

# **T-61.184**

## **Automatic Speech Recognition: From Theory to Practice**

`http://www.cis.hut.fi/Opinnot/T-61.184/`  
November 8, 2004

**Prof. Bryan Pellom**

Department of Computer Science  
Center for Spoken Language Research  
University of Colorado

`pellom@cslr.colorado.edu`

**T-61.184**

# Announcements

- **I still need 3 more volunteers to present their project topic on November 22<sup>nd</sup>**
- **The goal is to present to the class (and myself) your chosen topic area.**
- **Brief 10 minute presentation (project overview)**
- **Does not have to reflect your completed project (since that is due December 8<sup>th</sup>).**

# Today

- **Speech Processing & Recognition Toolkits**
- **Language Modeling Tools & Standards**
- **Review Speech Recognition Systems**
- **Industry vs. Academic Recognizers**
- **Trends in the Speech Recognition Field**
- **Hot Topics for the Future**

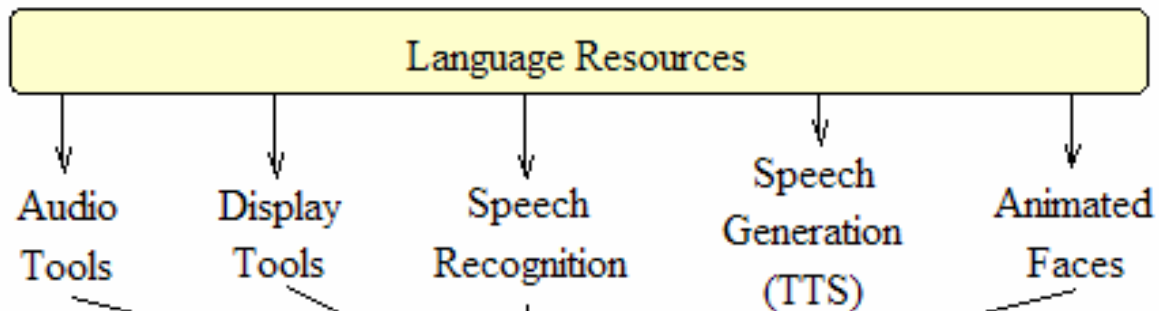
# Speech Toolkits

# CSLU Speech Toolkit

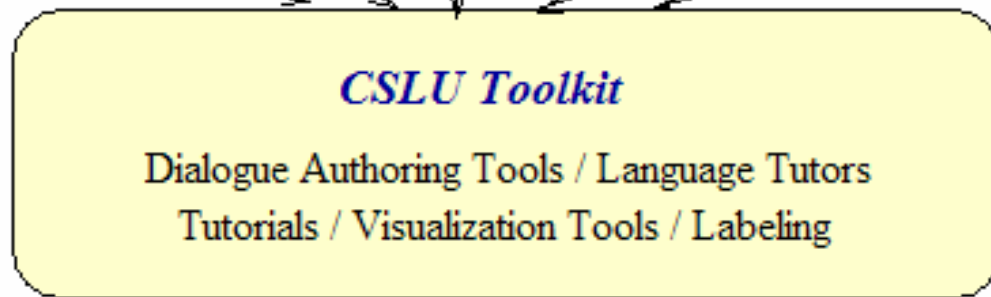
- **Speech Toolkit under development since 1992**
- **Oregon Graduate Institute (OGI)**
  - Oregon Health and Science University (OHSU)
- **C / C++ algorithms wrapped with Tcl/Tk language**
- **Contains**
  - HMM (Neural Network / Gaussian) Speech Recognition
  - Facial Animation (CU Animate from Univ. of Colorado)
- **Download,**
  - <http://www.cslu.ogi.edu/toolkit>

# CSLU Speech Toolkit

*Fundamental Components:*



*System Integration:*



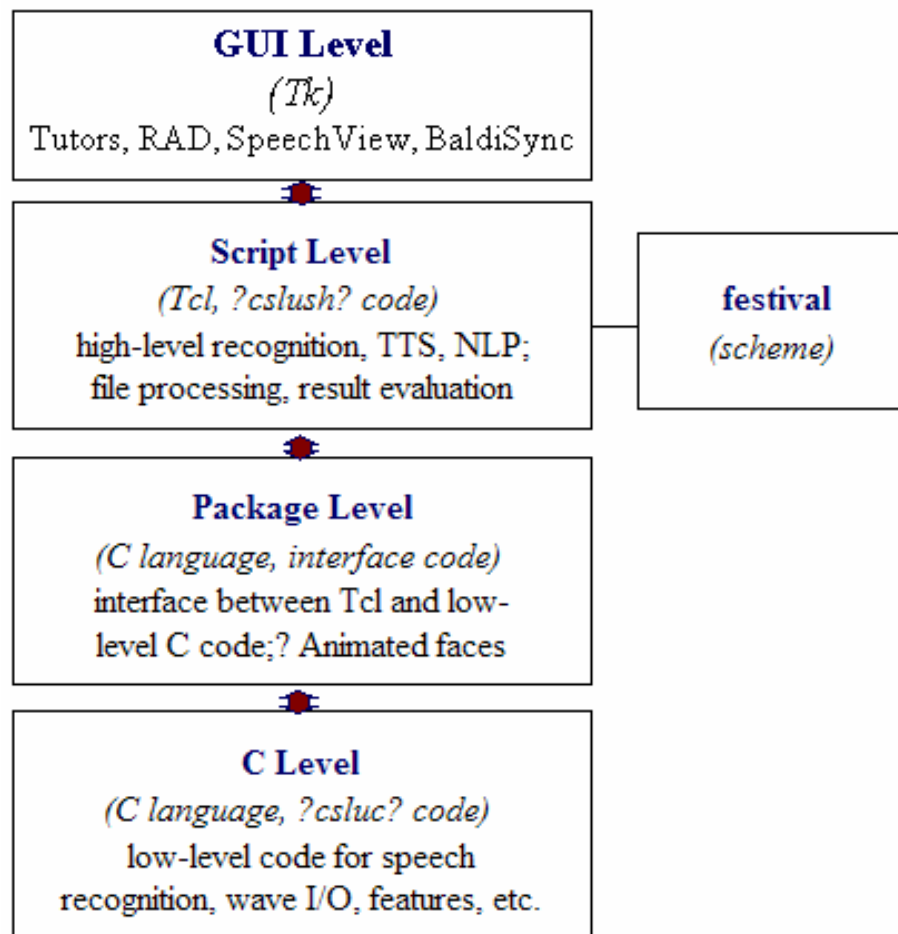
*Technology Transfer:*

High Schools??? Universities??? Researchers??? Industry

Four arrows point downwards from the bottom of the CSLU Toolkit box to the text "High Schools???", "Universities???", "Researchers???", and "Industry".

