DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

VOICE INTERFACE TO AN ON-LINE DICTIONARY

by

Mary E. Weber
weber@isip.msstate.edu
EE 4012 Senior Design Project
April 18th, 1996
Mississippi State University

ABSTRACT

In the era of natural language recognition machines, the access of electronic equipment through a speech interface will go a long way in making state-of-the-art technology available to a larger class of users. A typical application useful to a significant group of people (e.g. students) is an on-line dictionary that can be accessed using voice commands. Currently, no such dictionaries exist for UNIX-based computer systems. Some personal computers offer this feature to a limited extent, but these are constrained by the amount of memory required for a large vocabulary recognition system. In this project, we design an interface that uses a public-domain speech-recognition software to recognize the specified words and accesses a dictionary that is available on-line. The resulting system will be publicly available from the ISIP home page.

WHY VOICE INTERFACE TO A DICTIONARY?

MORE NATURAL TO SPEAK THAN TO PROGRAM

- Database query requires complicated programming languages
- Interface by speaking is natural
- Definition is found easier
- Writing process speeds up
- Test bed for other data base queries
  - Library Resources
  - Telephone Directory
  - Television Listings

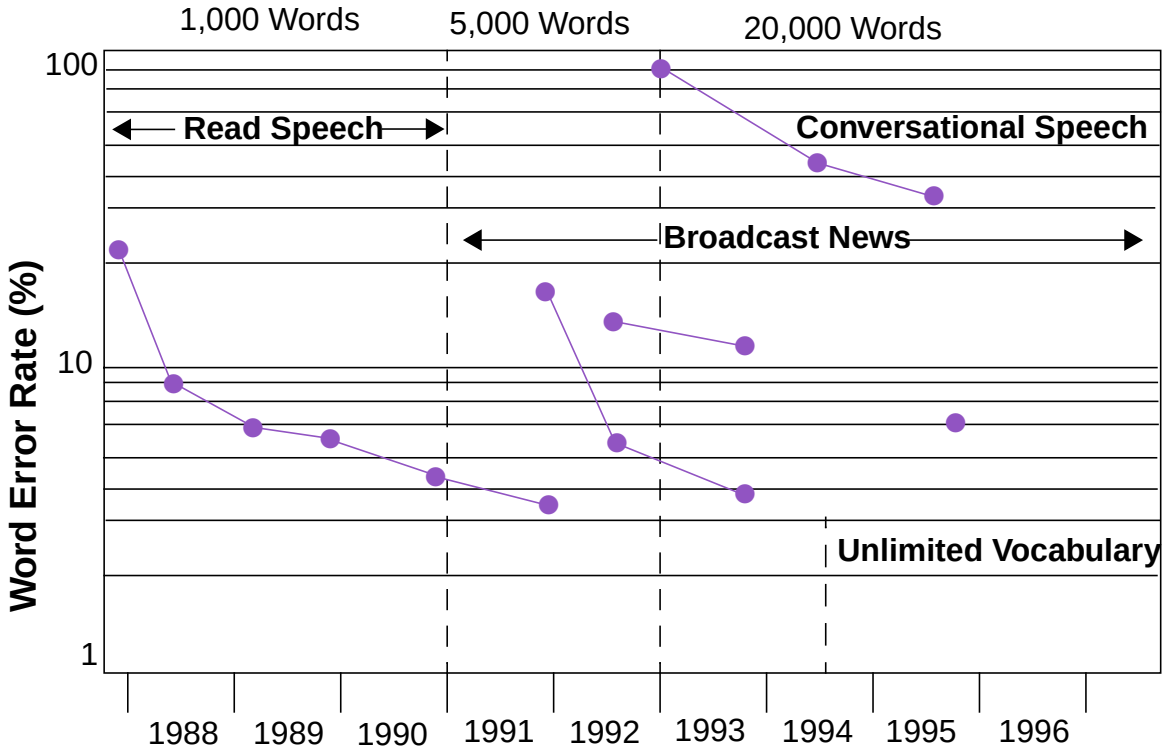
STATE OF THE ART
SPEECH RECOGNITION / UNDERSTANDING

When a system comprehends what is spoken

Challenges:

- Word spacing
- Coarticulation / Context
- Dialect
- Speaking rate / style

Performance:



- **Results - Speaker Independent**

WHY ABBOT?

