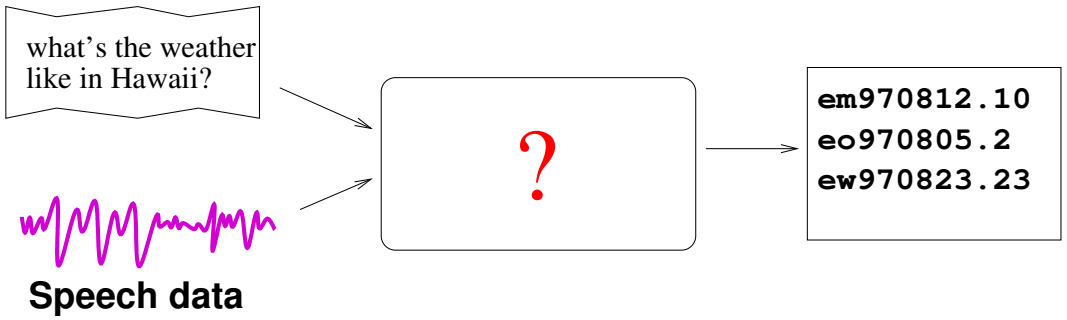
Text REtrieval Conference 7 (1998)

Spoken Document Retrieval at Cambridge University



Sue Johnson *
Gareth Moore
Phil Woodland

Pierre Jourlin *
Karen Spärck Jones *

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

Cambridge University Engineering Dept. and Computer Laboratory

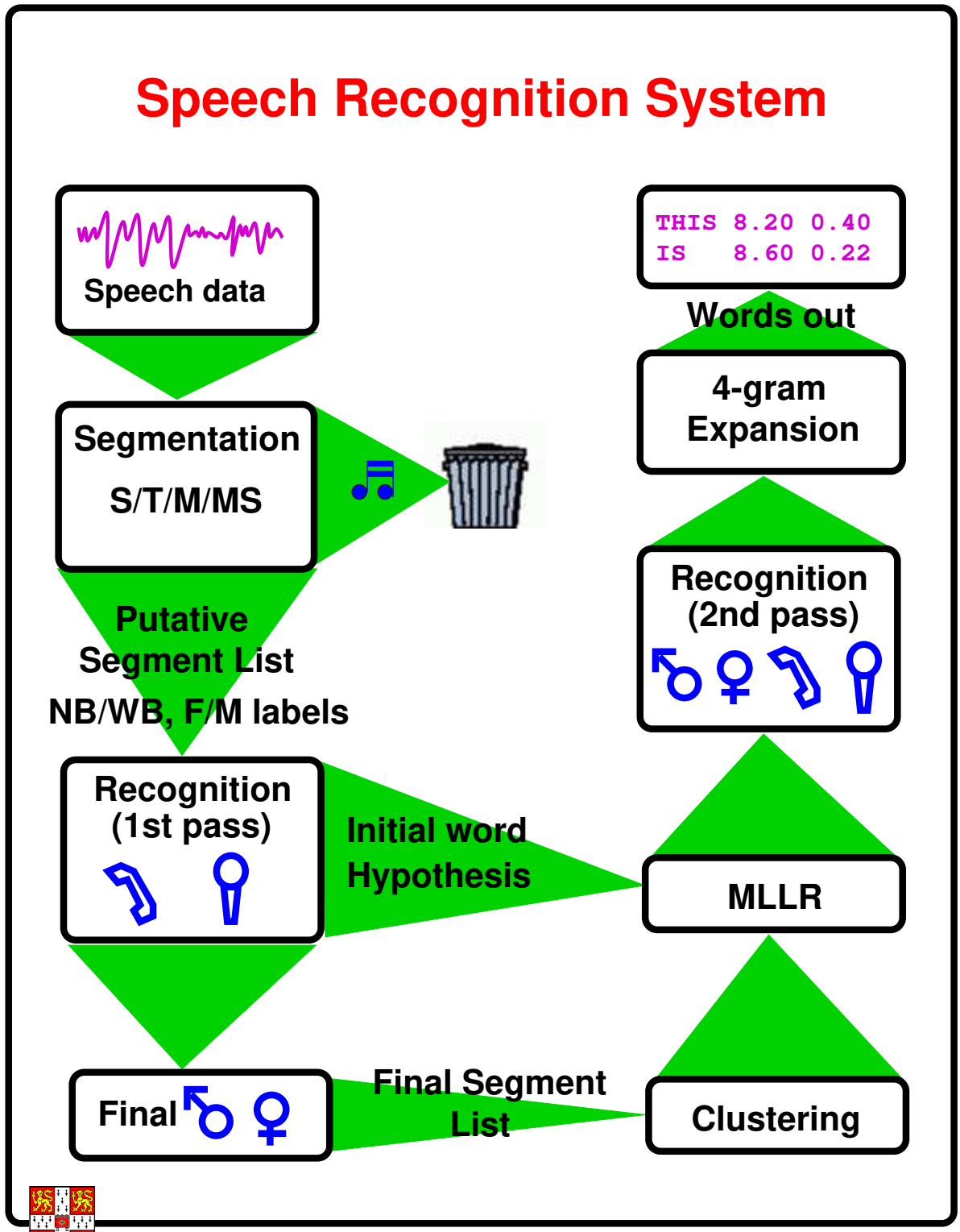
Overview of Talk

- Data Used in Experiments
- Speech Recognition System
- Basic IR engine
- Integration of Speech and IR
- SDR problems - discourse style
- Extensions I - stopping and mapping
- Extensions II - word pairs and POS
