Social Network Analysis (and More) in Multimedia Indexing: Making Sense of People in Multiparty Recordings

Alessandro Vinciarelli
IDIAP Research Institute - CP592 Martigny (Switzerland)
e-mail: vincia@idiap.ch

Outline

- Part I - Introduction

Outline

- Part I - Introduction
  - Making sense of people?

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.
  - The story segmentation problem.

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.
  - The story segmentation problem.
- Part III - What's Next?

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.
  - The story segmentation problem.
- Part III - What's Next?
  - Towards Social Signal Processing?

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.
  - The story segmentation problem.
- Part III - What's Next?
  - Towards Social Signal Processing?
  - The social status recognition problem.

Outline

- Part I - Introduction
  - Making sense of people?
  - A one-slide introduction to Social Network Analysis.
  - From SNA to Multimedia Indexing.
- Part II - Applications
  - The role recognition problem.
  - The story segmentation problem.
- Part III - What's Next?
  - Towards Social Signal Processing?
  - The social status recognition problem.
- Conclusions.

Part I
Introduction

Making Sense of People (I)

One of our most common activities is to make sense of people, i.e. to understand, predict and recall the behavior of persons we know little or even nothing about.

Making Sense of People (II)

The domain studying the way we make sense of people is called Social Cognition and relies on two major assumptions:

- Social Cognition is a form of categorical thinking, i.e. we tend to class others into predefined categories or stereotypes.
- Social Cognition is thinking about relationships, i.e. we make sense of people through the relationships they have with others.

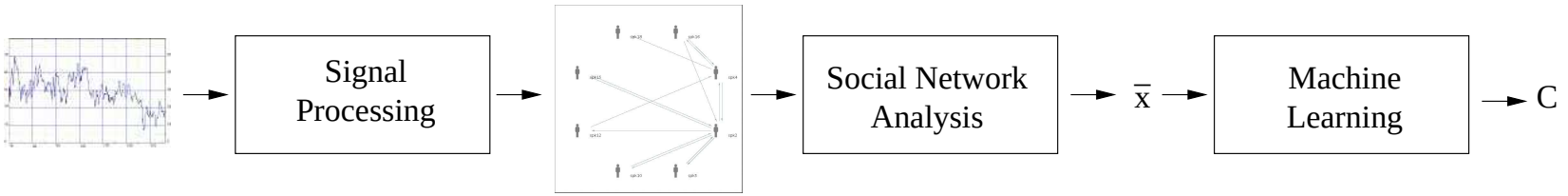
Technology has learnt from neurology (neural networks), genetics (genetic algorithms), physiology (speech processing), etc. Why not to learn from Social Cognition?

What is Social Network Analysis?

In very simple terms, Social Networks are graphs where each node corresponds to an individual and each link corresponds to a relationship. Social Network Analysis (SNA) is a corpus of mathematical techniques, mostly based on graph theory, that extract quantitative measures about social relationships:

- how much a person is central.
- how close two or more individuals are to each other.
- how many social groups are present.
- who belongs to which social group.
- etc.

From SNA to Multimedia Indexing



If a Social Network is extracted from the signal, then each individual can be represented with a vector of social features. The vector can be mapped into socially relevant high level information. This requires two main operations:

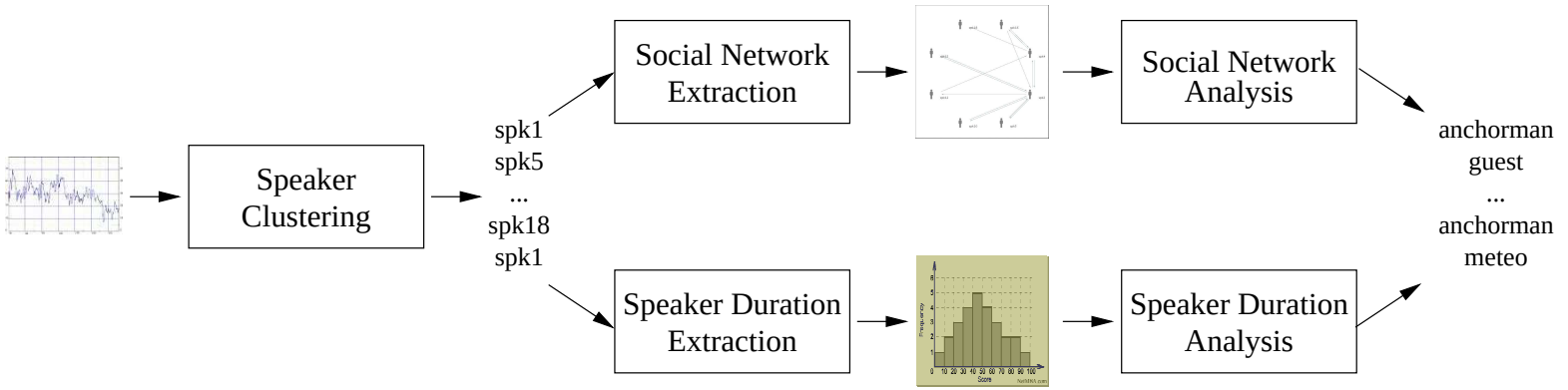
- Automatic extraction of Social Networks from data.
- Machine Learning techniques for people classification.

Part II.1
Role Recognition

The Role Recognition Problem

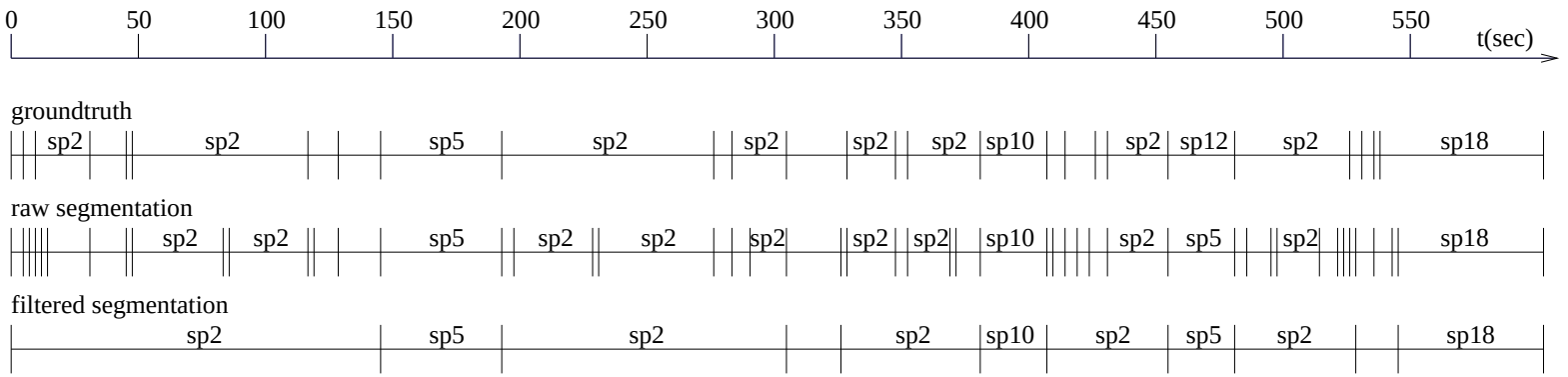
- The role recognition problem consists in assigning automatically each individual a role r belonging to a predefined set $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$.
- The experiments have been performed over corpora of radio programs and the roles are:
  - Anchorman (AM)
  - Second Anchorman (SA)
  - Guest (GT)
  - Interview Participant (IP)
  - Abstract (AB)
  - Meteo (MT)

