/feedback.html*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, May 2001.

[19] R. Duncan, R. King, and J. Picone, "Varmint: A Bug Tracking and Resolution Tool", *http://www.isip.msstate.edu/projects/speech/education/tutorials/isip_env/util/devel/varmint*, Institute for Signal and Information Processing, Mississippi State University, Mississippi, USA, May 2001.

08/15/99 — 08/14/00: RESEARCH AND EDUCATIONAL ACTIVITIES

In the second year of this project, we focused our efforts in five major areas:

- **Production System Release**: the first release of the production speech recognition system, based on our modular libraries, is scheduled for July 1.

- **Hosted two workshops**: a software design review held in January 2000, and a one-week training workshop held in May 2000.

- **Software engineering**: upgraded our distribution to use the autoconf facility, added an automated report tracking system to our on-line support, created a multi-platform support facility.

- **Foundation classes**: added algorithms and other signal processing building blocks, introduced classes to handle acoustic models, search algorithms, and knowledge sources, and released a front-end that allows arbitrary algorithms to be implemented using a graphical user interface.

- **Java Applets**: enhanced our pattern recognition applet with several important new features, including generation of arbitrary data sets, clustering, and visualization of decision surfaces.

The workshops appear to be extremely successful as demand has far surpassed our original estimates for enrollment (and taxed our facilities). The number of serious users of the recognition system is continually growing. It is becoming a challenge to provide same-day response to most support requests, particularly given the wide range of experience levels from the users.

A. Production System Release

An overview of a typical speech recognition system is shown in Figure 1. There are three main components to this system: signal processing, language modeling, and search. We have had a prototype system in release now for over one year. This system was recently evaluated as part of DoD's yearly evaluation cycle [1] — an important step towards gaining wider acceptance of the system as a state of the art system. We are now nearing the first major release of our production system that is built from the ISIP foundation classes. We currently have many of the core pieces implemented, including the signal processing section (described later), acoustic modeling, and a prototype hierarchical search engine that was demonstrated at our January workshop. Language modeling classes are currently under development and nearing completion. Integration of these classes into a system has begun, and is expected to be completed by mid-summer.

Novel aspects of this system include a generalized hierarchical search engine, shown in Figure 2, and a flexible approach to
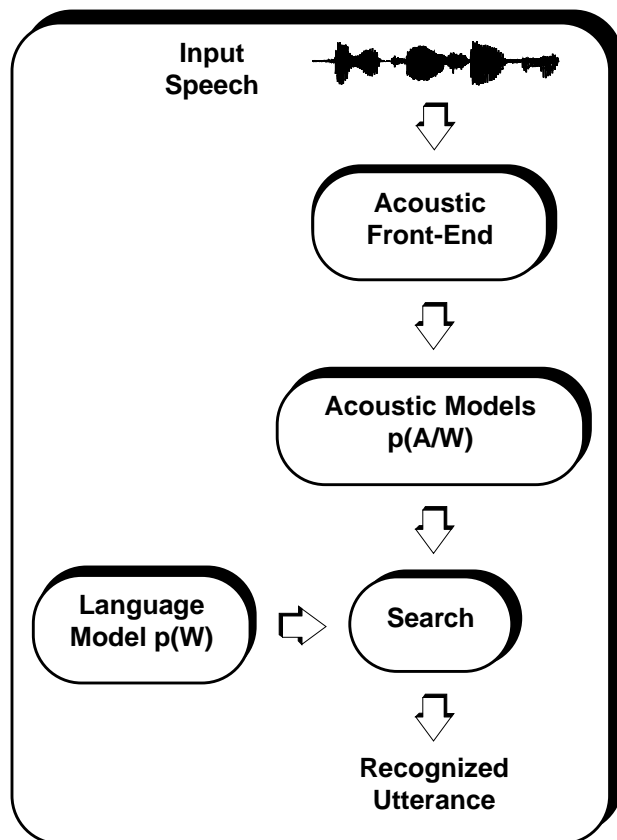
Figure 1. A typical speech recognition system.

signal processing that allows new algorithms to be implemented using a CASE-based approach involving a graphical user interface. The search engine is crucial to this system, in that it is by far the most complex and unwieldy component. A clean implementation that provides users reasonable programming hooks into all levels of the process is very important. The generalized approach below still requires significant work with respect to efficiency and memory resources. We are using the prototype system currently in release to develop these details, and then transferring that knowledge into the production system.

Pieces of the production system, particularly the foundation classes, have been in release since November 1999. The most current versions of the code are also available from our anonymous CVS server. The foundation classes are slowly stabilizing as we add more upper-level functionality and expose more bugs. The C++ language definition and implementation has recently begun to stabilize, making many things possible using the latest version of the compilation tools. This in turn has allowed us to change several aspects of our class designs. We believe we have reached convergence on most major aspects of the system design, and now plan to remain backwardly compatible with subsequent releases. We have also developed tools to automatically convert data formats between the prototype and production systems, thereby allowing users to leverage features of both systems while the latter is under development.

B. Outreach Via Workshops
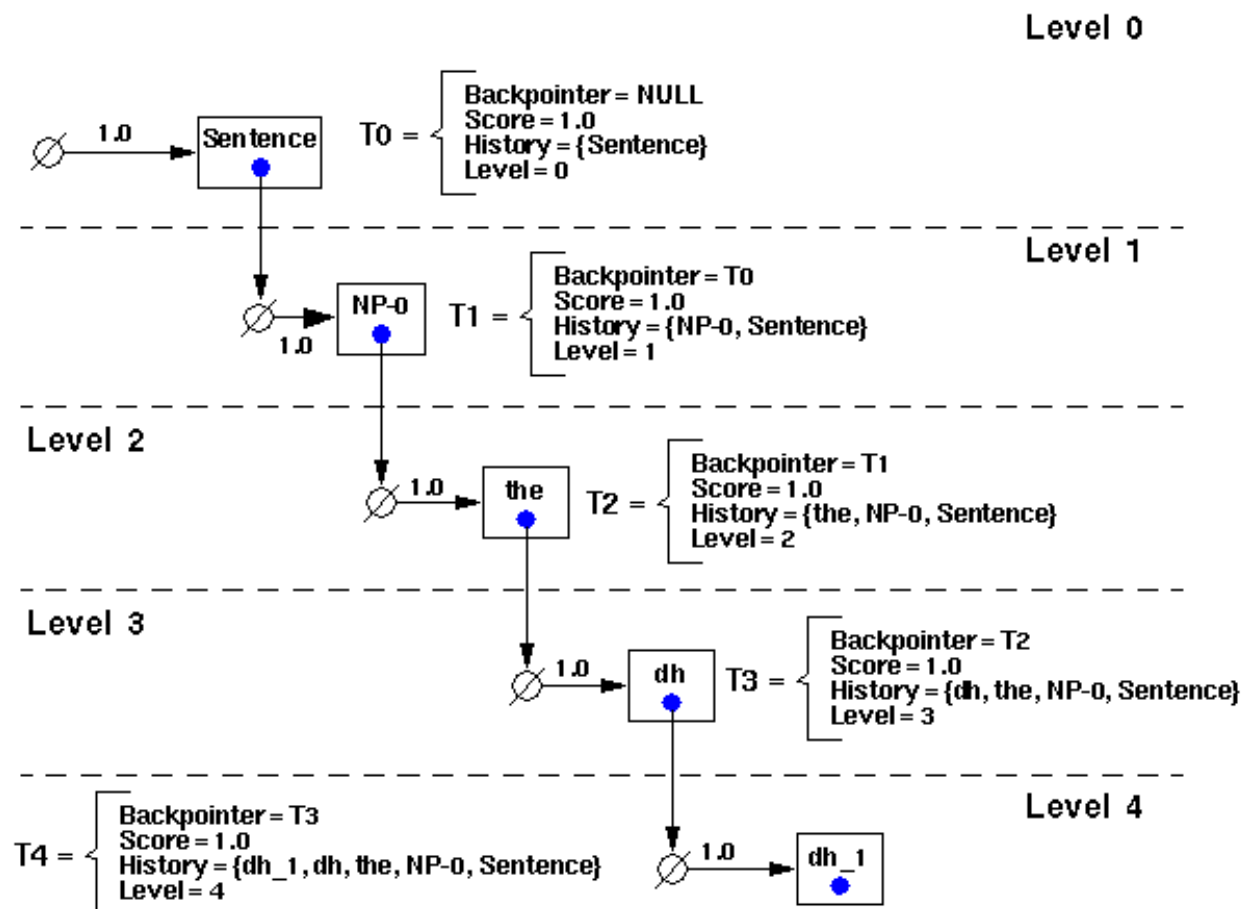
Figure 2. An overview of a generalized search engine that allows users to implement speech recognition systems as a hierarchy of knowledge sources.

In January 2000, we hosted our first annual design review workshop. The primary function of this workshop was to review our software design. A second important goal was to solicit the community for input on new features. We arranged this workshop to be a two-day event. The first day was devoted to an overview of the system, including demos, and general discussion. The second day consisted of a training session where we walked users through our on-line tutorial. A complete archive of the workshop, including presentation slides, is available on-line at the workshop web site [2]. Participants received notebooks containing handouts of the lecture notes, as well as a CDROM containing all software and instructional materials used at the workshop.

Attendance at this workshop was slightly lower than expected: 9 participants from several foreign countries (China, Finland, Korea), government agencies (FBI, DoD), and industrial sites (IBM, MITRE, Lincoln Labs). This was partially due to the time at which the workshop was held (January 5) and concerns about residual Y2K problems. Nevertheless, it was a very productive workshop in that users were able to build an entire large vocabulary continuous speech recognition (LVCSR) system during the training session, and left feeling very good about the software. Several collaborations resulted from the workshop, including invited talks at IBM, a collaboration with Georgia Tech on the classroom of the future [3], and potential collaborations with MITRE on various DoD-related speech recognition applications.

Feedback from this workshop was generally positive. Some examples are shown in Figure 3. We were particularly interested in thoughts about the summer workshop, and action items for the following year. A summary of the discussion about future plans is available [4] on the web. Given the diverse group of participants, it was hard to form a consensus on the priorities of these items. However, generally speaking, there were no surprises relative to our current plans.

In May 2000, we will offer our first extended training workshop [5], which is geared towards entry-level graduate students. Travel funds are provided to encourage graduate students from underrepresented institutions to attend. The program [6] for this week-long workshop combines morning lectures on theory with afternoon laboratories focused on skill-building. The morning lectures are split into two parts: fundamental theory and applications to speech recognition (explanations of how the theory is actually implemented in a system). The laboratories involve skill-building projects ranging from basic recognition foundation class programming to conversational speech recognition system development. Participants literally leave the workshop with a toolkit to run some of today's important research tasks, and should be able to make programming-level modifications to the system.

