Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences
Western Switzerland

MSE | MASTER OF SCIENCE IN ENGINEERING

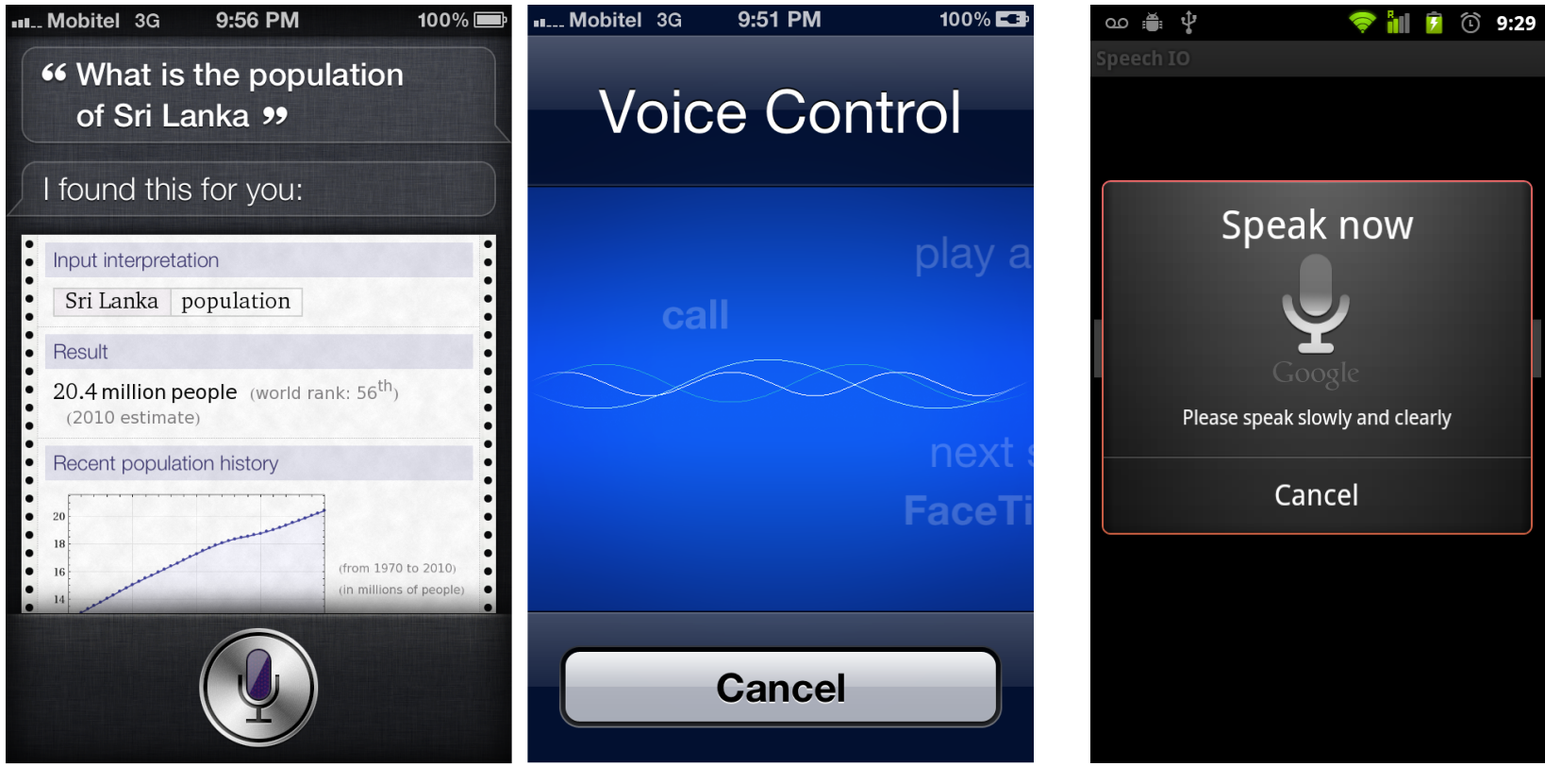
Multimodal Processing, Recognition and Interaction

Introduction

Elena Mugellini, Jean Hennebert, Stefano Carrino

Introduction aux Modèles de Markov Cachés (HMMs)