A Role Recognition Approach



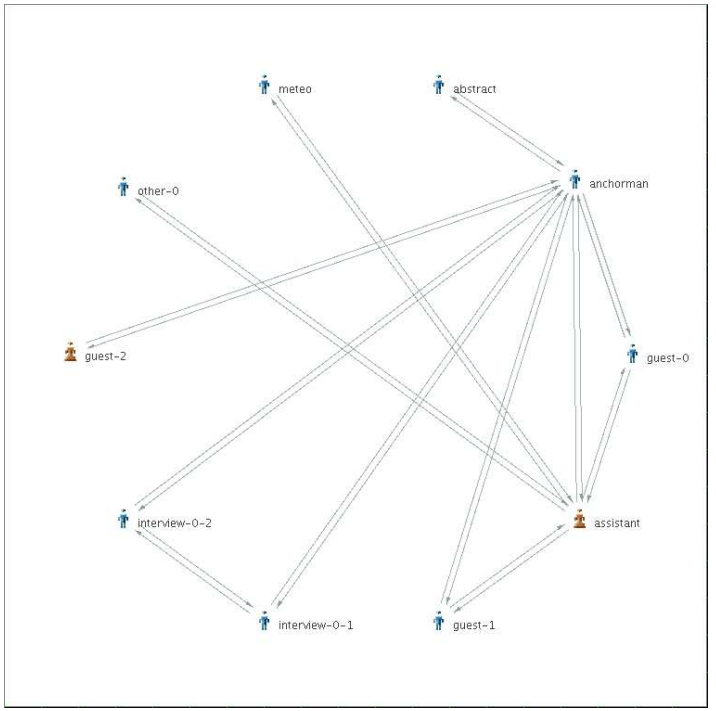
- The first step of the process is the application of an unsupervised speaker clustering approach.
- The segmentation resulting from the first step is used to extract information about:
  - the pattern of social relationships
  - the duration distribution of different speakers
- The two information sources are then combined into a single classification approach.

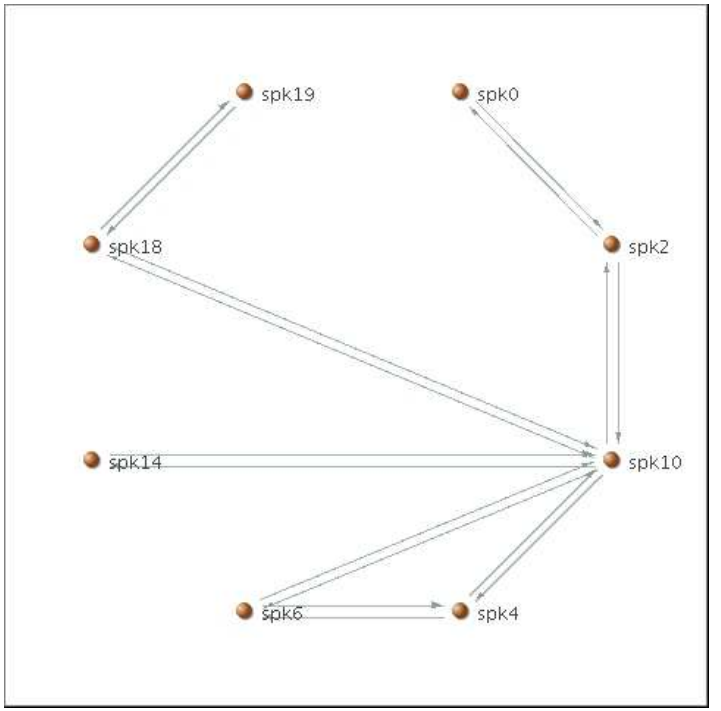
Social Network Extraction (I)



Speaker Clustering techniques enable one to split multiparty audio recordings into single speaker segments. The network can be extracted by connecting adjacent speakers.

Social Network Extraction (II)





The speaker clustering is not a perfect process, thus the resulting network is noisy, i.e. it involves spurious individuals and spurious relationships.

