

PART III
Automatic Speech
Recognition (ASR)

Contents (Part III)

- **Introduction**
- Instance-based approach (DTW)
- Model-based approach (HMM, HMM/ANN)
 - Acoustic model
 - Language model

ASR: What for?

- Telecommunications:
 - access to data or services over the telephone (e.g. AT&T's Maxwell personal telephone attendant)



ASR: What for?

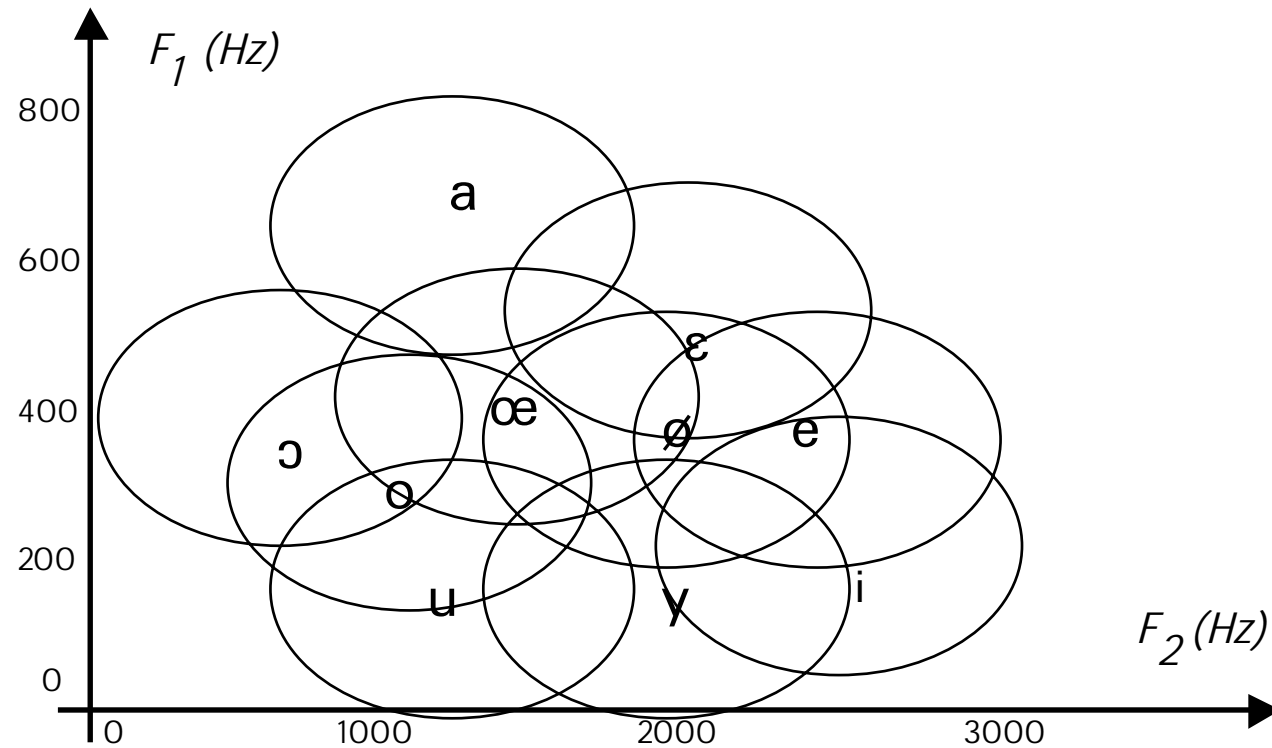
- Office/desktop:
 - voice control of PC/Workstations, of programs, dictation systems
- Manufacturing/Business:
 - aid in manufacturing process, quality control, stock control and management
- Medical/Legal:
 - creation of medical/legal reports, briefs, diagnostics...
- Others:
 - games, aid to handicapped, interactive kiosk information systems

Typology of ASR systems

- **Speaker**-dependent vs. -independent
 - **Language** constraints:
 - isolated word recognition
 - connected word recognition
 - keyword spotting
 - continuous speech recognition
- } +**vocabulary** :
small (100),
medium (5000),
large (50000)
- +**perplexity**
- **Robustness** constraints
 - laboratory (office) conditions: imposed microphone, no ambient noise
 - (quiet) telephone system (Δ mic., Δ noise in a given range)
 - real-life (human-like) ASR...

