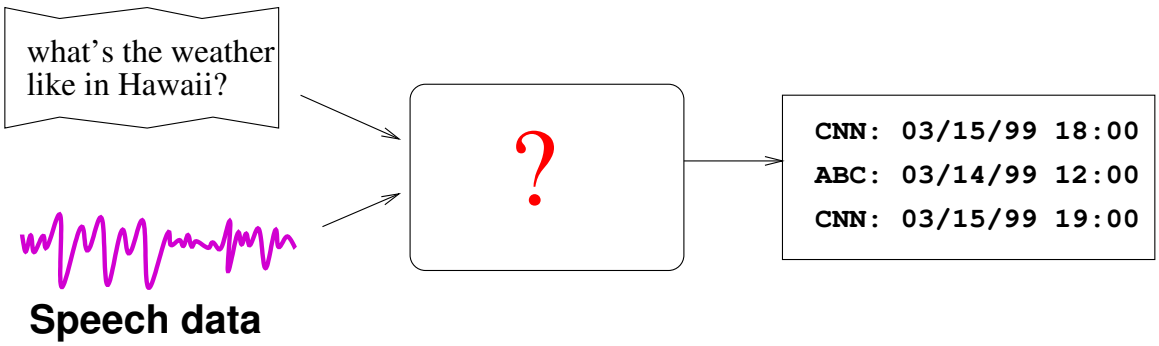
# ICASSP '99

# The Cambridge University Spoken Document Retrieval System




Sue Johnson
Gareth Moore
Phil Woodland

Pierre Jourlin
Karen Spärck Jones

Email : sej28@eng.cam.ac.uk

http://svr-www.eng.cam.ac.uk/Research/Projects/mdr/

Cambridge University Engineering Dept. and Computer Laboratory

# What is Spoken Document Retrieval ?

## SDR is :

- A combination of ideas from the fields of Speech Recognition and Information Retrieval
- Retrieving documents which may be relevant to a user's request.
- The source of the documents is speech
- The requests are made in natural language

## SDR is not :

- Answering the users request directly
- Summarising information from documents

## We use:

- Broadcast news stories as documents
- Requests written in text
- Static document set but varying queries

