Speech Processing

LIA : Laboratory of Computer Science, University of Avignon

Language Processing group :

- about 12 researchers/18 Phd Students
- Topics : rich transcription, speech analytics, dialogue systems, natural language processing

Georges Linarès (georges.linares@univ-avignon.fr)

Speech Processing

- Outline :
  - Introduction
    - What is Speech ?
    - Speech as a part of the Artificial intelligence project
    - An historical view of speech processing
  - Generalities about statistical speech processing
  - Speech recognition Systems : state of the art
  - Speaker identification
  - Practical work : structuring video database by analysing spoken contents

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Speech Processing

- What is Speech ?
  - Verbal mean of communication
  - Speech is not *writen language*
  - Technically :
    - Sounds produced by the vocal folds, the breathing, the articulatory system :
      - Source : vocal folds (pitch)
      - Modulation due to articulators

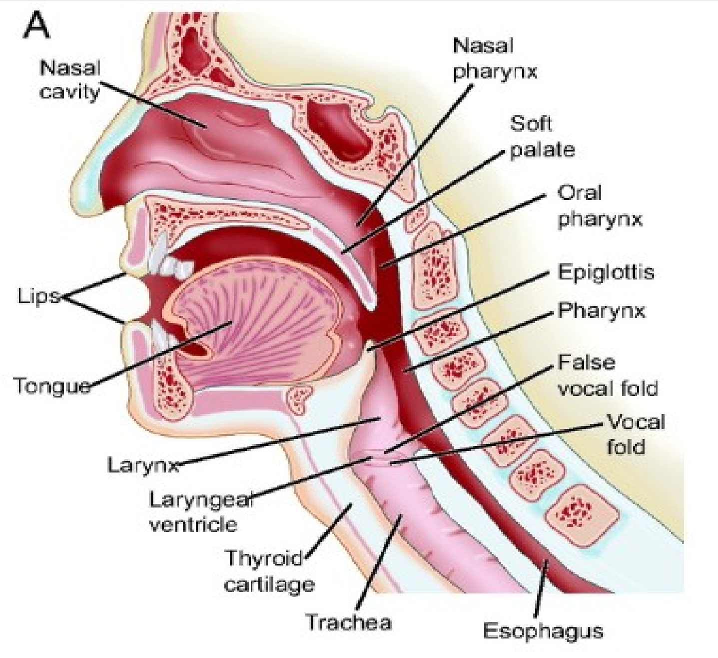
UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

What is Speech ?

***Phonological* point of view :** Speech is a production of the human vocal system

Articulators :



What is Speech ?

***Sociological* point of view :**
Speech is the main communication mean of human communities







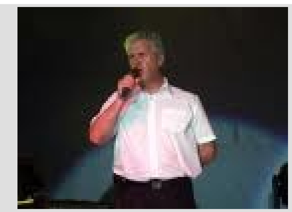
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Speech Processing : **the engineer point of view**

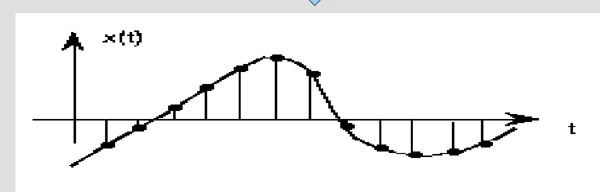
- Speech is a mean to exchange information
- Speech records contain information related to :
  - The semantic contents
  - The speaker identity, emotional state, intents, ...
  - The context
- Speech is useful to :
  - Driving machines by voice commands
  - Extract informations

Representation of speech signal

Analogic speech signal (acquisition)

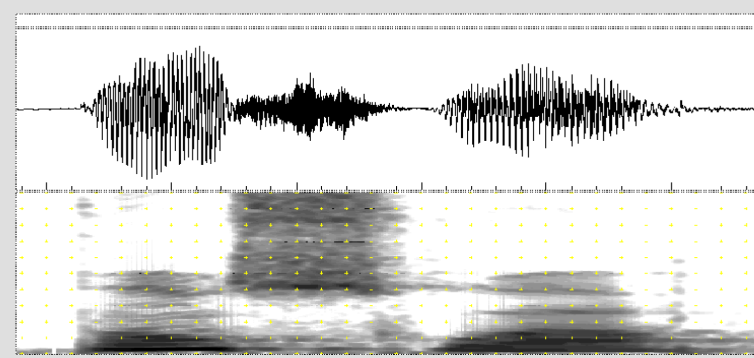


Digitalization, quantification



2D representation :

- Feature extraction
- Analysis in a slidding window



What is Speech ?

Speech is a complex communication mean :
- Human languages are complex (acoustic/linguistic structure)
- Human thinking is complex
- A spoken message resuts from :
  - the context of the discurse
  - acoustic environments
  - Who are the speakers and the listeners
- Speech understanding relies on :
  - Various level of knowledge (lexical, linguistic, pragmatic, semantic, ...)
- High variability of speech (knowledge sources, media, contents)

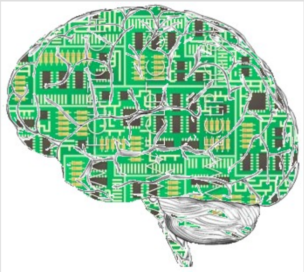
UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Speech processing and A.I.

A.I. project : an *intelligent* machine :

- android?
- a thinking machine ?
- a machine able to perform complex tasks ?
- Simulation of human capacities :
  - Perception (recognition), problem solving, decision making
- **<u>Industrial applications</u>**
- **<u>Building system tractable by computers</u>**



History of Speech processing

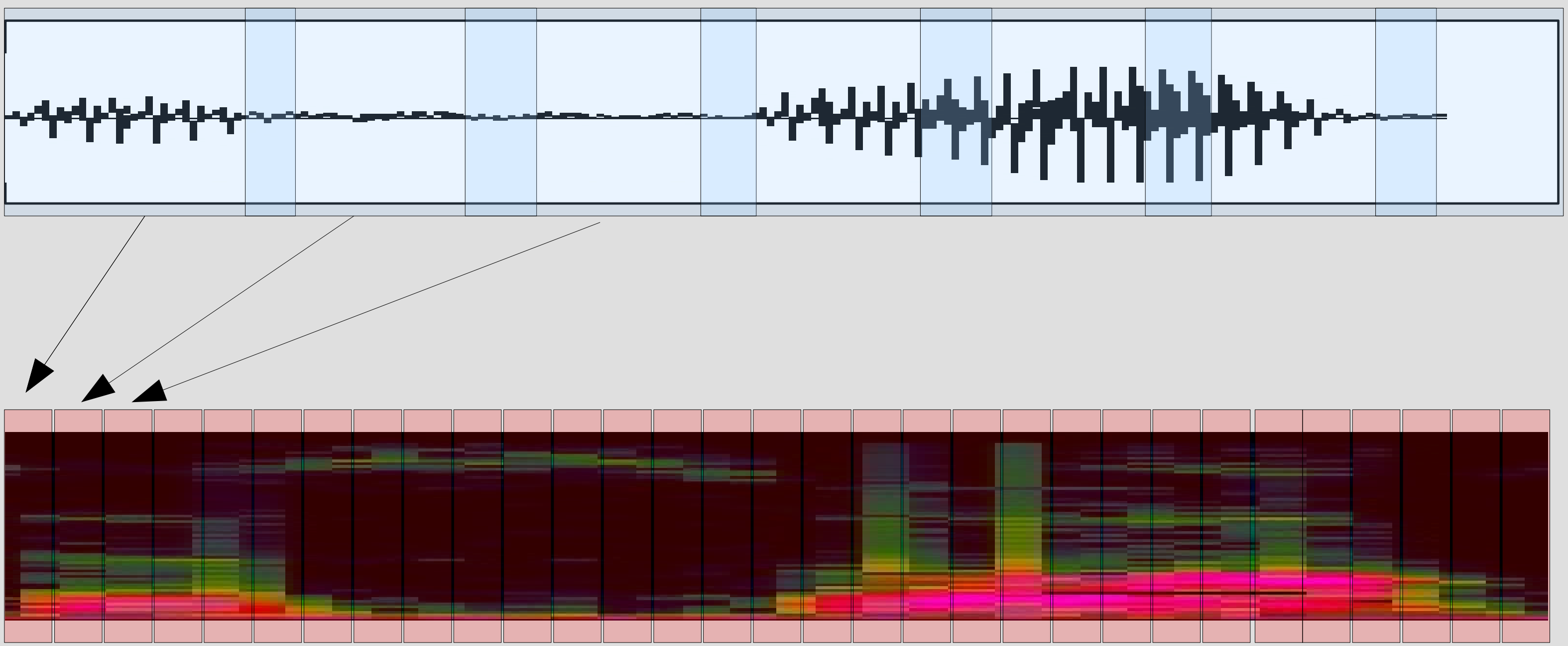
- Human-inspired approaches (1970-)
  - Knowledge-based approaches
    - But *humans are still mysterious for the science*
  - Neuromimetic approaches
    - Artificial brains seem easy to build (not so clear...)
    - **machine learning** *versus* **machine knowing**
- Nowadays (1990-):
  - *Machine Learning* but ***statistical modeling***

Statistical Speech modeling

- 4 key issues :
  - Features extraction
  - Spectral/cepstral models
    - pattern recognition problems
  - Modeling temporal structures
  - Modeling high level information
    - linguistic, semantic, pragmatic...

