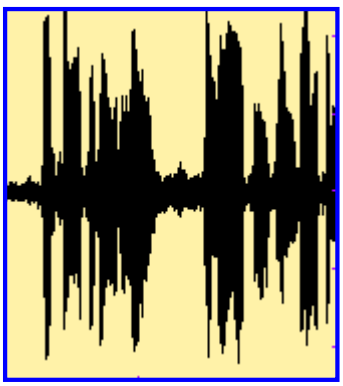
Speech and Language
Grant Holders Workshop

Speech Recognition in Multimedia Document Retrieval



BLAZING HEAT STAGGERING HUMIDITY AND DANGEROUS REPTILES THIS IS THE FLORIDA EVERGLADES

IT WAS IN THIS FRAGILE ECO SYSTEM ABOUT TEN MILES FROM WHERE I BOARDED THE FLIGHT

Sue Johnson

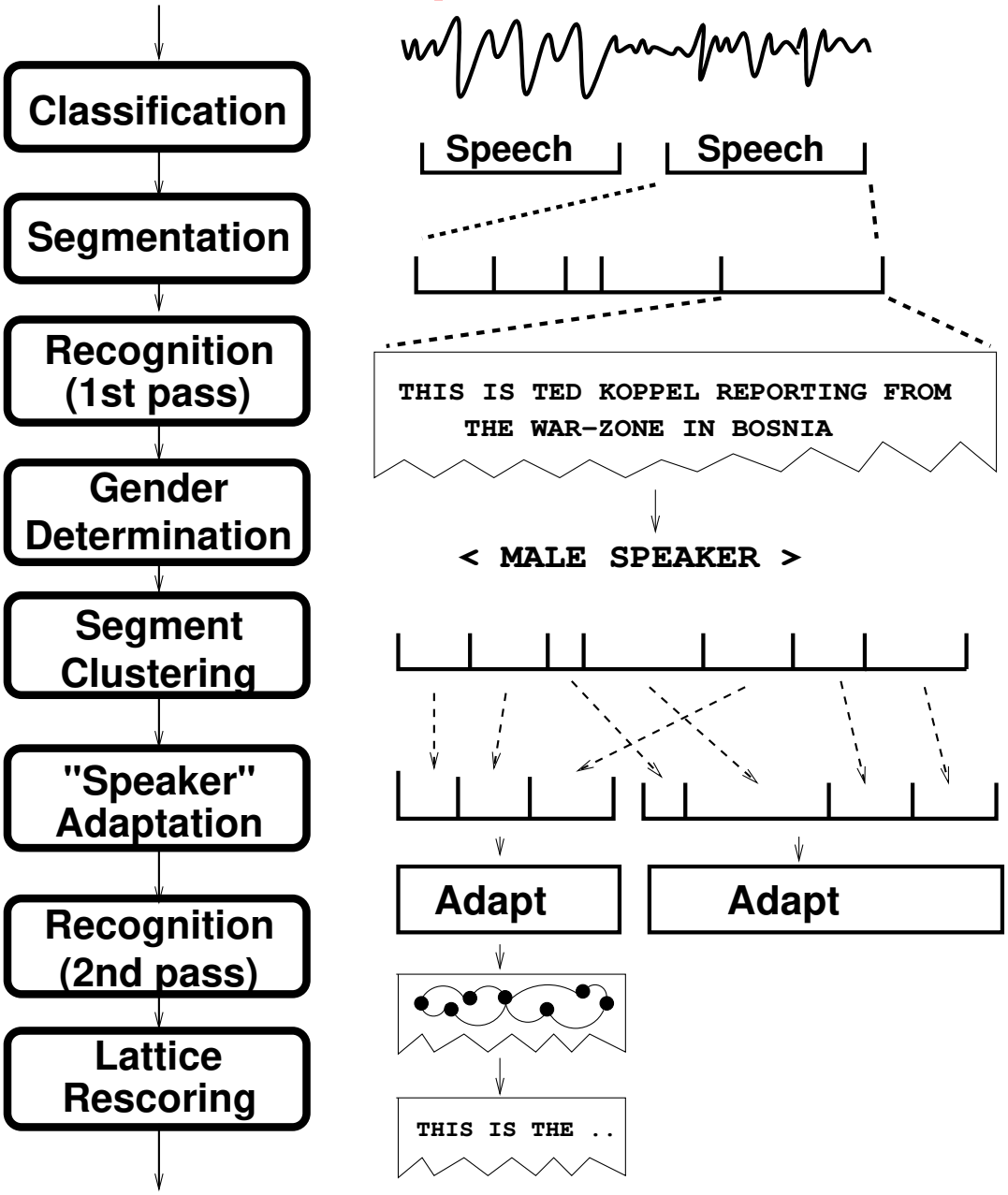
http://www-svr.eng.cam.ac.uk/~sej28/

Cambridge University Engineering Dept. and Computer Laboratory

Broadcast News Problems

- Continuous Audio Stream
- Multiple Unknown Speakers
- Large Vocabulary including many names
- Different audio conditions :
  - telephone bandwidth
  - background music
  - background noise and events
  - degraded acoustics
  - varying channel conditions
- Different speaking styles :
  - different accents
  - non-native speakers
  - spontaneous speech
- Overlapping speech