T-61.184

# CSLU Speech Toolkit Architecture



T-61.184

# What makes the CSLU Toolkit Appealing?

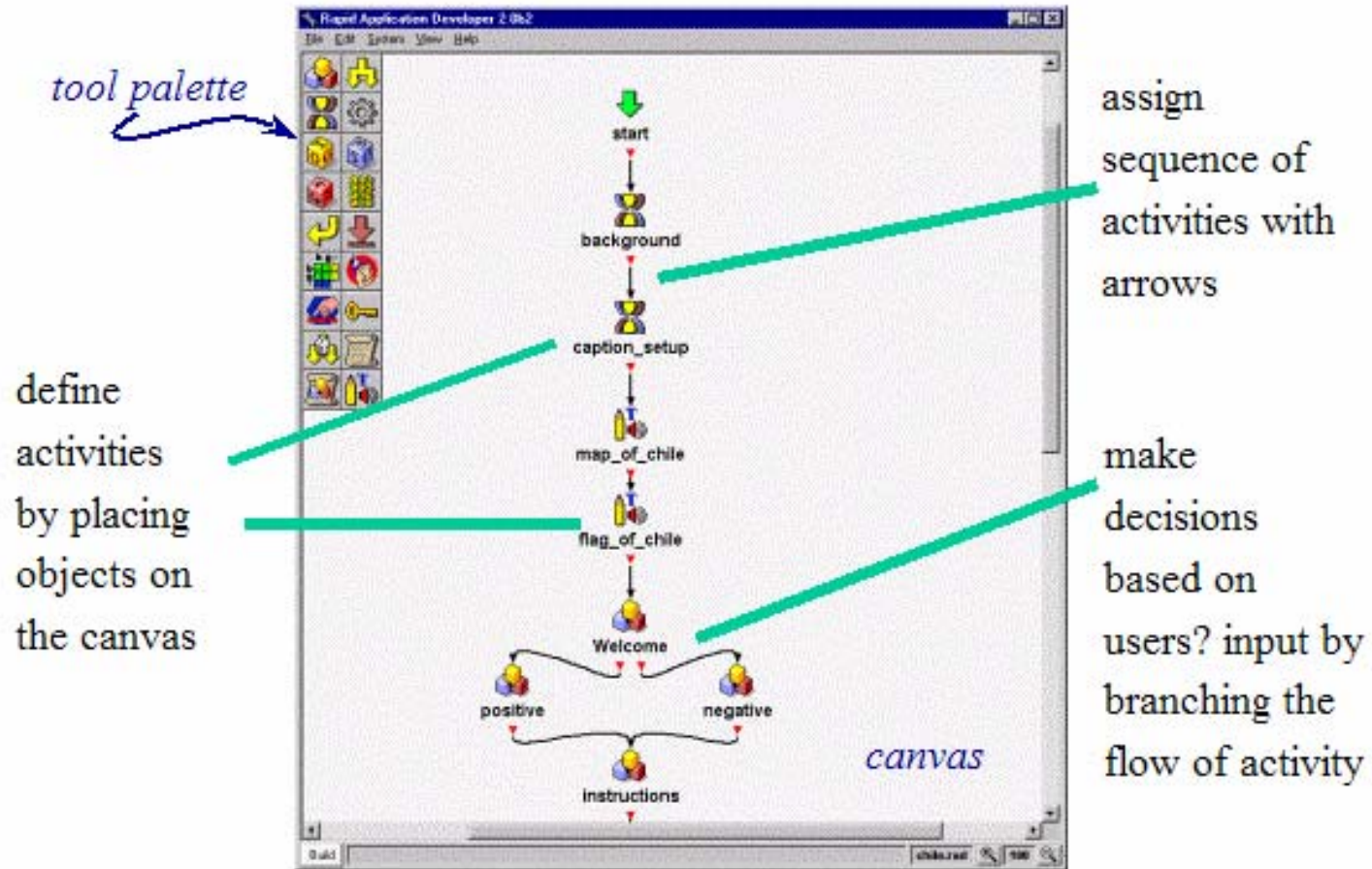
- **Ease of use**
  - Researchers have put time into a managed download
- **Modular and Extendable Framework**
  - Tk Language wonderful for GUI design, simple language to learn
  - Tcl provides scripting language to integrate core technologies
  - Core algorithms written in C for speed but accessed through Tcl functions
- **Tutorials provided with toolkit to assist researchers in designing and developing applications**
- **Works well under realistic conditions**



# What's Inside the Toolkit?

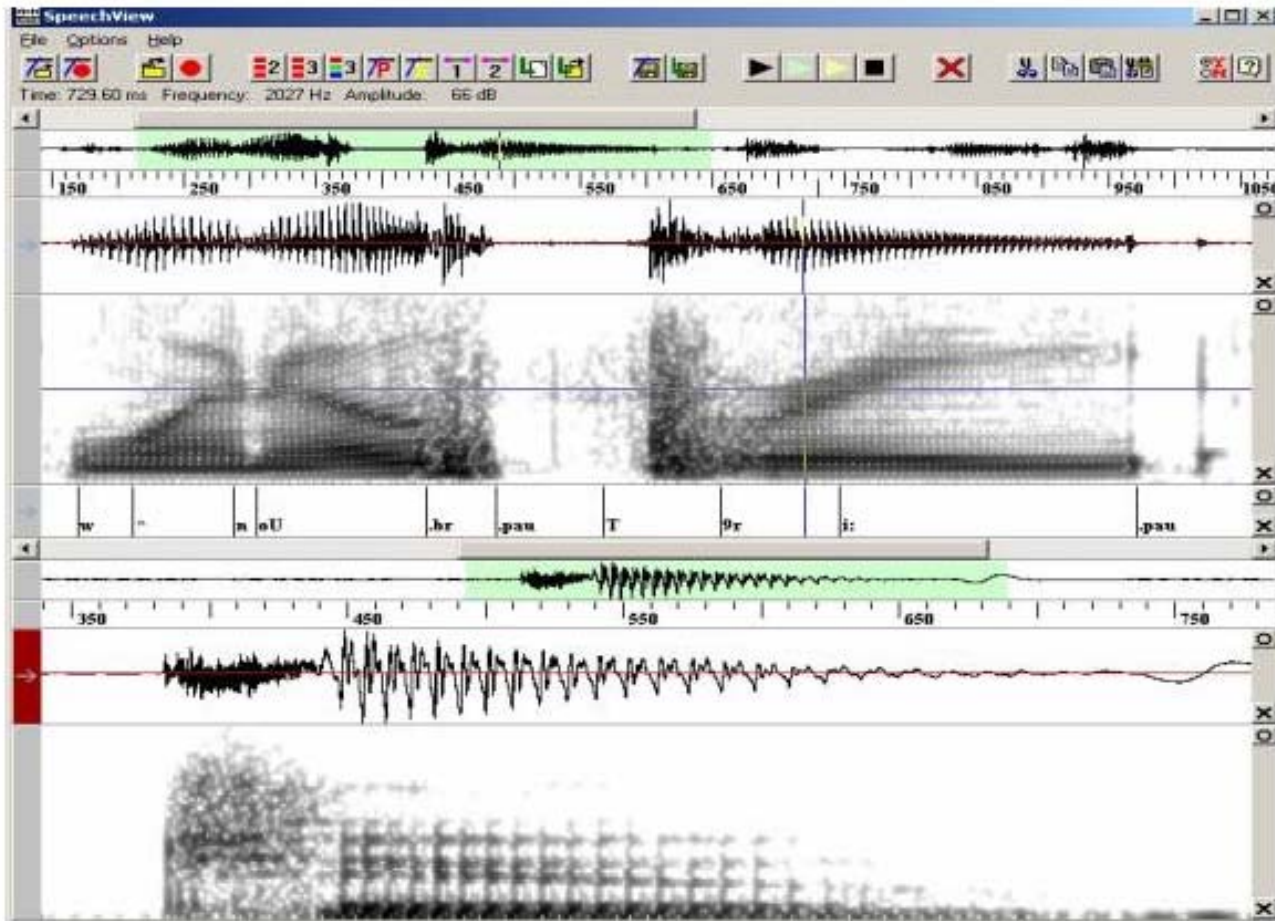
- **Audio Processing Routines & Feature Extraction**
- **Tools to record and display audio**
  - SpeechView
- **Speech Recognition**
  - Standard HMM with Gaussians or HMM/ANN hybrid
- **Text-to-Speech Synthesis**
  - Festival TTS system from University of Edinburgh / CMU
- **Facial Animation**
  - Baldi (University of California at Santa Cruz)
  - CU Animate (University of Colorado)
- **Natural Language Understanding**
  - “Profer” Parser
- **Directed Dialog Application Builder**
  - “Rapid Application Developer (RAD)”

# Rapid Application Developer (RAD)



T-61.184

# SpeechView (CSLU Toolkit)



T-61.184

# Snack Sound Toolkit

- **Center for Speech Technology, KTH, Stockholm Sweden**
  - Written and developed by Kåre Sjölander
- **Audio functionalities via Tcl/Tk and Python environment**
- **Download from**
  - <http://www.speech.kth.se/snack/>
- **Provides,**
  - High-level sound objects
  - Streaming support
  - Multi-platform
  - Multiple simultaneous record and playback threads
  - Real-time signal filtering
  - Real-time signal visualization (waveforms, spectrograms, etc)