Statistical Foundations (I)

The role recognition problem can be thought of as finding the vector $\vec{r}^*$:

$$\vec{r}^* = \arg\max_{\vec{r} \in \mathcal{R}^G} p(\vec{r}|\mathcal{Y}) \tag{1}$$

where

- $\mathcal{R}$ is the set of predefined roles
- G is the number of speakers a_i.
- $\vec{r} = (r_1, \ldots, r_G)$ is the vector of the speaker roles
- $\mathcal{Y} = \{\vec{y}_1, \ldots, \vec{y}_G\}$ is the set of the vectors representing the speakers
- $\vec{y}_i = (\tau_i, \vec{x}_i)$, where τ_i is the percentage of times for which speaker a_i talks.

Statistical Foundations (II)

By applying the Bayes Theorem and by taking into account that $\mathcal{Y}$ is constant, the problem can be formulated equivalently:

$$\vec{r}^* = \arg \max_{\vec{r} \in \mathcal{R}^G} p(\mathcal{Y}|\vec{r}) p(\vec{r}) \tag{2}$$

We assume that the roles of the different speakers are statistically independent:

$$\vec{r}^* = \arg \max_{\vec{r} \in \mathcal{R}^G} \prod_{i=1}^{G} p(\vec{y}_i|r_i) p(r_i) \tag{3}$$

We further assume that τ_i and $\vec{x}_i$ are statistically independent:

$$\vec{r}^* = \arg \max_{\vec{r} \in \mathcal{R}^G} \prod_{i=1}^{G} p(\tau_i|r_i) p(\vec{x}_i|r_i) p(r_i) \tag{4}$$

The Data

The experimente have been performed over two corpora of radio programs. The first (called C1) contains 96 news bulletins for a total of 19 hours and 56 minutes of material, the second (called C2) contains 26 talk-shows for a total of 26 hours of material.

Corpus	AM	SA	GT	IP	AB	MT
C1	41.2%	5.5%	34.8%	4.0%	7.1%	6.3%
C2	17.3%	10.3%	64.9%	0.0%	4.0%	1.7%

The table reports the percentage of data each role accounts for.

Results

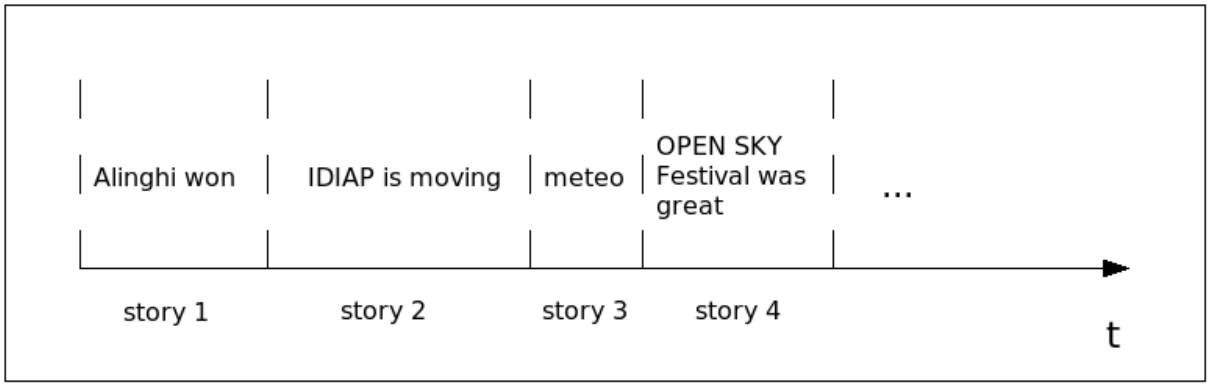
The results are reported in terms of accuracy, i.e. percentage of time correctly labeled in terms of role.

Corpus	all	AM	SA	GT	IP	AB	MT
C1	81.1	94.9	1.0	95.8	0.0	58.9	73.4
C2	81.3	70.2	88.3	89.8	18.3	29.7	5.0

The experiments are performed over the whole corpus using a leave one out approach.

