Spoken Document Retrieval at Cambridge University

S.E. Johnson†, P. Jourlin‡

P.C. Woodland† & K. Spärck Jones‡

November 17th, 1999

† Cambridge University Engineering Department

‡ Cambridge University Computer Laboratory

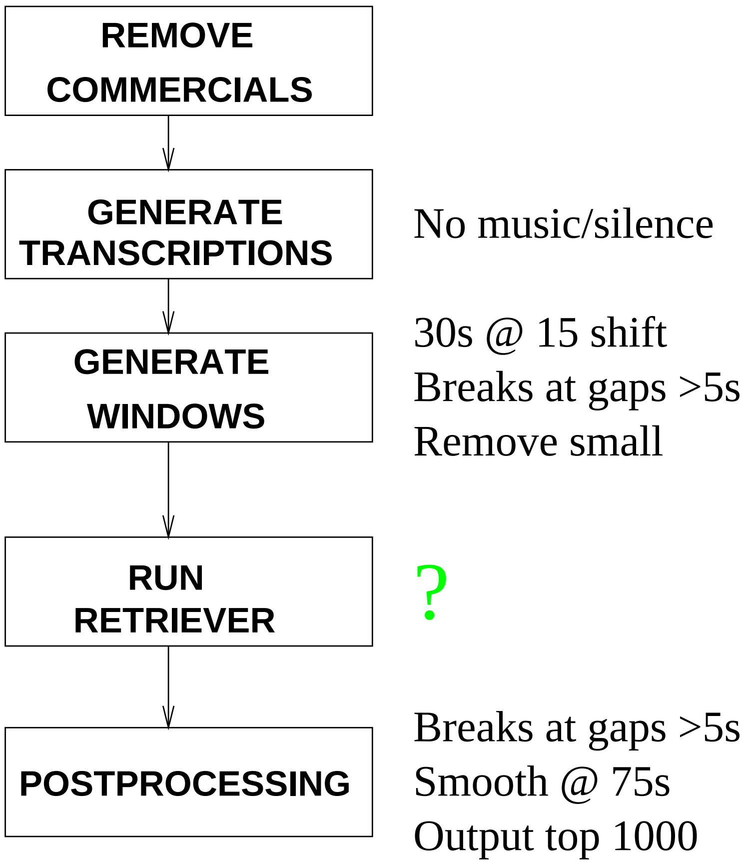
Outline

- Summary of Data
- Story Unknown System
- Commercial Elimination System
- Commercial Elimination Results
- Recognition System
- Cross Recogniser Results
- Retrieval (Pierre)

Summary of Data

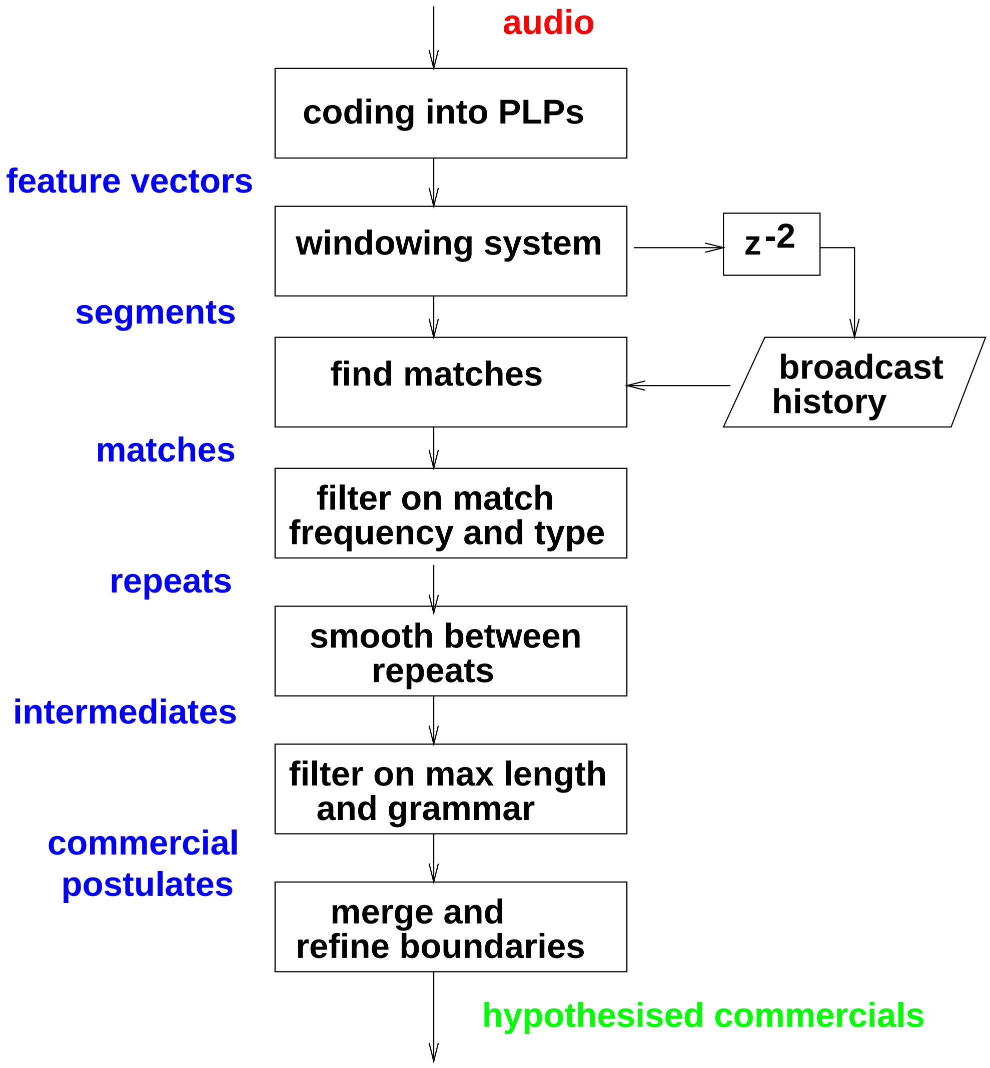
	TREC-7 (dev)	TREC-8 (eval)
Length of Audio	~ 70 hours	500 hours (SU)
Number of Docs	2,866	21,754 (SK)
Number of Words	770,000	3,900,000 (SK) 4,700,000 (SU)
Average Doc length	269 words	180 words (SK)
Number of Queries	23	50
Length of Query	14.7 words	13.7 words
# Rel Docs /Query	17.0 docs	36.4 docs (SK)