Challenges

- Coarticulation!



Challenges

- **Inter-speaker** variability
 - Vocal tract, gender, dialects
- **Language** variability
 - From isolated words to continuous speech
 - Out-of-vocabulary words
- **Noise**
 - Convolutional: recording/transmission conditions, reverberation
 - Additive: recording environment, transmission SNR
 - Intra-speaker variability: stress, age, humor, Lombard effect, ...

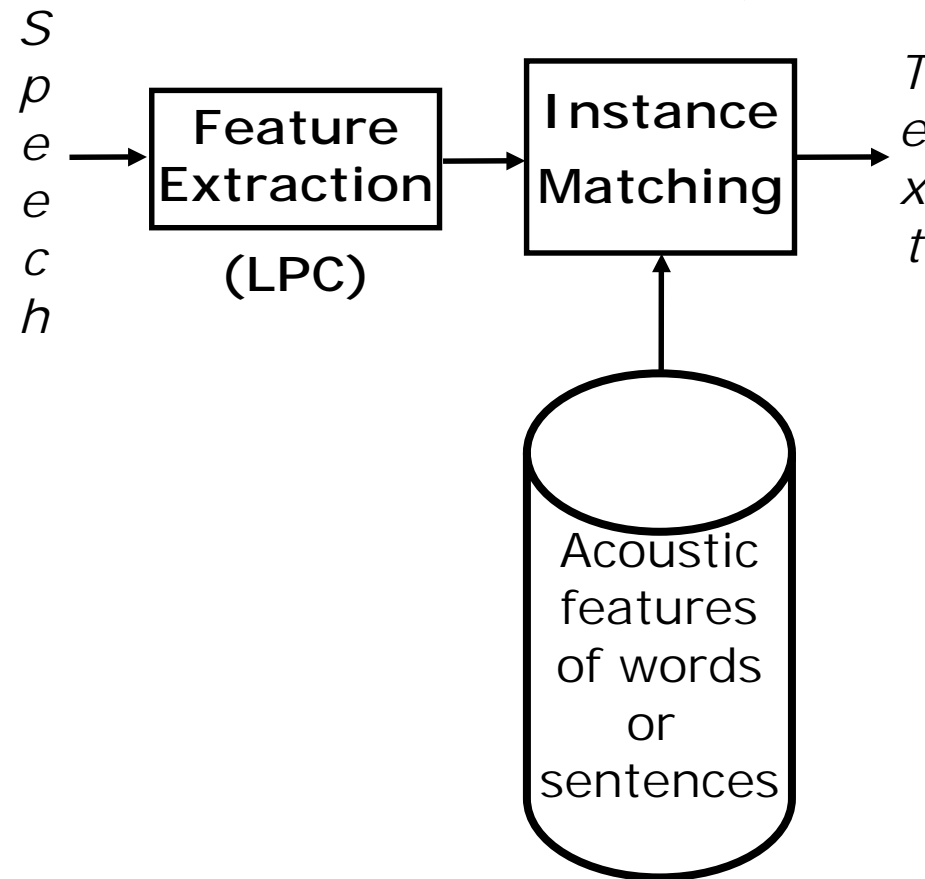
Levels of complexity

	<i>Isolated</i>		<i>Connected</i>		<i>Continuous</i>	
<i>Speaker dependent</i>	small	1	small	4	small	5
	large	4	large	5	large	6
<i>Multi speaker</i>	small	2	small	4	small	6
	large	4	large	5	large	7
<i>Speaker independent</i>	small	3	small	4	small	5
	large	5	large	8	large	10

Contents (Part III)

- Introduction
- **Instance-based approach (DTW)**
- Model-based approach (HMM, HMM/ANN)
 - Acoustic model
 - Language model

ASR flow-chart (70 's)