Transcription Process



Classification and Segmentation

Aims : Discard pure music

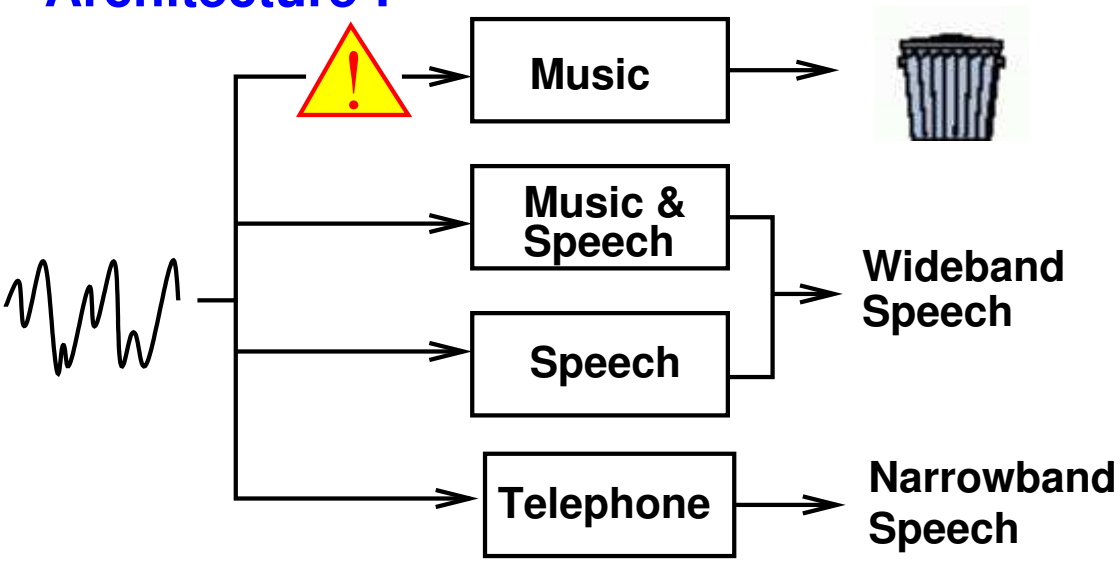
Classify into wideband vs telephone-band

Split up the different audio conditions

Split up the different speakers

Make segments < 30 s for recogniser

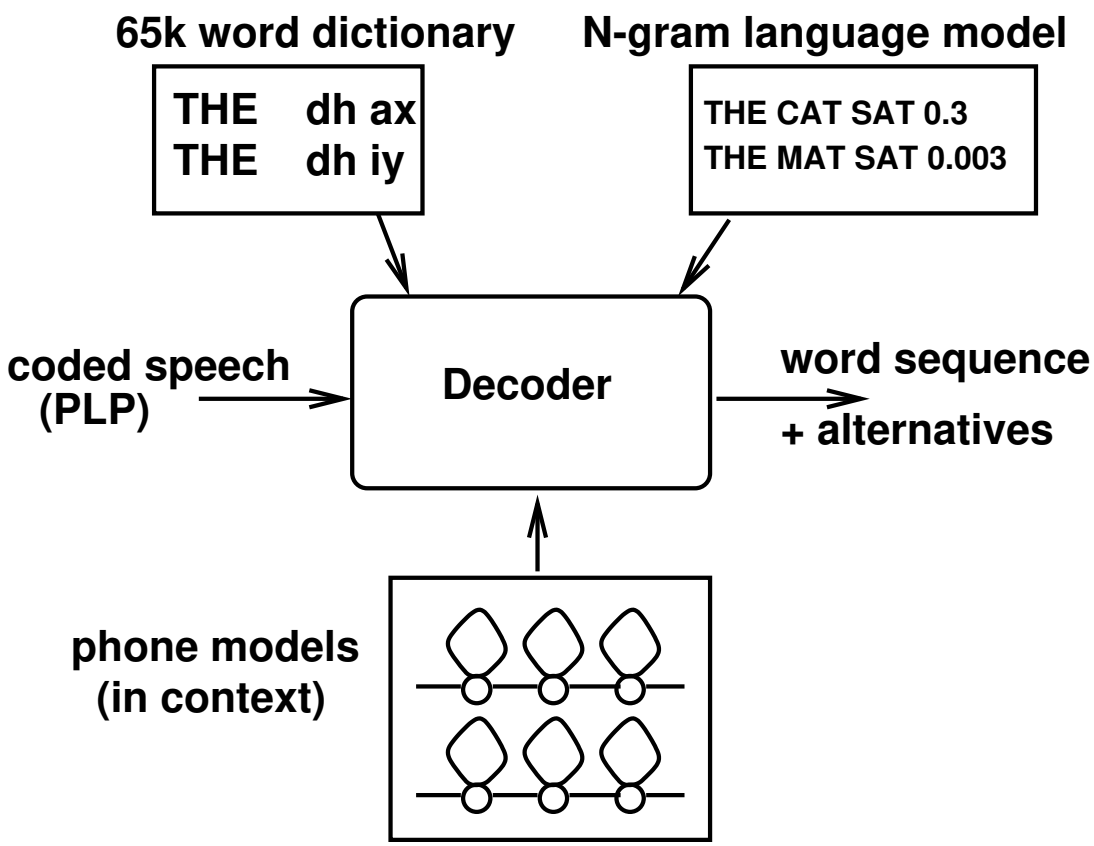
Architecture :