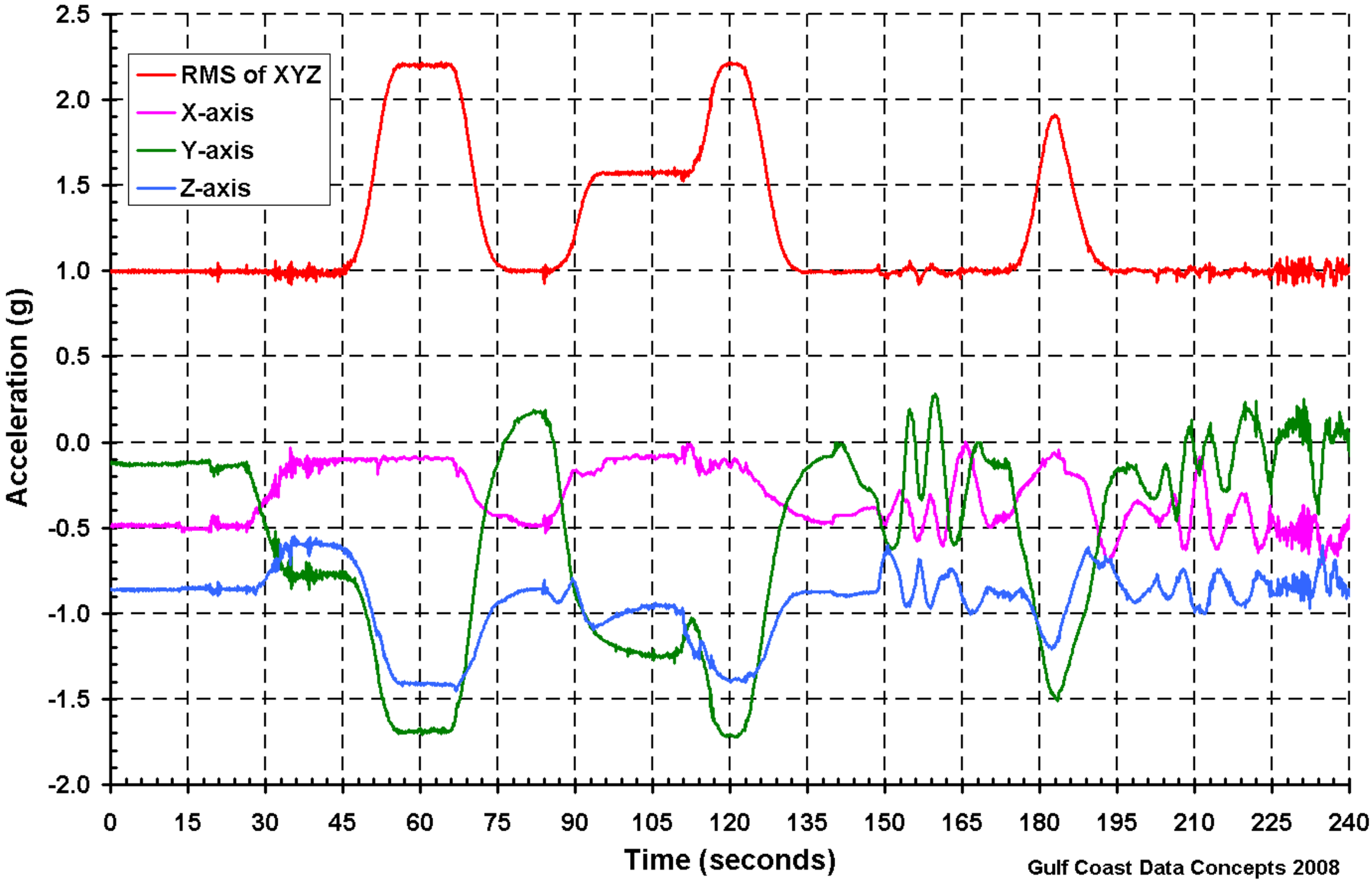
- Motivation:
  - Séries temporelles
  - Traitement de la parole