The **instance-based approach** (DTW)
OK for small vocabulary, speaker depdt

Instance-based ASR

- Unknown utterance $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$
(with $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$)
Known utterances $\mathbf{Y}^1 = \{y^1_{11}, y^1_{12}, \dots, y^1_{1J(1)}\}$
 $\mathbf{Y}^2 = \{y^1_{21}, y^1_{22}, \dots, y^1_{2J(2)}\}$
...
 $\mathbf{Y}^K = \{y^1_{K1}, y^1_{K2}, \dots, y^1_{KJ(M)}\}$
- Compute $D(\mathbf{X}, \mathbf{Y}^k)$ for $k = 1 \dots M$
- *Recognize:*

$$\mathbf{X} = \mathbf{Y}^{best}$$
with $D(\mathbf{X}, \mathbf{Y}^{best}) \leq D(\mathbf{X}, \mathbf{Y}^k)$ for $k = 1 \dots M$
- OK for **spkr-dpndt** isolated word reco.

Global distance $D(X, Y^k)$?

- Local distance: $d(x_n, y_j^k)$?

- Euclidian distance: $d(x_n, y_j^k) = (x_n - y_j^k)^T (x_n, y_j^k) = \sqrt{\sum_{i=1}^d (x_{ni} - y_{ji}^k)^2}$

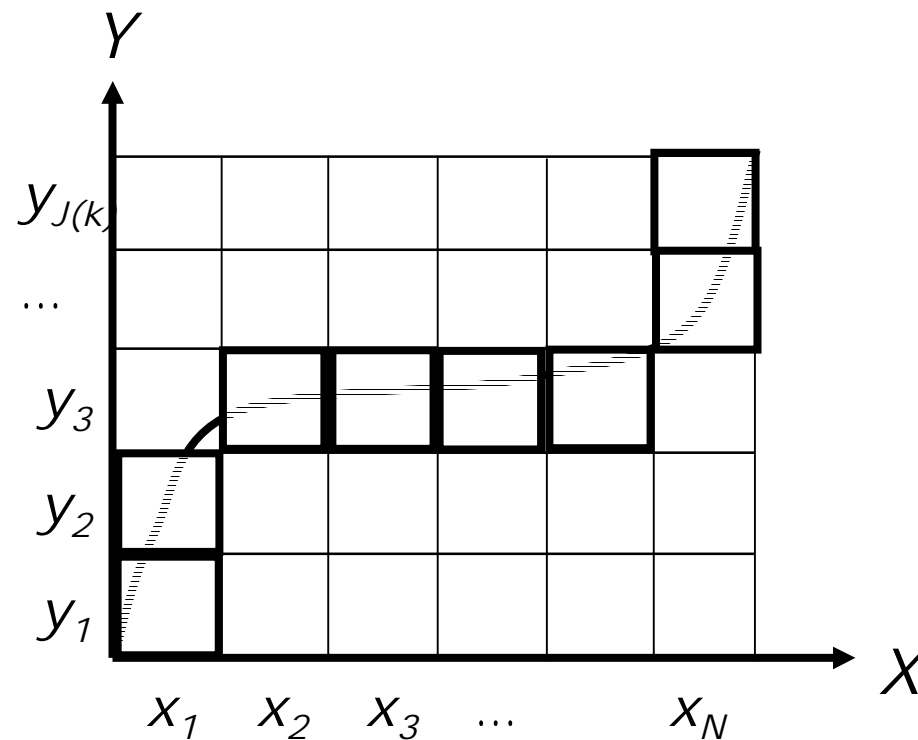
- Mahalanobis distance: $d(x_n, y_j^k) = (x_n - y_j^k)^T \Sigma^{-1} (x_n, y_j^k)$

- Itakura distance (LPC-based)

- ...

Global distance $D(X, Y^k)$?