Story Unknown System



- Commercials, music and silence removed
- Windowing system
- Broadcast structure enforced

Commercial Elimination System



- Search broadcast history audio directly
- Smooth between matches
- Safeguards against re-broadcast stories

Commercial Elimination Results

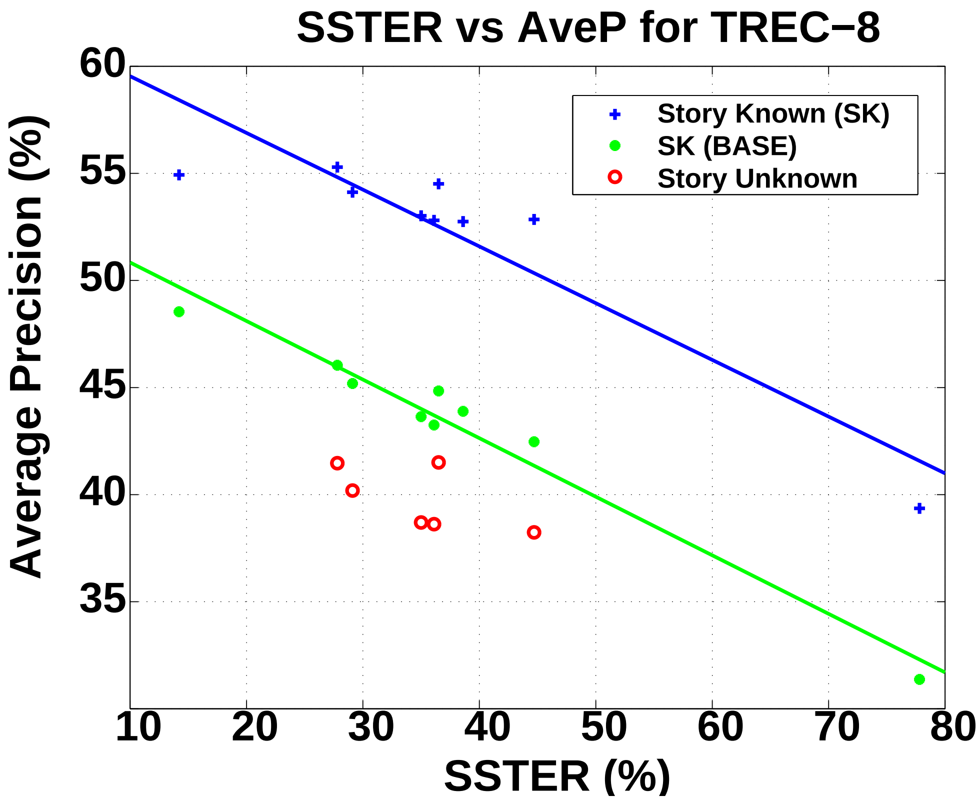
Broad.	Non-Stories	Stories	Total
CNN	26.2h=35.71%	2822s=0.46%	27.0h=11.03%
ABC	12.8h=65.49%	28s=0.02%	12.8h=20.48%
PRI	1.9h=16.56%	297s=0.10%	2.0h= 2.16%
VOA	0.5h= 4.95%	132s=0.04%	0.5h= 0.49%
ALL	41.4h=36.32%	0.9h=0.23%	42.3h= 8.42%

- 42.3 hours of data removed before recognition
- 97.8% of this is non-story
- AveP on `s1p1` increased by 1.2% relative
- AveP on LIMSI increased by 1.4% relative (filtering before post-processing)
- Contrast run removes 63.4hrs - 95% non-story

Recognition System

- Remove Commercials
- Segmentation
  - 1 to 30 seconds, single speaker, bandwidth labelled
  - Remove pure music and silence
  - Cepstral mean normalisation
- First Pass (`HTK-s1p1`)
  - Gender-independent, bandwidth-specific triphones
  - 60K vocab, 4-gram LM
  - 3xRT, WER = 26.6%
- Second Pass (`HTK-s1`)
  - Gender determination and clustering
  - MLLR adaptation
  - Gender-specific, bandwidth-specific triphones
  - **108K vocab**, 4-gram LM
  - 13xRT, WER = 20.5%
  - (WER on Nov. Hub-4 1998 eval = 15.7%)

Cross-Recogniser Results



- New TERs to model retrieval processes
- Removing commercials helped s1 and s1p1
- SU system lags SK benchmark by around 3.5% AveP
- SK system retrieval improved by around 9% AveP
- Drop in SK AveP 2.5% upto 45% SSTER (32% WER)