TC-STAR Final Review Meeting
Luxembourg, 29 May 2007

Speech Transcription

Jean-Luc Gauvain

What Is Speech Recognition?

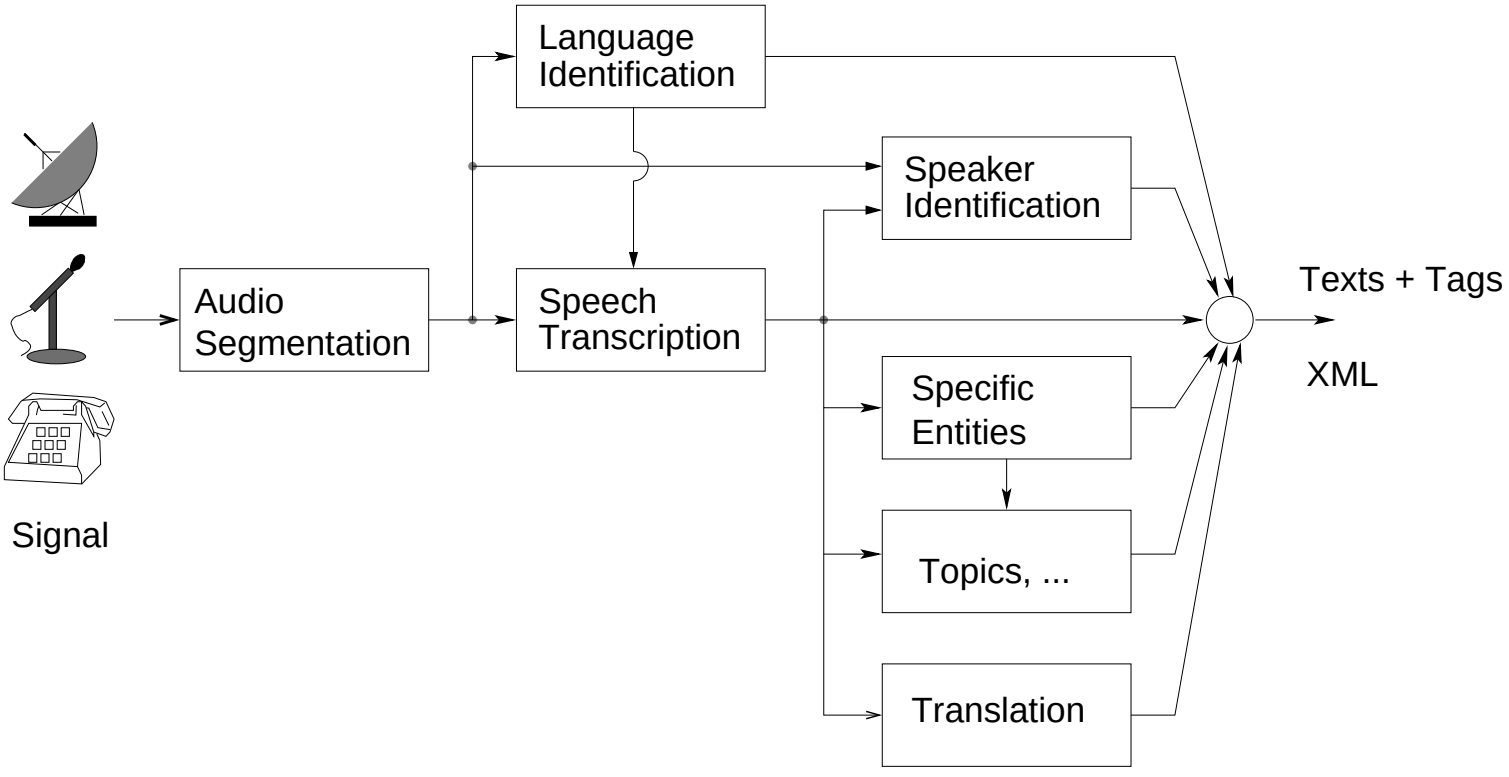
Def: Automatic conversion of spoken words in text (eg. read speech)

Other definitions: ability to understand spoken words, to respond to verbal commands, to interpret human speech, ...