- Non linear time warping

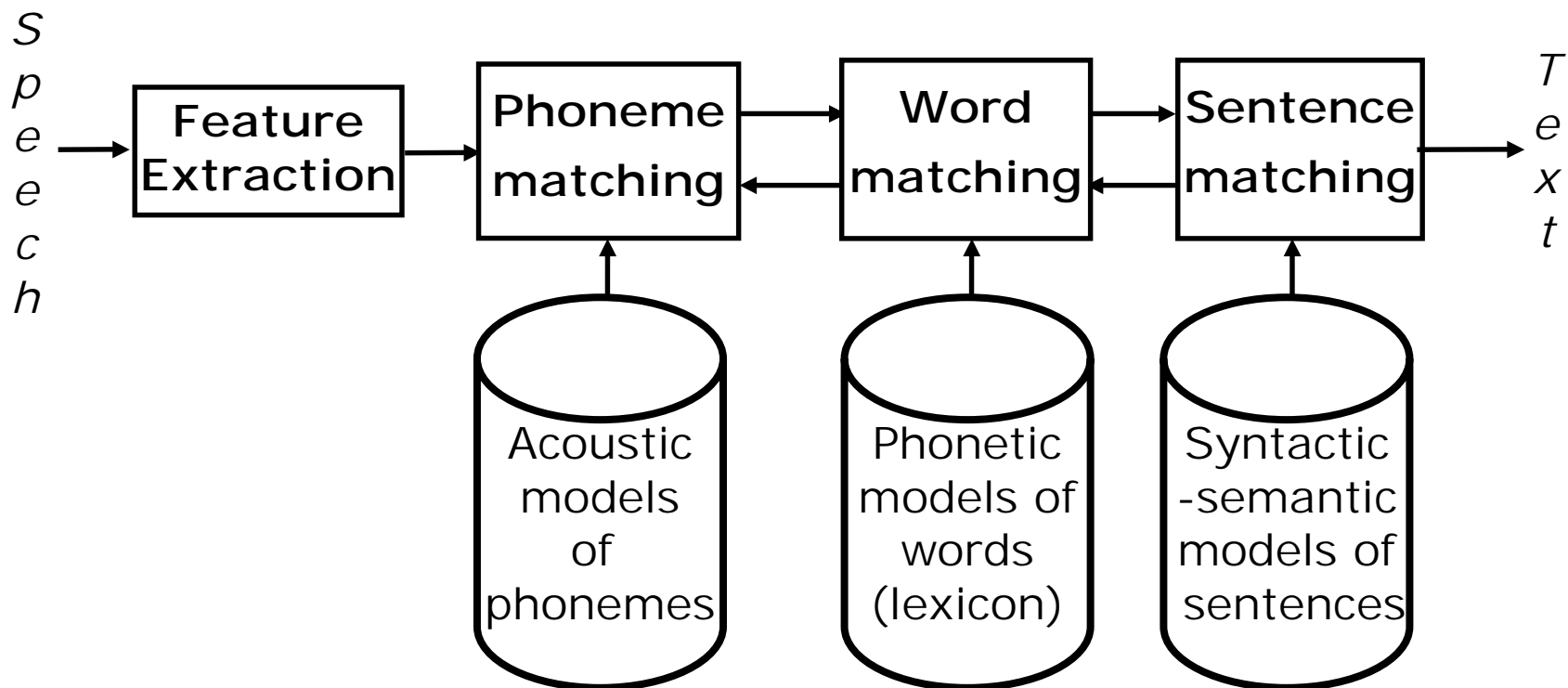


- Best path? **Viterbi algorithm**

Contents (Part III)

- Introduction
- Instance-based approach (DTW)
- **Model-based approach (HMM, HMM/ANN)**
 - Acoustic model
 - Language model

Today 's ASR flow-chart



Phoneme-based approach using statistical models (HMM or HMM/ANN) for acoustics and linguistics: Large vocabulary, speaker indepdtd

Model-based ASR

- Unknown utterance

$$\mathbf{X} = \{x_1, x_2, \dots, x_N\}$$

$$(with\ x_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T)$$

Models of known utterances:

$$M_1, M_2, \dots, M_J$$

- **BAYESIAN (MAP) CLASSIFICATION:**

- Compute $P(M_j|\mathbf{X})$ for $j=1\dots J$

- Recognize:

$$\mathbf{X} = M_{best}$$

with $P(M_{best}|\mathbf{X}) \geq P(M_j|\mathbf{X})$ for $j=1\dots J$

Model-based ASR

- $P(M_j|X)$ = « **posterior** probability of M_j »
not easy to compute
- Bayes rule:
$$P(M_j | X) = \frac{P(X | M_j) \cdot P(M_j)}{P(X)}$$
- $P(X|M_j)$ = « **likelihood** of X »
- $P(M_j)$ = « **prior** probability of M_j »
- $P(X)$ = constant