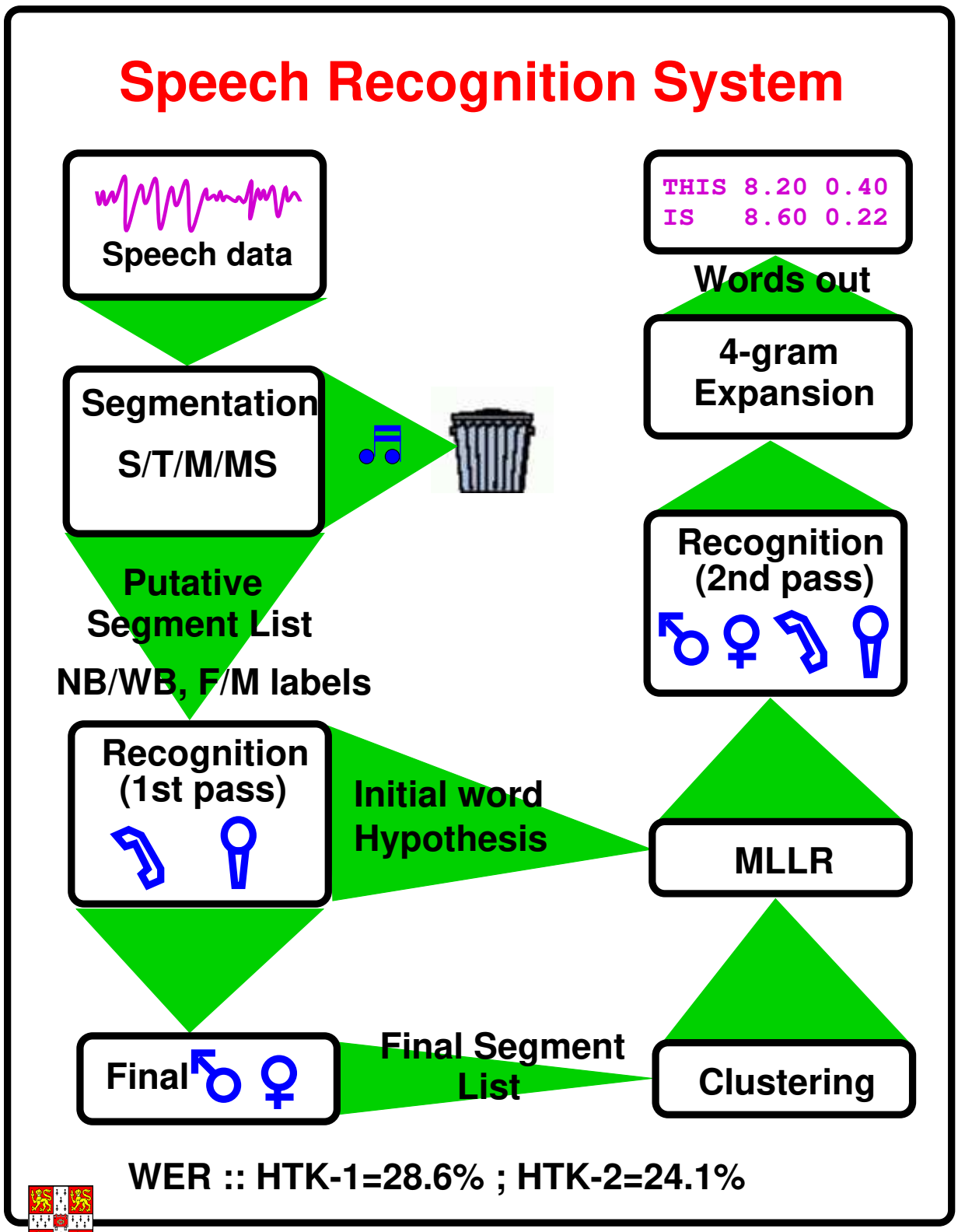
# Data Used in Experiments

- First international evaluation of SDR in 1997
- TREC-6 document set
  - 50 hours of audio
  - 11 news shows
  - 1451 documents
  - 283 words/doc average
- Query set not suitable as known item problem
- CU-60 Query Set
  - 60 queries
  - 7.1 words per query
  - 9.2 rel docs/query
- Evaluate on: Precision = # relevant & retrieved / # retrieved

  Recall = # relevant & retrieved / # relevant

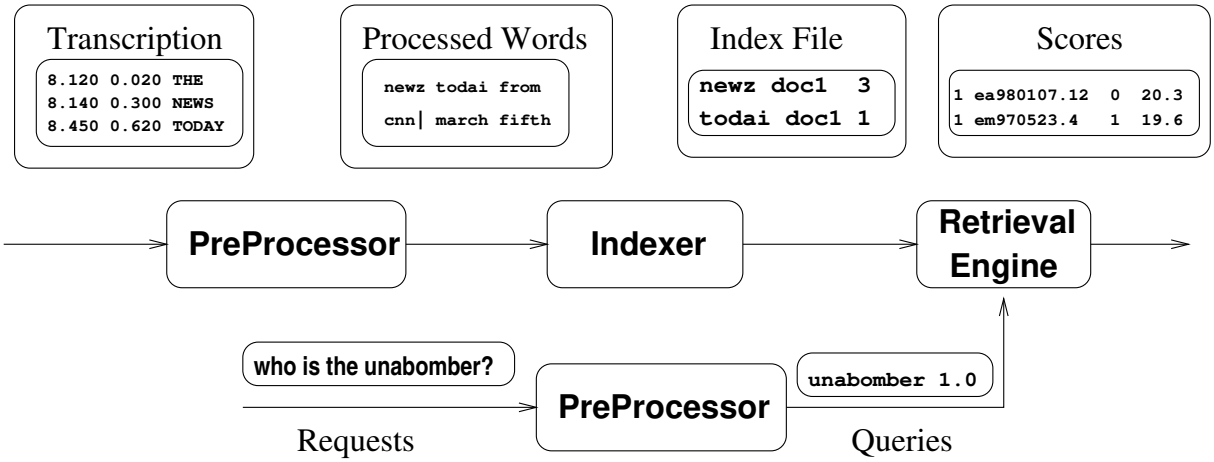
  Average Precision

# Speech Recognition System




WER :: HTK-1=28.6% ; HTK-2=24.1%

# Information Retrieval Engine




**PreProcessor:** Remove Stop Words
Deal with abbreviations (e.g. C.N.N.)
Map certain words/phrases
Suffix Strip (Porter Stemmer)