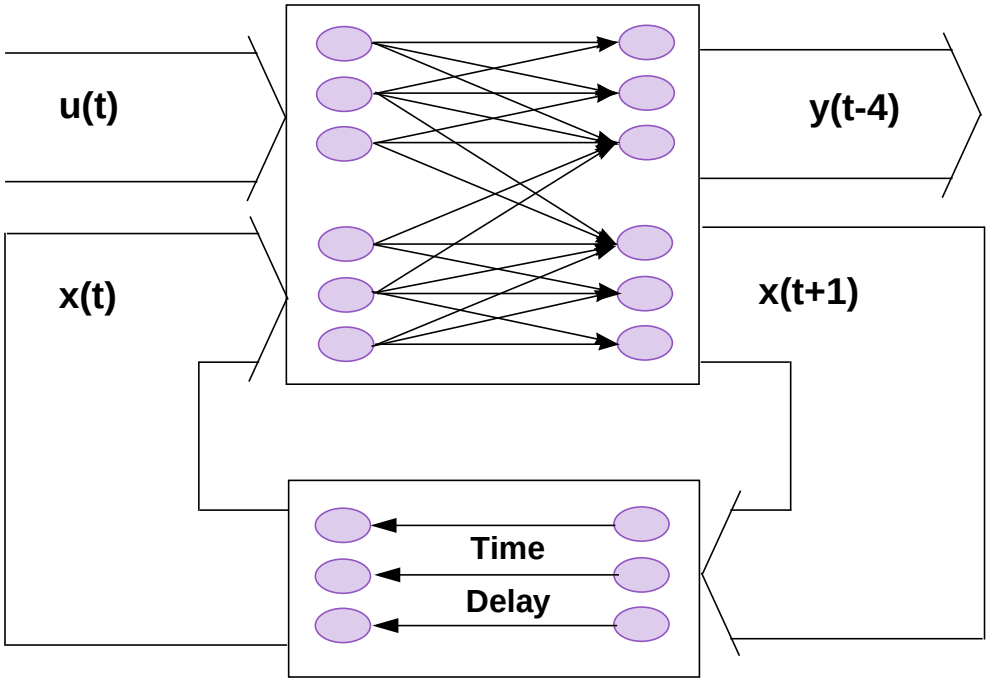
The competition - Cambridge HTK System

a generic HMM recognizer

ABBOT	**Cambridge HTK**
Hybrid connectionist hidden HMM	Gaussian HMM
Context - independent	Context - dependent
Recurrent Network System	Tied - State System
Cost - **Free**	Cost - **$100,000**

VS.

ACOUSTIC MODELING IN ABBOT



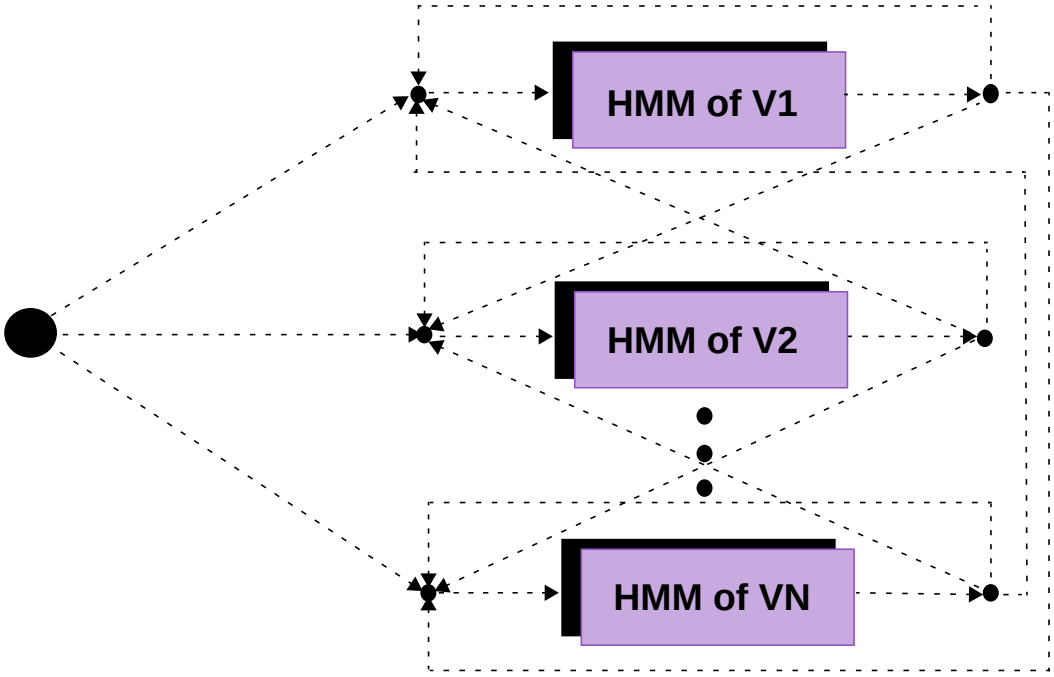
Phonetic Context-Independent Recurrent Neural Network