# Tcl/Tk Interface with Snack

- **To read and play a wav file,**

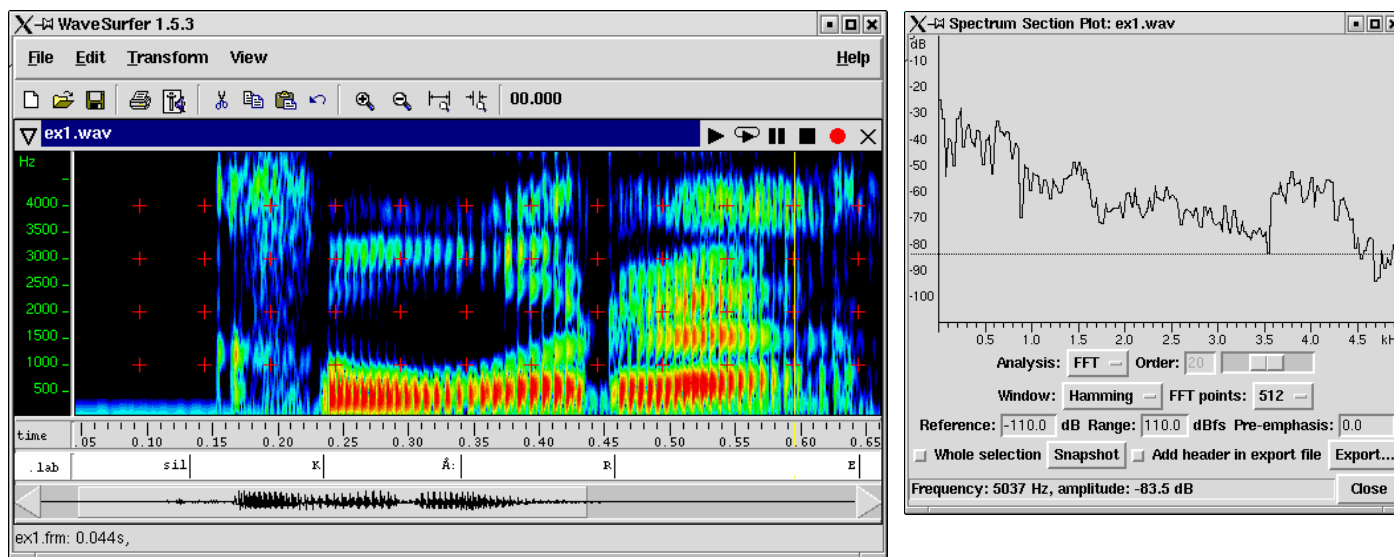
```
package require snack
snack::sound snd
snd read ex1.wav
snd play
```

- **To plot the waveform,**

```
canvas .c
pack .c
.c create waveform 0 0 -sound snd
```

# Wavesurfer

- An audio play/record utility written using the Snack toolkit interface to Tck/Tk
- Similar to “SpeechView” in the CSLU Toolkit



T-61.184

# Hidden Markov Model Toolkit (HTK)



- **Developed by Cambridge University (1989-)**
  - ❑ First version by Prof. Steve Young
  - ❑ Speech, Vision, and Robotics Group
  - ❑ Set of C libraries and tools for speech recognition research
- **Entropic Research Laboratories (1993-1999)**
  - ❑ Commercialized the software, maintained support
  - ❑ Mid 1990's – joint venture between Cambridge and Entropic
  - ❑ Microsoft bought out Entropic in 1999.
- **Open Source HTK: (2000-)**
  - ❑ HTK API (HAPI) developed and distributed with HTK
  - ❑ <http://htk.eng.cam.ac.uk>

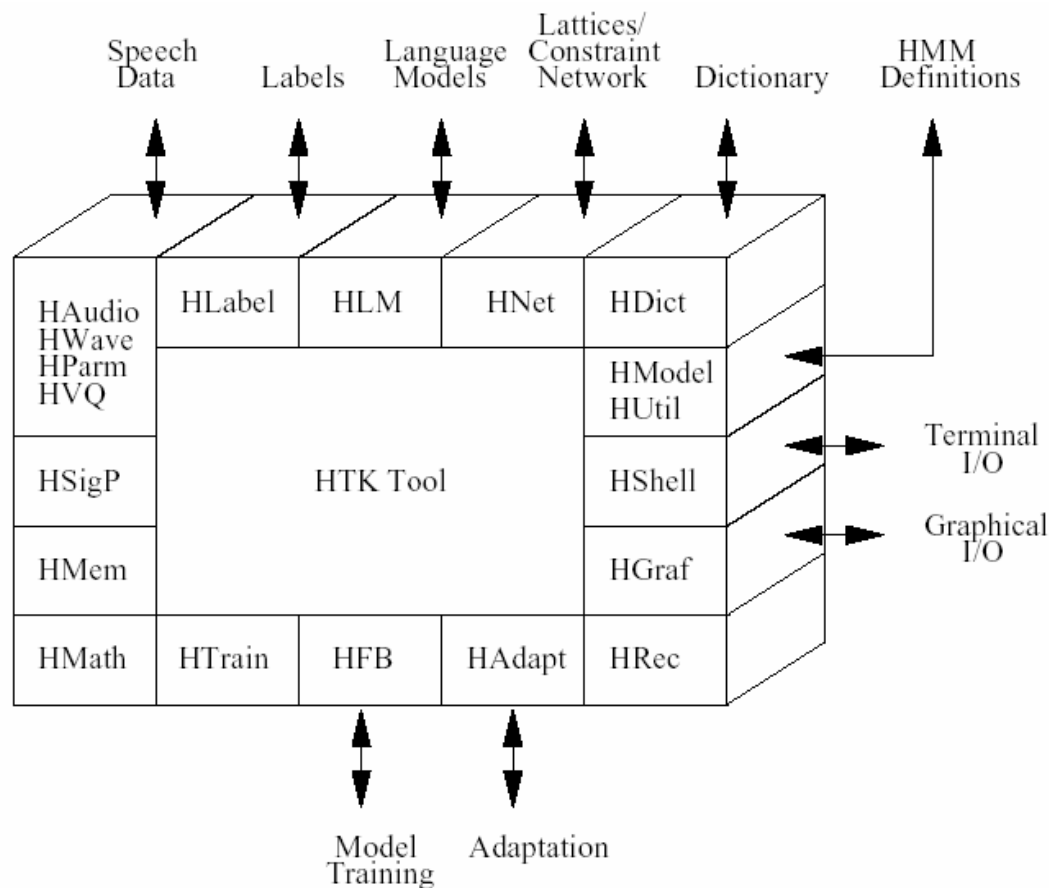
# Hidden Markov Model Toolkit (HTK)

## ■ Open Source HTK: (2000-)

- You can build a product with HTK, but can not redistribute the source code.
- Can use HTK to train acoustic models for commercial products
- HTK is a toolkit for speech recognition research, not a general-purpose dictation system