- Extensions III - Expansion
- Overall Results - Summary
- Overall Conclusions from TREC-7
- Recognition Errors and IR performance

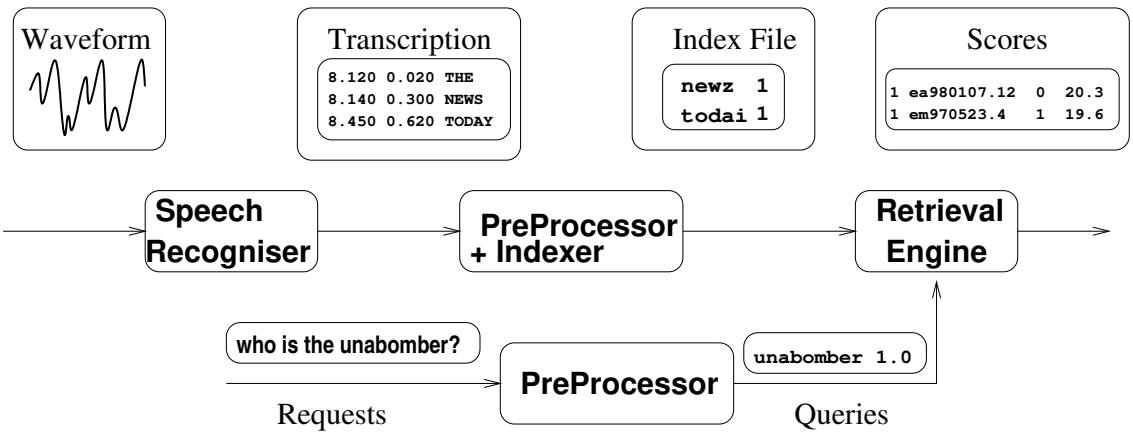
Data Used in Experiments

	Development	*Evaluation*
Document Set	**TREC-6 Test**	**TREC-7 Test**
Query/Relevance Set	**CU60**	**TREC-7 Test**
# Documents	**1451**	**2866**
# Words	**410,000**	**770,000**
# Queries	**60**	**23**
# Relevant Docs	**549**	**390**
Ave. Query Length	**7.1 words**	**14.7 words**
Ave. Relevancy	**9.2 docs**	**17.0 docs**
HTK WER	**24.1 %**	**24.8 %**
Baseline WER	**50.0 %**	**34.6 & 47.1%**