Input: acoustic vector u(t)
current state x(t)

$$y_i(t) \cong P_r\left(q_i(t) \middle| u_1^{t+4}\right)$$

Output: output vector y(t-4)
next state vector x(t+1)

LANGUAGE MODELING IN ABBOT

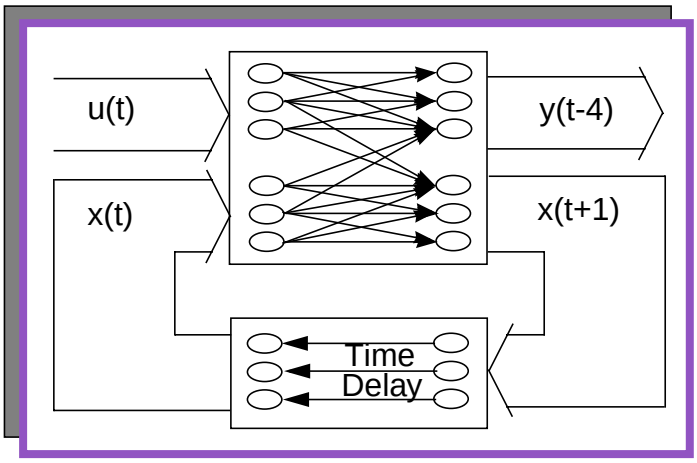


- Phone Set - 79 phone symbols, vowels have 3 levels of stresses
- Connectionist component - trained phone classifier
- Models - context & gender independent

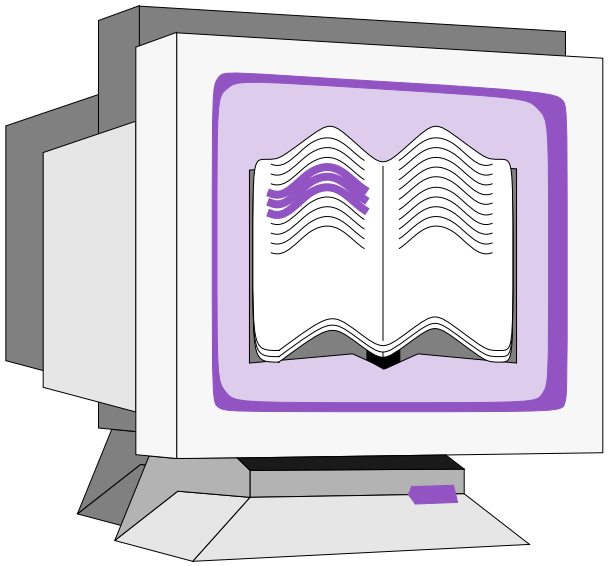
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

THE ABBOT DEMO

- Record at 16 kHz
- Determine the endpoints
- Convert to ASCII
- Normalize audio-gain