# HTK Architecture

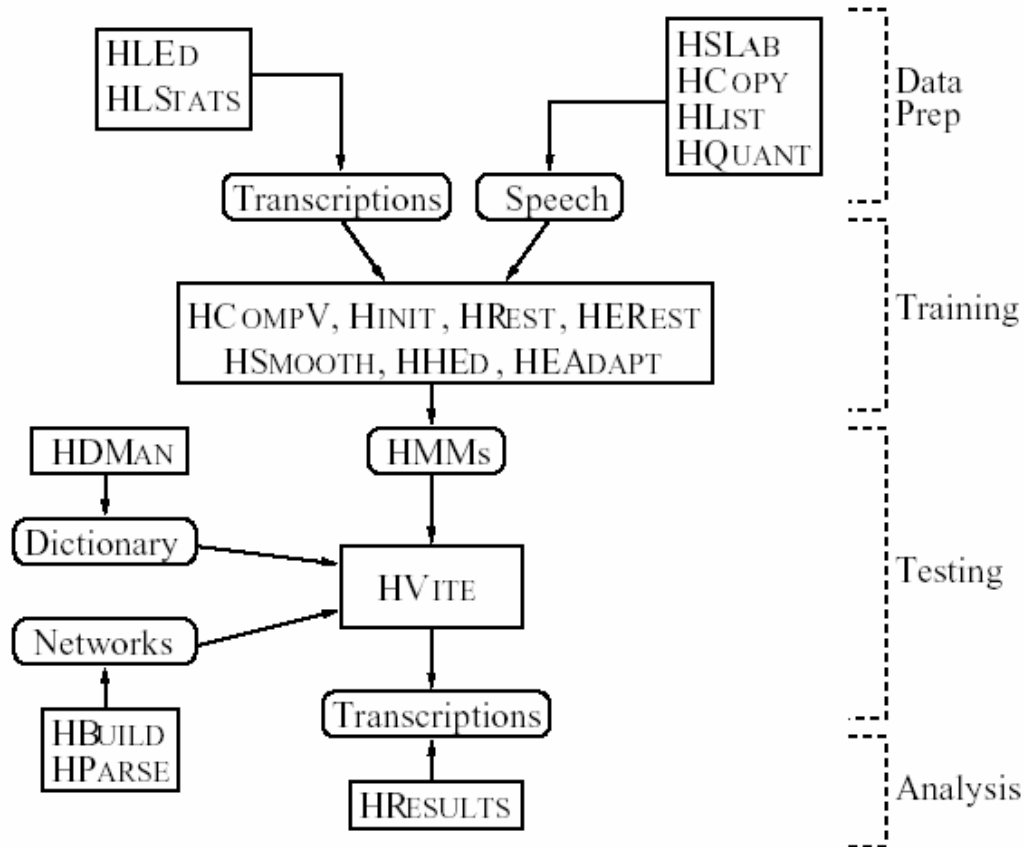


- **System divided into library modules**
- ***HShell* – user input and output**
- ***HMem* – memory management**
- ***HLM* – language model interface**
- ***HNet* – Finite state grammars**

T-61.184

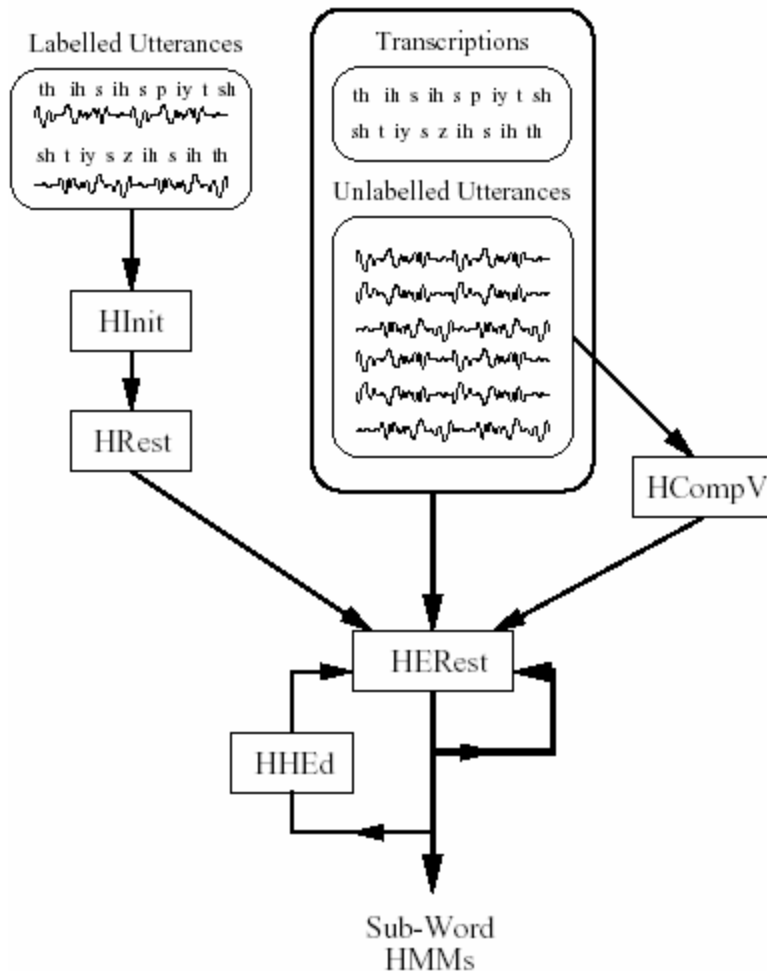
# HTK Architecture

- HTK functions are accessed from the Unix command-line
- Tools provided for estimating HMM parameters
- Testing can be done with Viterbi decoder (HVite)



T-61.184

# The HTK Model Training Process



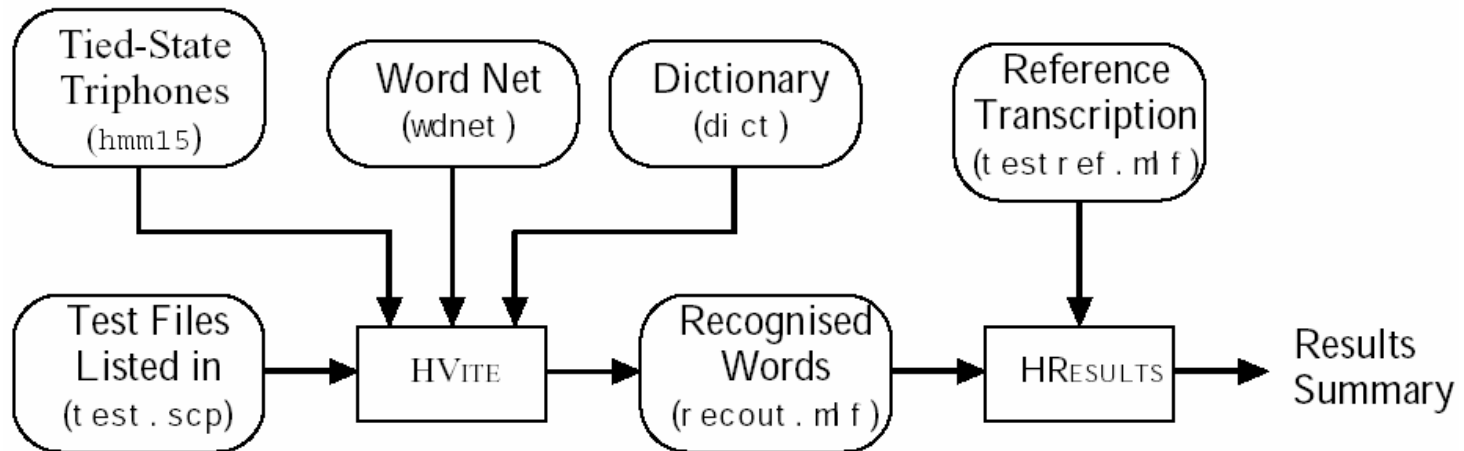
- Requires transcriptions and audio files. Data can be hand-aligned (shown left) or automatically aligned during training (right-branch)
- HRest and HERest perform Baum-Welch estimation of HMM parameters
- HHed allows for various parameter tying schemes and mixture incrementing

# The HTK Recognition Tool (HVite)

- **Performs Viterbi-based recognition using the Token-passing algorithm**
- **Batch-mode or Live-mode recognition**
- **Supports cross-word triphones**
- **Can generate lattices**
- **Language Models Supported:**
  - Word loops,
  - Finite State Grammars
  - Back-off bigram language model