Twenty-four participants, including 18 graduate students, will attend the first workshop. Nineteen institutions from seven countries are represented, including by design a broad range of U.S. universities. Established research groups such as MIT, Rutgers and University of Colorado at Boulder are represented. Underrepresented universities such as North Carolina A&T are also participating. Further, universities less prominent in speech recognition research, such as University of Houston, University of Denver, and Old Dominion University, are represented by graduate students early in their Ph.D. programs. Hence, we are well along towards our goal of increasing access to speech technology by incorporating underrepresented groups. In fact, workshop enrollment was three times what was originally expected, and our acceptance rate was approximately 66% of the applicants.

We plan to broadcast live still images from this workshop on the conference web site [5] this year using a networked camera. Next year, we will attempt a live Internet broadcast using facilities

1. Things I liked: Clearly structured presentation of the system. Presentations covered most aspects of the system. Code-testing (e.g. the diagnose() function) was stressed. Demo. Instrumental in inspiring confidence in the system.

2. Things that need improvement: In the demo, most of my time was spent typing in long pathnames. I suggest writing a (one-line) shell script for each step, and simply allowing the user to read the pathnames.

3. Things in a summer training workshop: Make a list of use-scenarios (e.g. word-lattice-construction; lattice-rescoring; 1-pass decoding; viterbi-training; EM; segmentation; alignment; speaker adaptation; system-extension with new acoustic models like SVMs, etc.) and go over how to do each one.

4. Things to do differently: I would spend a bit more time on some of the tougher algorithmic issues, e.g. how exactly right-word extensions work for cross-word context dependence, and what happens if you want to look (say) 5 phones to the right. And when traces are extended, presumably there are some circumstances where two traces can be merged; how exactly is this handled? Basically, I'd like to get a more explicit sense of what the toughest issues were, and how the system handles them. Perhaps a 1/2 hour on this.

Overall, I thought it was an excellent review.

---

1. Things you liked about the workshop:
The workshop went well, and met my goals:
a. get to know the ISIP group to facilitate collaboration
b. get the software up and running
c. get a feel for where the project is going
d. articulate my needs

2. Things that need improvement: Hmmm, having trouble thinking of practical changes....
Maybe a tutorial on "current issues in speech recognition"

3. Things you would like to see in the summer training workshop: I plan to send students not experienced in speech recognition, so tutorials would be useful.

4. Things you would like to see us do differently for the design review:
see 2.

I appreciate it very much that I had the opportunity to come to MSU to learn automatic speech recognition (asr).   I have done survey on the availability of asr software which I can apply LDC DARPA-TIMIT data (on cd-rom) for training an asr system. But I did not have any one until I talked Mr. Dave Graff of LDC who suggested that I talked to you.

The hospitality you extended to me is sincerely appreciated. I have done some paper-reading on speech recognition years ago. But this is the first time I am trying to use it for data systems. The need at my organization is speaker-independent asr.

Most of the telephone data is conversational, I will try to work on Switchboard-type of data, hopefully in the summer.

I tended to think that it would be a nice thing for a user if the input and modules can be simplified. I am an engineer at my job; the real end-users may not be engineers.

Figure 3. Examples of feedback collected from SRSDR'00. Comments about topics for the extended summer workshop we extremely helpful and have been incorporated into the program.

available at MS State (at no charge to the project). All materials developed for the workshop, including laboratory exercises, will be posted on the web site. We have already had several requests for these materials from people who cannot attend the workshop.

C. Software Engineering

With every release of our system, we accumulate more experience with the challenges of supporting research software in a Unix environment, where things tend to be less standardized. With the addition of a full-time software engineer this year, we were able to address these issues in a much more powerful manner. One of the most popular distribution mechanisms for software in Unix is an automatic configuration system developed by the GNU organization. This system, known by various names such as configure and autoconf, automatically searches the system during installation for required software packages, and configures software accordingly. A typical installation procedure consists of the following sequence of commands:

- tar xvf isip_proto_v5.3.tar   unpacks the software distribution
- cd isip_proto_v5.3:   enters top-level directory
- configure --prefix=/usr/local/isip configures the software and sets the installation directory
- make   compiles and links the software
- make install   installs the software

This deceptively simple procedure has taken years of refinement for Unix systems, and involves locating many important tools (gcc, perl, Tk/tcl, shells, etc.), and deciding how best to build the software given the local system's capabilities. In recent years, installation using this approach has become fairly smooth under a multitude of Unix systems.

The overhead cost in adopting this form of installation procedure is high. The tools to do this are not trivial. This year, we finally mastered this software and incorporated this facility into our releases of the prototype system. This should resolve most support issues we have dealt with involving system incompatibilities, and definitely minimizes the effort required by users to install the system (since everything is automatic). Migrating our previous installation procedure required a major overhaul, but was clearly well worth the effort.

As mentioned previously, support activities are requiring an increasing amount of time. Hence, it became clear that we needed to install a formal method of support request tracking. We have installed a public domain system called RT — Request Tracker [7]. This is a powerful system that has most of the standard features included in such packages: ticket numbers, time-stamping of requests, automatic acknowledgements, queues, and resolution tracking. RT is popular, particularly within our university. We are able to leverage other installations on campus, and enjoy excellent technical support on the package from other campus organizations. RT has been extremely useful in managing our support line. For example, we are now able to generate automated reports on the timeliness of our service. Any email to our support line, help@isip.msstate.edu, is automatically routed to the RT system and acknowledged. Our goal is to provide a reasonable response to each request within a 24 hour period when support staff are on-site. Our software engineering staff position manages this system as part of his job responsibilities.

A third step we have taken to improve the quality of our distributions involves the development of a multi-platform and multi-OS environment to check releases for compilation and run-time problems. We purchased a Pentium workstation and installed multiple operating systems: Sun

Solaris x86, Windows NT, and Linux. We also have Sun Sparc Solaris machines in our lab. Further, we recently acquired an IBM AIX machine as a donation from IBM. These machines are used to check every release before it is actually made available to the public. Though we tend to be very methodical in our debugging and validation methods, we have found instances where software will pass validation runs on all but one of the operating systems. Such problems are subtle and rarely flagged by compilers (even though we use a common compiler across all machines). Though such multi-platform checks are time-consuming, they are necessary if one wants to avoid problems. Linux support, in particular, has recently become a critical requirement.

Along similar lines, we have also recently acquired professional strength code development tools for Unix. Previously, we have been relying on a public domain code checker — dmalloc. Unfortunately, software of the complexity we are developing breaks most public domain tools. Such tools are not able to properly diagnose and isolate problems. Hence, we now have access to two professional quality development tools offered by Rational Software. Even these tools do not catch 100% of the problems observed in our code, primarily due to the complications that arise from the use of many levels of C++ templates. However, such tools are often able to help us resolve problems in minutes rather than hours, and have greatly increased productivity.

Using these tools, we were able to isolate and fix a number of memory bugs in our current releases. Some of these were quite subtle and took hours of run-time to reproduce. However, our current releases are now free of all known memory bugs, and are vastly improved over previous releases. The software has been checked on a much wider range of tasks as well.

D. Foundation Classes

The foundation classes, upon which all higher-level software is built, continue to grow in terms of their breadth and depth. We currently support the following libraries in our class hierarchy:

- system (i.e., Console, MemoryManager)
- input/output (i.e., Signal Object File, Sof Parser)
- math (i.e., Scalars, Vectors, and Matrices)
- data structures (i.e., Linked Lists, Hash Tables)
- shell (i.e., CommandLine, Filename)
- multimedia (AudioFile)
- statistics (GaussianModel, StatisticalModel)
- algorithms (Cepstrum, Linear Prediction)
- signal processing (FrontEnd, Features)
- pattern recognition (PCA, LDA)
- automatic speech recognition (Recognizer).

This year, we have focused on higher-level libraries such as Statistics, which provides statistical models for each state in our acoustic models, and Data Structures, which provides graph objects used in the search engine.

Our recent focus has been the development of the signal processing portion of the system. An overview of the tool we have developed to provide users an easy way to build signal processing systems is shown in Figure 4. The users have at their disposal any of the tools available in our Algorithms library. For example, an industry-standard front-end uses a Fourier Transform operation, a Cepstrum calculation, time derivatives of feature vectors, and a special type of normalization algorithm. Each of these modules is available as a class under the Algorithms library. Each class has a special set of methods that interface to the application builder, known as

Inputs   **Recipe Pool**   **Outputs**
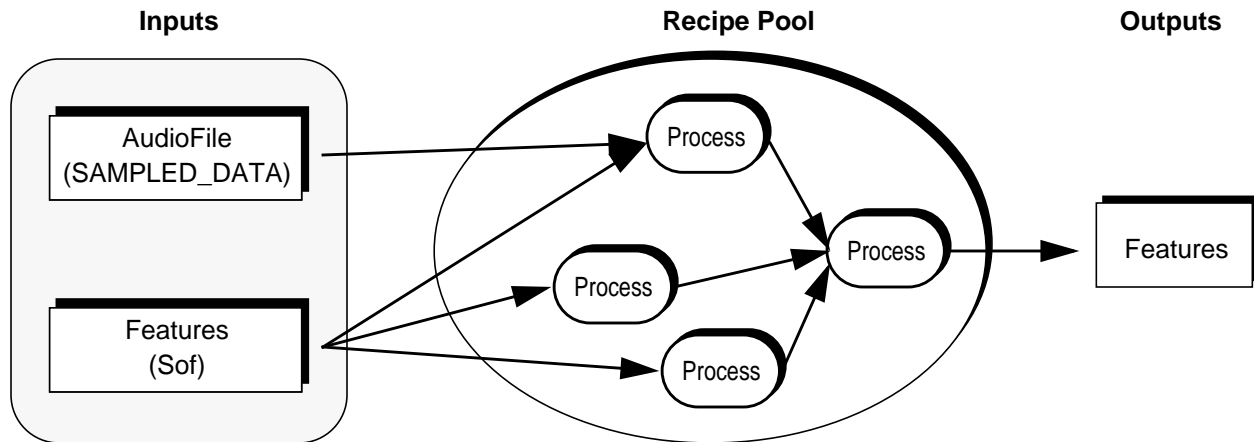
Figure 4. An overview of a CASE-based tool that implements signal processing algorithms using a GUI-oriented tool. Users essentially flow-chart an algorithm as a graph of recipes. The system automatically schedules operations to implement the desired algorithm.

the Transform class (the corresponding utility is isip_transform).

Our Transform class combines the user's algorithm specifications by building a graph depicting the sequence of algorithms to be applied to the signal, as shown above in Figure 4. Such block diagram type approaches to signal processing have been popular for a number of years in signal processing. Transform is a very powerful class in that it allows users to mask, combine, and postprocess measurements of the signal in arbitrary ways. Since the internal structure of this class is somewhat complicated, a graphical user interface is essential. Rather than have users edit a parameter file containing information about the algorithms and their interconnections, users can manipulate this representation using graphical tool. We decided to implement this in Java to make it as portable as possible. Our first release of this tool will coincide with the summer workshop.