Details : Gaussian Mixture Classifier

3 hours training per class

Speech Recognition



- Acoustic Models trained on Broadcast News Audio
  (35 / 70 hours)
- Word List and LM trained on Broadcast News text
  (100 - 200 Million Words)

Clustering and Adaptation

Purpose :

- Make models more appropriate for the data

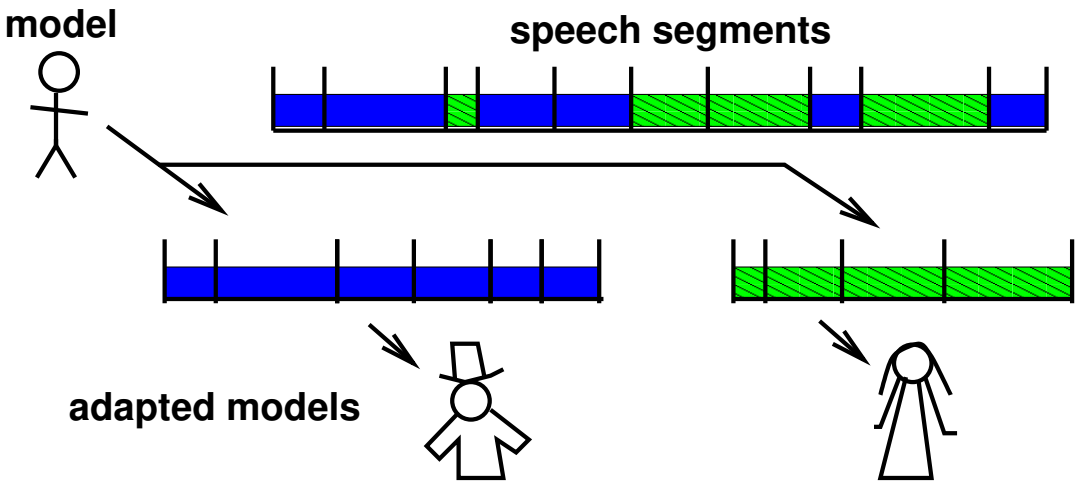
Clustering :

- Represent each segment mathematically
- Calculate the distances between the segments
- Group "close" segments together

Adaptation :

- Generate matrix for each cluster which increases the likelihood of the data
- Apply this matrix to the models
- Use the new models for recognition

Diagram :



1997 CSR HUB-4 Evaluation

Task:

- Transcribe 3 hours of continuous BN audio

Work:

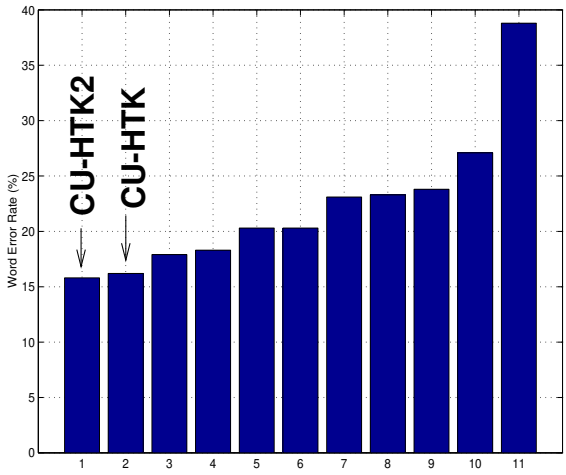
- Accurate audio segmentation algorithm developed
- General HMMs trained on 70 hours of data
- Multi-pass recognition with adaptation
- Complex acoustic models for final pass
- Improved language models (4-gram + class trigram)