$$\max P(M_j|X) = \max [P(X|M_j) \cdot P(M_j)]$$

Model-based ASR

$$\max P(M|X) = \max [P(X|M) \cdot P(M)]$$

- M is a sequence of words (W_1, W_2, \dots, W_K)
- (W_1, W_2, \dots, W_K) may have several phonetic transcriptions P_l ($l=1 \dots L$)
- $$P(X|M) = P(X|P_1)P(P_1|M) + P(X|P_2)P(P_2|M) \\ \dots + P(X|P_L)P(P_L|M)$$

$P(X|P_l)$ ← Acoustic model

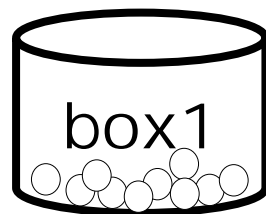
$P(P_l|M)$ ← Phonetic model

$P(M)$ ← Language model

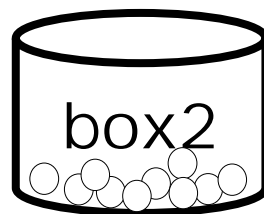
Contents (Part III)

- Introduction
- Instance-based approach (DTW)
- Model-based approach (HMM, HMM/ANN)
 - **Acoustic model**
 - Language model

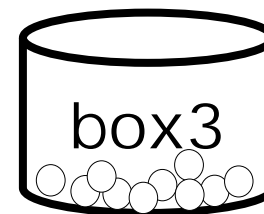
Hidden Markov Model (HMM)



$$\begin{aligned} P(r/B1) \\ P(b/B1) \\ P(g/B1) \end{aligned}$$



$$\begin{aligned} P(r/B2) \\ P(b/B2) \\ P(g/B2) \end{aligned}$$



$$\begin{aligned} P(r/B3) \\ P(b/B3) \\ P(g/B3) \end{aligned}$$

} Emission probabilities

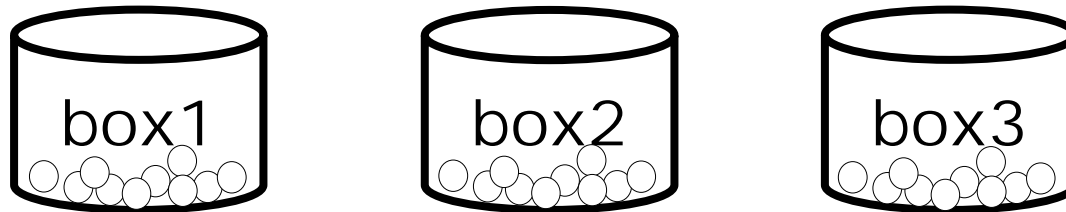
$$\begin{aligned} P(B1/B1), P(B2/B1), P(B3/B1) \\ P(B1/B2), P(B2/B2), P(B3/B2) \\ P(B1/B3), P(B2/B3), P(B3/B3) \end{aligned}$$

} Transition probabilities

Double, embedded stochastic process:

- choose box using transition probs.
- choose ball using emission probs.

Hidden Markov Model (HMM)

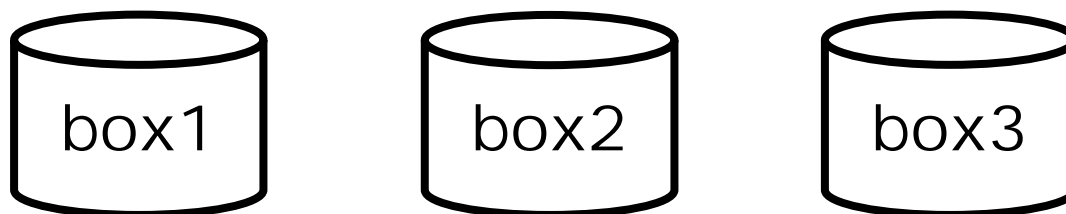



$$\begin{aligned} P(r,b,r|Model) = & P(r,b,r|B1,B1,B1) P(B1,B1,B1) \\ & + P(r,b,r|B1,B1,B2) P(B1,B1,B2) \\ & + P(r,b,r|B1,B1,B3) P(B1,B1,B3) \\ & + P(r,b,r|B1,B2,B1) P(B1,B2,B1) \\ & + P(r,b,r|B1,B2,B2) P(B1,B2,B2) \\ & \dots \\ & + P(r,b,r|B3,B3,B3) P(B3,B3,B3) \end{aligned}$$