We have also made significant progress towards the development of the production recognition system by implementing statistical modeling aspects of the system. The decoder portion of a speech recognition system can be regarded as a hierarchy of graphs [8]. The leaf nodes of the lowest level of this hierarchy are states in a hidden Markov model. Our StatisticalModel class implements a generalized state, which can be an arbitrary mixture of distributions — Gaussians, Exponentials, Laplacians, etc. Mixtures of Gaussian distributions are most popular in speech recognition today; exponential models are becoming increasingly popular for some aspects of the problem (they are rooted in maximum entropy theory).

We have begun tying these together to build a full-fledged recognition system. We have also created conversion utilities that transform outputs of the prototype system into formats accepted by the new system. This is an important capability since it allows us to interface the two systems and maintain some level of backward compatibility. It also makes the development cycle more efficient, since we can incrementally test isolated components of the new system on large enough tasks to expose subtle bugs.

E. Java Applets

As we described in our last report, the first year of the java programming phase of our project focused on stabilizing our approach to Java. Since the language was undergoing dramatic

changes, it was hard to modularize our applet development. The net result was that progress was slow. Fortunately, in the second year of this effort, Java has stabilized considerably, and we have been able to push ahead on using Java in our development environment. We have also had time to expand the capabilities of our existing applets.

A good example of this is our pattern recognition applet. This year, we were able to greatly enhance its educational value by augmenting basic capabilities with more visualization. An example of this is shown in Figure 5. Users can specify the parameters of a Gaussian distribution, and can view the support regions for these distributions in the original data space (along with the resulting decision regions). In Figure 6, we show two new data sets that were added by one of our sophomore undergraduate programmers. These data sets pose interesting challenges for classification algorithms since they are not linearly separable (they can't be separated by decision regions formed by hyperplanes). We have plans to include algorithms that can handle such data — for example, an algorithm based on support vector machines which is currently under development in our research group. The pattern recognition applet was used extensively in our Fundamentals of Speech Recognition course this semester.

Another novel feature added to this applet was the ability to classify the data using two popular clustering algorithms — KMEANS and Linde-Buzo-Gray (LBG). These algorithms iteratively reestimate cluster centers, and form decision regions based on nearest neighbor calculation with respect to these cluster centers. A Java applet is an ideal forum in which to learn about such algorithms because you can see the decision regions evolve with each iteration. This is demonstrated in Figure 7. Users select the number of clusters they want to use, and the number of iterations to be performed, and can then step through each iteration of the algorithm. Classification results are displayed in the description box to the right. In Spring '01, we plan to use this applet extensively in a pattern recognition course.

We have also begun implementation of several new applets. One which we are very excited about is an applet that helps students visualize search algorithms. This was motivated by a visit to MS State's 3D immersive visualization environment known as the COVE during the January design review, and viewing a demo where one walks along the edge of a mountain. We are attempting to create such a visualization of the search space during recognition using standard Java components (as opposed to some of the experimental virtual reality engines that are not quite standard yet). Further, we are developing a basic digital signal processing applet demonstrating the sampling process for signals. This is intended to be used in our undergraduate Signals and Systems course. Most of this work is being carried out by our undergraduate programmers, and represents a nice, non-critical path in which they can contribute to our research program.

Finally, we have begun some collaborations with Georgia Institute of Technology and MS State's College of Engineering to use our job submission applet to do audio indexing of classroom lectures. This application was presented at a recent Internet 2 conference [3], and is an application enabled by the vast bandwidth potential of Internet 2. It will become more feasible when we expand our compute serving resources in the third year of this project. In this application, audio from a lecture is shipped to our system for automatic transcription via recognition, and time-alignment. The results are then transferred back to a database which can be searched by students ("Show me all the lectures about Fourier Transform."). While we are in the early stages of the development of this capability, it is a good example of the burgeoning interest in audio indexing and audio mining.
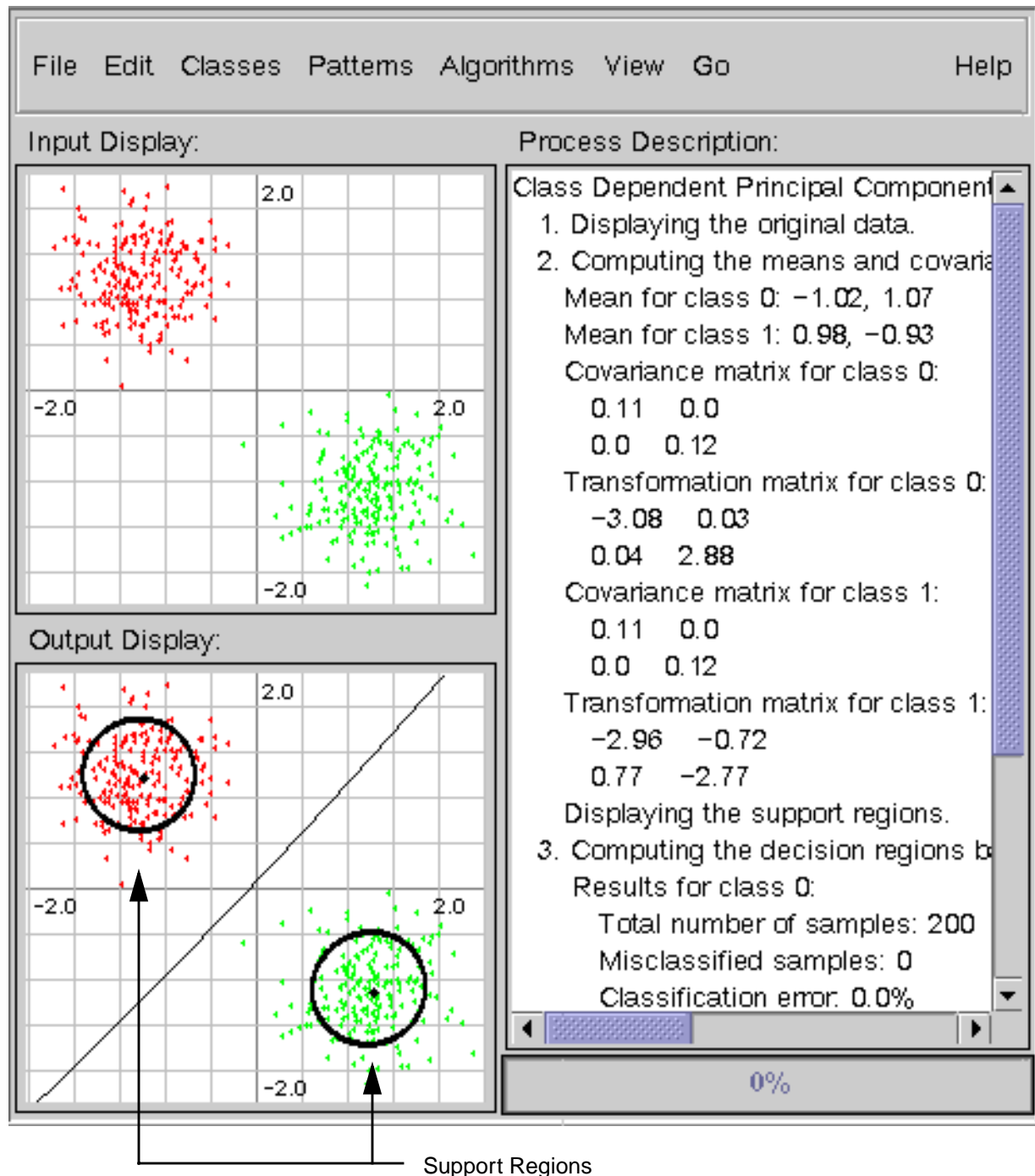
Figure 5. An example of enhanced visualization capabilities in our pattern recognition applet. We now allow users to specify the parameters of Gaussian distributed data through dialog boxes. In the case above, two Gaussian distributions were generated with means of [-1,1] and [1,-1] respectively. Principal Components Analysis (PCA) was chosen as the classification algorithm. Step-by-step output from the PCA approach is shown in the scrolling box on the right. Once the means and covariances are computed in Step 2, the support regions for the distributions are displayed in the lower left. These are crucial to understanding how algorithms such as PCA transform data. Other enhancements to this applet include the addition of new clustering algorithms, and an interactive status bar (lower right) that provides feedback for users when time-consuming operations are being performed.
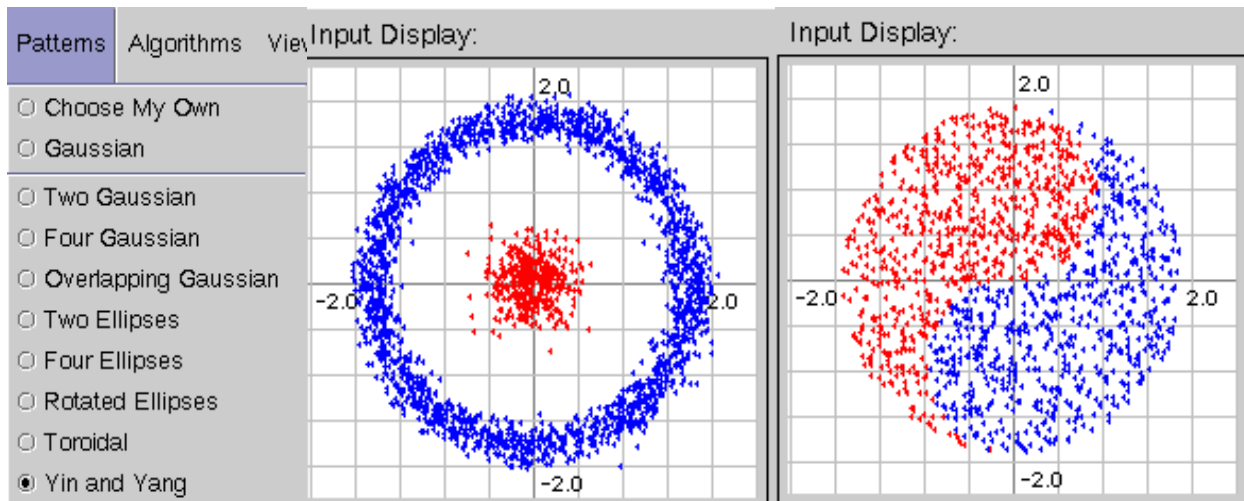
Figure 6. Two new data sets that pose challenging problems for classification algorithms have been added by one of our undergraduate programmers.
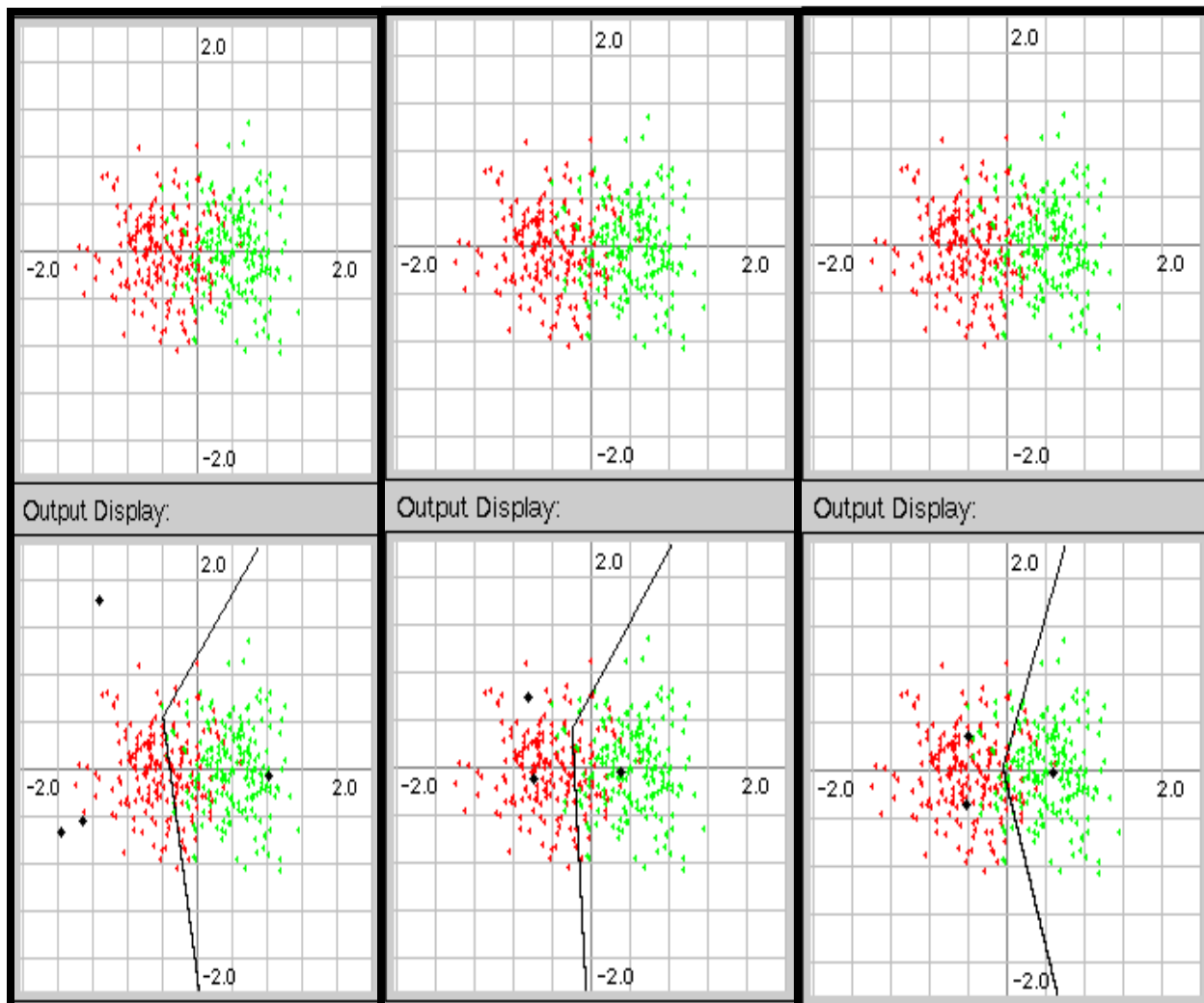
Figure 7. An example of a clustering algorithm in which users see the decision regions evolve.