# The HTK Recognition Tool (HVite)

- **Typical Testing Setup for HTK:**



- ***HVite* is the decoder and *HResults* provides word error rate metrics**

# Language Modeling Tools for Speech Recognition

# HTK's Language Model Toolkit

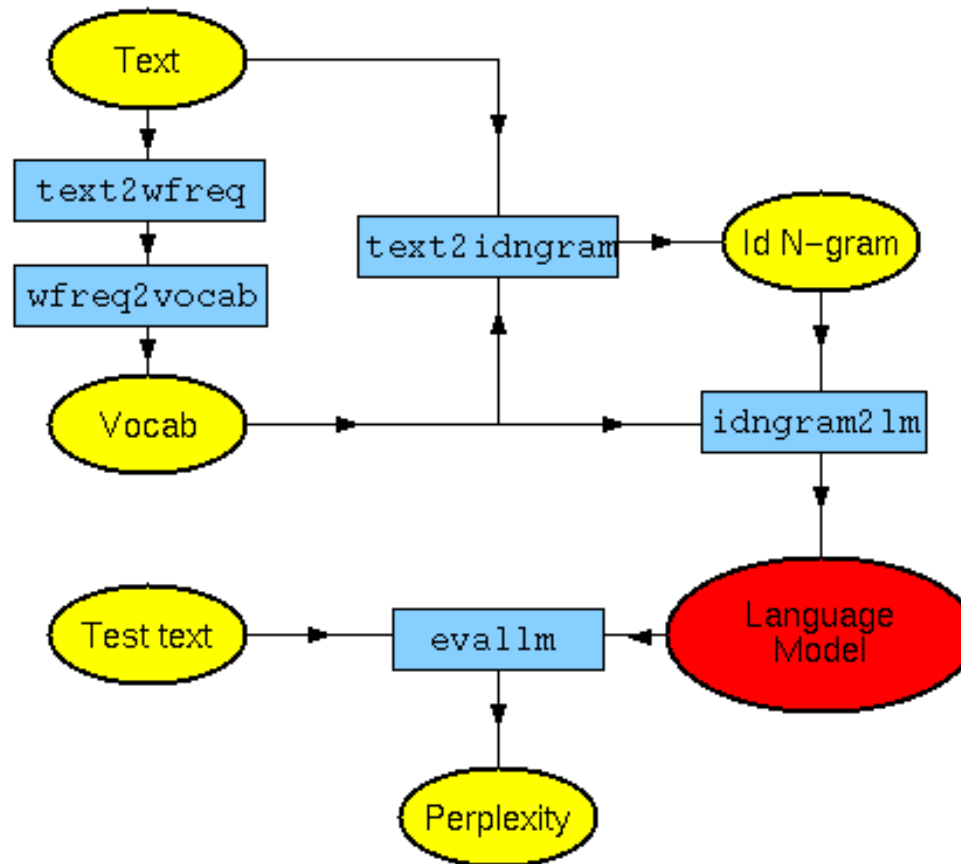
- Recently added with version 3.2 of HTK
- Supports n-gram language models and class-based n-gram models
- Allows for unsupervised determination of word-classes using “word-exchange” algorithm
- Allows for language models to be merged and also perplexity calculation from interpolated language models.

# CMU/Cambridge Language Model Toolkit

- **Developed by P.R. Clarkson and R. Rosenfeld**
  - `http://mi.eng.cam.ac.uk/~prc14/toolkit.html`
- **Estimates N-gram models (arbitrary N)**
- **64,000 word or less vocabularies**
- **Implements several discounting strategies**
  - Witten Bell, Good Turing, Absolute, Linear Discounting
- **Very easy to use, but no longer updated (1999-)**



# CMU/Cambridge Language Model Toolkit



T-61.184

# SRI Language Model Toolkit (SRILM)

- **Under development since 1995**
  - <http://www.speech.sri.com/projects/srilm/>
- **Mainly supports N-gram language models**
  - N-grams of arbitrary length
- **Vocabularies > 64k**
- **Implements several discounting strategies including modified Kneser-Ney**

# SRI Language Model Toolkit (SRILM)

- **Also provides,**
  - ❑ Ability to prune n-gram language models
  - ❑ Generate random sentences based on LM statistics
- **Language Model Types**
  - ❑ Skip-grams, Cache Language Models, Class-based language models
- **Other functionalities**
  - ❑ N-best list rescoring
  - ❑ Lattice rescoring
- **Worth noting that SRILM assumes single-byte character encodings for handling text data.**

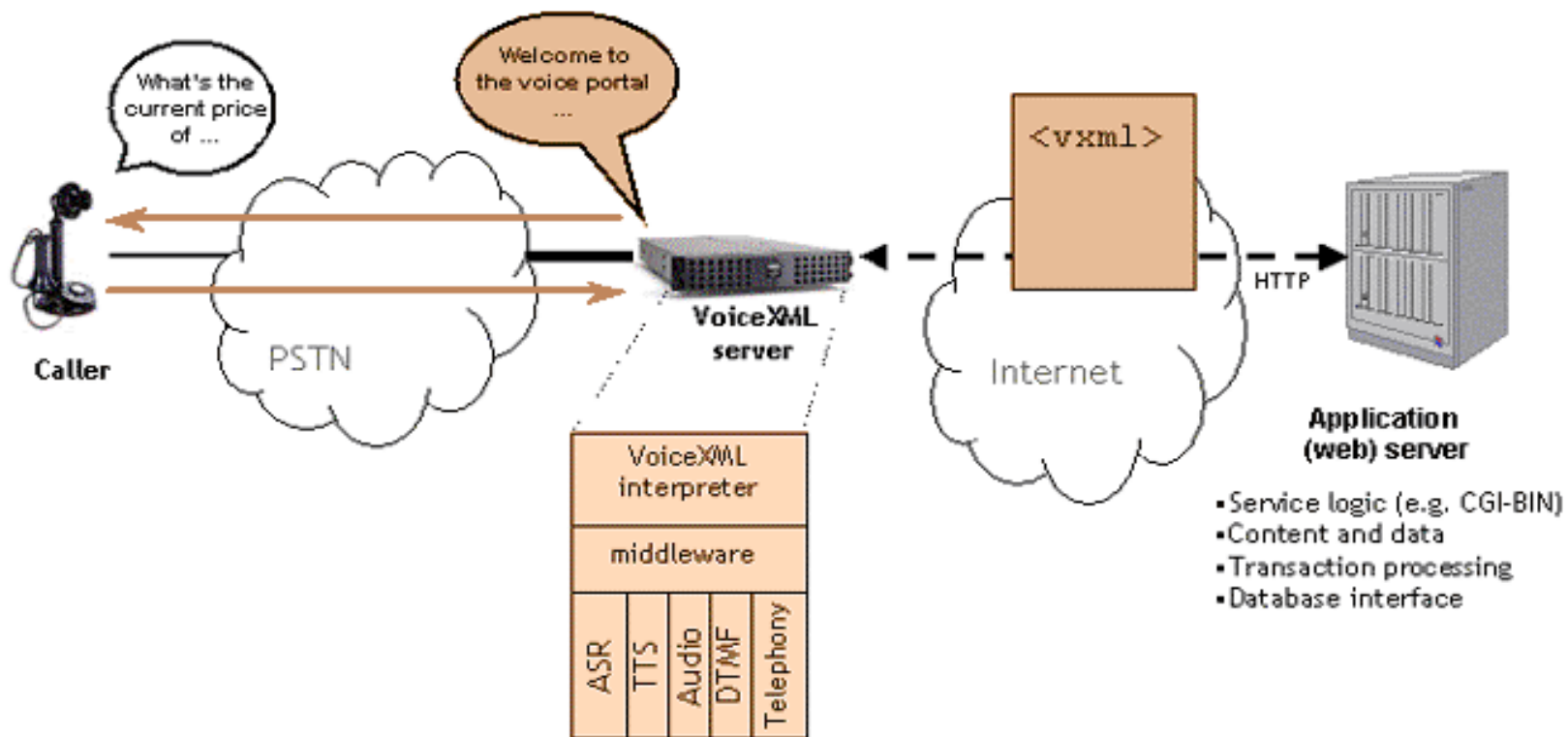
# Voice-Standards Impacting ASR Design

**T-61.184**

# Voice Extensible Markup Language (VoiceXML)

- **A markup language for creating voice-based applications**
  - Version 1.0 was first published in March 2000
- **Assumes a voice browser with:**
  - Audio and keypad input,
  - Audio output
- **The voice browser typically runs on a specialized voice gateway node**
  - Connected to the Internet and,
  - Connected to the public switched telephone network