Statistical Speech modeling : feature extraction

Speech parametrization



Georges Linarès – CERI-LIA – University of Avignon

Statistical Speech modeling : feature extraction

- Speech parametrization :
- LPC : Linear Predictive Coding
  - Principle :
    - to code the prediction errors
    - Auto-regressive models

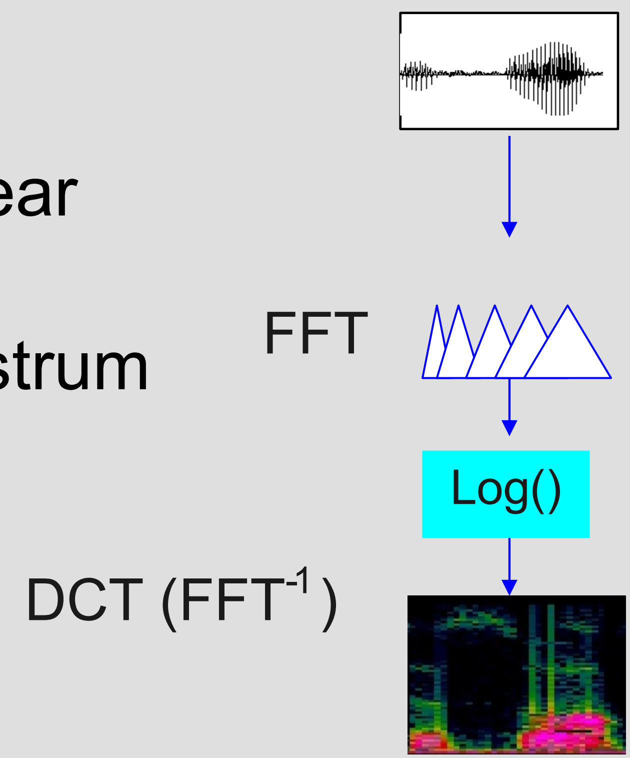
$$s(n) = \sum_{i=1}^{P} a_i s(n-i)$$

$$e(n) = x(n) - s(n) = x(n) - \sum_{i=1}^{P} a_i x(n-i)$$

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : feature extraction

- Speech parametrization :
  - PLP : perceptually-based linear prediction
  - MFCC : Mel Frequency Cepstrum coeffcients

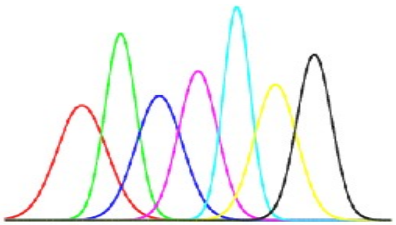


Statistical Speech modeling : feature extraction

- Speech parametrization : open issues
  - **Robustness**
  - Dimensionality reduction
  - Discriminative/generative approaches
    - LDA : Linear discriminant analysis
    - PCA : principal component analysis
    - ICA : independant componenent analysis
  - **Combination of audio features**
    - Complementarity of features

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling cepstral patterns



- Modeling cepstral features
  - Consensual approach : Gaussian Mixture Models
  - Principle : approximation of probability density function of cepstral patterns

$$l(x|\mu,\Sigma)=\frac{1}{(2\pi)^{m/2}||\Sigma||^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\,\Sigma^{-1}(x-\mu)\right)$$

$$l(x|Gmm_k)=\sum_{i=0}^{N} w_k\,.\,l(x|\mu_k,\Sigma_k)$$

Parameters : $\lambda_k=(\mu_k,\Sigma_k)$

Statistical Speech modeling : modeling cepstral patterns

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

- Gaussian Mixture Models
  - Allow us to estimate probabilities (likelihood) of observations knowing a model
  - Problems :
    - Estimation (training)
    - Integration to client systems

Statistical Speech modeling : modeling cepstral patterns

- Training Gaussian Mixture Models
  - Principle
    - Criterion: Maximun Likelihood
    - Optimization algorithm :
      - Estimate of the model parameters maximizing the Likelihood
      - Training strategy : Expectation-Maximisation
        - E : estimate of the likelihood
        - M : updating de parameters to maximize Likelihood.

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling cepstral patterns

- EM :
  - Iterative process (until convergence)
  - Updating functions (N-component Gmm) :

$$l(x|Gmm) = \sum_{k=0}^{N} w_k . l(x|\lambda_k)$$

Likelihoods

$$w'_k = \frac{\sum_X P(X|\lambda_k)}{\sum_X \sum_i P(X,|\lambda_i))}$$

Weights

$$P(X|Gmm_i) = \frac{l(X|Gmm_i)}{\sum_{k=0}^{N} l(X|Gmm_k)}$$

Probabilities

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling cepstral patterns

- Updating rules :

Means $\longrightarrow$

$$\mu_k^{t+1} = \frac{\sum_x P(x|\lambda_k^t).x}{\sum_x \sum_i^N P(x|\lambda_i^t)}$$

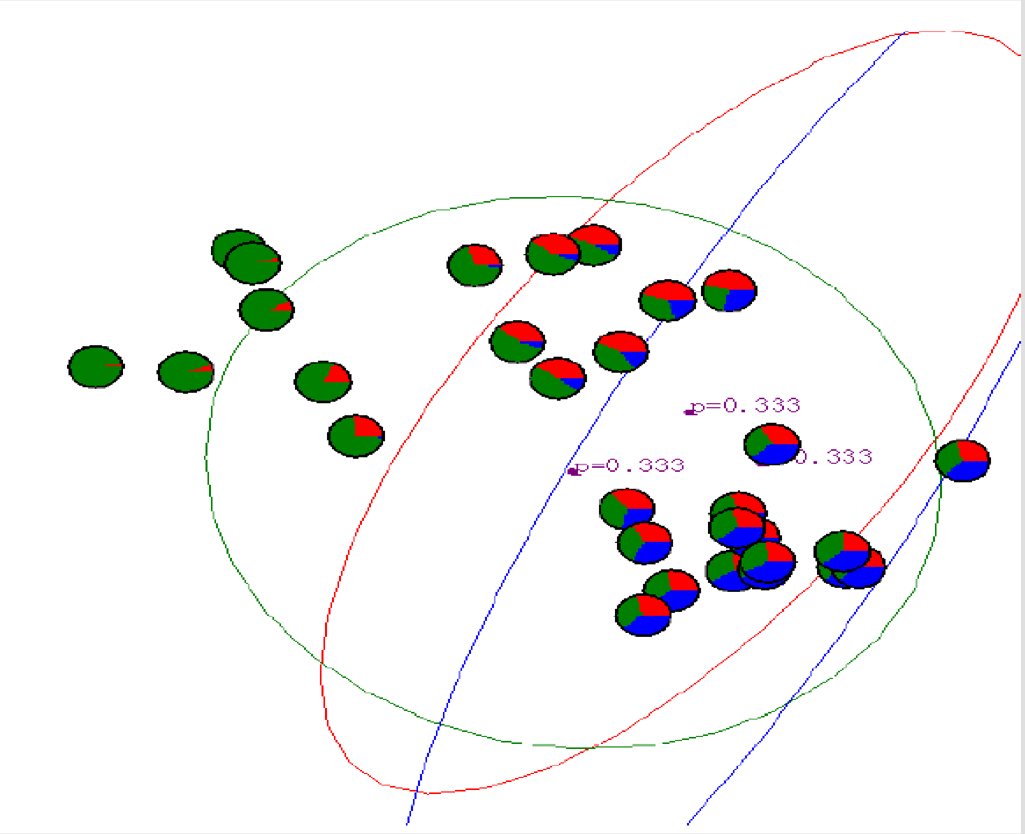
Variance $\longrightarrow$

$$\Sigma_k^{t+1} = \frac{\sum_x P(x|\lambda_k^t)[x-\mu_k^{t+1} i][x-\mu_i^{t+1}]^T}{\sum_x \sum_i^N P(x|\lambda_i^t)}$$