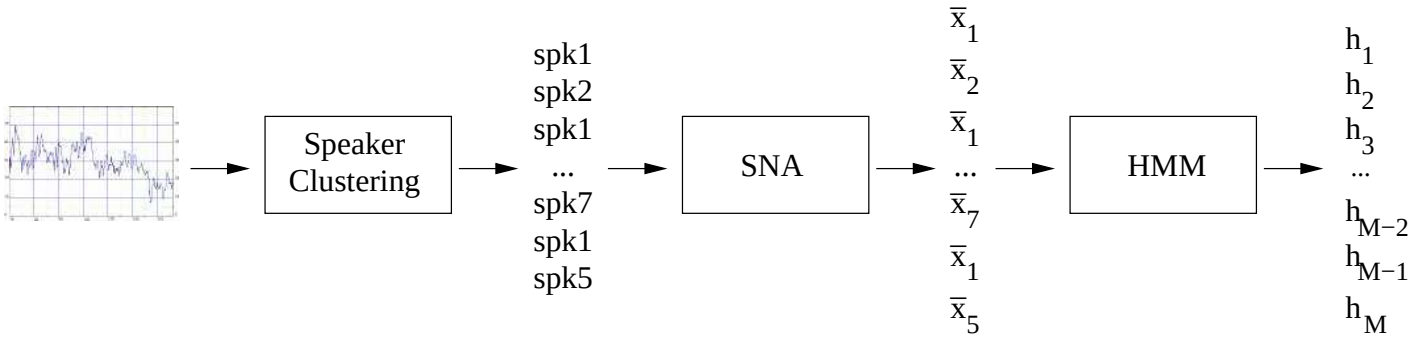
Part II.2
Story Segmentation

The Story Segmentation Problem



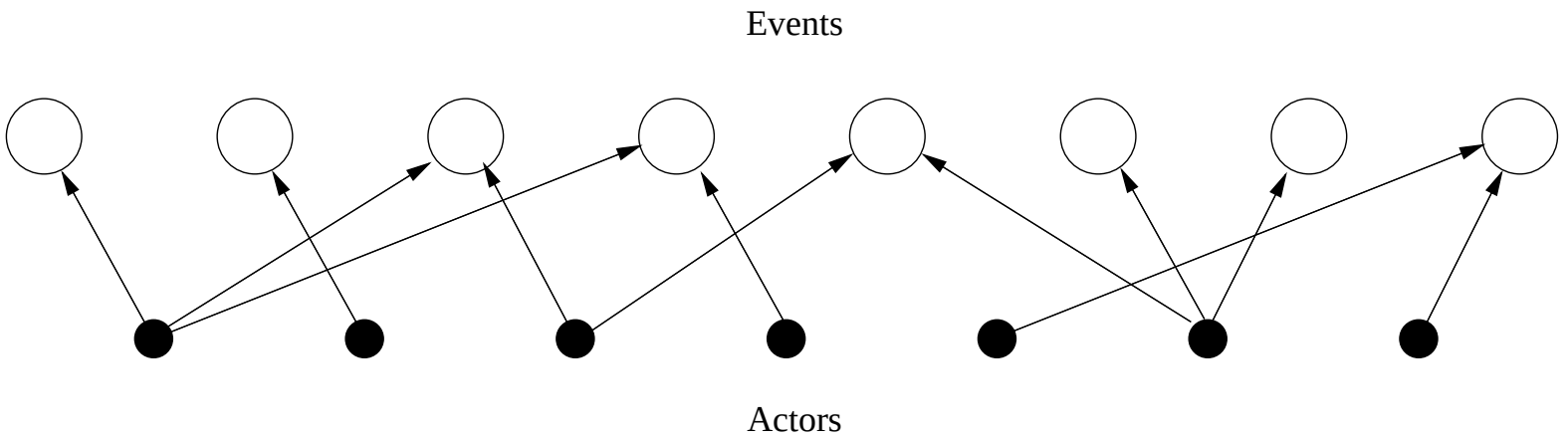
- The identification of semantically coherent segments makes the acces to the content easier.
- In the case of broadcast news, the segmentation is performed in terms of stories.