# VoiceXML Component Interface



T-61.184

# Example VoiceXML Script

```
<?xml version="1.0"?>
<vxml version="1.0">
  <menu>
    <prompt> Choose from <enumerate/></prompt>

    <choice next="sports.vxml"> sports </choice>
    <choice next="weather.vxml"> weather <choice>
    <choice next="news.vxml"> news <choice>

    <help>
      If you would like sports scores, say sports.
      For local weather reports, say weather, or
      for the latest news, say news.
    </help>

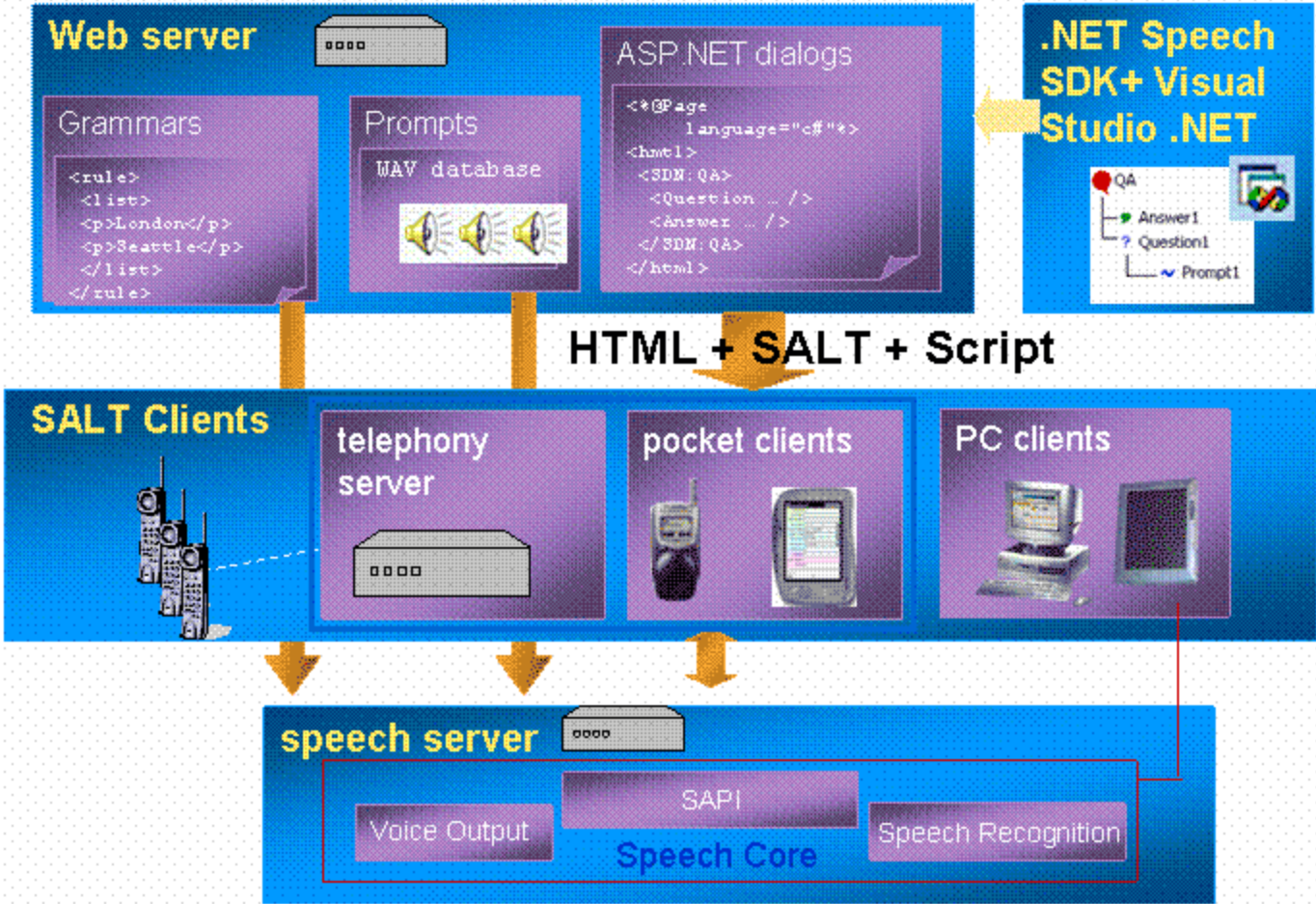
    <noinput>You must say something.</noinput>
    <nomatch>Please speak clearly and try again.</nomatch>
  </menu>
</vxml>
```

# Microsoft Speech Application Language Tags (SALT)

- **Extension of HTML**
- **Adds speech interface onto web pages**
- **Tags are designed for,**
  - Voice Only browsers (e.g., connected to a telephone)
  - Visual Browsers (to add a multi-modal capability)
- **Example Tags:**
  - `<prompt>`, `<listen>`, `<dtmf>`, `<smex>`,  
`<record>`, `<bind>`, `<grammar>`



# SALT Architecture



# Uses of SALT

- **Speech input/output augmented web pages**
  - Speech-driven form filling
- **Dialog flow-control for voice-only access**
  - Telephony applications
- **Multi-modal access from a wide range of devices**
  - PDA
  - Tablet PC
  - Telephone, Cellphone
  - Desktop PC

# CMU Open SALT Browser

- **Developed at Carnegie Mellon University**
  - <http://hap.speech.cs.cmu.edu/salt/>
- **Consists of,**
  - Festival                      text-to-speech synthesizer
  - Sphinx-II                      CMU speech recognizer
  - Mozilla                         web-browser
- **Worth trying out the download and playing with the system.**
- **Binaries are available for Windows**
- **Source code is provided on the website**

# Speech Recognition Systems and Useful Toolkits

**T-61.184**

# CMU Sphinx-II

- **Recognizer Developed at Carnegie Mellon University**
  - ❑ <http://cmusphinx.sourceforge.net>
- **Open Source, Real-time recognizer**
- **Uses 4 feature streams**
  - ❑ 12 MFCC,
  - ❑ 12 delta MFCC,
  - ❑ 12 delta-delta MFCC,
  - ❑ power, delta power, delta-delta power
- **Semi-continuous Hidden Markov Models (SCHMM)**
  - ❑ 4 codebooks containing 256 Gaussians

# CMU Sphinx-II

- **Senone Based Acoustic Clustering**
  - State-dependent output distributions shared across different phonetic models
  - Each state represented by a weighted set of Gaussians (Gaussians modeled by VQ codebooks)
- **Lexical Prefix Tree or Flat Search**
- **Cross-word acoustics in first pass**
- **N-gram ( $N \leq 3$ ) in first-pass of search**
- **Vocabulary size restricted to 64k words**
- **Prelude to Microsoft Whisper Speech Recognizer**

# CMU Sphinx-III

- **Extends on Sphinx-II by providing continuous density HMMs (CDHMMs) for acoustic models**
- **Also recently made open source through BSD Style license**
- **Supports back-off bigram and trigram language models**
- **Generally much slower than Sphinx-II, but more accurate.**

# Recent Work on CMU Sphinx-III

- **CMU Researchers are still updating Sphinx-III**
- **Focus is on real-time implementation and API**
- **Sphinx 3.5 improvements**
  - LDA/HLDA feature-space transforms
  - Continuous Listening Mode
  - Phoneme Lookahead
  - MLLR speaker adaptation (model-space transform)
- **More information on Arthur Chan's website,**
  - <http://www-2.cs.cmu.edu/~archan/>



# CMU Sphinx-4

- **Sphinx-III ported and extended to the Java programming language**
- **Joint collaboration between**
  - Sphinx group at Carnegie Mellon University,
  - Sun Microsystems Laboratories,
  - Mitsubishi Electric Research Labs (MERL),
  - Hewlett Packard (HP)
- **Batch-mode or Live Mode Decoding**
- **Supports n-gram language models and Java Speech Grammar Format (JSGF).**

# Java Speech Grammar Format (JSGF)

```
#JSGF V1.0
```

```
public <basicCmd> =<startPolite><command><endPolite>;
```

```
<command> = <action> <object>;
```

```
<action> = open | close | delete | move;
```

```
<object> = [the | a] (window | file | menu);
```

```
<startPolite> = (please | kindly | could you | oh  
mighty computer) *;
```

```
<endPolite> = [ please | thanks | thank you ];
```

# Sphinx-4 Performance

- **TIDIGITS**                      **0.55% WER**                      **0.07x RT**
  - Digits 0-9
- **WSJ5k**                              **6.97% WER**                      **1.22x RT**
  - 5k vocabulary Dictation task
- **HUB4**                                      **18.7% WER**                      **4.4x RT**
  - 60k vocabulary broadcast news transcription

***WER = Word Error Rate (lower is more accurate)***

***RT = Real-Time Factor (lower is faster)***

# CSLU/OGI Speech Recognizer

- **Designed within the CSLU Speech Toolkit (packages in C/Tcl)**
- **Supports HMMs with Mixture Gaussians**
- **Also supports HMM / Neural Network Hybrid Recognition**
  - ❑ 3-layer MLP
  - ❑ ~200 hidden nodes typical
  - ❑ Typically uses biphone unit clustering (determines # of output nodes in NN)
- **Word-internal acoustic modeling only (at least as of 1999)**
- **Tutorial on Training/Testing a recognizer with the CSLU Toolkit is available**
- <http://cslu.cse.ogi.edu/research/asr.htm>

# Mississippi State (ISIP) Speech Recognizer

- **Institute for Signal and Information Processing (ISIP)**
  - ❑ <http://www.isip.msstate.edu/projects/speech/software/>
- **Open Source license**
- **Modular C++ based speech recognition system**
  - ❑ Complete end-to-end toolkit with tutorials for retraining and testing
- **Developed over several years (by many students)**
  - ❑ Strict Programming Style enforced
  - ❑ Strong project management principles
- **Implements many state-of-the-art methods**
  - ❑ May be lacking speaker adaptation?
  - ❑ Speed issues?