Statistical Speech modeling : modeling cepstral patterns

EM - Example

3 Gaussian components

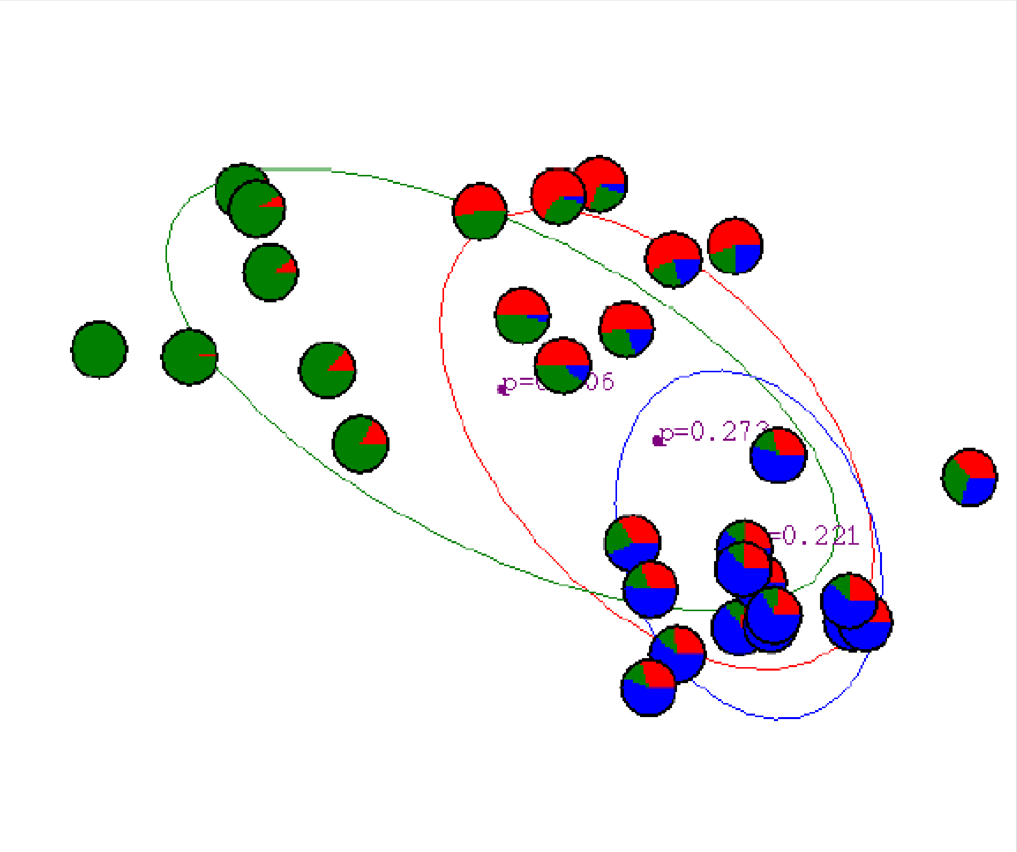


A.W. Moore

Statistical Speech modeling : modeling cepstral patterns

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Iteration 1

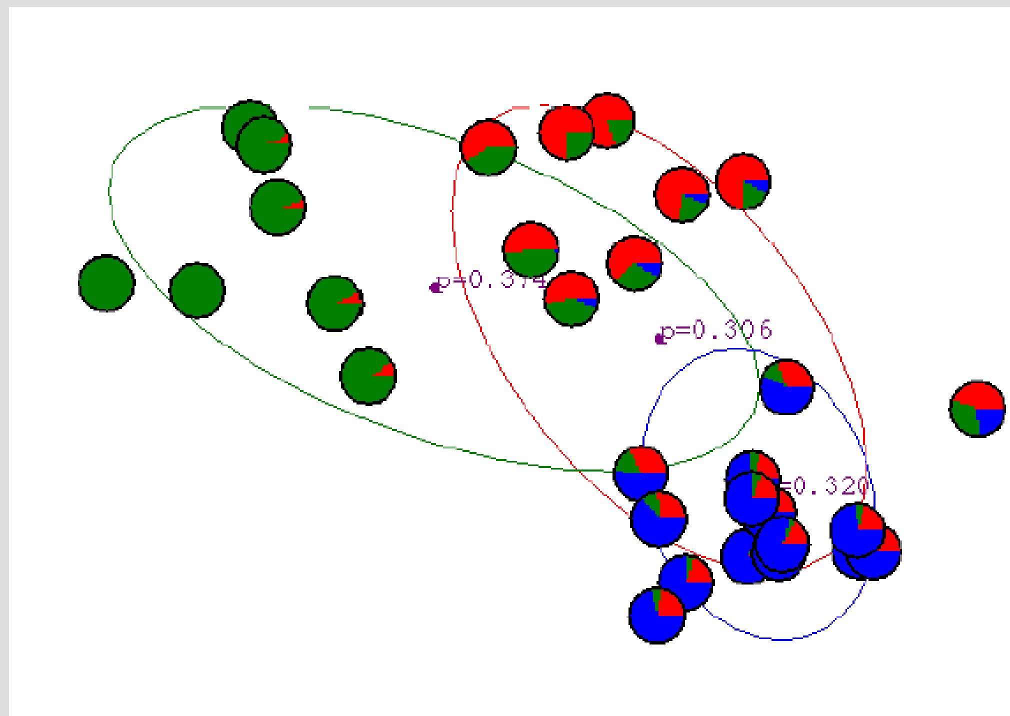


A.W. Moore

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling cepstral patterns

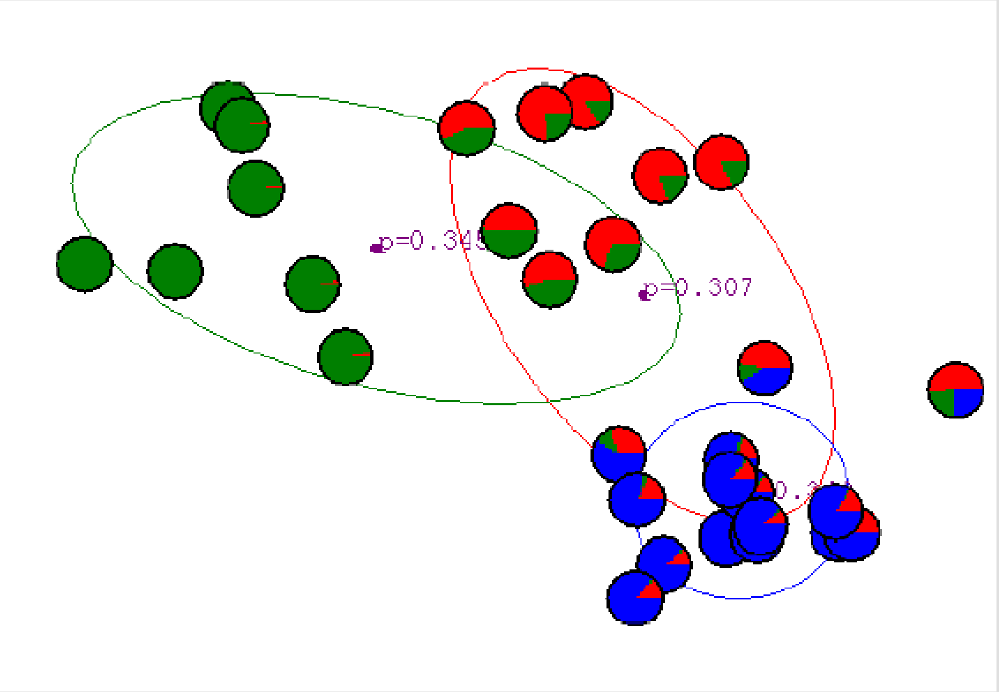
Iteration 2



A.W. Moore

Statistical Speech modeling : modeling cepstral patterns

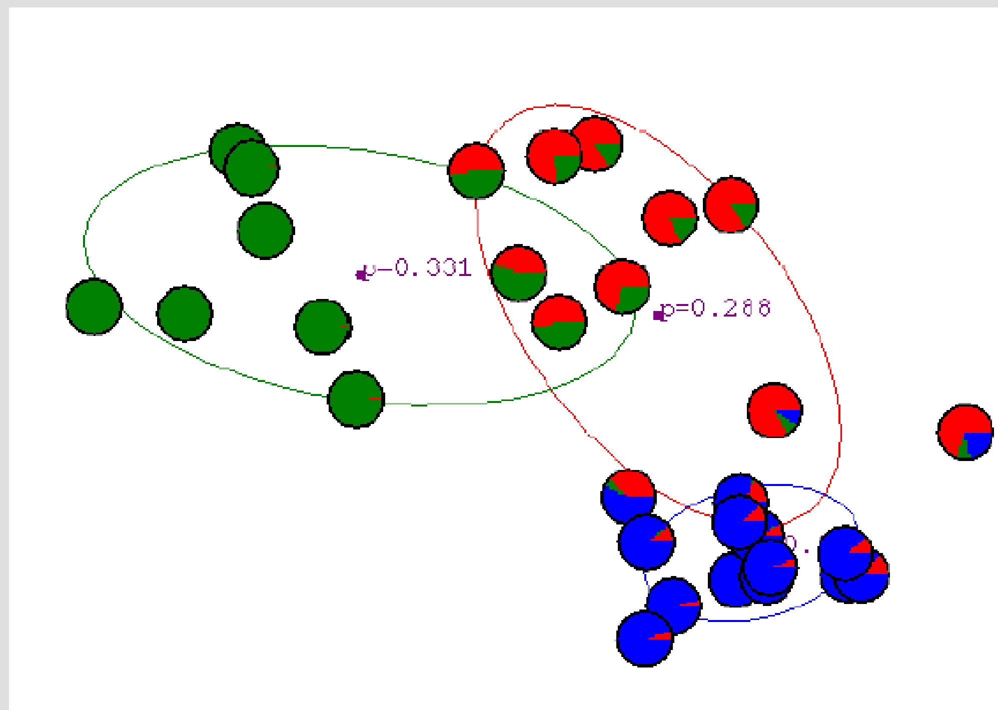
Iteration 3



A.W. Moore

Statistical Speech modeling : modeling cepstral patterns

Iteration 4

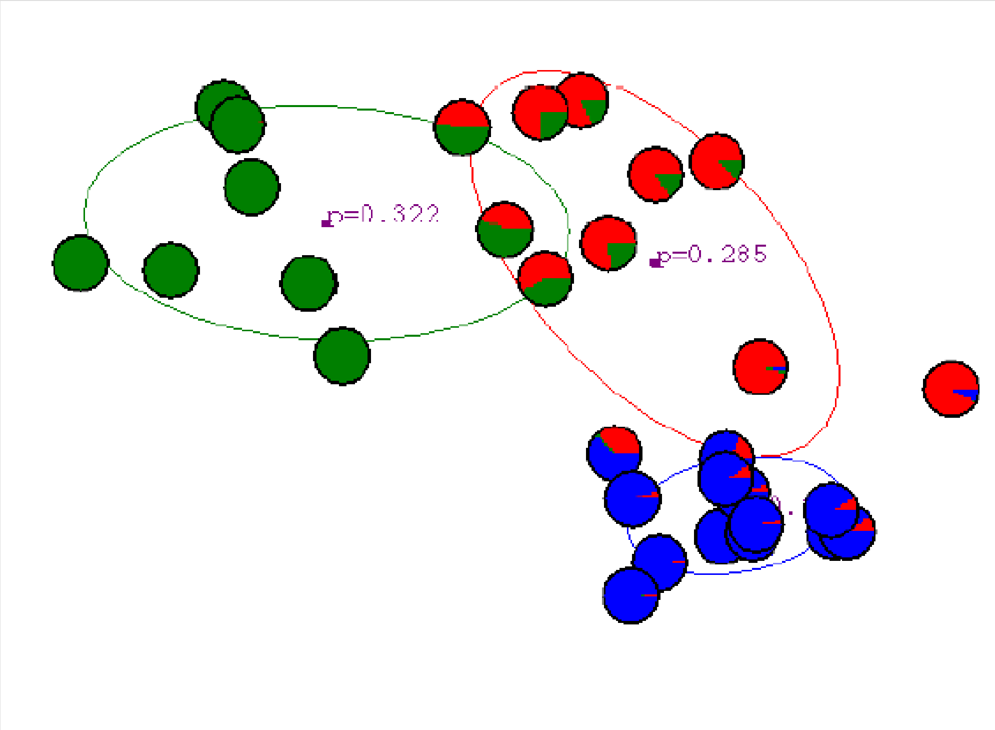


A.W. Moore

Statistical Speech modeling : modeling cepstral patterns

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

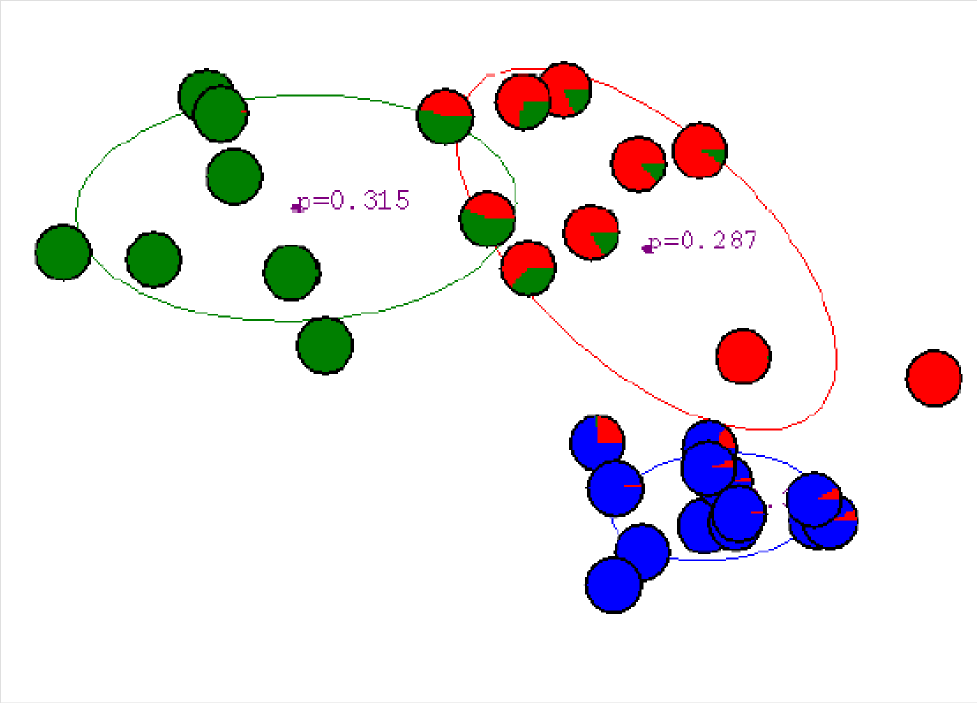
Iteration 5



A.W. Moore

Statistical Speech modeling : modeling cepstral patterns

Iteration 6

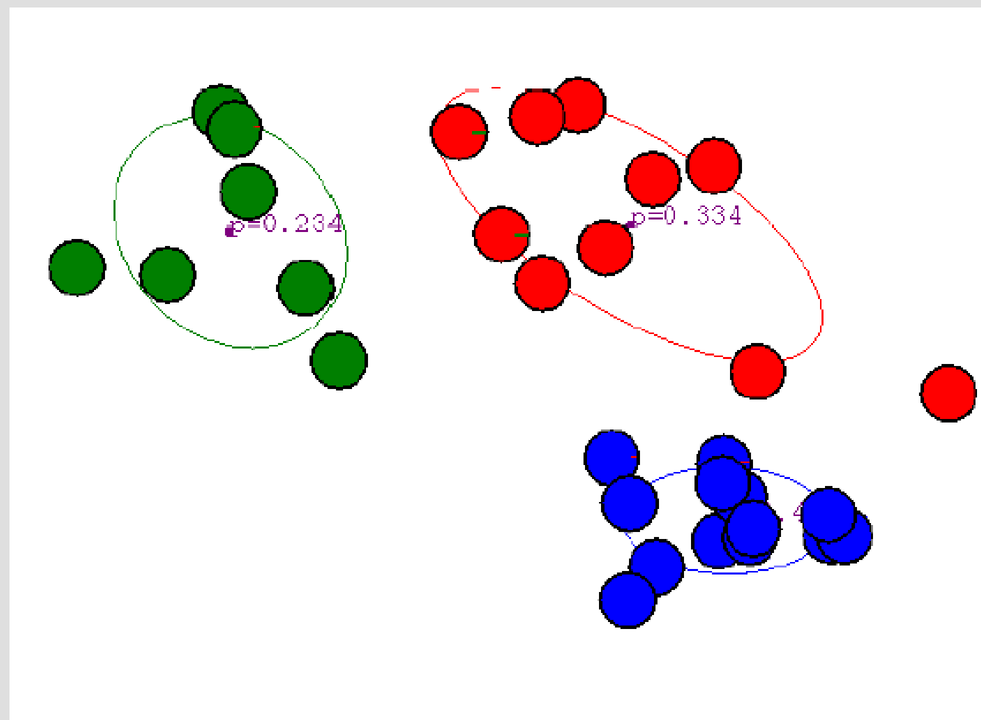


A.W. Moore

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling cepstral patterns