The Story Segmentation Approach



- The main idea of the approach is that people involved in the same story are more likely to interact with each other, thus stories are expected to correspond to social groups.
- SNA is used to extract feature vectors accounting for the social groups and HMMs are used to map the vector sequences into story sequences.

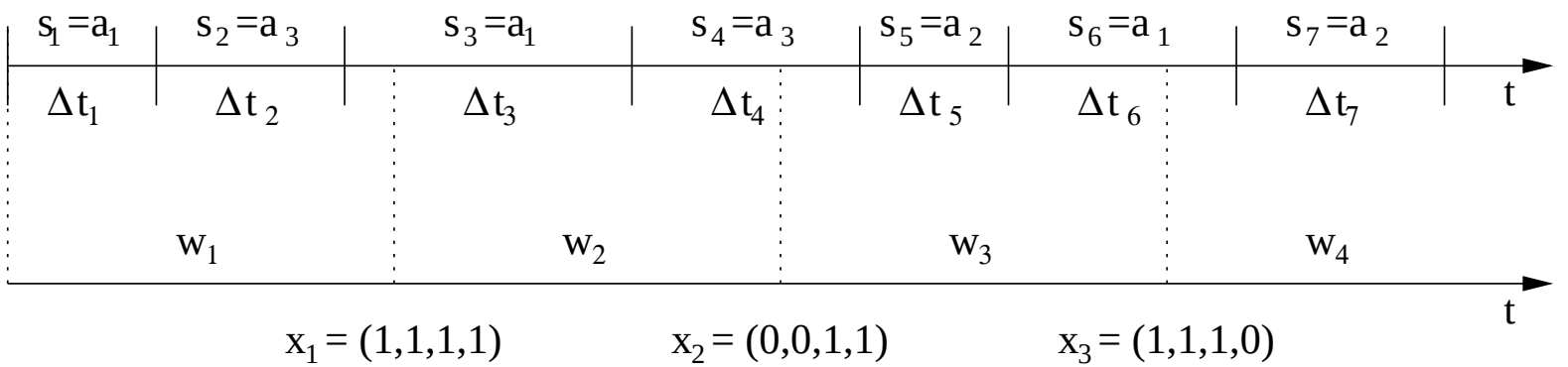
Affiliation Network Extraction (I)



An Affiliation Network is a bipartite graph, i.e. with two kinds of nodes: actors and events. Links are allowed only between nodes of different kind. There are two major approaches to define the events:

- Gatherings: meetings, parties, etc.
- Proximity in time and/or space.

Affiliation Network Extraction (II)



In the case of broadcast news, the events are defined using the proximity in time. Each speaker a_i is represented by a vector $\vec{y}_i = (y_{i1}, \ldots, y_{iN})$, where N is the number of events and $y_{ij} = Z$ when a_i talks during event e_j (and 0 otherwise).
The dimension of $\vec{y}$ is reduced using the PCA and the resulting vectors are $\vec{x}_i$.

Statistical Foundations (I)

- The goal of the story segmentation is to assign each vector $\vec{x}_i$ a label h_i which can be either the number of a story or the *anchormen* role.
- The story segmentation problem can be thought of as finding the sequence $H^* = (h_1, ..., h_M)$ which maximizes the following *a-posteriori* probability:

$$H^* = \arg\max_{H \in \mathcal{H}} p(H|X)p(H) \tag{5}$$

where

  - $\mathcal{H}$ is the set of all possible H sequences.
  - M is the number of single speaker segments detected at the speaker clustering step.