**Indexer:** Store words each document contains
documents each word is in & doc lengths

**Retriever:** Uses Combined Weights Formula

$$cw(i,j) = \frac{(\log\{N\} - \log\{n(i)\})\, tf(i,j)\, (K+1)}{K(1 - b + b\, ndl(j)) + tf(i,j)}$$

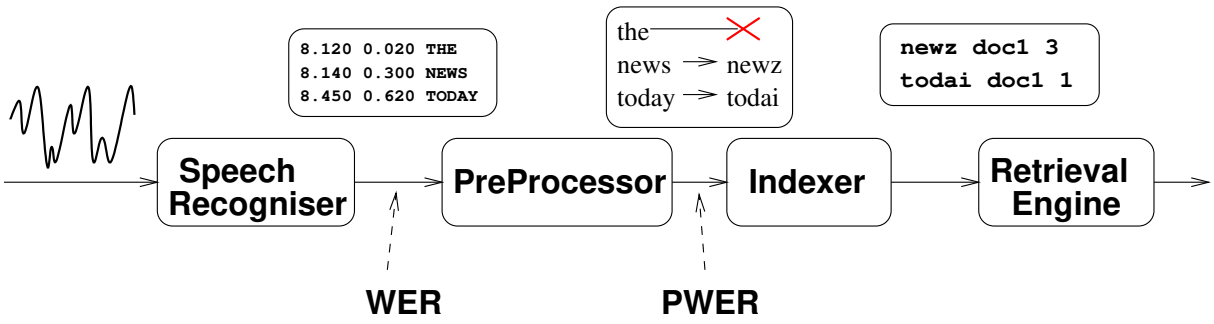
tf(i,j) = term frequency
ndl(j) = normalised document length
n(i) = collection frequency

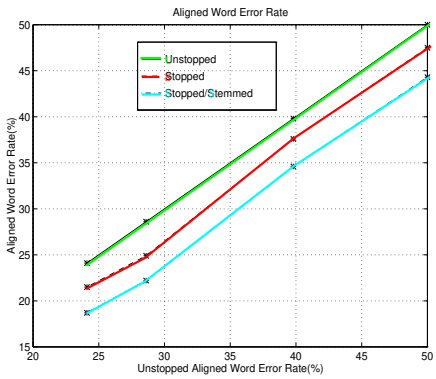
# Integration of ASR and IR

## Properties of Speech Transcriptions

+ Fixed and Known Vocabulary
+ No Spelling Mistakes
? Confidence Scores
? Lattices
- Transcription Errors

## Effect of Processing on WER



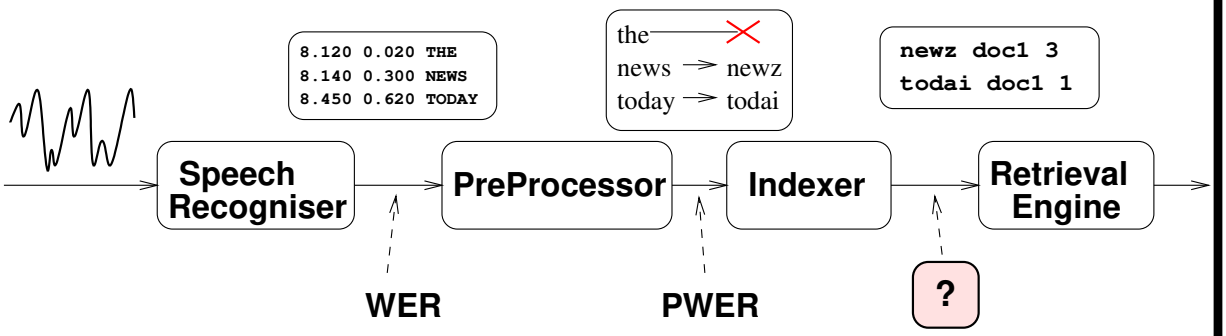




**Stopping** reduces WER (on these transcriptions)

**Stemming** reduces WER (always)

# Modifying WER for use in SDR

## But PWER does not model input to retriever




1) Substitutions are 1 error (unlike retrieval)