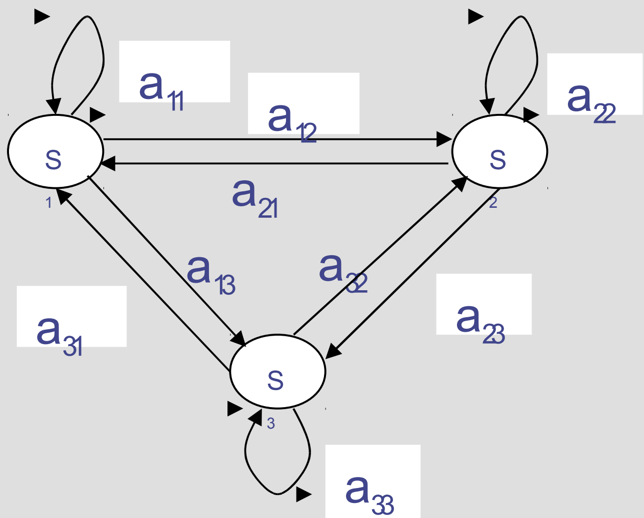
Iteration 20



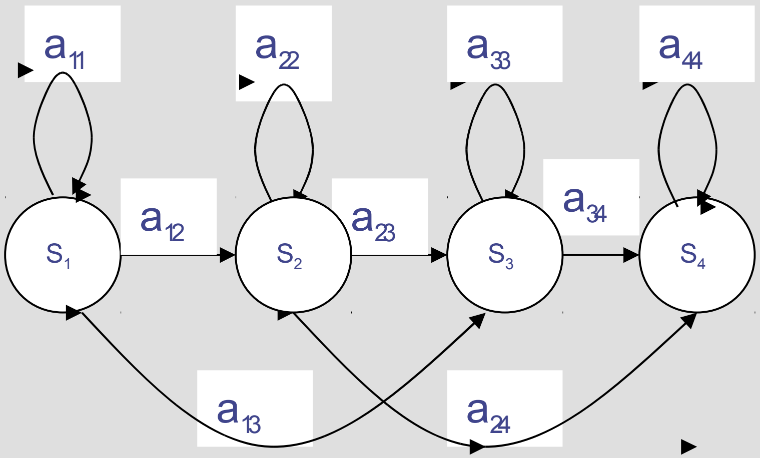
A.W. Moore

Statistical Speech modeling : modeling **temporal** patterns

Hidden Markov Models (HMM)



Ergodic HMM



Left-right HMM

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling **temporal** patterns

- Model topology:
  - State number
  - Links (transitions)
- Transition probabilities
- States :
  - density probability functions
    - GMM
    - Neural networks
    - others....

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Statistical Speech modeling : modeling **temporal** patterns

HMM is well defined tool to :

- estimate $P(O|\lambda)$, for an observation sequence O and a HMM λ,
- search of the state sequence Q maximazing $P(Q|O, \lambda)$ (decoding stage, Viterbi Algorithm)

•Training stage : find the optimal λ (Baum-Welsh/EM algorithms)

Speech Processing **systems**

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

- Automatic Speech Recognition :
  - Extracting the linguistic content of audio/audiovisual documents
  - Many applications
  - Research since 1970