play

```
<Audiofile filename="20060614_1500_1730_OR_SAT">
<SpeechSegment lang=eng spkr=MS4 stime=2751.13 etime=2846.84>
<Word stime=2751.64 dur=0.31 conf=0.980> Mister </Word>
<Word stime=2751.95 dur=0.80 conf=0.974> President, </Word>
<Word stime=2752.98 dur=0.18 conf=0.996> in </Word>
<Word stime=2753.17 dur=0.22 conf=1.000> these </Word>
<Word stime=2753.39 dur=0.61 conf=0.995> questions </Word>
<Word stime=2754.01 dur=0.15 conf=1.000> we </Word>
<Word stime=2754.16 dur=0.25 conf=0.999> see </Word>
<Word stime=2754.42 dur=0.36 conf=1.000> the </Word>
<Word stime=2754.78 dur=0.37 conf=1.000> European </Word>
```

Processing Speech



Why Is Speech Recognition Difficult?

Text:	I do not know why speech recognition is so difficult
Continuous:	Idonotknowwhyspeechrecognitionissodifficult
Spontaneous:	Idunnowhyspeechrecnitionsodifficult
Pronunciation:	YdonatnowYspiCrEkxnISxnIzsodIfIk∧lt
	YdonowYspiCrEknISNsodIfxk∧l
	YdontnowYspiCrEkxnISNsodIfIk∧lt

Important variability factors:

Speaker	*Acoustic environment*
physical characteristics (gender, age, ...), accent, emotional state, situation (lecture, conversation, meeting, ...)	background noise (cocktail party, ...) room acoustic, signal capture (microphone, channel, ...)

Audio Samples (1)

- WSJ Read stop

  PREVIOUSLY HE WAS PRESIDENT OF CIGNA'S PROPERTY CASUAL GROUP FOR FIVE YEARS AND SERVED AS CHIEF FINANCIAL OFFICER

- WSJ Spontaneous

  ACCORDING TO KOSTAS STOUPIS AN ECONOMIST WITH THE GREEK GOVERNMENT IT IS IT IS NOT YET POSSIBLE FOR GREECE TO TO MAINTAIN AN EXPORT LEVEL THAT IS EQUIVALENT TO EVEN THE THE SMALLEST LEVEL OF THE OTHER WESTERN NATIONS

- Broadcast News

  ON WORLD NEWS TONIGHT THIS WEDNESDAY TERRY NICHOLS AVOIDS THE DEATH SENTENCE FOR HIS ROLE IN THE OKLAHOM A CITY BOMBING. ONE OF THE WORLD'S MOST POPULAR AIRPLANES THE GOVERNMENT ORDERS AN IMMEDIATE SAFE TY INSPECTION. AND OUR REPORT ON THE CLONING CONTROVERSY BECAUSE ONE AMERICAN SCIENTIST SAYS HE'S READY TO GO.

Audio Samples (2)

- Conversational Telephone Speech

  Hello. Yes, This is Bill. Hi Bill. This is Hillary. So this is my first call so I haven't a clue as to - - what we're supposed to do yeah. It's pretty funny that we're Bill and Hillary talking on the phone though isn't it? Ah I didn't even think about that. {*laughter*} uhm we're supposed to talk about did you get the topic. Yes. So we're supposed to talk about what changes if any you have made since ...

- EPPS stop

  Mister President, in these questions we see the European social model in flashing lights. The European social model is a mishmash which pleases no one: a bit of the free market here and a bit of welfare state there, mixed with a little green posturing.

Audio Samples (3)

