

# Audio challenges for enhancing transcriptions: the JUMAS case study

Daniele Falavigna

FBK - Fondazione Bruno Kessler, Trento - Italy

Thessaloniki, October 24 2008

# Overview

- The automatic transcription problem
- ASR system components
- Typical issues in ASR
- Speech processing issues related to Jumas

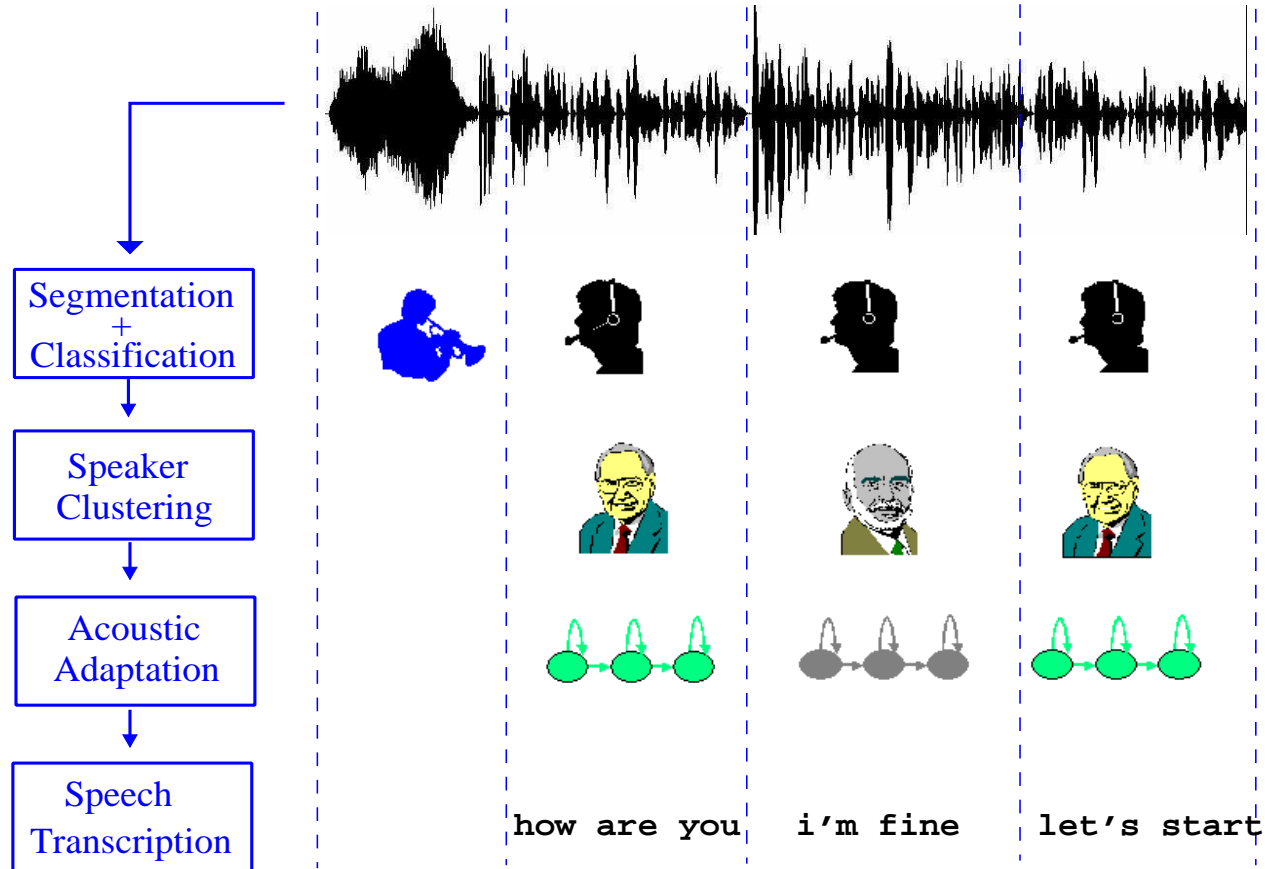
# Automatic Speech Recognition

- Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words uttered by a person into a microphone or telephone
- The ultimate goal of ASR research is to develop a system capable to recognize in *real-time with 100% accuracy all words that are intelligibly spoken* by any person, *independent of vocabulary size, environment noise, speaker characteristics and accent, or recording conditions*, etc.
- The most significant measure for evaluating ASR performance is the *Word Error Rate*, that is the frequency of errors made by the system:

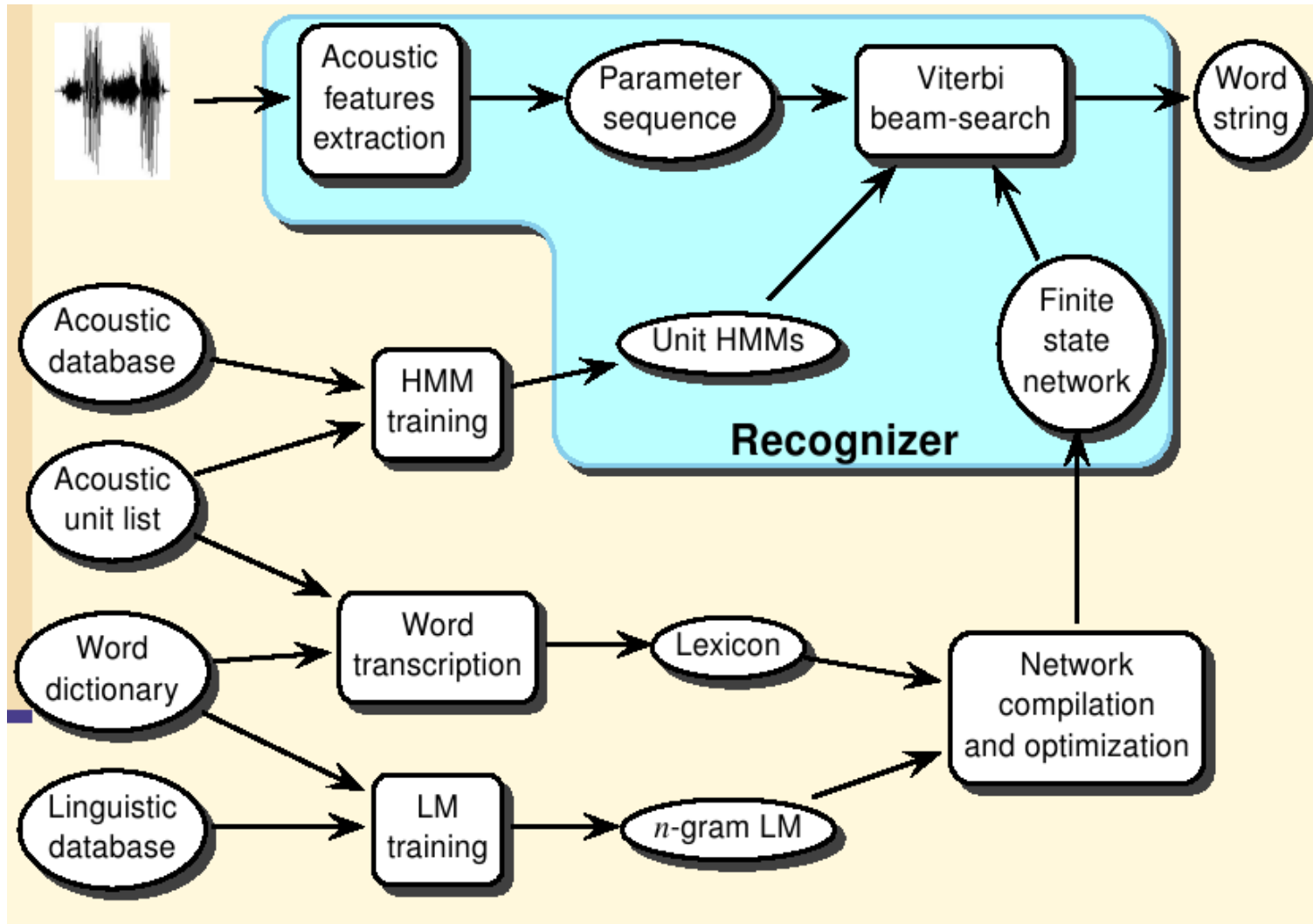
$$\text{WER} \equiv 100 \times \frac{\text{number of errors}}{\text{number of reference words}}$$

$$\text{number of errors} = \# \text{substitutions} + \# \text{deletions} + \# \text{insertions}$$

# Automatic Transcription



# ASR System Overview: Training + Run Time



# Capabilities of ASR Systems

Parameters that characterize the capabilities of ASR systems:

Parameters	Range
Speaking Mode	Isolated word to continuous speech
Speaking Style	Read Speech to spontaneous speech
Enrollment	Speaker-dependent to speaker-independent
Vocabulary	Small (< 30 words) to large(> 60000 words)
Language Model	Finite-State to context-sensitive
Perplexity	Low (< 20) to high (> 200)
SNR	High (> 30dB) to low (< 10dB)
Recording equipment	Noise-cancelling microphone to cell phone
Response Time	Quasi-Instantaneous up to tens times real time

# ASR Performance

- To evaluate the WER we need a representative sample of data, called *evaluation set*, for which a *reference transcription* is available
- The output of the automatic recognition on these data is then compared to the reference transcription to verify system performance
- The errors are computed by means of the *edit distance*, or *Levenshtein distance*
- Typical WERs for different domains are:

Connected digits	$\leq$	0.5 %
Continuous dictation	$\leq$	5 %
Studio broadcast news	$\leq$	10 %
Telephone news reports	$\leq$	20 %
Telephone conversations	$\leq$	30 %
Meetings (head mounted microphone)	$\leq$	30 %
Meetings (distant microphone)	$\leq$	50 %

## Acoustic Variability in Speech

- Acoustic variability has a strong impact in determining the difficulty of the ASR problem
- Sources of variability can (not only) be related to:

<b>Source</b>	<b>Recording equipment</b>	<b>Transmission channel</b>	<b>Environment</b>
Speaker	Head-mounted microphone	Analog line	Noise
Style	“Far” microphone	Digital line	Reverberation
Mood	Microphone arrays	Bandwidth limitations	
Coarticulation			

Automatic recognizer are still *much more sensitive to variability than humans*

## How do We Cope With Acoustic Variability?

**Acoustic features:** features extracted from the speech waveform are aimed at reducing irrelevant characteristics, while preserving phonetically relevant information

**Speaker/condition dependent training:** when targetting a particular speaker/condition, acoustic models can be trained on corresponding data. Examples are: *speaker-dependent/gender-dependent* systems

**Degrees of freedom:** statistical parameters must be added to the acoustic models, to be able to cope with the different situations. This also increases the need of training data, to allow robust estimation

## How do We Cope With Acoustic Variability? (cont)

**Acoustic normalization:** different conditions are somehow identified in the incoming feature stream, and for each of them a specific transformation is computed and applied to the features themselves, in order to reduce the cross-condition differences

**Model adaptation:** the statistical parameters of the acoustic models are *adjusted* so as their “distance” from the acoustic features is reduced

**Adaptive training:** Model adaptation to each specific condition is performed during training as well and, in the end, a *canonical* model is built as a combination of many condition-adapted models

## Statistical Language Models

- In *spoken language* processing, given a vocabulary, we need to be able to recognize *any* word sequence that can be built with those words
- But we cannot afford to *remove* linguistic constraints from the recognition process, given the weaknesses of the acoustic model, and the inherent ambiguity of spoken language
- So we use a *statistical language model (LM)*, that assigns different *a priori probabilities* to the different word sequences
- LM probabilities are combined with acoustic probabilities during the search of the “most probable” word sequence, among all possible word sequences
- The most used statistical LM for ASR is based on the *n*-gram approximation
- Usually  $n \geq 3$  is used (trigrams, fourgrams, ...). Unigram and bigrams are used when *training data do not allow to reliably estimate higher-order models*

## Issues in Jumas: Acoustic Variability

In Court rooms critical sources of acoustic variation could be:

- Spontaneous speech
- Speaker accent/dialect
- Non-native speech
- Emotional state of the speaker
- Overlapping speech
- Microphone position (speaker movements)
- Reverberation (distant microphone)
- Environmental noise (steps, doors, paper rustle, etc)

Scientific literature reports: up to 9% absolute WER increase due to “cross-talk”, up to 30% absolute WER increase due to distant microphone

In a pilot experiment, we also measured  $\approx 30\%$  WER increase between audio recordings carried out with head-mounted or distant (3m) microphone

## ASR Issues in Jumas: Language Variability

Judicial debates concern an unrestricted domain of discourses and difficulties could come from:

- Mismatch between training and testing conditions
- Spontaneous speech
- Use of dialect (single words and whole sentences)
- Non-native speech
- Out of vocabulary words (e.g. names of persons or institutions, technical terms, etc.)

Also in this case LM adaptation can be applied by interpolating probabilities of  $n - grams$  estimated on huge “out-of-domain” text corpora with probabilities of  $n - grams$  estimated on “in-domain” text corpora  $\Rightarrow$  *need training data*

## ASR Issues in Jumas: Language Variability

The language used by the actors of a trial can differ according to both the role and the state of the debatements. A preliminary analysis of some trial recordings showed that:

- the languages used by both the judge and the prosecutor are more formal than that of the witness: sometimes the judge dictates parts of the law code
- the language used by the judge at the beginning of a session, e.g. during the rollcall, is more schematic than that used during the debatement
- also the language used when pronouncing the sentence is formal
- in general, the witness make many hesitations, false starts, repetitions, etc
- lawyers (often absent) use a formal language and tend to minimize their interventions

To model the above scenario from the point of view of statistical language model *training data*, i.e. transcriptions (possibly exact or, at least, approximate) of trial debates, are needed

## Demos

Two videos showing:

- output of a transcription system, developed within the framework of the EU Project TC-STAR, addressing spoken language translation of European Parliament Plenary Speeches
- a system, developed within the framework of the EU Project CHIL, capable to detect multiple speakers who are overlapping one each other during a meeting