F. REFERENCES

[1]   R. Sundaram, A. Ganapathiraju, J. Hamaker and J. Picone, "ISIP Public Domain LVCSR System," to be presented at the Speech Transcription Workshop, The University of Maryland University College, College Park, Maryland, USA, May 2000.

[2]   J. Picone and W.C. Chapman, "Speech Recognition System Design Review," *http://www.isip.msstate.edu/conferences/srsdr00*, Mississippi State University, Mississippi State, Mississippi, USA, January 2000.

[3]   J. Picone, C. Atkeson and I. Alphonso, "Harnessing High Bandwidth: Applications in Speech Recognition," presented at the Spring 2000, Internet2 Member Meeting, Washington, DC, USA, March 2000.

[4]   J. Picone, "Summary of SRSDR'00," *http://www.isip.msstate.edu/conferences/srsdr00/technical_program/session_08/index.html*, Mississippi State University, Mississippi State, Mississippi, USA, May 2000.

[5]   J. Picone, "Speech Recognition System Training Workshop," *http://www.isip.msstate.edu/conferences/srstw00*, Mississippi State University, Mississippi State, Mississippi, USA, January 2000.

[6]   J. Picone, "Workshop Program," *http://www.isip.msstate.edu/conferences/srstw00/html/program.html*, Mississippi State University, Mississippi State, Mississippi, USA, May 2000.

[7]   J. Vincent, "Request Tracker," *http://www.fsck.com/projects/rt/*, March 2000.

[8]   N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 84-107, September 1999.

08/15/98 — 08/14/99: RESEARCH AND EDUCATIONAL ACTIVITIES

In the first year of this project, we focused our efforts in three major areas:

- **Core Technology**: extensions of the speech recognition system required to enhance its appeal to our customer base (driven by customer feedback);
- **Foundation Classes**: building blocks such as vectors, matrices, and data structures that simplify and standardize the development of higher-level classes;
- **Web-Based Information**: a comprehensive and informative web site that constitutes a central point of contact for everything related to the project.

We have seen interest in the project grow as evidenced by the fact our mailing list has grown to 150 participants, and we have received several serious inquiries about collaborations based on our system (one of which resulted in participation in a joint NSF/EU proposal [1]). Major milestones for the first year of the project included the release of a fully functional speech recognition system (including feature extraction and training), and the development of a remote job submission capability that lets users submit jobs to our system over the Internet.

A. Core Technology

State-of-the-art speech recognition technology is the foundation upon which this project is built. We must not loose site of the importance of this core technology to this project. This technology is being developed in a parallel effort funded by the Department of Defense. The system has improved dramatically since the start of our NSF project in August'98. We briefly review the enhancements made to the system in the first year of this project, and then discuss the impact this has made on our efforts within this project. Next, we describe some enhancements made to the system to broaden its appeal to our customer base, and review some initial attempts at cross-platform portability.

A.1. System Status

In the past year, three important capabilities have been added to the speech recognition system we have been developing: feature extraction, word graph generation, and Hidden Markov Model (HMM) training. Feature extraction is the process by which the speech signal is converted to a sequence of vectors that serve as input to the recognition system (a continuous density HMM system). This is often called the front-end. Our approach was to initially replicate an industry-standard front-end consisting of mel-spaced cepstral features [8] and their first and second-order derivatives. This is one step that allows users to duplicate results obtained with other commercial and proprietary systems. This capability was delivered as part of a general front-end capability summarized in Figure 1.

A second key feature added to the system was the ability to generate word graphs. Speech recognition experiments are time-consuming primarily because of the large language models used in decoding portion [8] of a system (these large language models are desirable because they maximize performance). Hence, to save time on subsequent experiments, a word graph is constructed that represents most plausible, or highly probable, hypotheses that can be generated by this network. This graph is then rescored with new acoustic or language models depending on the nature of the research. This network can be quickly rescored — a process that often runs at

least ten times faster than the process required to generate the network. Because of these time savings, word graph rescoring is an extremely popular method of doing speech research for conversational speech recognition, and is therefore a modality that must be supported for a system to be widely accepted.

This feature was added to the system in early 1999, and required significant changes to the way the decoder manages the search process. Fortunately, the current implementation represents a vast improvement in the architecture and performance of the system [8]. Though it took longer than expected to add this feature to the system (we struggled with this for 6 months), the final implementation is extremely clean and efficient, and will have a positive impact on subsequent generations of the system. The system now supports more decoding modes than any commercially available system, and more than most proprietary systems as well.

A third essential feature that was added to the system this year was HMM training [11]. This essentially provided closure on the speech recognition system in that users are now able to build real systems from scratch (previously, one had to borrow some component from the system from another source). We first implemented an algorithm known as Viterbi decoding, in which only the best state sequence through a network is considered. This is an elegant algorithm in that it is efficient, fast, and consistent with a formal languages view of the speech recognition problem. This allowed us to develop the proper control structures and code infrastructure quickly. Once this was complete, we added Baum-Welch training [12], which is more popular in state-of-the-art systems.

The current system is described in great detail in an upcoming publication [8]. This is one of the first such publications to provide a tutorial on the details of search algorithms, and is being published in a journal that emphasizes education of entry-level graduate students in signal processing. It is our hope that the time spent on this publication will pay off as more researchers are made aware of the availability of this system and project. Our system is also represented at an upcoming major speech conference [13] that will include a panel discussion on the merits of public domain speech technology.

A.2. System Enhancements

In any such project intended to develop public domain code, it is important to adapt to the changing needs of your user community. In the past few years, several major sites in speech research have shifted to the use of features based on an algorithm known as perceptual linear prediction [14]. This technique in some cases has been shown to deliver small improvements in performance. It is an important feature to include if such sites are attempting to replicate their best systems with our public domain system. We completed an initial implementation of this algorithm this year, and are in the process of integrating it into our front-end architecture. A summary of this approach is shown in Figure 2.

Another feature that was requested by several sites this year was the ability to switch language models in the middle of the recognition process. This feature is useful for two reasons: (1) it enables the development of command and control applications that switch between sub-grammars depending on what words are recognized (context-senstive language models or menus); (2) it allows the recognition system to dynamically recurse through language models at runtime, rather than compile all language models into one big network. The first point impacts the development of voice-driven menu systems and real-time demonstrations. As a word or phrase is recognized, a

new language model can be loaded, providing a small set of words or phrases as the next choice. This is the strategy used by most systems that allow you to accelerate your desktop menus with voice commands.

A second, and equally compelling reason to consider this feature, is the development of systems that consist of a hierarchy of grammars (for example, sentences in terms of words, words in terms of syllables, syllables in terms of phones, etc.). Many existing systems will compile such a system into one large network. While this is sometimes attractive for efficiency reasons, for research flexibility it is often better to process this hierarchy of networks at runtime ("on the fly"). This allows users to change one small module of the system (for example, allowing a word to be represented multiple ways) without recompiling the entire system.

Implementation of this capability requires use of an approach called "caching." The system must only activate those portions of the network that represent active words, or words used in the recent past, and leave the remainder of the language model stored on disk. Fortunately, by implementing this strategy, we can handle much larger language models, such as those used in the broadcast news [15] and audio data mining research fields. Since there is currently a shift towards such applications in various research communities including NSF and DARPA, we feel it is important to provide a solution to this problem. Broadcast news language models tend to be large (because they are dealing with lots of words that occur infrequently) and cannot be managed in memory as a single unit. Our conversational speech recognition system could not handle such a large language model due to the number of bigram entries contained in this model.