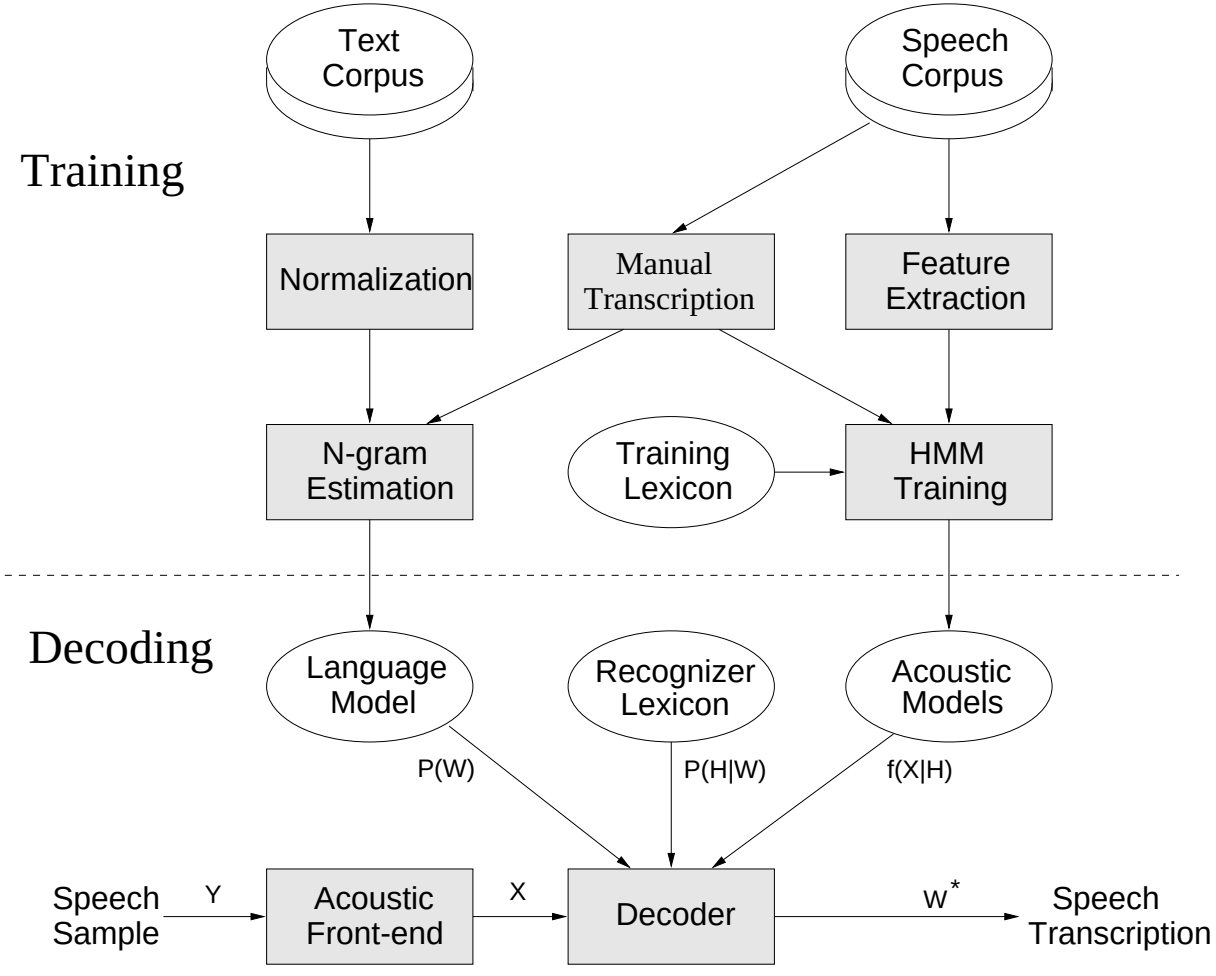
- EPPS, spontaneous stop

  Thank you very much, Mister Chairman. Hum When I came to this room I I have seen and heard homophobia but sort of vaguely ; through T. V. and the rest of it. But I mean listening to some of the Polish colleagues speak here today , especially Roszkowski , Pek , Giertych and Krupa,...

- EPPS, non native

  My main my main problem with the environmental proposals of the Commission is that they do not correspond to the objectives of the Sixth Environmental Action Plan. As far as the traffic is concerned, the sixth E. A. P. stressed the decoupling of transport growth and G. D. P. growth.