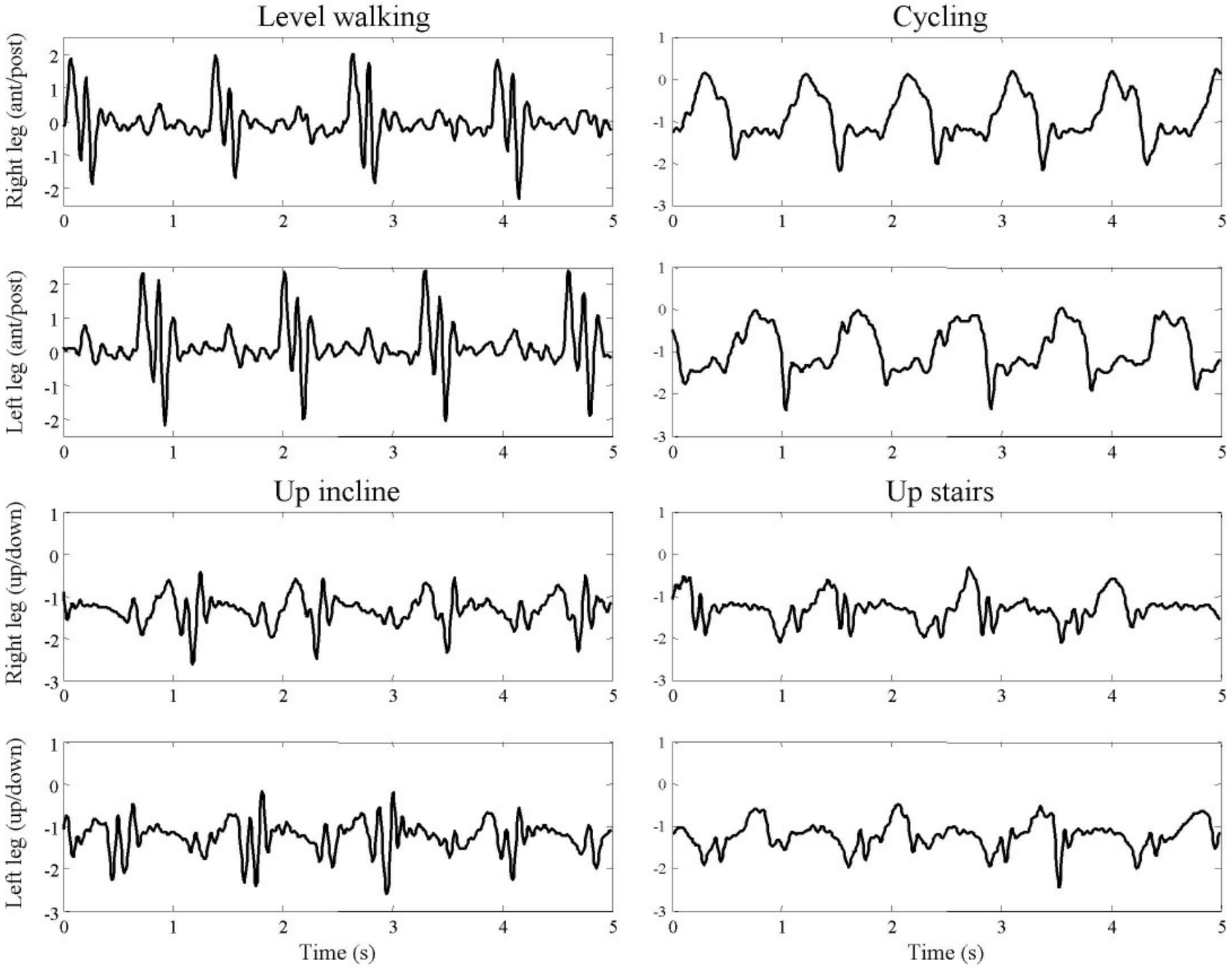
Siri **Voice Control**

Motivations et défis

SÉRIES TEMPORELLES

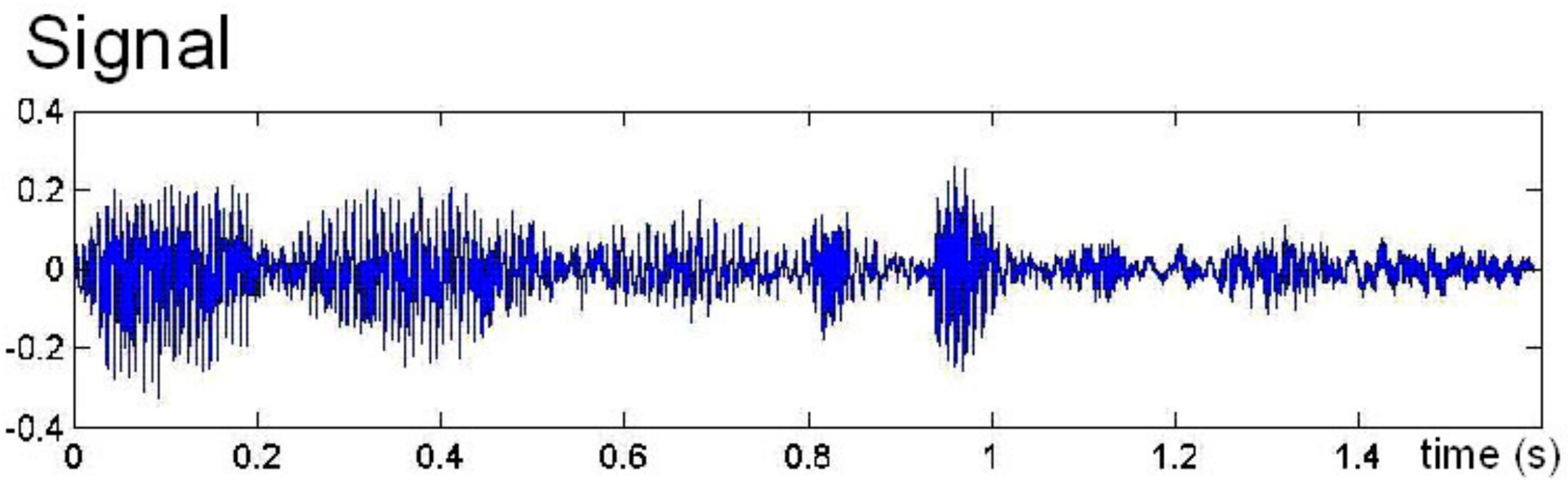


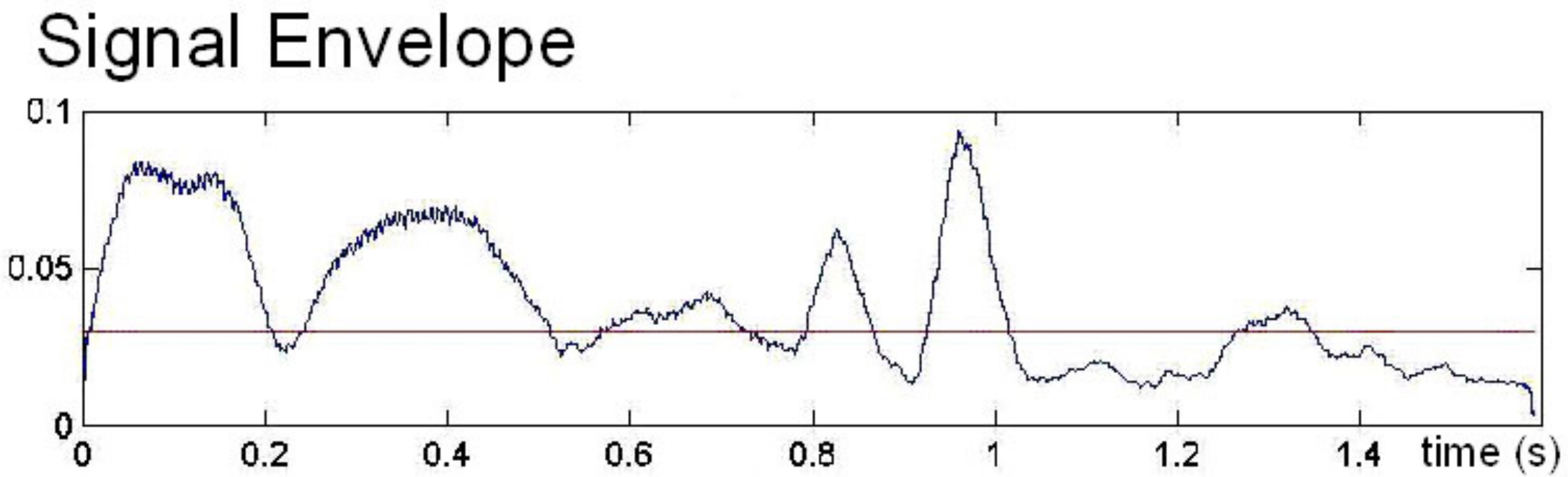
Source: http://www.gcdataconcepts.com/wdwxlr8r.html