Statistical Foundations II

- The term p(H|X) is estimated using a fully connected Hidden Markov Model (HMM) with S+1 states, where S is the maximum number of stories that can be observed and the "+1" state is for the anchormen role.
- The term p(H) is estimated using a tri-gram statistical language model:

$$p(H) = \prod_{k=3}^{M} p(h_k | h_{k-1}, h_{k-2}) \tag{6}$$

The Story Segmentation Results

The table reports the purity as a function of the number of windows and the amount of variance retained.

	variance fraction			
win	70%	80%	90%	100%
10	0.74	0.76	0.76	0.78
14	0.74	0.76	0.76	0.77
20	0.75	0.77	0.78	0.79

- the purity is always around 0.75.
- the average number of stories detected by the system is 16.5.

Part III
What's Next?

Towards Social Signal Processing?

Human-human communication involves Social Signals, an array of nonverbal behaviors, mostly unconscious, which convey socially relevant information.

- Vocal Social Signals are often unconscious and cannot be easily controlled: the information they carry is thus reliable, language independent and, to some extent, culture independent.
- Vocal Social Signals can be analyzed through robust and established Signal Processing techniques.

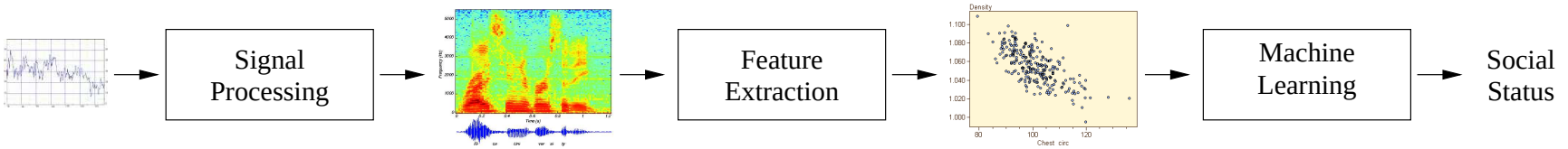
This is the basis of a potential new domain: Social Signal Processing.

The Social Status Recognition Problem

Social psychology tells that how we say things is as important as what we say:

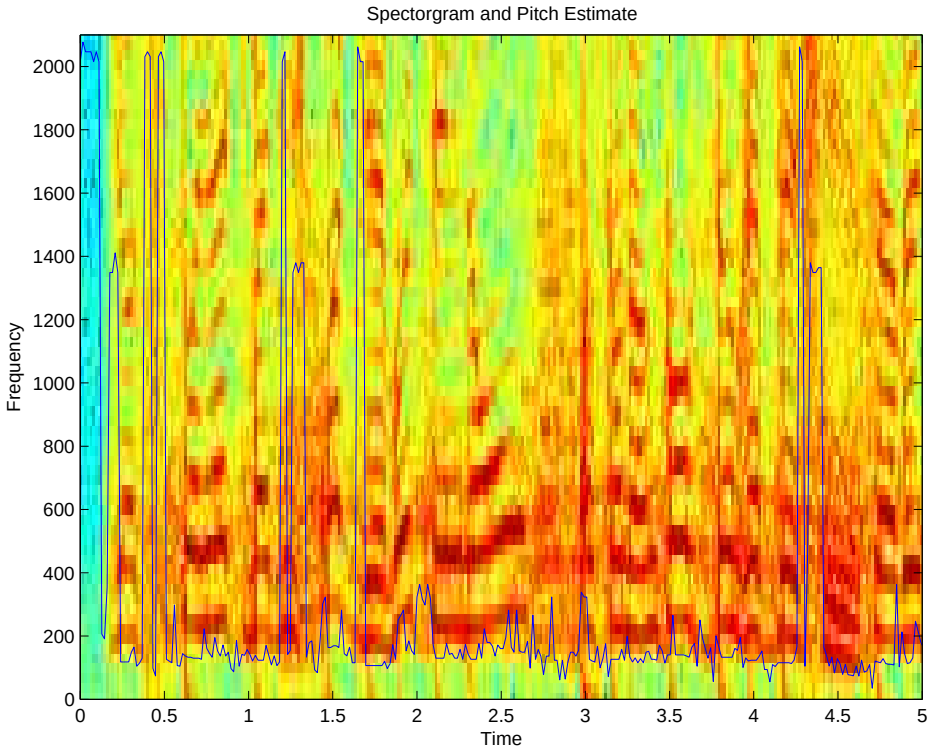
- The delivery, i.e. the non-verbal characteristics of the way people talk, conveys important information about social status.
- Social Signal literature suggests some characteristics that can be extracted through Signal Processing.
- In the case of broadcast news, the main social statuses are journalist and non-journalist. The goal of this work is to automatically recognize the status of each speaker using only the voice.