In addition to these algorithmic enhancements, several supporting tools were developed to facilitate language modeling. These include a grammar compiler that accepts regular expressions as input and outputs a finite state machine description used by our system, and a grammar compaction algorithm that reduces the size of a word graph without compromising performance. We also developed several display utilities [16] that allow users to look at speech data and analyze its frequency content. We have interacted with several industrial sites on the development of these tools. In fact, one commercial site is making extensive use of this tool in a production data collection capacity, and hired one of our undergraduate programmers for the summer to customize this tool to better suite their unique needs.

A.3. Cross-Platform Portability

We have begun to address issues related to portability across different computing platforms. Our plan remains to develop code under the Solaris operating system on two platforms — Sun Sparc and Intel PC (Solaris x86). Within these environments, we use a common compiler, gcc, provided by the Free Software Foundation (GNU). This compiler is supported across a wide range of platforms, including Microsoft Windows'9X and Windows NT. Since an ANSI standard for C++ has only recently been adopted, and most implementations of C++ are still not compatible (this likely will continue for a few years even after the adoption of the standard), using a common compiler across all platforms is the best way to guarantee portability. Currently, we periodically release a Windows version by porting the GNU environment to Windows, and compiling the code under gcc and the bash shell. This vastly simplifies the Windows porting effort.

Perhaps the fastest growing user population for our system is the Linux community. Fortunately, the backbone of the Linux system is GNU software and the gcc compiler in particular. In fact, GNU recently turned over development of its compiler, gcc, to an organization known as

EGCS [17], which is a collection of free software advocates who have been informally developing compiler extensions to gcc for years. EGCS is responsible for all subsequent releases of gcc (the two have effectively merged) and is the compiler delivered on most production releases of the Linux operating system (RedHat for example). Hence, conformance to gcc standards covers a large portion of both the Windows and Unix markets, yet minimizes portability issues.

However, the Linux version of gcc distributed by EGCS currently lags the Sun Solaris version in one important dimension: wide character support. One of the founding principles of our system is that languages be handled in native format using Unicode character encoding. The current ANSI C/C++ standards support a wide character encoding with a collection of new functions. These must be supported by a runtime library, which Sun Solaris provides, but production releases of Linux do not. Hence, the current release of our code will not compile under Linux. We expect this problem to be resolved within the next three months. It doesn't make sense to provide an interim work around, because anything we do will be quickly obsoleted. It is expected EGCS will do a much better job in the future of tracking standards in C++.

B. Foundation Classes

The most critical part of the first year of this project was to lay the proper foundation upon which all other software can be written. This is perhaps the most difficult part of any technology-specific project — balancing the efficiency and simplicity of the target application with extensibility and flexibility needed by future research. We believe this is the major strength of our project — the environment is object-oriented from the ground up and designed to be neutral to any particular algorithmic approach. Fortunately, we have been able to leverage preexisting code developed for a much less ambitious project [18] involving speech recognition. However, most of this code needed revamping based on the changes in the C++ language definition and gcc compiler capabilities.

The hierarchy of classes is shown in Figure 18. Our goal in the first year of this project was to deliver everything through the DSP libraries, which we refer to as the ISIP foundation classes (IFCs). In the second year of the project, we begin the "Great Convergence" as we rewrite the speech recognition system using these IFCs. We expect this task to be completed by the end of 1999. This latter version of the system we will be the basis for our training workshops which will begin in the summer of 2000.

B.1 Integral Types

At the bottom of our class pyramid is a header file that defines all low-level data types available to the user. These are known and the integral types, and in many ways mirrors the syntax of the C programming language. Our current integral types file is shown in Figure 4. This deceptively simple file was the result of much hand-wringing about two competing design philosophies. The C programming language was designed to be somewhat architecture independent. For example, the datatype "int" could be 16-bits long on one machine, and 32-bits long on another. A C program written properly would work on both machines regardless of the specific implementation. On the other hand, in signal processing, we often need access to specific data types. For example, speech signals are stored as 16-bit integers, and need to be represented as such in C. The problem in C is that a 16-bit integer doesn't really exist. Instead, you can use a "short int" and must assume that an integer is 32 bits long.

There is a growing movement within the C++ community to support data types such as int32 for a 32-bit integer. In fact, Microsoft seems to be leaning this way with its Visual C++. The integral types file shown in Figure 4 represents a synthesis of the two approaches. Programmers used to constructs such as "long" have these available. However, the scalar classes, to be discussed later, are defined to be specific sizes, and hence use the size-specific data types. In the future, as architectures and compilers expand to 64-bits and 128-bits, our code will be backwardly compatible with no modifications, because the types are tied to a specific number of bytes. By providing new types, such as "double double" or "long long," we can also accommodate the new wider formats.

B.2. System and I/O Classes

Our goal in the design of our system is to abstract the user from details of the operating system. The system library serves this function by encapsulating all operating system specific activities, such as file I/O and character string processing. Both libraries represent a significant improvement over the previous implementation of this environment [18]. The System library supports low-level operations such as file management (opening, closing, reading, writing), character processing (Unicode support), error handling and error notification. All of these require interacting with operating system-specific C functions, and hence must be centralized and abstracted from user-level code.

The I/O library contains a novel approach to file formats. All files in our environment are represented as Signal Object Files (Sof) [18]. Sof alleviates the need for users to read and write data manually (formatting of data tends to be one of the most time-consuming aspects of speech research). All objects know how to read/write from/to an Sof file. In fact, higher-level objects just recurse through their hierarchy of objects for I/O. Sof is simply a smart indexing scheme that keeps track of what objects are stored in a file. It allows multiple instances of an object (a String object can be written several times to the file), and handles ASCII or binary files transparently. It is also machine-independent in that users need not worry about byte-ordering or floating point representations across platforms — this is handled automatically within Sof.

A centralized data storage strategy is essential in speech research, and other data-intensive research areas as well. There is nothing speech-specific about Sof. Sophisticated database management strategies have yet to provide a clean solution for researchers, so the tendency has been to use proprietary formats or unstructured formats (ASCII). Most database packages don't want to deal with byte-formatted, such as an MPEG audio or video stream, and don't allow partial I/O of these types of data (retrieve only the middle three seconds of this audio file). There are some public domain file formats being supported within the community [19], but these do not interface well to C++ and have certain limitations in terms of the flexibility (only one instance of an object can be written to a file). Sof is an important part of our overall strategy to ease the programming burden of our users.

B.3 Math Classes

As shown in Figure 3, the next step up from our low-level IFCs are the math classes. This is actually the first level we expect application developers to interact with — users should not use the system classes directly, and should only use the I/O classes for programming I/O into new class definitions. The math classes consist of scalars, vectors, and matrices. Their implementation

follows that of the C++ Standard Template Libraries (STL), with two important exceptions. Our types are allowed to be size-specific. For example, a Long is always a 32-bit integer, a VectorLong is always a vector of 32-bit integers, etc. This gives the programmer complete control of data sizes. Also, our matrix class is a vector of vectors, where each vector can have a different dimension. This costs very little in overhead, and makes the matrix class much more useful as a container class.

We began implementing the math classes with a simple strategy that did not rely on templates. This was mainly due to a legacy of gcc not supporting a useful model of templates. Gcc was based on an inclusion model in which all code related to a template had to be included in the header file. Obviously, for the system of the scale we are talking about, this is impractical. Fortunately, with the release of version 2.8.X of the gcc compiler, appropriate hooks have been provided to allow source and header files to be separated for template definitions. A header file for a template version of a scalar class is shown in Figure 5. Our benchmarks on code implemented with and without headers seems to show that the template approach is every bit as efficient as the non-template version, and requires much less time to compile. It is a win-win situation thus far. This is summarized in Table 1 below.

Description	Non-Template	Template

Scalar Long:
    executable (kB)
    memory usage (kB)
    runtime (ms)
    compilation time (secs)
	564
	1484
	10.9 ms
	31.8
		566
		1484
		11.4 ms
		18.6

VectorLong:
    executable (kB)
    memory usage (kB)
    runtime (ms)
    compilation time (secs)
	708
	1592
	12.9
	76.8
		713
		1596
		13.1
		66.1

MatrixLong:
    executable (kB)
    memory usage (kB)
    runtime (ms)
    compilation time (secs)
	4590
	1640K
	17.6
	99.4
		4605
		1648K
		17.9
		76.9

Table 1. A comparison of template and non-template implementations of the math classes. Templates appear to finally be competitive with traditional code.

In the first year of this project, we have completed implementation of the math classes based on our new template approach. This was somewhat of a paradigm shift for us, and hence required some additional training of our programmers. This approach will pay great dividends when we move to higher-level libraries such as data structures, where one expects a template capability. With templates, user can build generic lists, hash tables, etc. of whatever object they desire. This is an extremely important capability for C++ programming. The only drawback of our approach is that the template implementation is specific to gcc, since there no clear standard for how to implement templates in the ANSI C++ specification. We continually monitor standards activities to see how we can improve our portability.

B.4 Concurrent Versions System (CVS) and Anonymous CVS Servers

One of the interesting aspects of our project is that truly concurrent development is being done by a large number of programmers (between six and eight people work on the system continuously). This places great stress on most non-commercial software management systems. Commercial packages, on the other hand, are expensive (typically $1K per seat) and hamper our ability to do concurrent development in a distributed fashion as we describe below. Hence, we spent some time this year converting our software management environment from the popular revision control system (RCS) [20] to a state-of-the-art package called Concurrent Versions System (CVS) [20].

CVS allows multiple users to check out the same code, make revisions, and check it back in. The tool automatically merges the code to produce what it thinks is the correct version. CVS also allows you to manage utilities, classes, libraries, etc., all within the same framework. It allows users to check out an entire software tree as a single unit, rather than manage this as individual files as is done in RCS. Unfortunately, this is not as simple as it sounds, and there is a steep learning curve for CVS. Hence, we spent some time developing wrappers for common CVS functions so that users were protected from the details of CVS [22]. This is important because CVS is unwieldy at times, and can corrupt files if not used carefully. However, we are now routinely using it in production, with satisfactory results.

One of the strongest arguments for using CVS is the capability it has for doing distributed software development. We have implemented an anonymous CVS server that functions much like an ftp server. Users can log into this server, and grab a snapshot of the code currently under development, and do concurrent development if necessary. This is a fairly new capability introduced into CVS in the last year, and is being used by several public domain software projects. In our case, it is an extremely useful capability because it allows users to update a portion of the environment as we make incremental changes, rather than download the entire environment each time we make a small change. Since our final environment will be large, the anonymous CVS distribution strategy allows users to maintain a current copy of the environment without repeatedly doing massive downloads. It also offers the potential for others to contribute to the environment. An on-line [23] tutorial on how to use our anonymous CVS server is available on the web.