Accelerometer data samples. Accelerometer signals are shown over a window of 5s corresponding to a few cycles of the following motor tasks: level walking, cycling, walking up an incline, and walking up stairs. Data are shown for the accelerometers positioned on left and right thigh with axes oriented in the antero-posterior and up and down directions.

Sherrill *et al. Journal of NeuroEngineering and Rehabilitation* 2005 **2**:16 doi: 10.1186/1743-0003-2-16


Signal
0.4
0.2
0
-0.2
-0.4
0
0.2
0.4
0.6
0.8
1
1.2
1.4
time (s)


Signal Envelope
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
time (s)

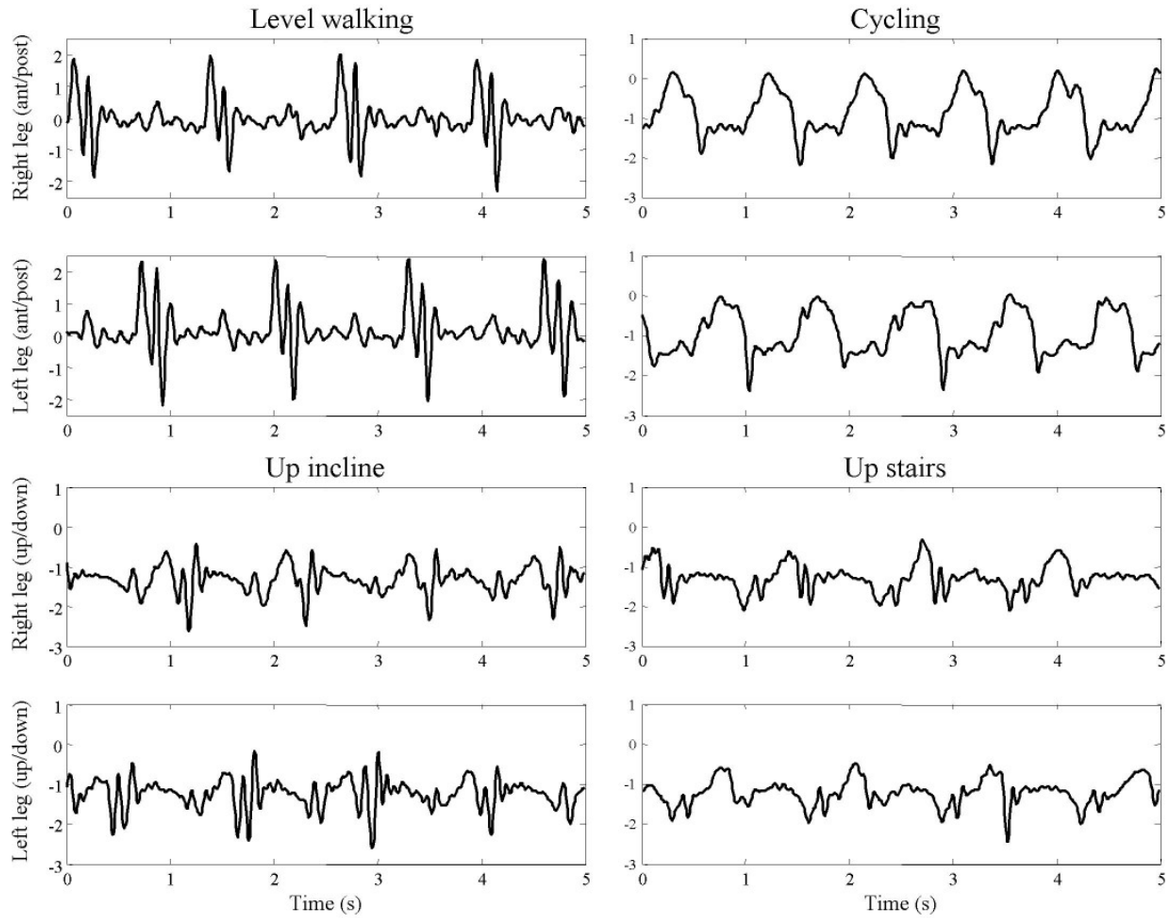
Séries temporelles - Définition

Séries temporelles *: la suite d'observations x_t (avec $t \in T$) d'une variable x à différents temps est appelée* ***série temporelle****. Habituellement, T est dénombrable, de sorte que t = 1,…, T.*