Results:

- Overall WER of 16.2 %
- Lowest WER by a statistically significant margin
- 15.8 % WER system also submitted

Results on Hub-4 1997 Eval

Overall Word Error Rate



Error Rate by Condition

	HTK	HTK2
clean, planned	9.9	9.4
clean, spontaneous	15.4	15.2
reduced bandwidth	20.1	19.5
background music	27.9	26.9
degraded acoustics	19.4	19.4
non-native speakers	24.1	22.1
all other combinations	29.9	29.1

ASR work for TREC Evaluations

TREC-6 (35 hours training and test data)

First system

- Gender Independent HMMs
- Single Pass decode with trigrams
- 45 times realtime
- 28.6 % Word Error Rate

Second system

- Gender Dependent Pass 2
- Clustering and adaptation
- 4-gram lattice rescoring
- 50 times realtime
- 24.1 % Word Error Rate

TREC-7 (70 hours training and test data)

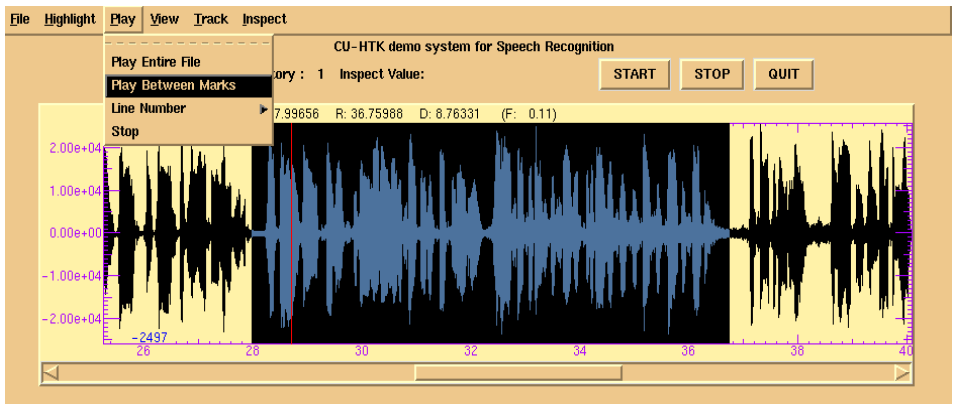
- Used second system
- Revised dictionary and LM
- Better acoustic models
- 25.5 % Word Error

Additional Speech Work

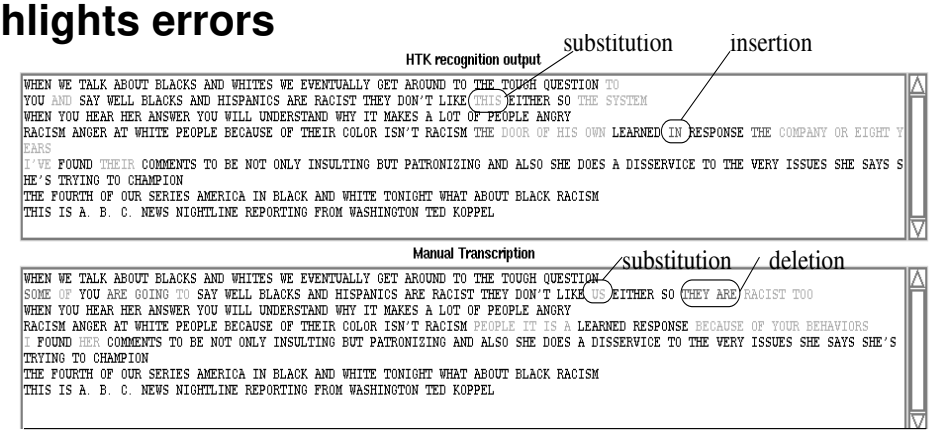
- Improvement of Adaptation Code
- New Implementation of Speaker-Adaptive-Training
- Speeding up of first-pass recognition
  (in conjunction with Entropic)
- Investigation of Speaker Clustering Techniques
  including a new one based on adaptation methods
- Development of tool for display and analysis
  of errors in Broadcast News (xpert)
- Demo of recognition set up with xpert

Error Analysis Demo

- **Graphical User Interface**
- **Shows waveform, automatic and manual transcripts**
- **Plays audio simultaneously**



- **Highlights errors**



- **Provides error statistics**