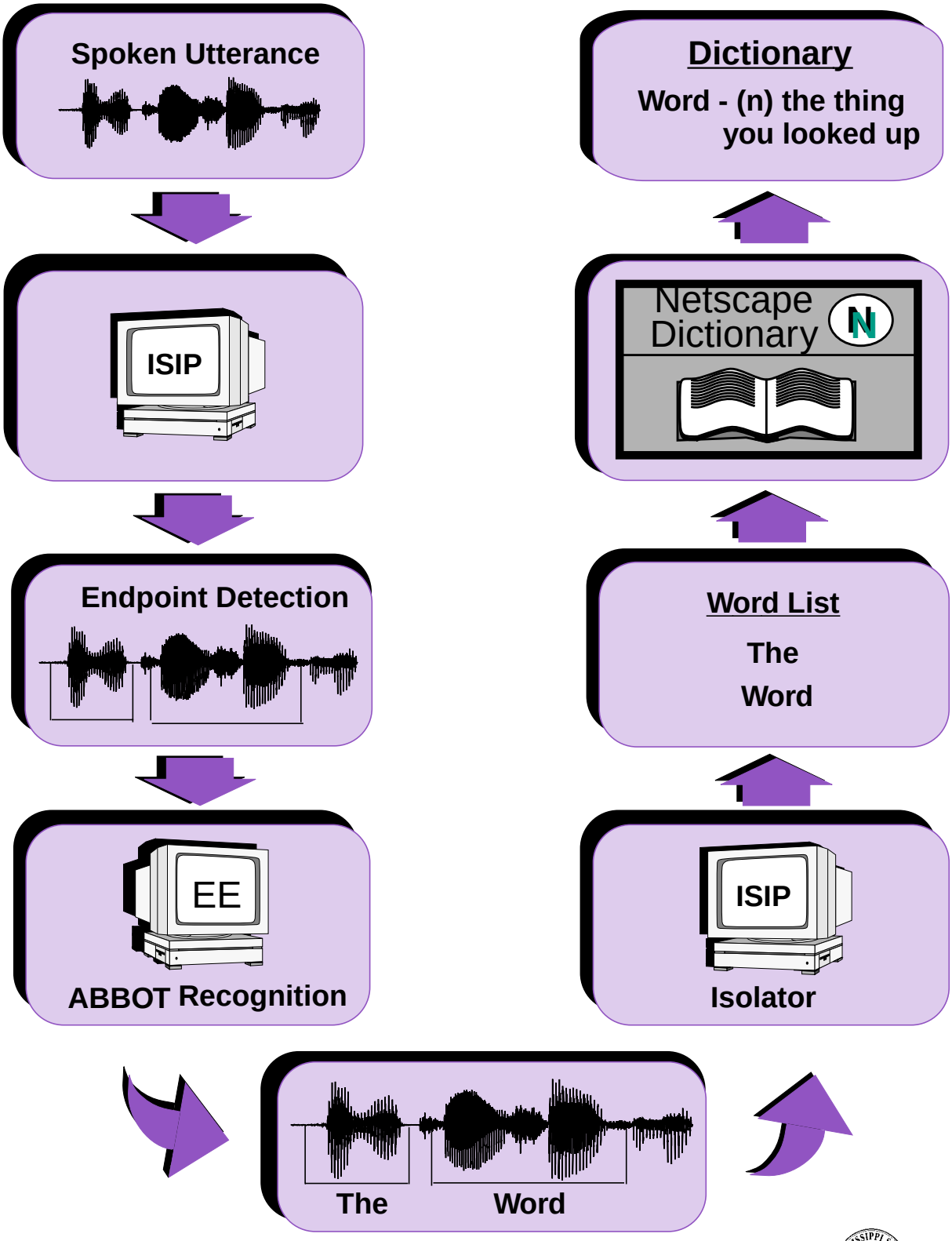
- Prints best guess to word & recognition continues
- The recognized word comes at the end



- Strip the recognized word from the end of the process
- Look up the word in the on-line dictionary

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

ARCHITECTURE OF CURRENT SYSTEM



WEBSTER DICTIONARY - WEB BASED SOURCE

Dictionary

- Systems of makeup
  - lexicon
  - grammar
  - semantic
  - phonology
- 160,000 entries
- Pronunciations

The Web

- Limited release of access
- Service for a fee
- Current version is first attempt
- CD Rom limited interface control

INTERFACING TO THE WEBSTER DICTIONARY

The Ultimate Interface

Natural Language Interface to the Dictionary

The Netscape Version
Point - and - click interface

- Type the word
- Hit return
- Retrieve definition

LANGUAGE MODELING ISSUES IN THE INTERFACE (GUI DESIGN)