Séries temporelles - Motivation

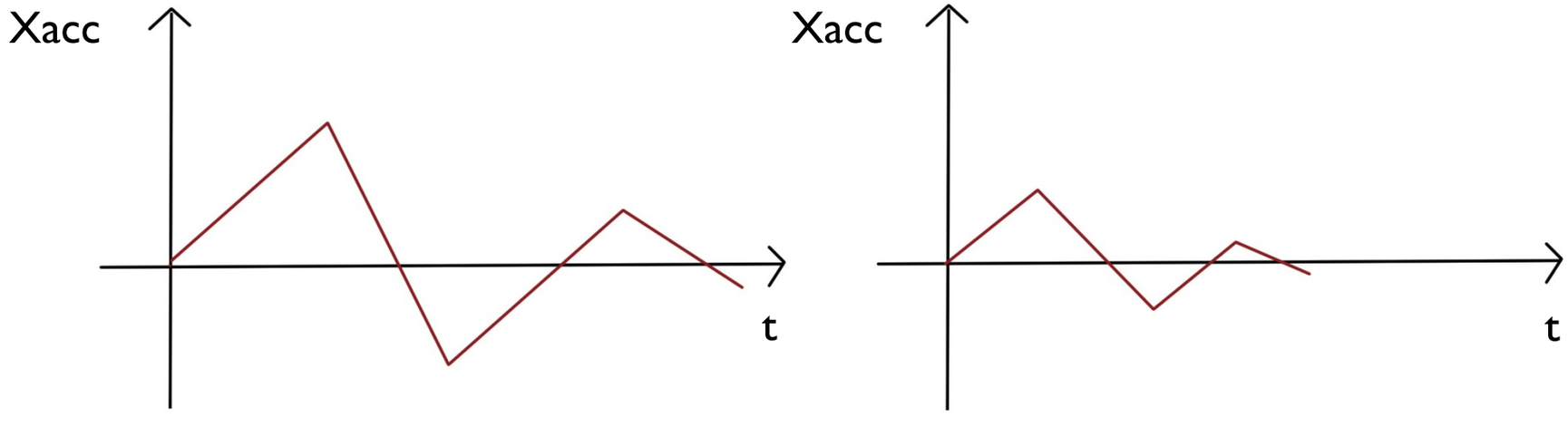
- Prévision

- Classification



Séries temporelles - Défis

- Pre-processing pour extraire les informations significatives
  - Feature extraction
- Durée variable pour un même type de input
  - Ex. gestes, mots, etc.



1s = 100 échantillons

0.7s = 70 échantillons

Séries temporelles - Défis

- Solutions
  - Approches holistiques
    - Caractéristiques génériques du signal: max, min, durée, etc.
  - Ré-échantillonnage
  - Utilisation d'algorithmes d'apprentissage automatique pour traiter directement avec les séries temporelles
    - HMM, CRF (conditional random fields), etc.

Motivations et défis

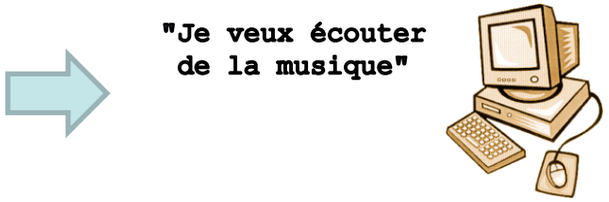
TRAITEMENT DE LA PAROLE

Voice-Based Interaction

- Pourquoi?
  - Parler est un moyen naturel de communiquer
  - Rapide
  - Si autres types d'interactions ne sont pas possibles
    - Free-hand interaction

Voice-Based Interaction

Speech recognition



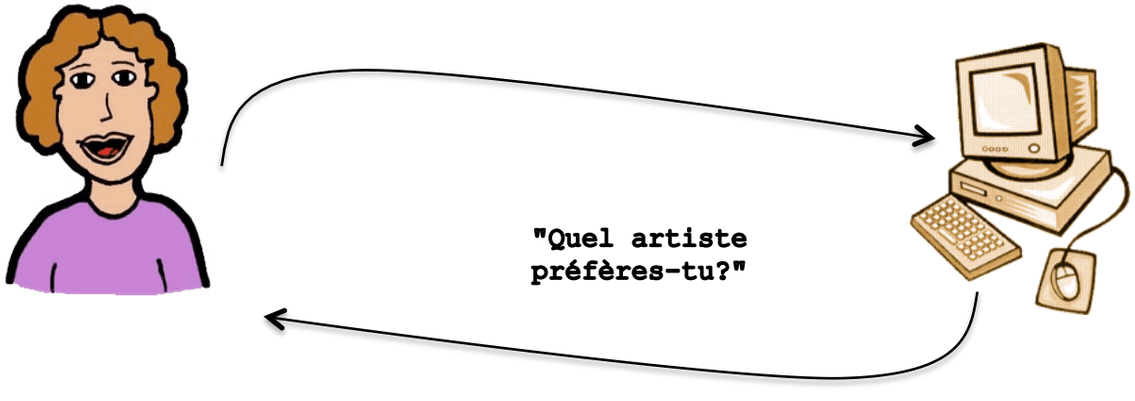
"Je veux écouter
de la musique"