2) Word-order dependent (unlike retrieval)

## For 1: Introduce Quasi-Term Error Rate (QTER)

$$\text{QTER} = \frac{(2S + I + D)}{W}$$

**however - still order-dependent**

## For 2: Remove Alignment Information

Access the terms directly

Introduce the concept of a

**Term Error Rate (TER)**

# Term Error Rate (TER)

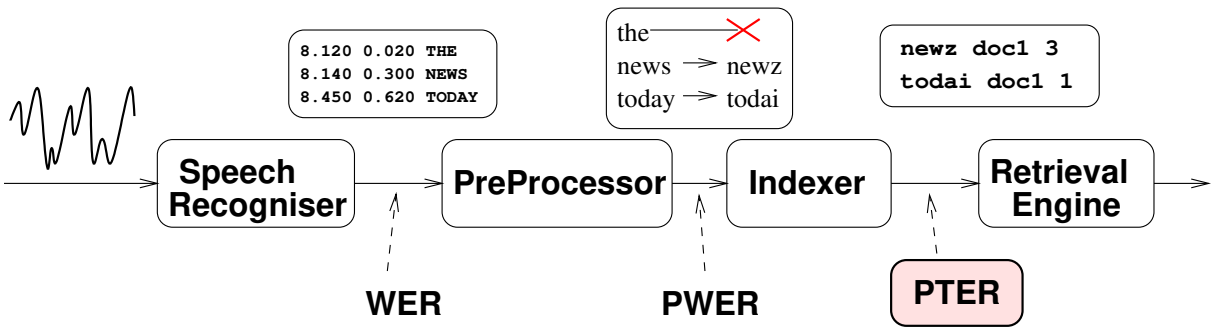
$$TER = \frac{\sum_w |Trans(w) - Ref(w)|}{W}$$