B.5 Software Quality Control

Maintaining quality software releases in a rapidly developing environment is always a challenge, especially when dealing with student programmers. In fact, this is one reason we will add a staff member to the project next year. We have been continually refining our software quality control process. Three important facilities were added to our environment to enhance our process. First, we adopted a public domain memory checking system known as dmalloc as a routine part of our software development cycle. Dmalloc checks for memory leaks and other such programmer errors. Though not an industrial strength product (such as Pure Software's Purify), dmalloc runs on all our supported platforms and does a good job of catching memory problems. It has made a big difference in the quality of our code. Typical released code that used to have two or three memory problems per class now are released with virtually no defects.

A second important step we have taken to improve the quality of our software is to introduce a diagnostic method in each class. As a programmer implements a class, a diagnose method is

provided that exercises all methods in the class, and produces predetermined output. We have integrated this into our make facility as well: "make diagnose" automatically generates a test program that uses this method. This facility, coupled with dmalloc, allows the programmer to do a fairly complete test and verification of the class before it is released.

Finally, we have instituted a web-based checklist facility that programmers use to make sure they complete all steps required of a release. As a programmer works through a class, the checklist is updated with each major step. The checklist includes items such as initial design, design review, implementation, diagnostics, debugging, dmalloc, documentation, cross-platform check, and release. Pertinent information, such as the programmer's name and date of completion, are automatically generated and stored in database. Everything is done via the web and an SQL database interface, which makes it extremely easy for the project manager to track progress.

C. Web-Based Information

Dissemination of information via the web is a critical part of this project. We have overhauled our web site to showcase this project and to make information more readily accessible to our users. The URL for the project is:

    http://www.isip.msstate.edu/projects/speech

This web site contains some novel features that are described below.

C.1. Project Web Site

We have designed and implemented a uniform look and feel for the web site, as shown in Figure 6. The hierarchy is designed to make it easy for users to access the software, educational resources, and on-line job submission facility. Most of the web pages are implemented using server-side includes that provide a uniform look and feel for all pages. We have also implemented a search capability using a public domain SQL database package. Records in this database are currently entered manually using a web-based interface. Our attempts at generating the database automatically produces unacceptable results (personal AltaVista was the best tool we looked at, but does not exist as a Unix package currently).

We have made it easy for people to contact us for support by providing a single point of email contact: help@isip.msstate.edu. We typically have been able to respond in less than one hour to most requests for help, though traffic has been fairly light thus far. Incoming requests to help are reviewed by the project manager and assigned to the appropriate student worker for resolution.

We have also added a facility for archival of all mail messages sent to our project-specific email alias: asr@isip.msstate.edu. The URL for this archive is *http://www.isip.msstate.edu/data/mailing_lists*. Any message to this list is archived and added to the web page by a process that runs nightly using mail processing tool (Monarch) that generates a threaded display. This archive has proven to be extremely useful when new members join the list. Eventually, we will add an FAQ to the web site to complement the information in the mailing list archives.

We also automatically track downloads of our software. The statistics on who is downloading our software can be viewed at the following URL: http://www.isip.msstate.edu/data/statistics/web. This page tracks hits on a user-specified set of web pages, allowing us to do a thorough analysis of who is accessing our web pages and downloading our software.

C.2. Documentation

We have begun building on-line documentation for our foundation classes. An example of this documentation [24] is shown in Figure 7. To the left, we have the entire class index in a scrollable window. To the right, we have the documentation for a particular class. Each major heading, such as "MAIN," has an overview of the library or set of libraries. The classes are grouped by their position in the hierarchy. Eventually, we will need to supply a search engine for random access to these pages.

Each individual web page is organized similar to a Unix man page, with the appopriate modifications to account for the fact that these are classes instead of library functions. The source code for the class is directly linked to the web page, making it easy for users to study the source code and the documentation simultaneously. The pages are structured as follows:

Section	Description	Links Provided
Name	class name	class header file
Synopsis	broad overview of the class	N/A
Quick Start	a working example	N/A
Description	brief description of the goals we had in designing the class	N/A
Dependencies	other classes included in the header files and required for compilation	corresponding classes on which this class is dependent
Public Methods	user interface (also shows methods required for all classes)	source code for each method
Public Constants	constants available for general use	N/A
Protected Data	information for programmers on the internal data	class header file
Private Methods	methods used internally in the class	source code for each method
Examples	simple examples how to use the code	N/A
Notes	other information relevant to users and programmers	N/A

Table 2. An overview of the information contained in a typical page documenting a class.

Pages for utilities, applications, and toolkits will follow the same format. A searchable database is also under development to support random access to these pages.

C.3. Educational Java Applets

We have made a strategic commitment to developing Java applications because of Java's inherent portability. However, this has been a mixed blessing because the Java language and associated toolkits are constantly changing. On top of that, each release of Java seems to have serious bugs that get in the way of developing robust applications. The net effect is that our programming efficiency in Java has been quite low, and our progress on educational applets has been hampered

by these problems. Java's lack of portability seems to be an industry-wide problem at the moment.

There are two strategic issues with Java programming. First, there is the GUI, or application interface. Java previously provided the Abstract Window Toolkit (AWT) as its standard interface. We developed a number of applications around this interface [25]. Unfortunately, this interface was recently obsoleted, and replaced with Swing [26]. We spent time this year training our Java programmers on the Swing interface, and porting our existing applications to this new interface. Swing is still somewhat buggy and working around these bugs has been a time-consuming process. We have, however, managed to release several new applets under Swing. An example of one such applet, which teaches the principles of digital filter design, is shown in Figure 8. We would like all our applets to have a common interface. Hence, some amount of retooling of existing applets was necessary.

To make matters worse, Netscape's latest releases of its browser are not fully compliant with Swing. Netscape recently seems to consistently lag Sun Microsystems on support of Java. Hence, users must download a plug-in from Sun to get true Java and Swing compliance. This appears to be the best solution at the moment, Netscape's commitment to retooling its browser appears to be questionable. We provide installation instructions on our web site [27] for how to download the package and install it in several different configurations. More importantly, we have also programmed our applets to probe the user's browser, detect this plug-in is missing, and prompt the user with a message indicating what to do to download the plug-in.

Despite our Java retooling problems, we have been able to develop two new applets. The first is the digital filtering applet mentioned previously, and shown in Figure 8. In this applet, a user can design a filter using several predefined algorithms involving well-known filter prototypes. The user can also draw a desired frequency response, and let the applet design the corresponding filter. The applet provides details on the actual design, including filter coefficients, frequency and phase response, and a pole/zero analysis. This applet is targeted towards split-level DSP courses, and undergraduate signals and systems classes.

A second applet involves demonstration of fundamental concepts in pattern classification. Users can select prestored data sets that highlight the differences between common classification schemes such as principle components analysis, linear discriminant analysis, and Euclidean distance. Users can also optionally enter their own data sets. Classifiers can then be trained on this data, and the results depicted in terms of classification regions. This applet demonstrates several statistical normalization principles used in signal to feature vector conversion process in speech recognition. It will be useful for graduate courses in pattern recognition, speech recognition, and digital signal processing. It has not been formally released because there are several Java problems with the user interface. We expect this applet to be released in the first quarter of the second year of this project.

We have also begun building a much more ambitious demo that is essentially a port of our Tk/Tcl demonstration of the search algorithm used in the recognizer. This demo has been available for some time as part of the recognition toolkit. It requires the Tk/Tcl toolkit on the platform running the demo, as well as a port of the recognizer. We have demonstrated this application on Windows as well as Unix machines. Our approach in Java was to port the C++ code for the recognizer, and retool the interface. Unfortunately, this turned out to be a much more ambitious effort than planned, for some of the reasons described above. Hence, we decided to better learn how to implement more straightforward applets in Swing first. In the second year of this project, we will

return to the problem of providing a Java-based graphical tutorial of how a speech recognition search engine works.

C.4. Remote Job Submission

One of the truly unique capabilities that we added to the web site this year was the ability to submit a speech recognition job over the Internet to our servers. The interface for this facility is shown in Figure 9. The page can be reached by clicking on experiments on the project main page, or directly from the following URL: *http://www.isip.msstate.edu/projects/experiments*. The page contains a CPU monitor on the upper left that shows the status of our compute servers, a dialog box on the bottom that is used to interact with the user, and windows to the right that provide status on active jobs and access to the results produced by the job. After a job is submitted, users can view the results on-line via a URL, or have the results transferred via email.

The current implementation is an initial prototype that will be refined in the coming year. It is certainly not robust and not as graphical as we would like. There are two important features included in the current system. First, users can run a canned experiment and obtain detailed information about how the recognizer analyzed the data. This is useful for comparing performance, replicating well-known results, or learning how the algorithm processes data. Second, users can supply their own audio file via a URL. This is useful if you want to compare the performance of several systems on the same data. Eventually, we will provide more support for editing data graphically, and interacting with parameters of the models. For the moment, most interactions are done via text boxes, and only a limited set of parameters can be modified.

D. Summary

The first year of this project has been productive in the much of the groundwork to support the subsequent years of the project has laid. From a human resources standpoint, the funding from this project has allowed the recruitment, training and development of four promising undergraduate students (two of whom plan to pursue graduate degrees in our department under ISIP's direction), two M.S. students (who plan to continue for a Ph.D.), and one Ph.D. student. All but one of these students will remain on this project until its conclusion. In addition to making fundamental contributions to the project in the first year, all have been trained on our strict software engineering paradigm. Since this project has strong synergy with a related project focusing on core technology development, we are able to leverage many resources and much infrastructure from that project. One of our most senior graduate students has transitioned from the core technology project to this project, and will manage technical aspects of this project in the second year.

The second year of the program provides for a professional staff position to manage the routine operations of the project, particularly support and web site development. We have recruited a senior engineer for this position. This individual has an MS in Computer Science, and over 20 years of experience in software engineering and computing systems in both industry and academia. For the past several years, he has been the computer systems administrator for another college on campus. Prior to that, he has operated a small consulting company that developed business management software for hospitals. Since he is already a university employee, he has begun interacting with our group on a volunteer basis so that he can familiarize himself with our operation.

This staff position has three primary duties: quality control, web site development, and support. He will directly supervise the students working on the project, and be responsible for software releases, bug fixes, and updates. A near-term goal is to implement the GNU configure [28] distribution paradigm into our system, so that our software will automatically configure itself upon installation. A secondary immediate goal will be to implement problem-tracking software so that all messages to our help line will receive proper prioritization and attention.

The second year of this project is a pivotal year in that we will hold our first set of workshops. In January 2000, we will hold a one-day industrial forum in which we conduct a formal design review of the system, and solicit feedback from the participants on desired enhancements for the coming year. This workshop is tentatively scheduled for January 6-7. We hope to have a mixture of senior professionals from industry and academia (with more of an emphasis on industrial participation). The program will most likely consist of a half-day of design reviews and demos, followed by a half-day of discussion about recommended enhancements to the system. We expect to develop a clear plan of action from this meeting in an attempt to focus our development towards things of interest to the general community.