Reconnaissance de la parole

Speech synthesis

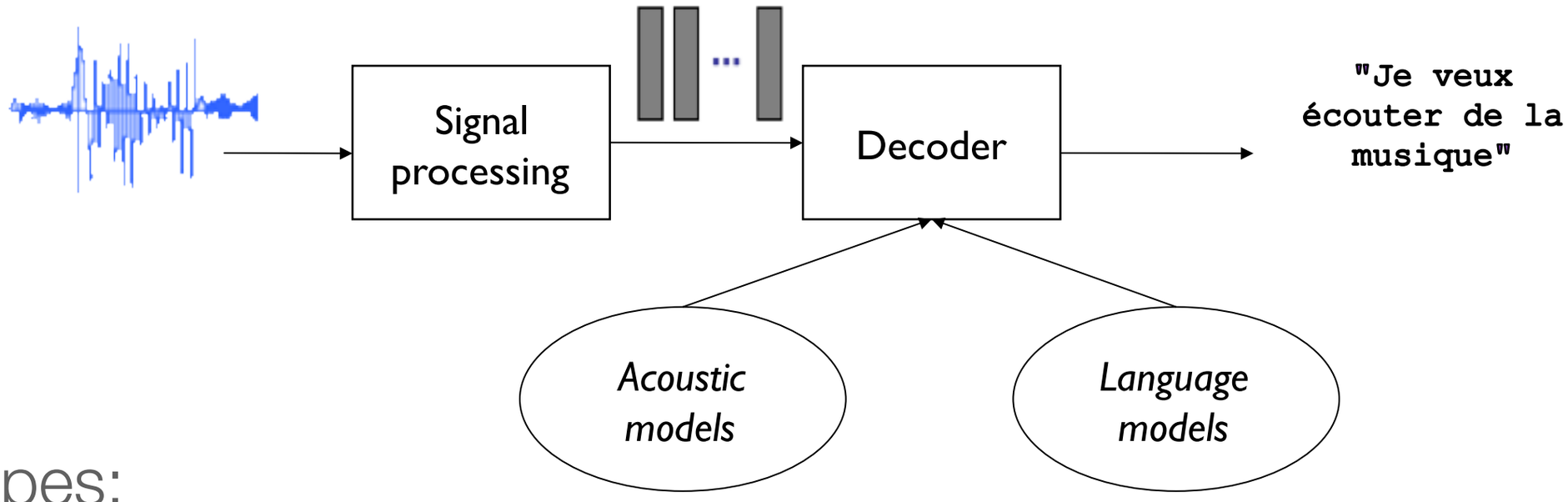
Synthèse vocale

"Je veux écouter
de la musique"



Dialogue

Reconnaisance de la parole



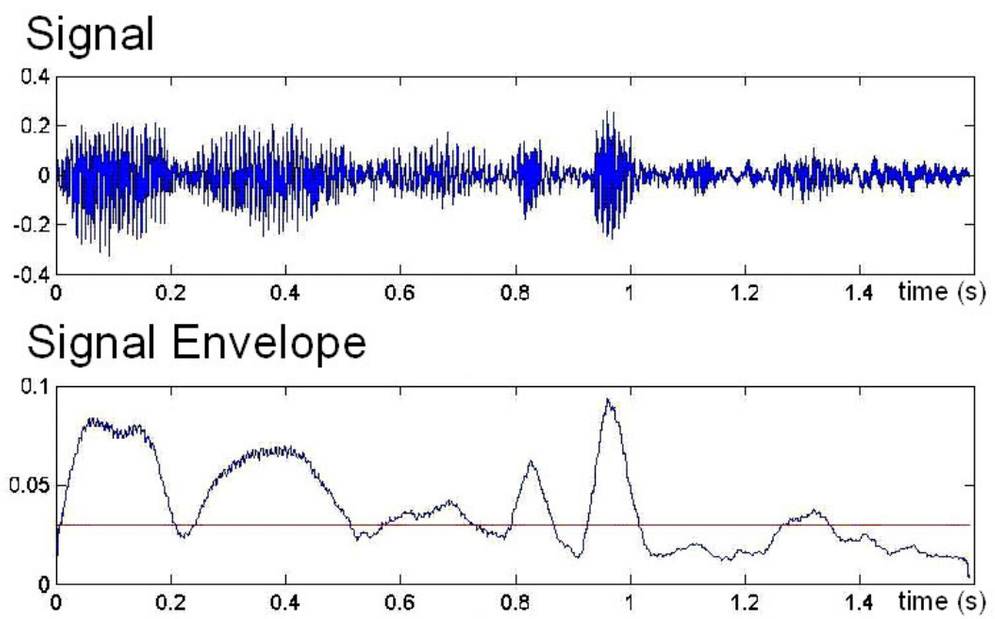
- Etapes:
  - Dans une application simple utilisant uniquement le modèle acoustique, l'application va traiter le mot prononcé en **phonèmes**, qui sont la partie fondamentale de la parole.
  - Ces phonèmes sont convertis en un format numérique.
  - Ce format numérique, ou pattern, est alors classifié utilisant une approche de machine learning

Reconnaisance de la parole - Paramètres

- Speaker-Dependent Vs Speaker-Independent
  - S.D.:
    - Avantage: meilleurs résultats
    - L'inconvénient majeur de ce système est qu'il est dédié à un seul utilisateur, et qu'il doit être entrainé avant son utilisation.
  - S.I.:
    - Avantage: il est indépendant du locuteur ☺
    - Les inconvénients majeurs avec un système indépendant du locuteur, en plus de la complexité accrue et la mise en œuvre difficile, sont son faible taux de précision globale et le temps de réponse plus lent.
  - Speaker Adaptive