ASR : some applications

Rich transcription

You are listening BBC news...

Large vocabulary
→Continuous speech
→Dealing with speaking styles
→Extracting meta-data
  →Speaker, topics, etc..

ASR : some applications

Voice command

Voice command :
- ➔small/medium vocabulary
- ➔isolated words
- ➔Embedded systems
  - Hardware constraints

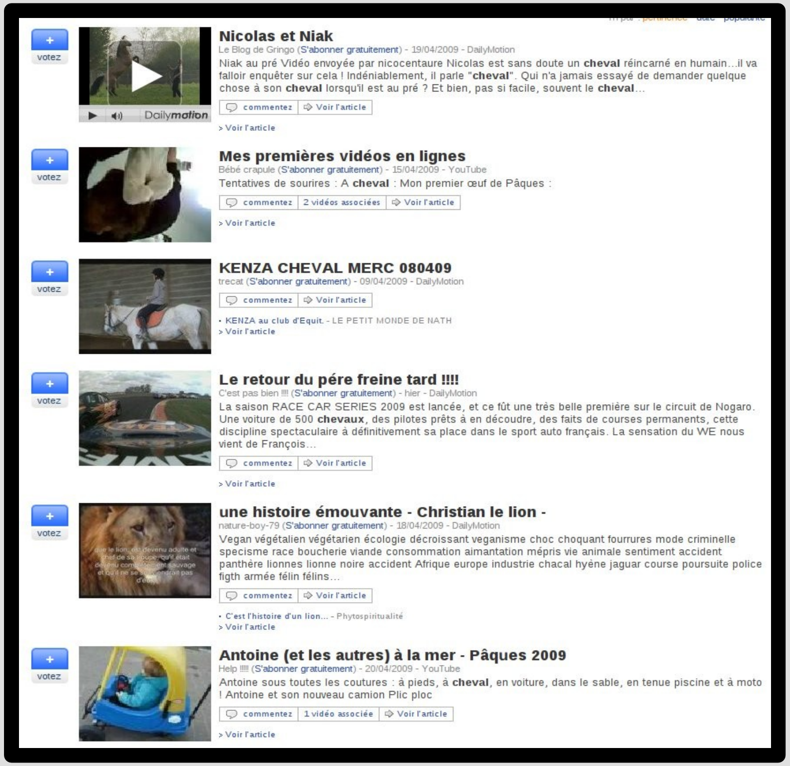
ASR : some applications

Voice command

→**ROBUSTNESS MAY BE CRITICAL!**

ASR : some applications

Audio search

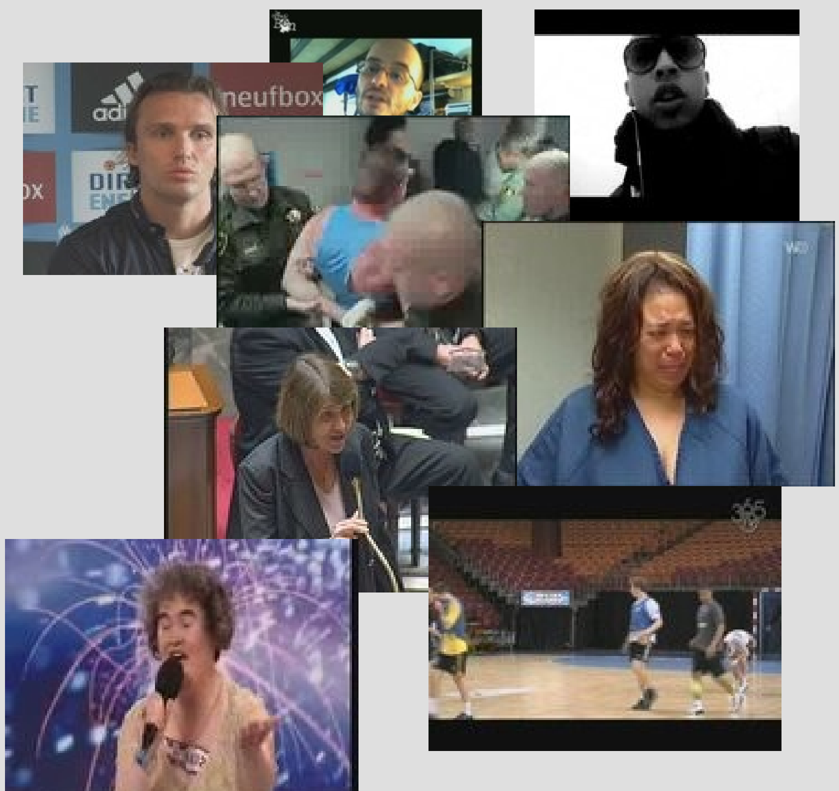


- Audio search
- →Spoken term detection
- →Topic detection
- →Entities search ?

ASR : some applications

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

Structuring audiovisual databases

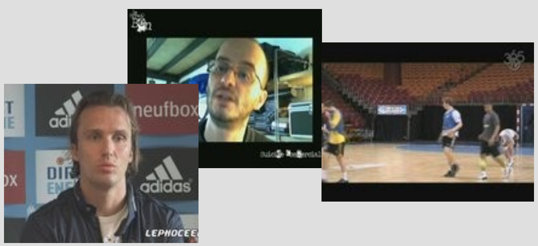


By-content structuring for efficient archiving and access
- ➔LVCSR
- ➔Extra-linguistic contents
- ➔Unexpected conditions

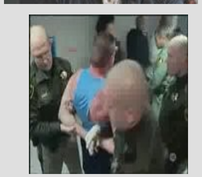
LABORATOIRE INFORMATIQUE D'AVIGNON

ASR : some applications

Example : video genre identification

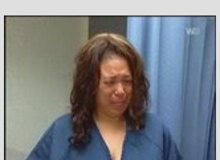


Sports



News

Musics



Commercials ?