Our first summer workshop is also tentatively scheduled for May 21-27. For this workshop, we will invite approximately 12 graduate students (and perhaps senior undergraduates) to spend one week in our lab learning about our system. Travel expenses will be paid for these students. The agenda will most likely consist of morning lectures and demonstrations followed by afternoon laboratories. Initial feedback on this has been very positive, with several sites suggesting they would subsidize attendance by their professionals rather than have their people miss the event. Our facilities can accommodate 12 students comfortably, with a reasonable ratio of students to staff, and adequate access to computing equipment. If interest exceeds this limit, we will investigate alternative facilities. However, our tendency for the first training workshop is to keep it small and focused, so that it can proceed as smoothly as possible.

E. REFERENCES

[9]   R.A. Cole, *et al*, "Multilingual Access and Retrieval using Communicative Interface Agents (MARCIA)," submitted to Multilingual Information Access and Management: Call for International Research Cooperation, National Science Foundation, June 1999.

[10]   N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," to appear in *IEEE Signal Processing Magazine*, September 1999.

[11]   J. Picone, "Continuous Speech Recognition Using Hidden Markov Models," *IEEE ASSP Magazine*, vol. 7, no. 3, pp. 26-41, July 1990.

[12]   Y. Wu, A. Ganapathiraju, and J. Picone, "Baum-Welch Reestimation of Hidden Markov Models," *http://www.isip.msstate.edu/publications/reports/isip_lvcsr/1999/baum_welch/ report_061599.pdf*, Mississippi State University, Mississippi State, Mississippi, USA, May 1999.

[13]   N. Deshmukh, A. Ganapathiraju, J. Hamaker, J. Picone and M. Ordowski, "A Public Domain Speech-to-Text System," to be presented at the 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 1999.

[14]   H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech." *Journal of the Acoustical Society of America*, vol. 4, pp. 1738-1752, 1990.

[15]   W.M. Fisher, *et al,* "Data Selection for Broadcast News CSR Evaluations," presented at the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, U.S.A., February 1998.

[16]   I. Alphonso, N. Deshmukh, and J. Picone, *http://www.isip.msstate.edu/projects/speech/ software/transcriber/index.html*, Mississippi State University, Mississippi State, Mississippi, USA, May 1999.

[17]   P. Bothner, *et al*, "Welcome to the GCC Project!," *http://egcs.cygnus.com*, June 1999.

[18]   J. Picone, "Managing Software Complexity in Signal Processing Research," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. III-41-III-44, Minneapolis, Minnesota, USA, April 1993.

[19]   J. Fiscus, *et al,* "SPeech Quality Assurance (SPQA) Package Version 2.3 AND Speech File Manipulation Software (SPHERE) Package Version 2.5," *ftp://jaguar.ncsl.nist.gov/pub/ spqa_2.3+sphere_2.5.tar.Z*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, June 1999.

[20]   W.F. Tichy, "RCS--A System for   Version   Control," *Software--Practice & Experience*, vol. 15, no. 7, pp. 637-654, July 1985.

[21]   "Concurrent Versions System (CVS)", *http://www.cyclic.com/cyclic-pages/howget.html*, Cyclic Software, Washington, D.C., USA, June 1999.

[22] R. Duncan, "Software Version Control System," *http://www.isip.msstate.edu/projects/speech/education/tutorials/cvs/index.html*, Mississippi State University, Mississippi State, Mississippi, USA, June 1999.

[23] I. Alphonso, "CVS Anonymous Download Instructions," *http://www.isip.msstate.edu/projects/speech/support/info/cvs_instructions.html*, Mississippi State University, Mississippi State, Mississippi, USA, June 1999.

[24] S. Balakrishnama and N. Deshmukh, "ISIP Software Documentation," *http://www.isip.msstate.edu/projects/speech/education/tutorials/isip_env,* Mississippi State University, Mississippi State, Mississippi, USA, June 1999.

[25] ICASSP applets paper

[26] E. Eckstein, M. Loy, and D. Wood, *Java Swing*, O'Reilly and Associates, Cambridge, Massachusetts, USA, 1998.

[27] R. Duncan, "Java Plug-In Installation Instructions," *http://www.isip.msstate.edu/projects/speech/support/info/java_instructions.html,* Mississippi State University, Mississippi State, Mississippi, USA, June 1999.

[28] D. MacKenzie and B. Elliston, *http://www.gnu.org/manual/autoconf-2.13/html_chapter/autoconf_toc.html*, Free Software Foundation, Boston, Massachusetts, USA, July 1999.
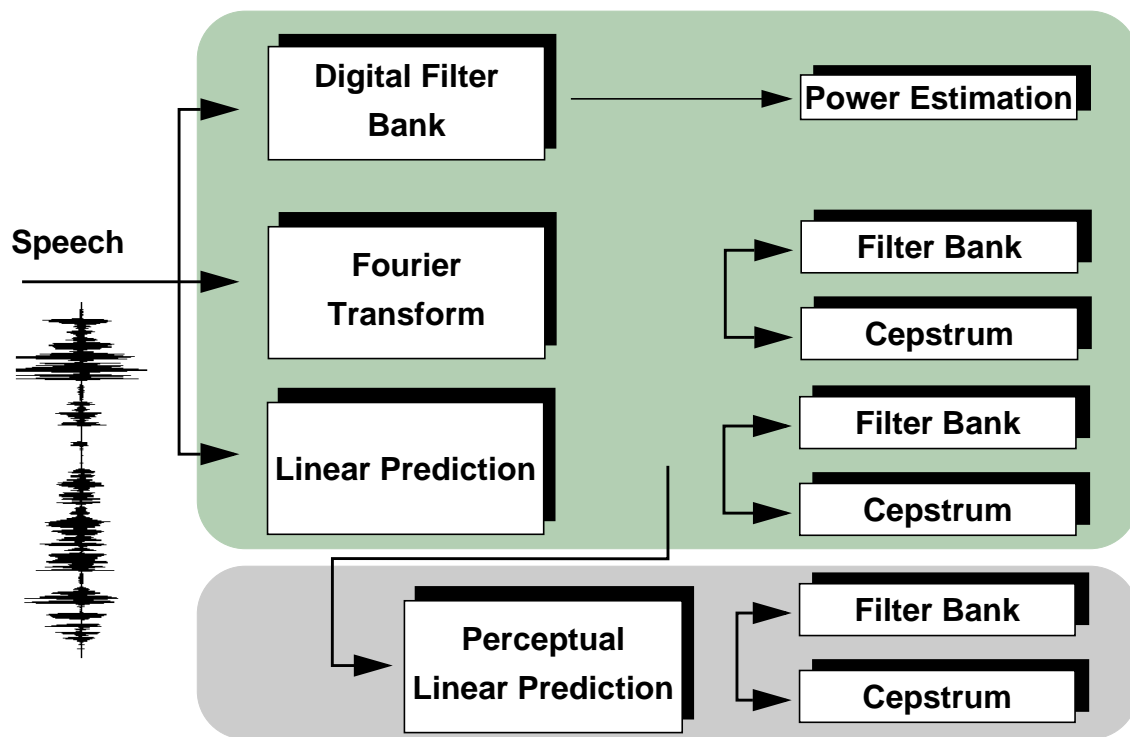
Figure 1. An overview of the front-end portion of the speech recognition system. Two popular analysis techniques, mel-spaced cepstrum and perceptual linear prediction, are supported in the system. Other approaches based on frame-based analysis techniques can be easily added.
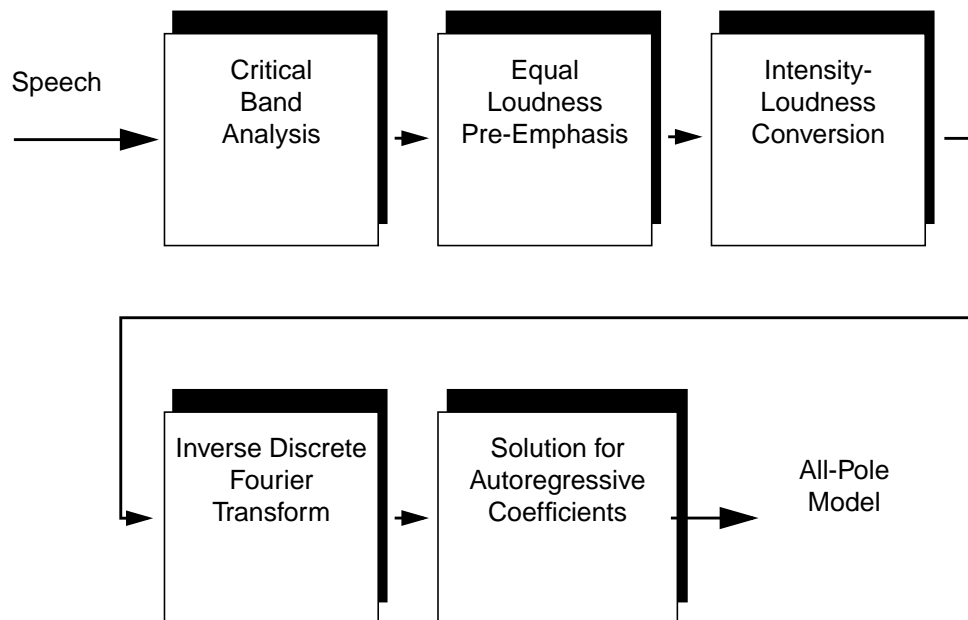
Figure 2. An overview of perceptual linear prediction (PLP) analysis. This front-end has become increasingly popular in recent years.

classes - utilities - scripts - speech - search - up

Any research in the area of signal processing requires the development of large applications in a relatively short period of time. Unfortunately, research commonly suffers from a creative backlog due to rewriting of common functions and the time spent in debugging such things as file I/O. It would be ideal to have a large, hierarchical software environment which can support advanced research in signal processing. The ISIP Foundation Classes (IFCs) and software environment are designed to meet this need, providing everything from complex data structures to an abstract file I/O interface. This software environment starts from the lowest, system level classes, and culminates in a state of the art public domain large vocabulary speech recognition system.