## Story Term Error Rate (STER)

**Average the TER for each story**

## Processed Term Error Rate (PTER)

**Calculate after pre-processing**

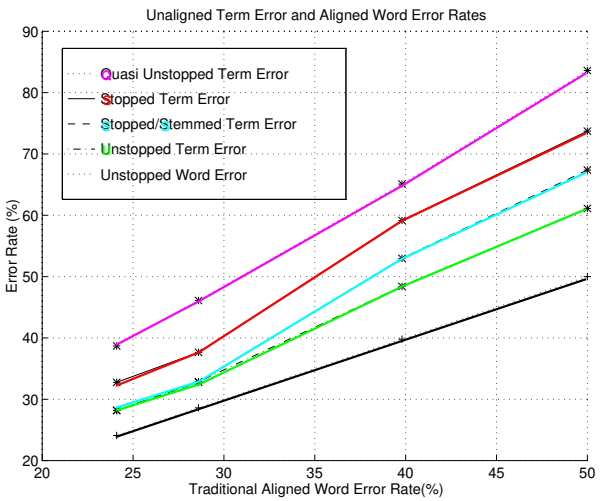



**QTER > TER (always)**

**QTER > WER (always)**

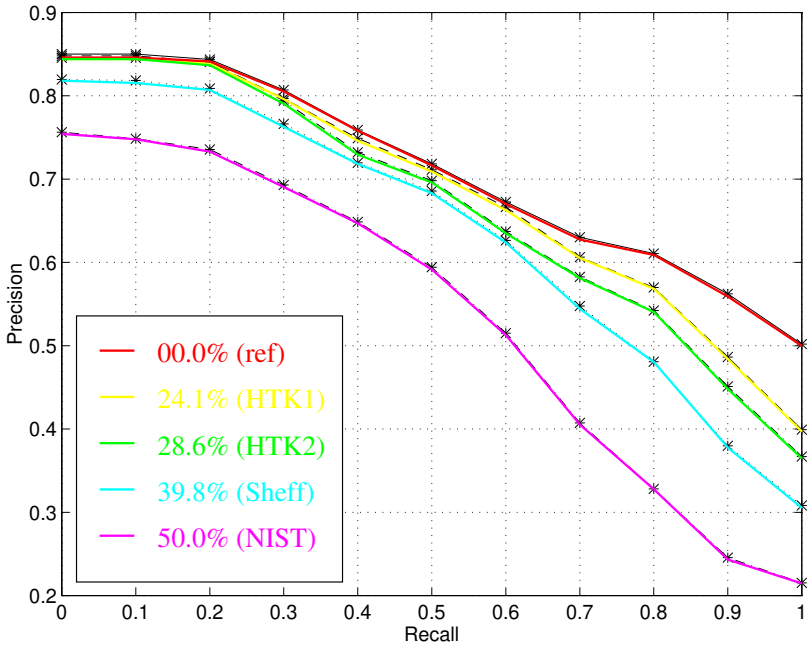
**Stopping increases TER (these transcriptions)**

**Stemming decreases TER (always)**

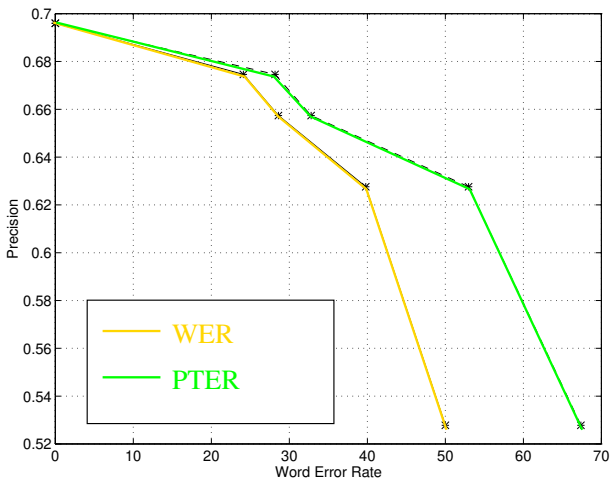


# Results on TREC-6/CU-60

## Interpolated Precision-Recall




## Error Rate vs Average Precision

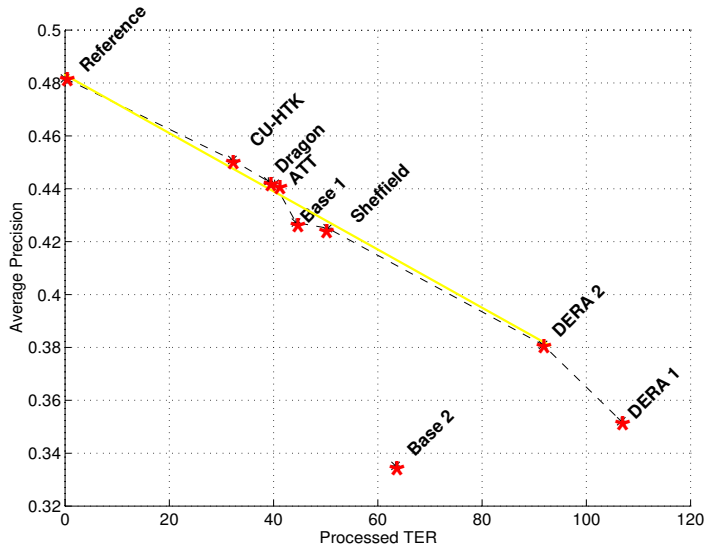


# TREC-7 System

**Data:**
- 100 hours
- 2866 stories
- 23 Queries

**System:** As described +
- POS weights on query terms
- Word-Pair modelling
- Query Expansion

## Effect of PTER on Average Precision



# Conclusions

- IR methods can be applied to speech
- System performs well on TREC task
- 25% WER for both TREC-6 and 7 data
- PWER models pre-processing
- PTER also models indexing
- IR performance degrades as PTER increases (linearly for TREC-7)

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