Need of extra-linguistic features
- →Spontaneity
- →Interactivity
- →Emotional load ?

Genre

Topics

Environnement

Fundamentals of statistical ASR

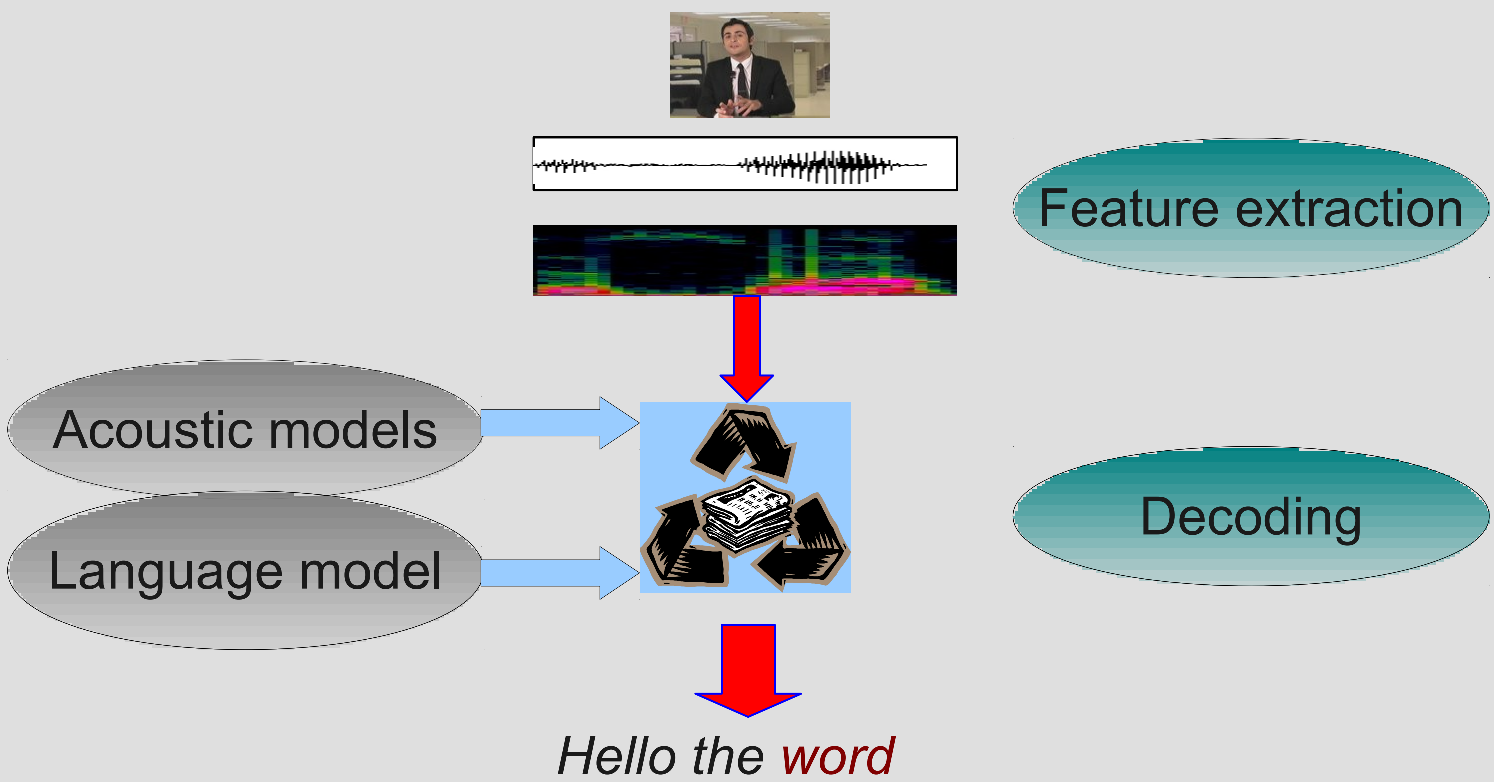
Search

$$\bar{W} = \underset{W}{argmax}\, P(W/X)$$

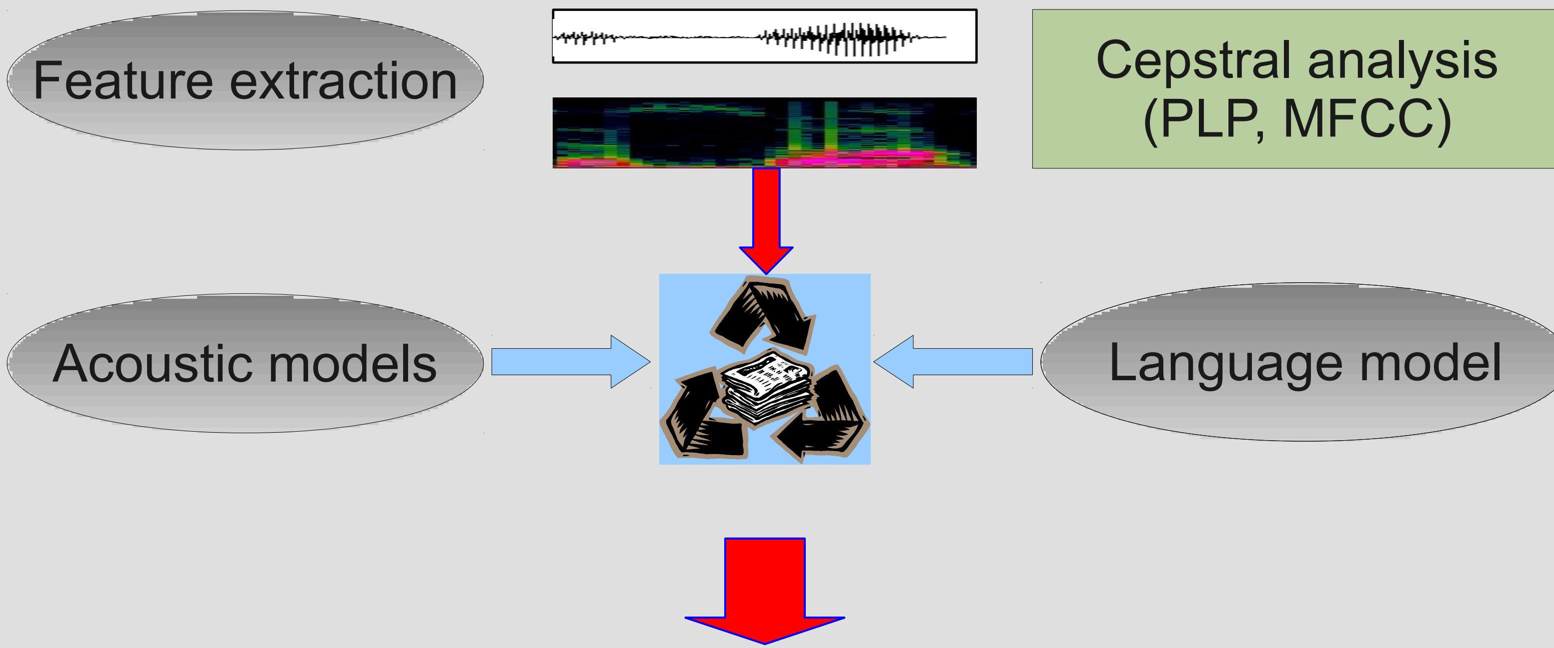
Linguisitic modelling

$$P(W/X) = \frac{P(X/W).P(W)}{P(X)}$$

Fundamentals of statistical ASR

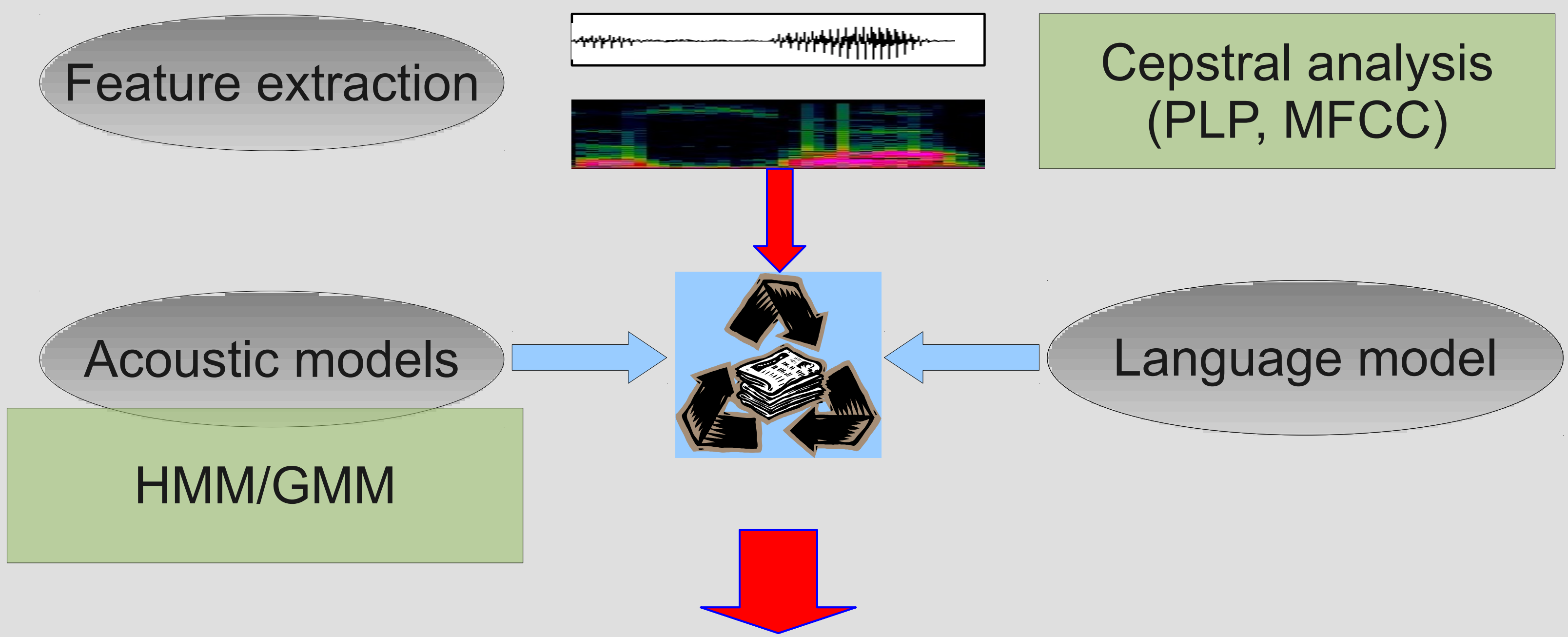


Fundamentals of statistical ASR



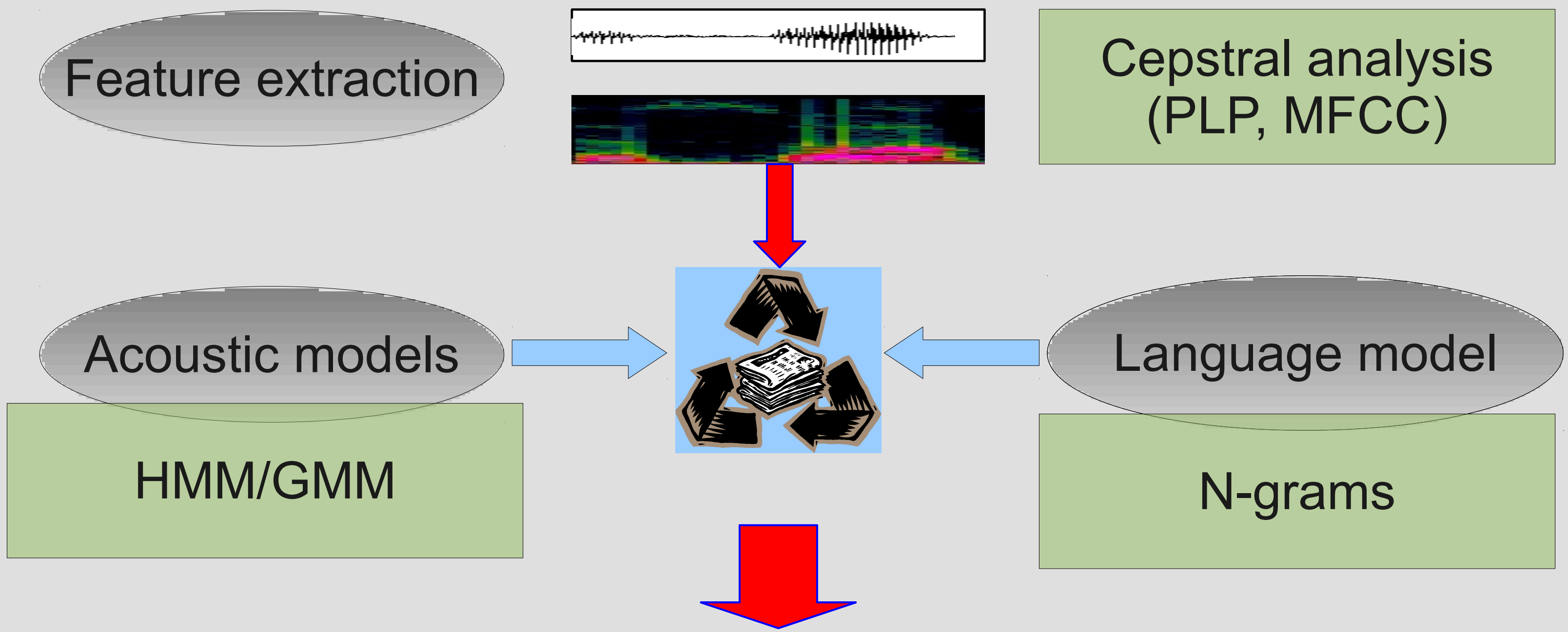
Hello the word

Fundamentals of statistical ASR



Hello the word

Fundamentals of statistical ASR



Hello the word

Fundamentals of statistical ASR

- Acoustic modeling :
  - Features in cepstral domain
  - Hiden Markov Models (HMM) for phoneme modeling
  - Gaussian Mixture Models (GMM)
  - Training on large annotated corpora
  - Generative models / discriminative learning : MLE+MMIE/MPE

Fundamentals of statistical ASR

- Main issues in acoustic modelling :
  - Models estimate
    - Cost of the training corpora
    - Tunning the training algorithms
    - Feature/models combination
  - Robustness
    - to acoustic conditions
    - to speaker, speaking styles (read/spontenous/conversational speech)

Fundamentals of statistical ASR

- Language models
  - N-grams statistics :

$$P(W)=P(W_i/W_{i-1},\dots,W\,i-n)\dots P(W_{i-k}/W_{i-2},\dots,W_{i-n-k})$$

- N=**3**,**4**,5
- Require very large corpora (several million words)
- Modelling the unseen events ?
  - interpolation or back-off to (n-1)grams

Fundamentals of statistical ASR

- Main issues in language models :
  - Exhaustive coverage of topics and speaking styles
  - Dealing with unseen events
    - OOV discovering and integration to LMs
  - Lack of semantics
    - Long-term dependencies
    - Semantic relationships