Obstacles Encountered

Every word in dictionary recognized

ABBOT is a CSR

Dictionary takes only word roots

Data transported between machines

Word transported to Netscape Dictionary

Practical Solutions

Triphone based recognizer

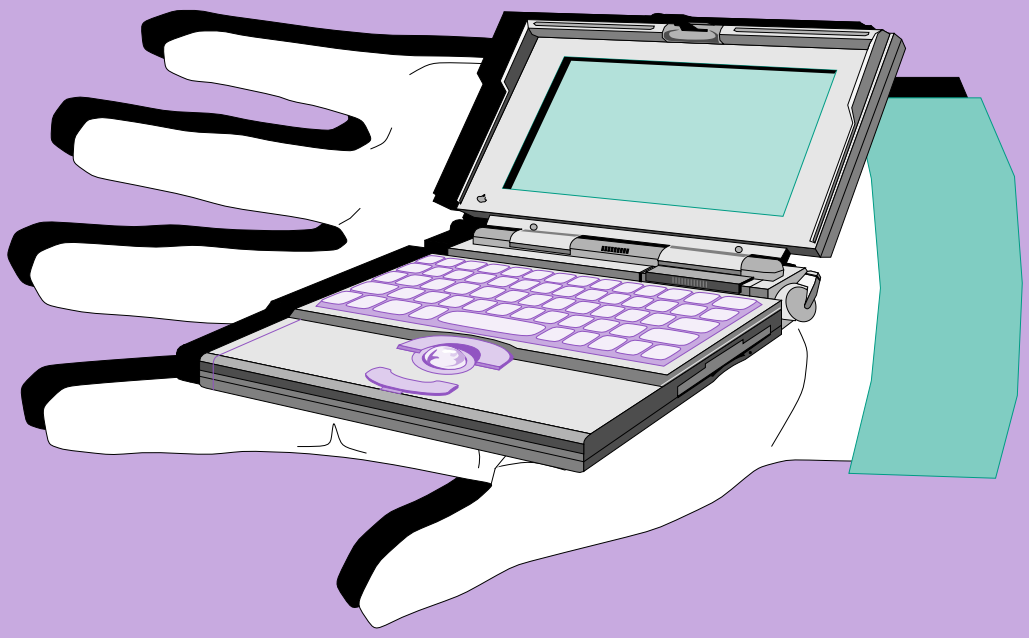
Language model changed to ISR

Portion recognizes prefixes / suffixes

Recognizer available locally

Dictionary available locally

BUILDING THIS DESIGN IN HARDWARE



Hand-Held Computer with DSP Chip and A/D Converter

- **Smaller than a credit card**

- **Plenty of memory for large recognition vocabulary**

SUMMARY

Designing a Voice Interface Dictionary

- An endpointed, spoken word
- A compatible speech recognizer
- An accessible, on-line dictionary
- A way to make all three work together

Future Enhancements:

- More adaptable recognizer (ISIP recognizer)
- Local dictionary access
- Cut down on real-time errors

REFERENCES

1. A.J. Robinson, *An Application of Recurrent Nets to Phone Probability Estimation*, in *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298-305, March 1994.
2. D.B. Roe and J.G. Wilpon editors, *Voice Communication Between Humans and Machines*, National Academy Press, Washington D.C., USA, 1994.
3. J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete Time Processing of Speech Signals*, MacMillan, New York, New York, USA, 1993.
4. L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.
5. V.V. Digalakis, Mari Ostendorf, and J.R. Rohlicek, *Fast Algorithms for Phone Classification and Recognition Using Segment-Based Models,* in *IEEE Transactions on Signal Processing,* vol. 40, no. 12, pp. 2885-2896, December 1992.
6. J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, *2nd Edition*, Macmillan, New York, New York, USA, 1992.
7. Kai-Fu Lee and Hsiao-Wuen Hon, *Speaker-Independent Phone Recognition Using Hidden Markov Models*, in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641-1648, November 1989.
8. Douglas O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley Publishing Co., Reading Massachusetts, USA, 1987.
9. Sadaoki Furui, *Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum*, in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 1, pp. 52-59, February 1986.
10. L.R. Bahl, F. Jelinek, and R.L. Mercer, *A Maximum Likelihood Approach to Continuous Speech Recognition,* in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, pp. 179-190, March 1983.
11. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1978.

ACKNOWLEDGEMENTS

A special thanks to the following people for their help with this project.

Dr. Joseph Picone
Rick Duncan
Neeraj Deshmukh
Sean Lauderdale
Arvind Ganapathiraju
Daniel Williams

and Dr. Anthony J. Robinson of CMU