with

$$\begin{aligned} P(r,b,r|B_i,B_j,B_k) &= P(r|B_i) P(b|B_j) P(r|B_k) \\ P(B_i,B_j,B_k) &= P(B_i) P(B_j|B_i) P(B_k|B_j) \end{aligned}$$

Training HMMs



- Data: « brbrrbgbrbbggbggbrbgbrrggbrbogg... »
- Emission and transition probabilities?
If states were known: counting
- **EM** (expectation-maximization) **algorithm**
 - Initialize Probs. (first guess if possible): M^0
 - Decode the data with $M^0 \rightarrow$ states
 - Re-estimate Probs. by counting: M^1  until $M^j \approx M^{j+1}$

Contents (Part III)

- Introduction
- Instance-based approach (DTW)
- Model-based approach (HMM, HMM/ANN)
 - Acoustic model
 - **Language model**

Language model

$$\begin{aligned} P(\mathbf{M}) &= P(\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_K) \\ &= \prod_{k=1}^K P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, \dots, \mathbf{W}_1) \end{aligned}$$

- ???
- $\rightarrow P(M)$ is actually $P(M \mid \text{a language model})$
- Most widely used : *n-grams*

n-gram models

- *Hyp*: the probability of having a word in a sentence does not depend more on all the words of the sentence than on the *n* previous words:

$$\begin{aligned} P(\mathbf{M}) &= \prod_{k=1}^K P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, \dots, \mathbf{W}_1) \\ &= \prod_{k=1}^K P(\mathbf{W}_k \mid \mathbf{W}_{k-1}, \mathbf{W}_{k-2}, \dots, \mathbf{W}_{k-n}) \end{aligned}$$

- ex: estimation with a trigram ($n=2$):

$P(\text{the weather is nice}) = P(\text{the} \mid _, _)$

$P(\text{weather} \mid \text{the}, _)$ $P(\text{is} \mid \text{weather}, \text{the})$

$P(\text{nice} \mid \text{is}, \text{weather})$

n-gram results

“That this simple approach is so successful is a source of considerable irritation to me and to some of my colleagues. We have evidence that better language models are obtainable, we think we know many weaknesses of the trigram model, and yet, when we devise more or less subtle methods of improvement, we come up short.”

F. Jelinek, « Up from trigrams », 1993

Conclusion

<i>Type</i>	<i>Task</i>	<i>Mode</i>	<i>Vocabulary</i>	<i>error rate</i>
<i>Isolated words</i>	Equiprobable words	Sp. Depdt	10 digits	0%
		Sp. Indepdt	39 ascii	4.5%
		Sp. Indepdt	1109 basic English	4.3%
		Sp. Indepdt	10 digits	0.1%
		Sp. Indepdt	39 ascii	7.0%
		Sp. Indepdt	1218 names	4.7%
<i>Connected words</i>	Sequence of digits id.	Sp. Depdt	10 digits	0.1%
		Sp. Indepdt	11 digits	0.2%
	Flight reservation	Sp. Depdt	129 words	0.1%
<i>Continuous speech</i>	Ressource management (perplexity 60)	Sp. Indepdt	991 words	3.0%
	Airline travel information system (perplexity 25)	Sp. Indepdt	1800 words	3.0%
	Wall street journal (perplexity 145)	Sp. Indepdt	20000 words	12.0%

Today 's error rates

- Importance of the language model:
 - Ressource Management:
without LM: 85% words; with LM: 97%
 - ⇒ The last 3% might still be a *language* problem
- These are laboratory systems, working on read speech!
Real life systems, spontaneous speech: -30%
:-(
- Current issues :

Robustness Spkr adaptation Language models