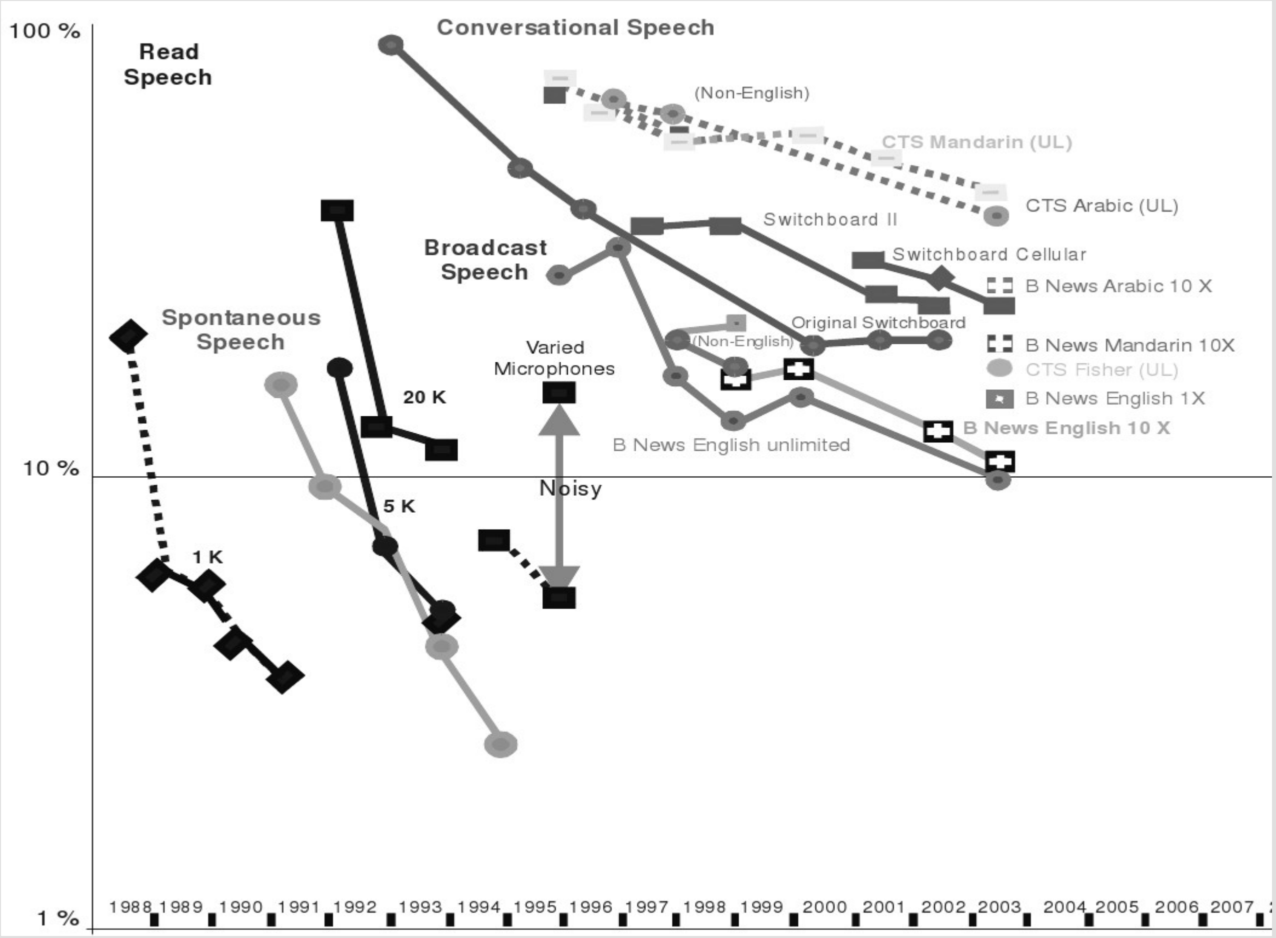
Fundamentals of statistical ASR

- Search algorithm
  - 2 main approaches :
    - Beam search (Viterbi)
    - Depth-first search (A*)
  - Issues :
    - Dealing with hardware constraints
      - Embedded systems/Large scale ASR
    - Fast decoding

Fundamentals of statistical ASR

- Recent advances in Search
  - System combination :
    - A postriori combination (ROVER)
    - Integrated approaches (Lecouteux & al, 2008)
    - Requires system complementarity, but similar accuracy (Bresdin & Gales, 2007)

Where we are in LVCSR ?



What about the cost ?

UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

Detail of Improvement	% WER
0. Baseline (RT-03 system)	13.4
1. 843-hour acoustic training	12.1
2. 1700-hour acoustic training	11.3
3. + MMI SAT PTM	11.2
4. + MMI SI PTM, SCTMs	11.0
5. + duration modeling	10.9
6. + online speaker clustering	10.8
7. + longer utterances	10.5
8. + new lexicon, LM	10.4

From n-Guyen & al, 2004

ASR: conclusion

- Is it a success story?
  - Tradeoff beetwen the expected gains and the costs is not so good
  - Are we on the limits of the HMM/N-gram framework ?
  - Is it the good/best paradigm ?
  - Can we do a better usage of ASR ?
  - What is the goal ?

Speech Analytics

- Analyzing speech to extract high level information :
  - Topics
  - Opinions
  - Roles...
  - speech understanding/interpretation.

Speech Analytics

- Methods :
  - 2 steps :
    - (1) automatic transcription
    - (2) ASR outputs processing
  - Critical points :
    - Feature extraction from rich transcription
    - Classification tasks
  - Tools : SVM, Neural Nets, Boosting,...
  - Application-oriented speech processing

UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

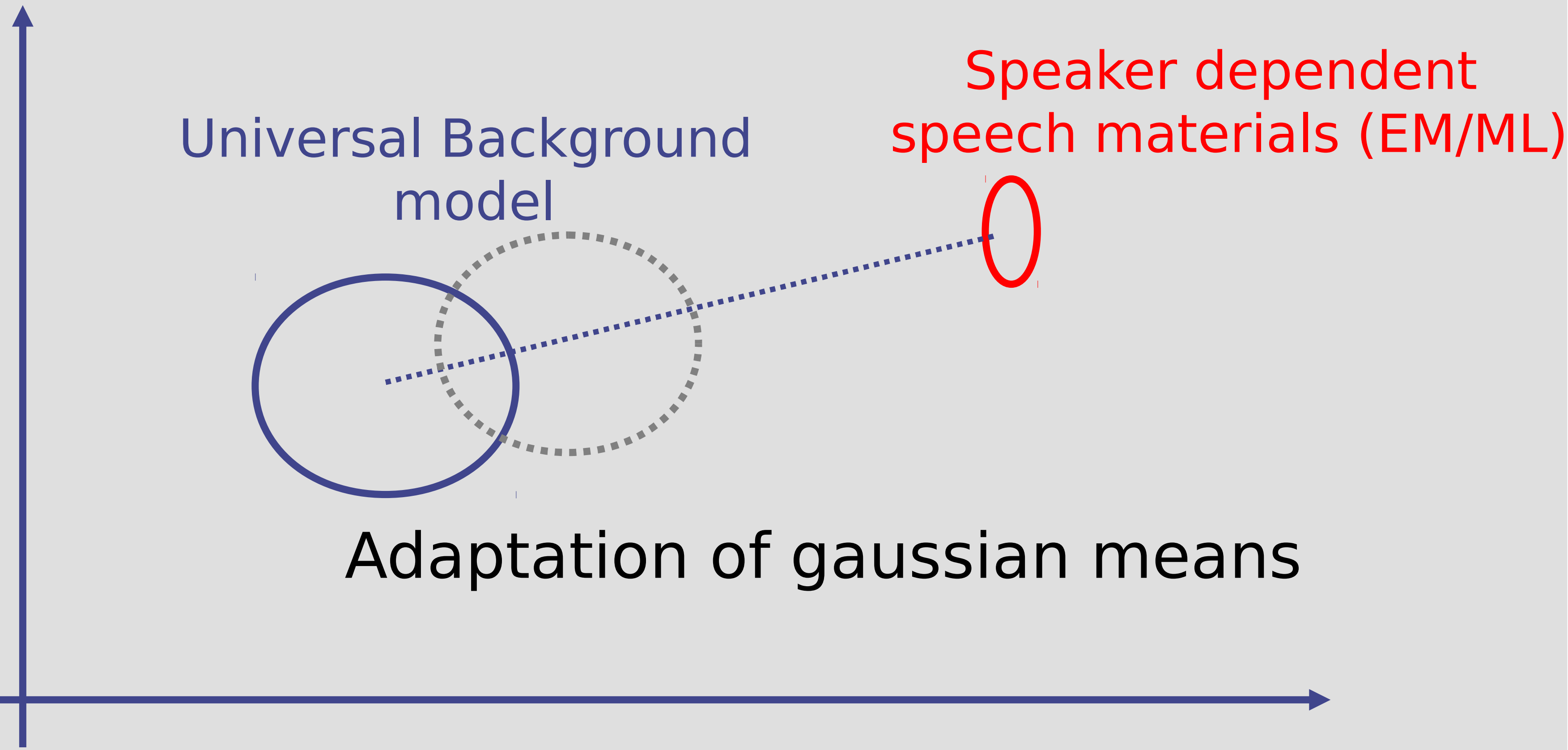
Speaker Recognition

- **Motivation :**
  - speech used to determine the true identity of the speaker
- **Speaker identification**
  - Who is speaking?
- **Speaker verification**
  - is the claimed identity true?
- **Speaker segmentation** : speaker turn detection, speaker tracking
- **Constraints :**
  - Open/close speaker set
  - Text dependent/independent

Speaker Recognition

- **General approach :**
  - Pattern recognition problem
  - Training phase : speaker-dependent models are enrolled on some speech samples
  - GMM-based approaches
  - Temporal structure of speech is not used

GMM based speaker identification



Adaptation of gaussian means

Performance of state-of-the art speaker identification systems

- NIST evaluation campaigns (speaker id)
  - American speakers, conversational speech
    - 5% error rate for the best system (2m30s)
- Identification on closed set
  - < 1% on studio data, 630 speakers (6s enrol., 3s test)
- **Error rates increase strongly on sponteneous speech and adverse conditions**
- New advances :
  - factor analysis for variability reduction
  - System combination

Statistical speech processing : conclusions

- Statistics is well defined framework to formulate speech processing problems
- ...and to build efficients systems
- Speech processing systems frequently rely on machine learning methods
- Research efforts mainly focused on the best way to apply mathematical tools to SP problems
- Many SP systems require other kind of information/other modeling paradigms to obtain significant improvments