# Neural Inference Computation (NICO) Toolkit

- **Developed by Nikko Ström**
  - Department for Speech, Music, and Hearing at KTH, Stockholm, Sweden
  - <http://www.speech.kth.se/NICO/>
- **Neural Network Toolkit for Speech technology applications**
- **Focuses on Recurrent Neural Network (RNN) for modeling phoneme probabilities**
- **Fast Back-propagation learning algorithm**
- **Toolkit has not been updated in quite some time (2000-)**

# University of Washington

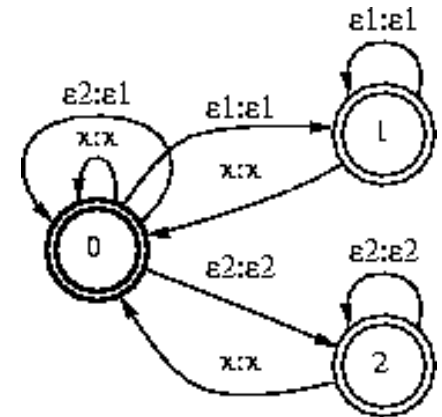
## Graphical Model Toolkit (GMTK)

- **University of Washington**
  - <http://ssli.ee.washington.edu/~bilmes/gmtk/>
- **Toolkit for developing graphical-model and dynamic Bayesian network (DBN) based speech recognition and general time series systems**
- **GMs represent statistical processes using a graph (a set of nodes and edges)**
  - Nodes are random variables
  - Edges encode conditional independent properties

# AT&T

## Finite State Machine (FSM) Toolkit

- Developed by AT&T and provided for non-commercial use
  - <http://www.research.att.com/sw/tools/fsm/>
- Software for building, combining, optimizing, and searching *weighted finite-state acceptors and transducers*
- Finite-state transducers are automata for which each transition has an output label in addition to the more familiar input label.
- Weighted acceptors or transducers are acceptors or transducers in which each transition has a weight as well as the input or input and output labels.



T-61.184



# AT&T

## Finite State Machine (FSM) Toolkit

- **Useful since it provides an efficient search network for speech recognition**
- **Complex search structures and acoustic model topologies can also be constructed using the toolkit**
- **Tutorial about FST's online at,**
  - `http://www.research.att.com/sw/tools/fsm/tut.html`

# University of Colorado

## SONIC Speech Recognizer

- A complete end-to-end recognition engine
- Continued development since March 2001
  - ❑ Binaries / Libraries Available
  - ❑ [http://cslr.colorado.edu/beginweb/speech\\_recognition/sonic.html](http://cslr.colorado.edu/beginweb/speech_recognition/sonic.html)
- Designed for efficient performance
- Modular, to support research needs
- Implements state-of-the-art techniques for search, adaptation, language modeling

## Current Capabilities

- **Keyword, Grammar, & Continuous Speech**
- **Phonetic Alignment**
- **Speech Detection**
- **Speech Compression / VoIP**
- **Speaker Adaptation**
- **Batch or Live-Mode Recognition**
- **Client / Server Interface**
- **Portable to new languages**
- **API**

# Interesting Aspects of SONIC

- **Supports VoIP & Client/Server Interface**
- **Rapidly portable to new languages**
  - French, German, Italian, Spanish, Japanese, Turkish, Russian, Arabic, Croatian, Portuguese, Korean, Polish, Finnish (recently)
- **Efficient in many respects**
  - Training is fast & simple (and parallel!)
  - Decoding is equally fast
- **Phoenix Semantic Parser Integrated into Recognizer**
  - Allows for “concept” language modeling directly within decoder
- **State-of-the-Art performance**
  - Uses new PMVDR feature representation

# SONIC Performance

- **TIDIGITS**                      **0.40% WER**                      **0.05x RT**
  - Digits 0-9
  - (0.16% WER after adaptation)
  
- **WSJ5k**                              **4.2% WER**                              **0.68x RT**
  - 5k vocabulary Dictation task
  - (2.8% WER after adaptation)
  
- **HUB4**                                      **14.4% WER**                                      **< 3.0x RT**
  - 60k vocabulary broadcast news transcription
  - (12.3% WER after adaptation)

# Industry and Academic Trends in Speech Recognition & Interesting Application Areas

**T-61.184**

# Industry Trends

- **Less Emphasis today on Dictation**
- **Focus on grammar-based applications “over the phone”**
  - Call Center applications
- **More recently a trend to support statistical language models and “say anything” technologies**
  - Nuance, ScanSoft (formerly SpeechWorks)
  - Natural Language Call-routing (AT&T)
- **Movement towards “Speech Servers” with a single priced licensing model**
  - Microsoft
- **Speech recognition embedded on the cell phone (with increasing complexity)**
  - Voice Signal Technologies

# Academic Research Trends

- **Generally driven by large-scale government sponsored programs**
  - ❑ DARPA Communicator (Spoken Dialog Systems)
  - ❑ DARPA Babylon (Two-way speech-to-speech translation)
- **Weighted Finite-State Transducers (WFSTs)**
- **Speaker-Adaptive Acoustic Training**
- **Discriminative Acoustic Training**
  - ❑ Minimum Phone Error (MPE) training,
  - ❑ Maximum Mutual Information (MMI) training
- **Novel Features, Robust ASR**
- **Seems to be less emphasis on Neural Network approaches (HMM/ANN hybrids)**
- **Rapid portability to new languages, handling data-sparse tasks**

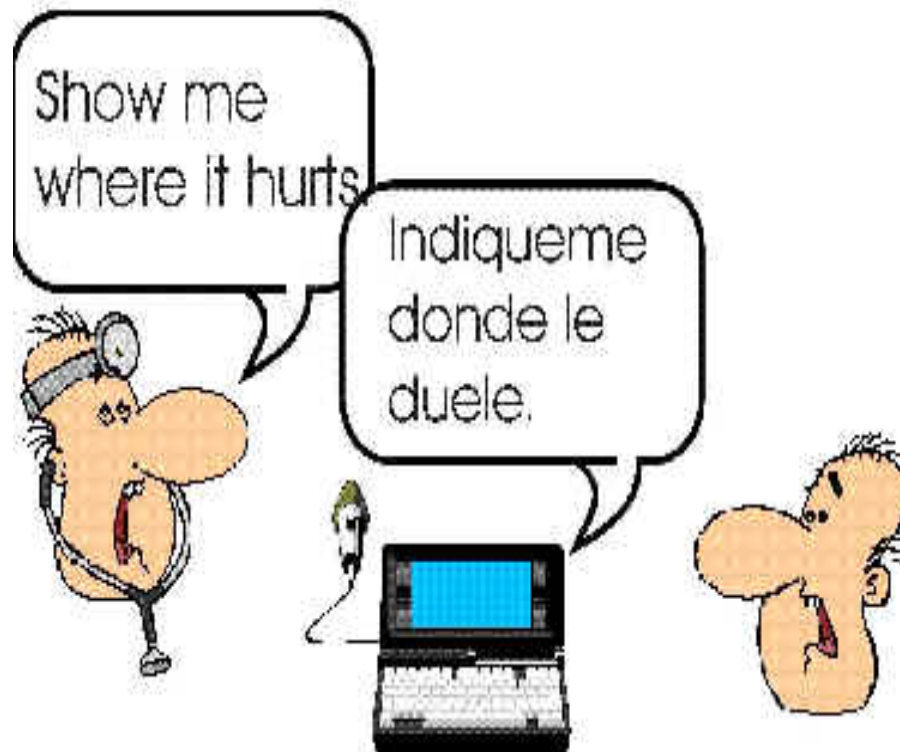


# Academic Research Trends

- **In USA, research driven by application domains**
- **Voice-Interfaces for Question/Answering**
- **Transcription of Broadcast News**
  - Speaker Segmentation
  - Rich Transcription (who is speaking, when)
  - Disfluency detection and modeling
- **Real-time recognition for difficult and large vocabulary speech tasks**
  - DARPA Ears program