Reconnaisance de la parole - Paramètres

- Keyword spotting Vs Continuous speech
  - Système de reconnaissance de mots isolés sont plus faciles à mettre en œuvre car le système connaît l'ampleur exacte de chaque mot
  - Pas besoin de segmentation



Reconnaisance de la parole - Paramètres

- Grammaire
  - La grammaire est utilisé non seulement pour définir les mots qui sont valables pour le système, mais aussi la syntaxe.
  - La grammaire, constitué par un ensemble de règles syntaxiques et sémantiques, est généralement spécifiée sur la base d'un ensemble de conditions.
- Vocabulaire (petit Vs grand)
  - Typiquement dépendant de la tâche.
  - Les petits vocabulaires sont plus faciles à reconnaître
    - La difficulté réelle de la reconnaissance est mieux indiquée par la complexité d'une tâche.
  - Tailles de vocabulaire: 10, 100, 1000, 10000 ou 64000 mots

Classification des systèmes

Signal Quality ↑

- **Application control**
  - Speaker (in)dependent
  - Runs on PDA, mobile phones
  - Web navigation, ...
  - A couple hundred words

- **Dictation System**
  - Speaker dependent
  - Runs on laptop, smart phones
  - Word processing, ...
  - Up to 50K words

- **Keyword recognition**
  - Word-print based
  - Runs on mobile phones
  - Voice dialing appl., ...
  - A few dozen words

- **Server-side ASR**
  - Speaker independent
  - Runs on big servers
  - Dialog based appl.
  - Up to 10K words

CPU →

Reconnaisance de la parole - Paramètres

Ex:

Dial three three two six five four

Dial nine zero four one oh nine

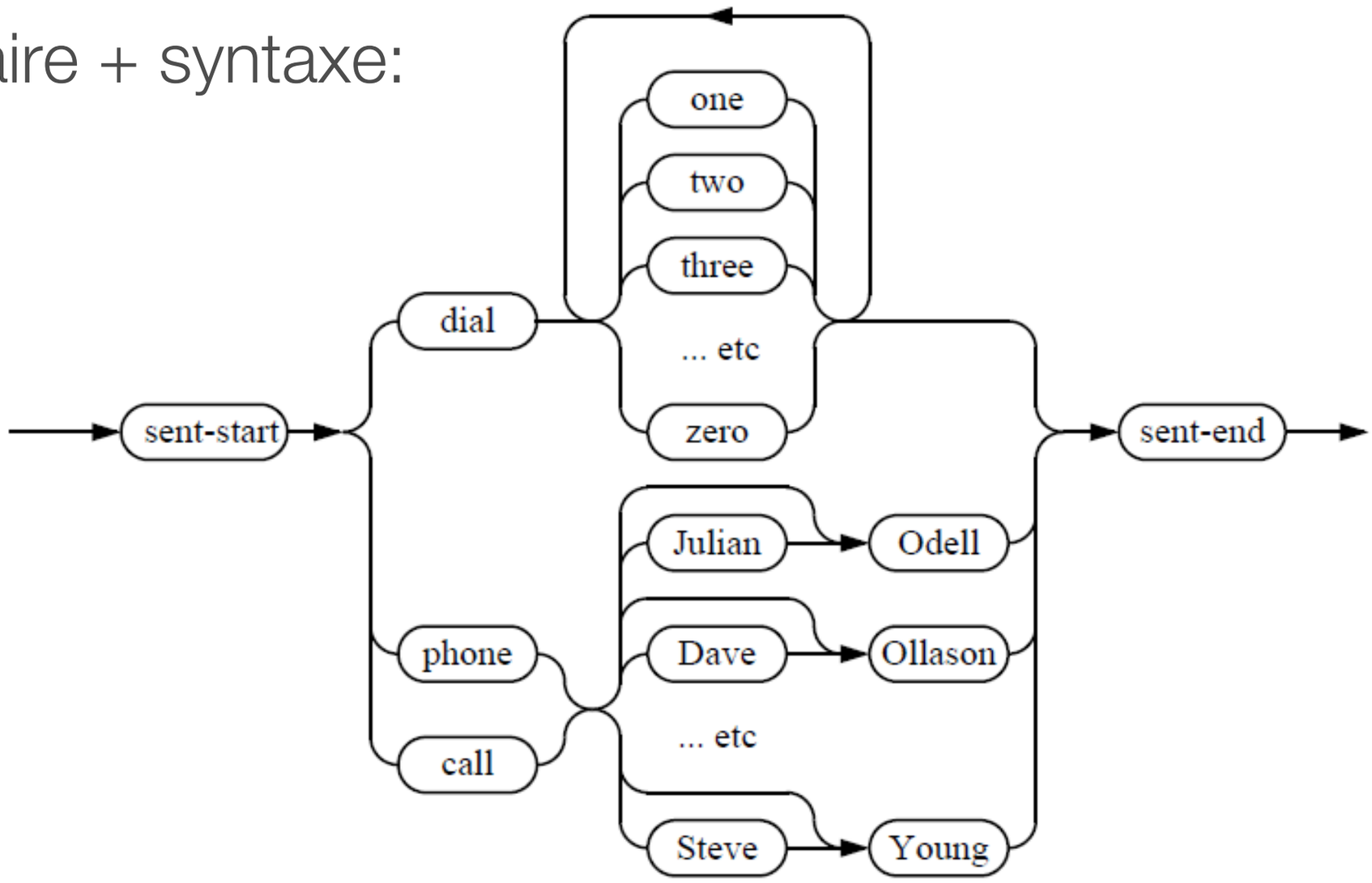
Phone Woodland

Call Steve Young

```
$digit = ONE | TWO | THREE | FOUR | FIVE |
         SIX | SEVEN | EIGHT | NINE | OH | ZERO;
```