Some significant features of the ISIP software environment include

- unicode compatibility and wide character support to allow multilingual applications
- abstract interface for file i/o
- well–equipped library of DSP functions
- advanced mathematical classes to provide linear algebra and matrix operations

The hierarchical structure of the software environment is as follows:
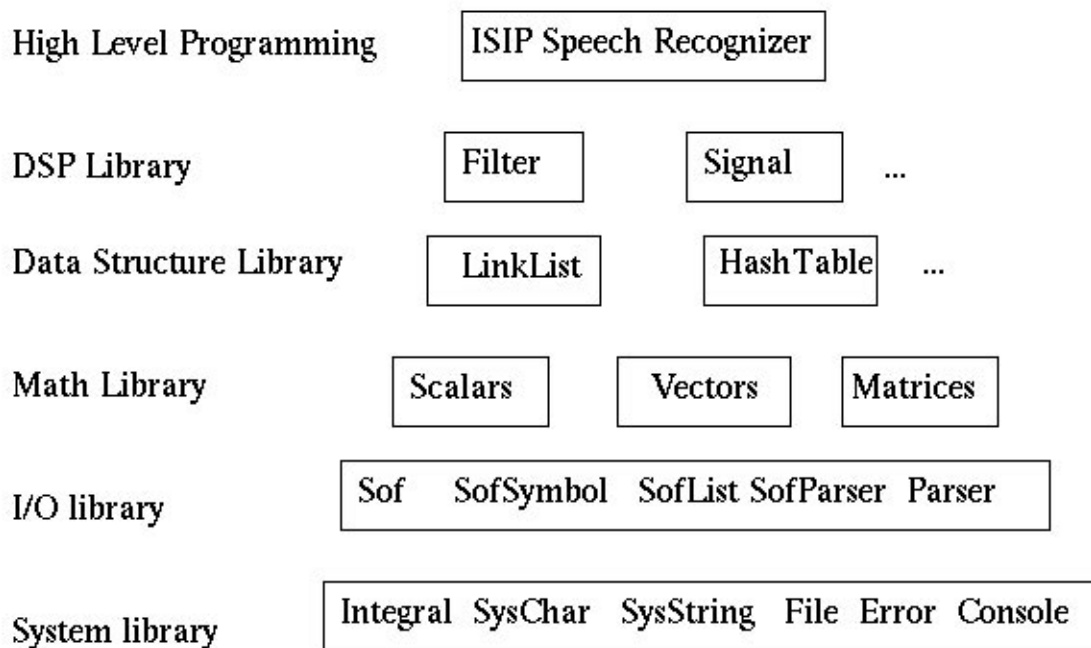

Figure 3. An overview of the hierarchy of ISIP classes. The system and I/O libraries are new additions to the class structure. The math and data structure libraries make extensive use of templates.

```
// file: $isip/class/system/Integral/IntegralTypes.h
// version: $Id: IntegralTypes.h,v 1.4 1999/07/12 18:29:29 duncan Exp $
//

// system include file
//
#include <wchar.h>

// this is the basic isip environment include file. all Integral types
// are defined in this file. these are also implemented as C++ classes.
// all software must be built upon these basic types.
...
typedef void* voidp;
typedef signed char boolean;
typedef unsigned char byte;
typedef wchar_t unichar;
typedef unsigned short int ushort;
typedef unsigned long int ulong;
typedef unsigned long long int ullong;
//typedef short int short;
//typedef long int long;
typedef long long int llong;
//typedef float float;
//typedef double double;
typedef unsigned char byte8;
typedef unsigned short int ushort16;
typedef unsigned long int uint32;
typedef unsigned long long int uint64;
typedef short int int16;
typedef long int int32;
typedef long long int int64;
typedef float float32;
typedef double float64;
```

Figure 4. The integral types define the fundamental building blocks of the ISIP environment. We have taken an approach that requires these types to be a fixed number of bytes.

```
// Scalar: our template scalar class              T max(T arg_1, T arg_2);
//
template<class T>                                 T abs();
class Scalar {                                     T abs(T arg);

protected:                                         T sign();
                                                   T sign(T arg);
  // internal data
  //                                               T factorial();
  T value_d;                                       T factorial(T arg);

public:                                            // useful for DSP
                                                   //
  // required static methods:                      T limit(T min, T max);
  //                                               T limit(T min, T max, T val);
  static String& name();
                                                   T limitHard(T thresh, T new_val);
  // required methods:                             T limitHard(T thresh, T new_val, T arg);
  //  no setDebug required
  //                                               T centerClip(T min, T max);
  boolean debug(unichar* message);                 T centerClip(T min, T max, T arg);
  T size();
                                                   private:
  // initialization and release methods.
  boolean init();                                  public:
  boolean release();
                                                   // define the class name
  // destructors/constructors                      //
  //                                               static const unichar CLASS_NAME[] = L"Scalar";
  ~Scalar();
  Scalar();                                        // define the default value(s) of the class data
  Scalar(Scalar& arg);                             //
  Scalar(T arg);                                   static const T DEF_VALUE = (T)0;
                                                   static const T DEF_RAND_MIN = (T)0;
  // former in-line methods
  //                                                 // default arguments to methods
  operator T();                                    //
                                                   static const long NEGATIVE = (long)-1;
  Scalar& operator= (T arg);                       static const long POSITIVE = (long)1;

  // get methods                                     static const long ERR = (long)20666;
  //
  boolean get(Scalar& arg);                        };
  boolean get(T& arg);
                                                   // all classes need to inherit Scalar
  // assignment methods                            //
  //                                               template class Scalar<long>;
  boolean assign(T arg);                           //template class Scalar<short>;

  // mathematical functions                        // end of include file
  //                                               //
  T min(T arg);                                    #endif
  T min(T arg_1, T arg_2);

  T max(T arg);
```
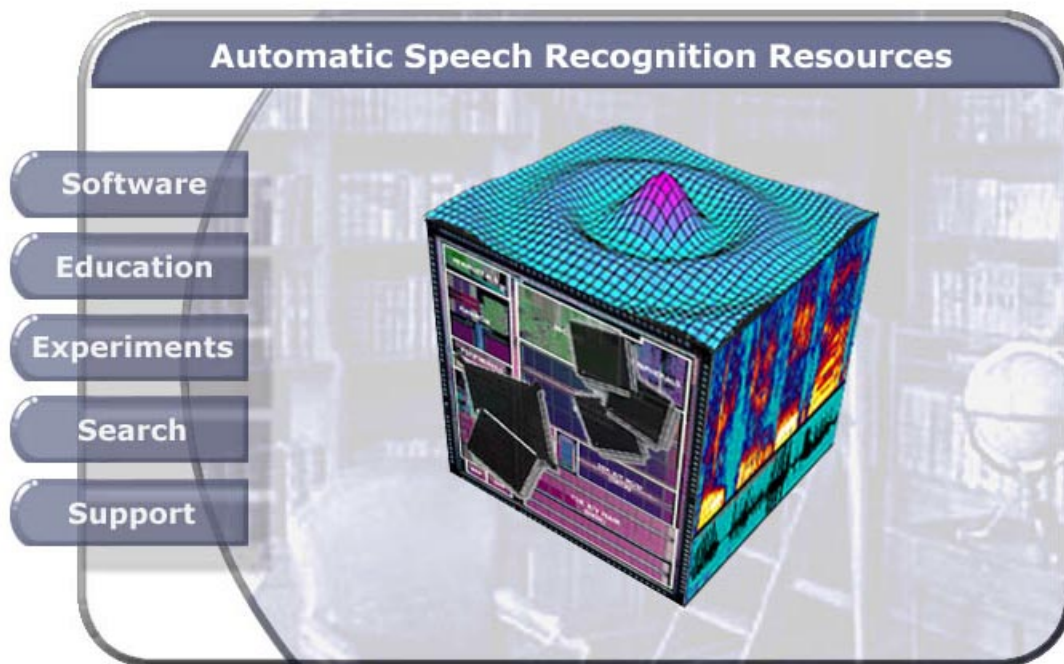
Figure 5. A template class definition for a scalar object. This template is used to build classes such as Long, Short, and Float.

Figure 6. A new set of web pages have been created to support the project. These have been designed to provide easy access to the web site. The choices to the left of the image mirror the physical organization of the web site.

Figure 7. An example page of documentation for the IFCs. The code is directly linked to the page, making it easy for users to view the code while studying the documentation.

FILTER DESIGN TOOL

Control Panel

| Kaiser ▼ | Plot Resolution: | 500 | Compute | Reset |

Enter Parameters for Kaiser Filter

Sample Frequency (Hz)	8000.0
Lower Critical Frequen...	100.0
Upper Critical Frequen...	500.0
Attenuation (dB)	10.0
Transitional Bandwidt...	100.0
Filter Order	0

Filter Coefficients

Numerator (7)   Denominator (0)

H(−3) = 0.0733
H(−2) = 0.0876
H(−1) = 0.0968
H(0) = 0.1
H(1) = 0.0968
H(2) = 0.0876
H(3) = 0.0733

Modify Coefficients

Frequency Response: Magnitude

Frequency Response: Phase

- Source Code

- A simple tutorial on using this applet.

Figure 8. An example of a Java Swing applet that demonstrates the concept of digital filter design. Swing has been a mixed blessing. While some aspects of GUI programming are nicely abstracted, other aspects, such as interactions between grid boxes and event handlers, have been problematic.
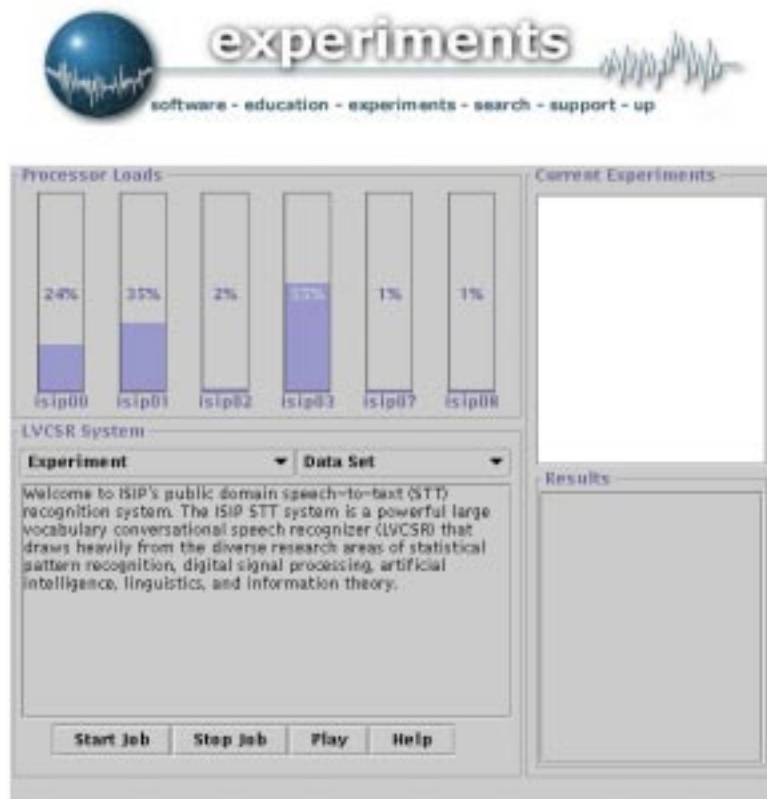
Figure 9. A Java applet that allows users to submit speech recognition jobs remotely to a bank of compute servers. Users can run canned experiments, or supply their own audio data. Parameters for the experiment can be specified via dialog boxes. Results are emailed to the user, and can be examined directly on the web site via links provided in the dialog boxes to the right.