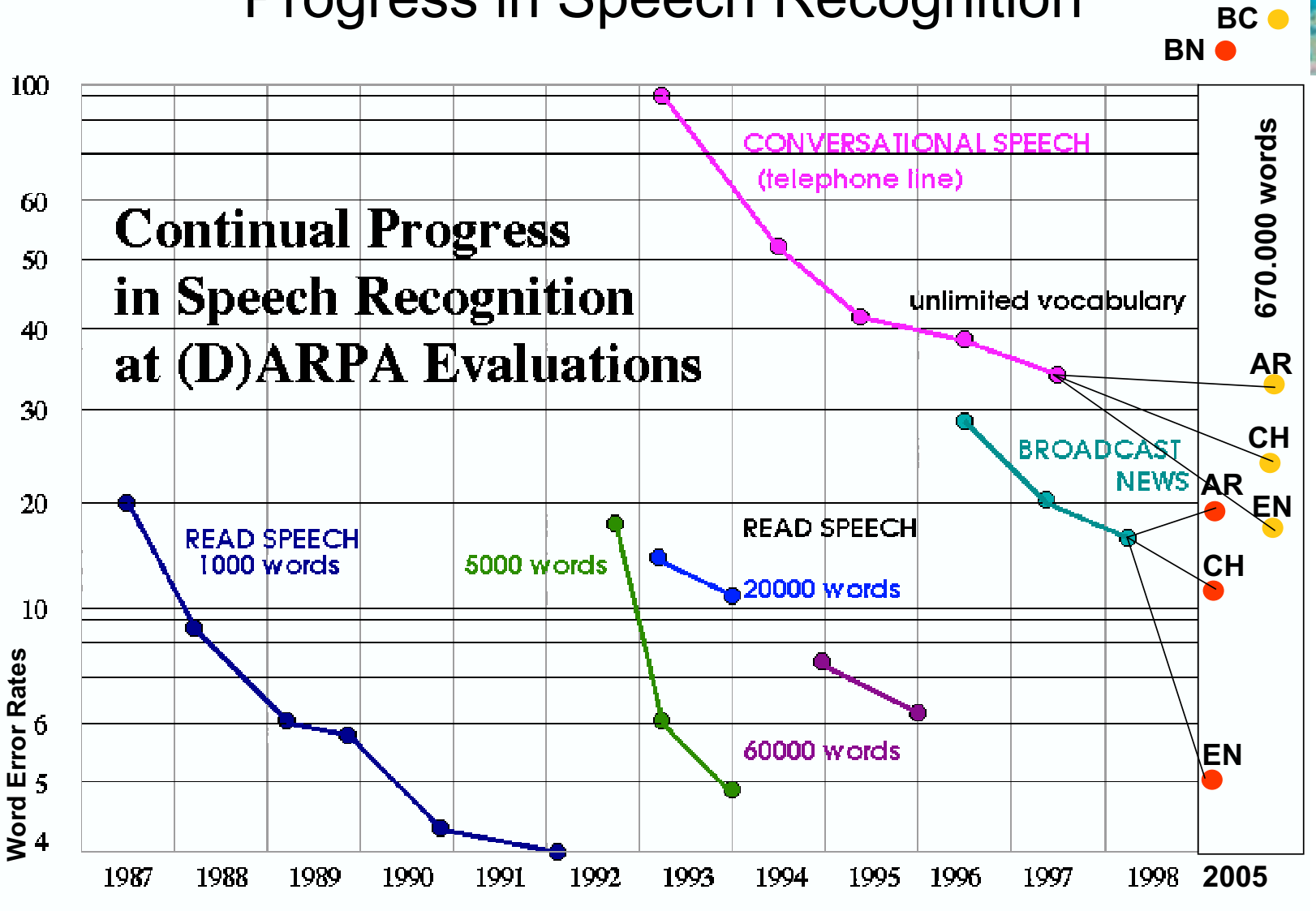
Inside ASR Systems



Indicative ASR Performance

Task	*Condition*	*Word Error*
Dictation	read speech, close-talking mic.	3-4% (humans 1%)
	read speech, noisy (SNR 15dB)	10%
	spontaneous dictation	14%
	read speech, non-native	20%
Found audio	TV & radio news broadcasts	10-15% (humans 4%)
	TV documentaries	20-30%
	Telephone conversations	20-30% (humans 4%)
	Lectures (close mic)	20%
	Lectures (distant mic)	50%
	EPPS	8%

Progress in Speech Recognition



(from T. Schultz)

ASR Activites In TCStar

- Three languages (EN, ES, MAN), three tasks (EPPS, Cortes, BN)
- Improving modeling and decoding techniques
  - audio segmentation, disctiminative features, duration models,
  - automatic learning, discriminative training,
  - pronunciation models, NN language model, decoding strategies, ...
- Model adaptation methods
  - speaker adaptive trainining, acoustic and language model adaptation,
  - cross-system adaptation, adaptation to noisy environments, ...
- Integration of ASR and Translation
  - transcription errors, confidence scores, word graph interface
  - punctuation for MT
  - transcription normalisation (98 vs. ninety eight, hesitations, ...)

Three Years Of Improvments in TCStar

WER for EPPS English