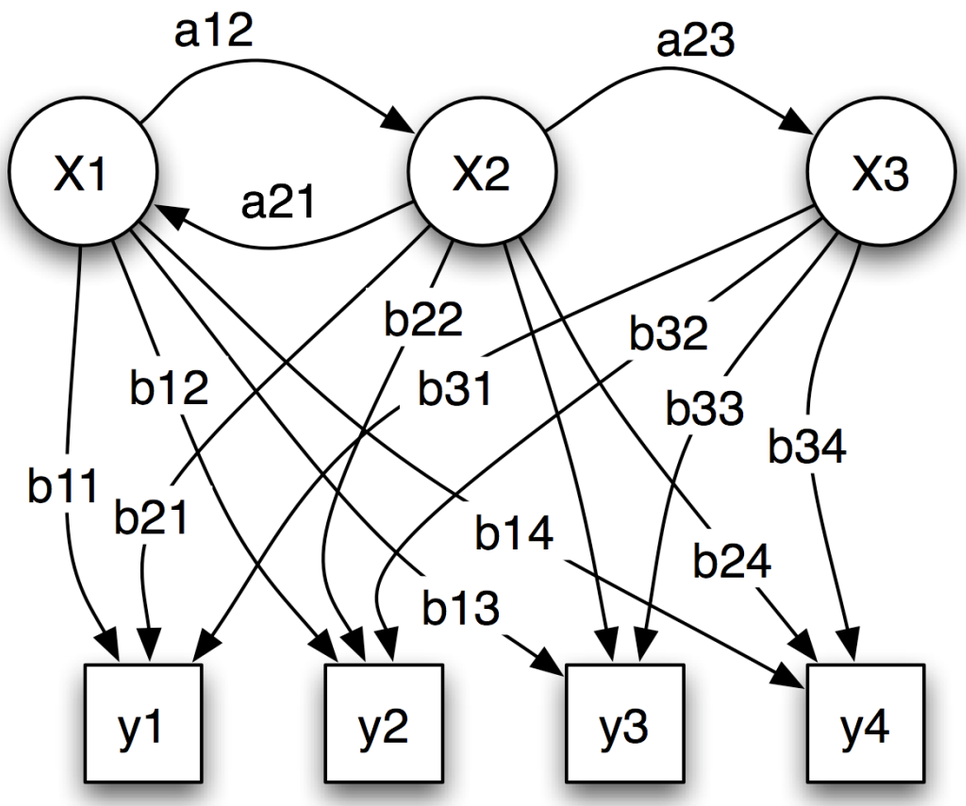
Reconnaisance de la parole - Paramètres

Grammaire + syntaxe:



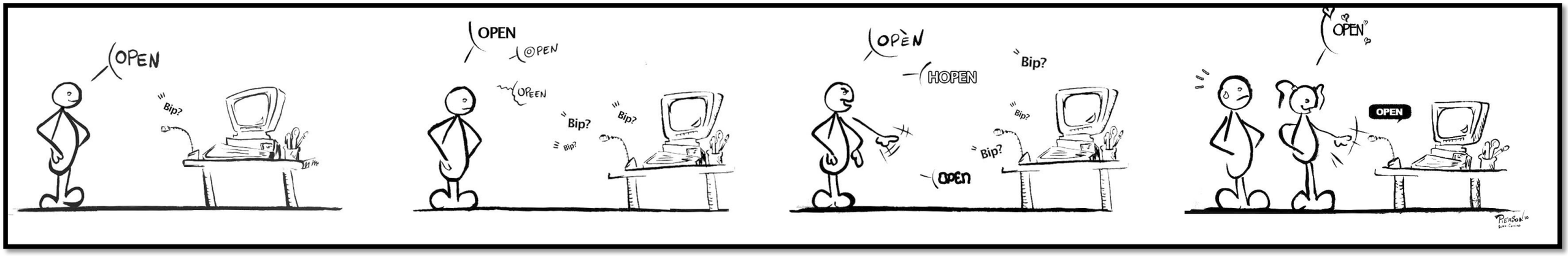
Algorithmes - HMMs

- Reconnaisance avec Modèles de Markov Cachés
- Avantages
  - User dependent ou independent
  - Pas nécessaire re-entrainer tout le système s'il y a des changements partiels
- Inconvénients
  - Dépendent du langage
  - Peut être lent (training)



Traitement de la parole: défis et inconvénients

- Feedback
  - No feedback
  - Yes/No feedback
  - Difficile fournir un feedback continu dans le temps



Traitement de la parole: défis et inconvénients

- Bruit
  - Très sensible au bruit dans l'environnement
  - Génère du bruit
    - Interaction vocal dans un lieu de travail?
- Interactions privées

What you should know

- Séries temporelles
  - Pourquoi sont-elles intéressantes?
- Traitement de la parole
  - Motivation
  - Importance des paramètres
  - Défis