# One-Way Phrase Translation System

- A device containing a set of phrases
- Each phrase is associated with an audio clip in the target language
- User speaks phrase and recognizer matches speech to list of known phrases
- Audio played for the translation (pre-stored audio files)



T-61.184

# Example

## Humanitarian Assistance Phrases

- **You will be safe here**
- **We have food for you**
- **We have clean water**
- **We have clothing**
- **We have blankets**
- **We have shelter**
- **We have shelter materials**
- **We have medical care available for you**
- **We have medical supplies**
- **We may have information about your family**

# Uses for One-Way Translation

- **Military and Peacekeeping Uses**

- Intelligence screening
- Civil Affairs
- Language Training Aid
- Ship boardings and inspections
- Border / Passport Control checkpoints

- **Police and Law Enforcement**

- **Coast Guard inspections and safety**

- **Refugee registration and Humanitarian Assistance**

- **Medical diagnostics and treatment**

# Intel Strong-Arm II Processor (~ iPaq)



**T-61.184**

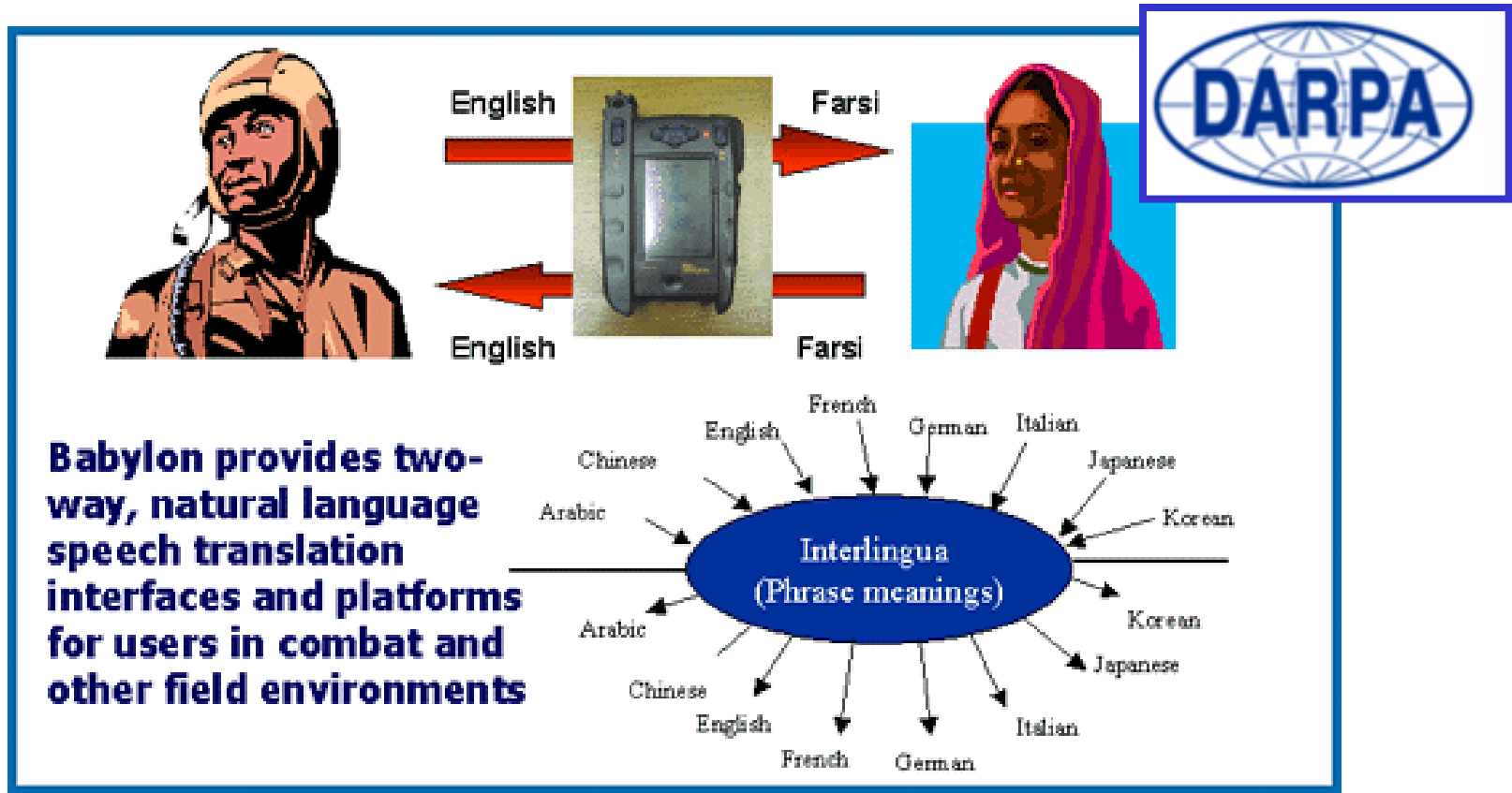
# The “Phraselator”



- Recently tested in Afghanistan under support from DARPA in USA ([www.sarich.com](http://www.sarich.com))

T-61.184

# Two-Way Speech-to-Speech Translation DARPA Babylon Program



**Babylon provides two-way, natural language speech translation interfaces and platforms for users in combat and other field environments**

T-61.184

# Reading and Language Training Systems

The Backyard Zoo (Demo)

Chapter 1

**THE Backyard ZOO**  
by Candy Carlile  
Illustrated by Ann Tosa

**A Great Idea**

It was the first day of summer vacation. Sue and Billy were eating breakfast.

"What can we do today?" Billy asked.

"I don't know about today, but I have an idea for tomorrow," said Sue. "We can go to the zoo!"

"But the zoo is far away," said Billy.

"Who will take us?"


"I don't mean the zoo in the city," said Sue. "We can make our own zoo in the yard! All our friends can bring their pets."

"What a great idea!" said Billy. "We can have balloons, peanuts, and lemonade. It will be lots of fun!"

"I'll call Grandpa," said Sue. "He worked at the zoo. He can help us make our zoo look just like the real one."

Billy ran to the door. "I'll tell everyone to bring their pets here tomorrow."

**1** **2** **NEXT**



**Again**

**Read for me**

**Summarize**

**Change to EDIT mode**

**EXIT**



# Reading and Language Training Systems

- **Detection and recognition of disfluent speech**
- **Pronunciation / Accent Monitoring**
  - Acoustic models tend to allow flexibility, but accent and pronunciation verification systems require discrimination power
- **Conferences,**
  - InSTIL/ICALL2004 Symposium on Computer-Assisted Language Learning
  - Eurocall

# Spoken Document Retrieval

- **Use of ASR to transcribe spoken audio documents**
  - Broadcast News
  - Radio Programs
  - Voice Mail?
- **Access spoken documents using IR techniques**
- **Take into account the probability of word correctness (confidence)**
- **Example,**
  - HP Speechbot --  
<http://speechbot.research.compaq.com/>

# Next Week

## ■ **Compensating for Speaker and Environment**

- Speaker Adaptation
- Environment Adaptation
- Speech Enhancement / Noise Reduction Methods
- Noise Robustness Methods
  - Feature-Space
  - Model-Space