Speech Recognition System



Basic Retrieval Engine



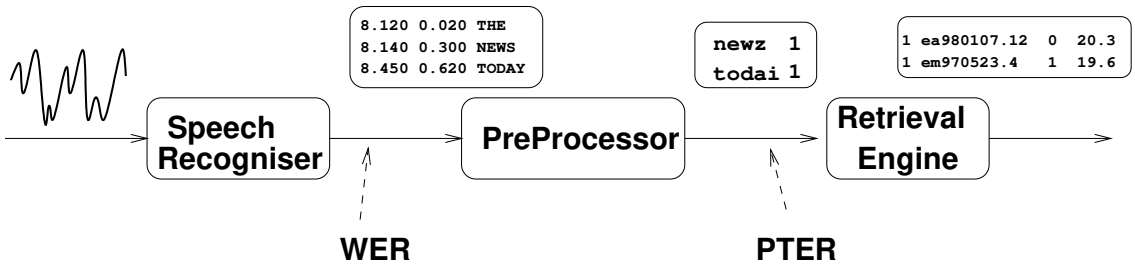
Okapi-based System

Uses Combined Weights Formula

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

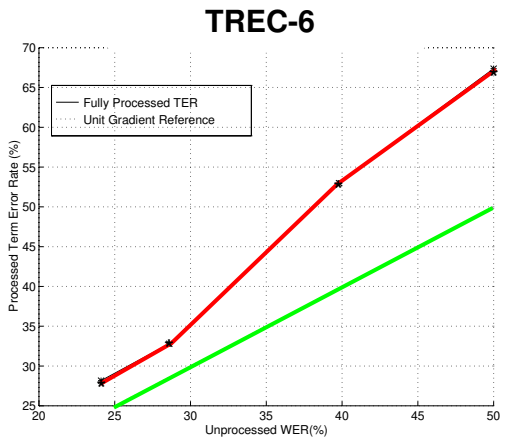
tf(i,j) = term frequency
ndl(j) = normalised document length
n(i) = collection frequency

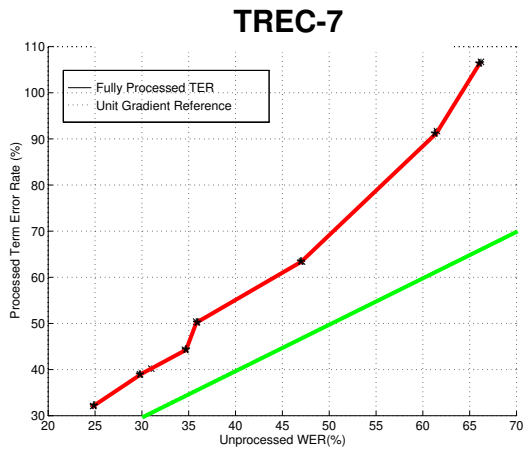
Integration of Speech and IR



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

WER vs Processed Term Error Rate





SDR Problems - Discourse Style

- **Words**

  `hmm, uh-huh, aaah`

- **Phrases**

  `Good morning sir`

- **Sentences**

  `This is A. B. C. News Nightline`

  `Reporting from Washington Ted Koppel`

- **Paragraphs**

  `I lost the voice ca- l- would you try and see if we've got the connection oh concerned about what he said about legal status I can't hear nothing. on just a second Mr. Gutierrez can you hear me sure can you hear me Mr. Gutierrez.....`

Extensions I - Stopping and Mapping

Stopping

Make Separate Stoplist for Queries

Add Broadcast-News Specific Terms

Remove Additional Function Words

Mapping

Include Stemming Exceptions List
(e.g. news vs new, governmental vs government)

Correct Bad Spellings in Reference
(e.g. Zuganauf, Zuganov, Zyuganov)

Standardize 2-Word Spellings
(e.g. baby sit, baby-sit, babysit)

Effects

	Evaluation				Develop
	AveP ref	R-Prec ref	AveP HTK	R-Prec HTK	AveP HTK
Stopping	+0.28	+1.36	+1.67	+1.43	+1.91
Mapping	+1.08	+2.13	+0.77	+1.02	+4.59

Extensions II - Word Pairs and POS

Word Pairs

TPIs added to Inverted File

Queries Tagged by Brill Tagger

Noun Compounds and Adjective-Noun Added

(e.g. north_ireland, gun_control)

POS Weighting

Weight Query Terms by POS

Proper Noun > Common Noun > Adj/Adv > Rest

Beware: Stemming Affects POS in Documents

Effects

Evaluation | Develop

	AveP ref	R-Prec ref	AveP HTK	R-Prec HTK	AveP HTK
Word-Pair	-0.64	-2.35	-0.58	-1.45	+0.84
POS	+0.34	-1.53	+1.01	+0.68	+0.22

Extensions III - Expansion

Statistical Expansion

Derives Statistics from Inverted File

Original Stem Must Exist in < 2% Documents

Expanded Stem (E) More Likely With Original (O)

Expanded Term More Common Than Original

Expanded Term Weights = P (O | E)

Original and Expanded Term Weights Normalised

(e.g. netanyahu -> isra)

Roots Expansion

Stems With > 4 First Letters in Common Added

(e.g. basketball -> basket)

Effects

	Evaluation				Develop
	AveP ref	R-Prec ref	AveP HTK	R-Prec HTK	AveP HTK
Statistical	+0.07	+0.00	+0.28	+0.00	-0.01
+ Roots	+0.26	-0.56	+0.34	-0.50	+0.40

Overall Results - Summary

Separate Effect

	Evaluation				Develop
	AveP ref	R-Prec ref	AveP HTK	R-Prec HTK	AveP HTK
Benchmark	46.61%	44.81%	43.45%	42.42%	62.87%
Stopping	+0.28	+1.36	+1.67	+1.43	+1.91
Mapping	+1.08	+2.13	+0.77	+1.02	+4.59
Word Pair	-0.64	-2.35	-0.58	-1.45	+0.84
POS	+0.34	-1.53	+1.01	+0.68	+0.22
Expansion	+0.07	+0.00	+0.28	+0.00	-0.01
Tuning	+0.26	-0.56	+0.34	-0.50	+0.40

Combined Effect

	Evaluation				Develop
	AveP ref	R-Prec ref	AveP HTK	R-Prec HTK	AveP HTK
Stopping	+0.28	+1.36	+1.67	+1.43	+1.91
Mapping	+0.80	+0.77	-0.90	-0.41	+2.68
Word Pair	-0.55	-1.45	+0.01	-1.45	+0.14
POS	+0.94	+0.87	+0.76	+1.73	+0.42
Expansion	+0.60	+0.37	+0.66	+0.36	+0.02
Tuning	+0.67	-0.34	+0.07	+0.85	+0.28
Overall	49.35%	46.39%	45.72%	44.93%	68.32%

Overall Conclusions

Overall AveP = 49.35% (ref) 45.72% (HTK)

Total Increase in AveP = 2.74% (ref), 2.27% (HTK)

Performance Degrades Slightly With WER

Processed Term Error Introduced

Individual Increases Not Significant

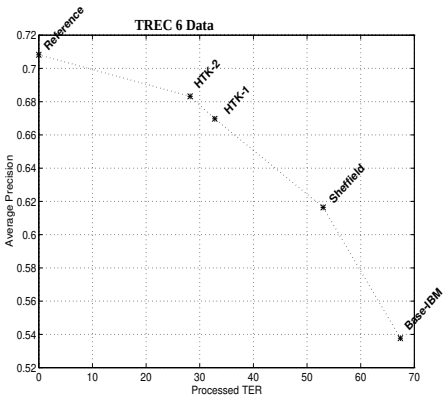
All Devices Interact With Each Other

Problems with Some Mappings and Word-Pairs

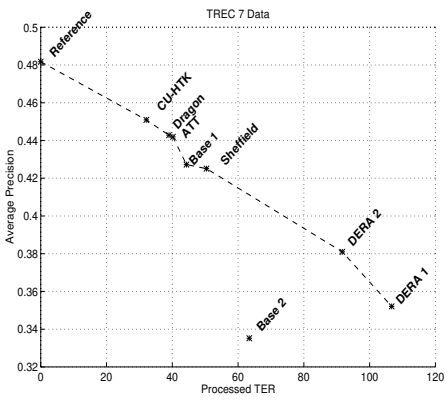
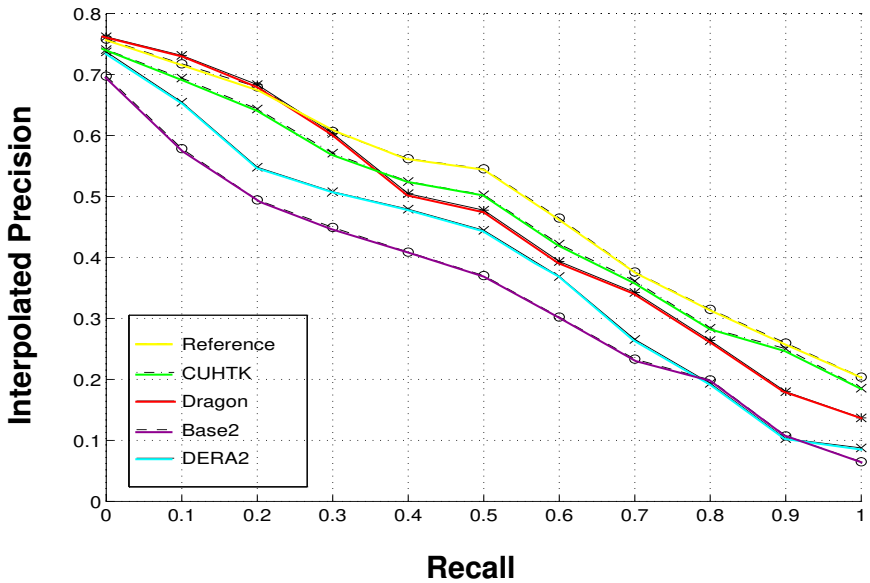
Recognition Errors and IR Performance

TER and Average Precision

TREC-6



TREC-7



TREC-7 Interpolated Precision-Recall