A Social Status Recognition Approach



- The voice of the speakers is analyzed using common Signal Processing techniques (in particular pitch tracking).
- The pitch is used to distinguish between voiced and non-voiced segments.
- Features extracted from pitch variations are fed to a linear classifier.

Pitch Tracking



The pitch curve is the average between the frequency at the first peak of the autocorrelation and the first peak of the Fourier transform.

Feature Extraction (I)

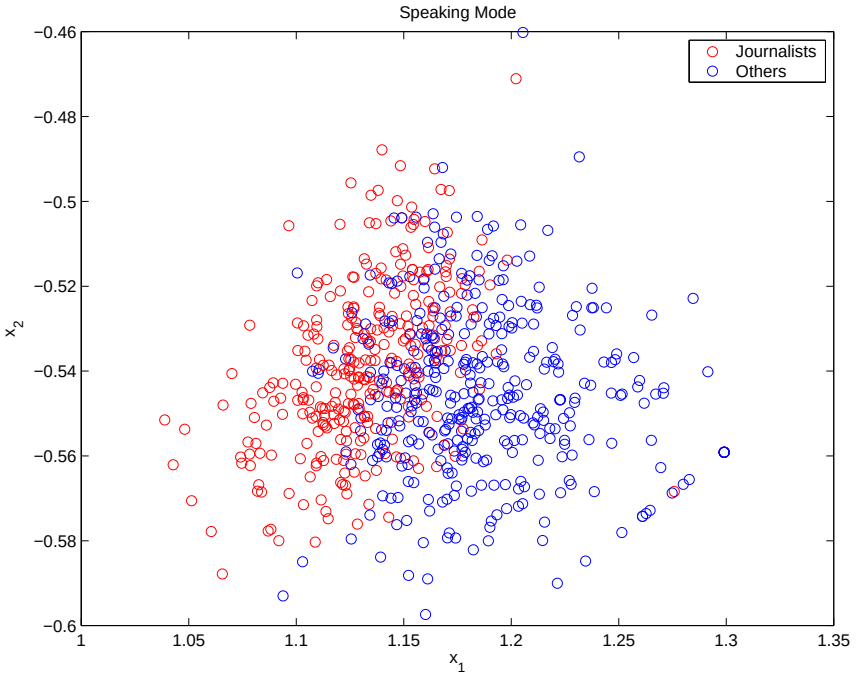
Consider the set $V = \{v_1, \ldots, v_M\}$ contaning the lengths of the voiced segments. The lengths are quantized because the pitch is measured at regular time steps and $T = \{t_1, \ldots, t_N\}$ is the set of the represented lengths. The relative entropy of the V elements distribution is:

$$H_v = \frac{-\sum_{i=1}^{N} p(t_i) \log p(t_i)}{\log N} \tag{7}$$

The same process can be applied for the non-voiced segments leading to an entropy H_s. As a result, each intervention can be represented using a feature vector:

$$\vec{x} = (H_s, H_v) \tag{8}$$

Feature Extraction (II)



After projecting the vectors onto the principal components, the plot shows that, although the overlapping, journalists and non-journalists occupy different regions of the feature space.

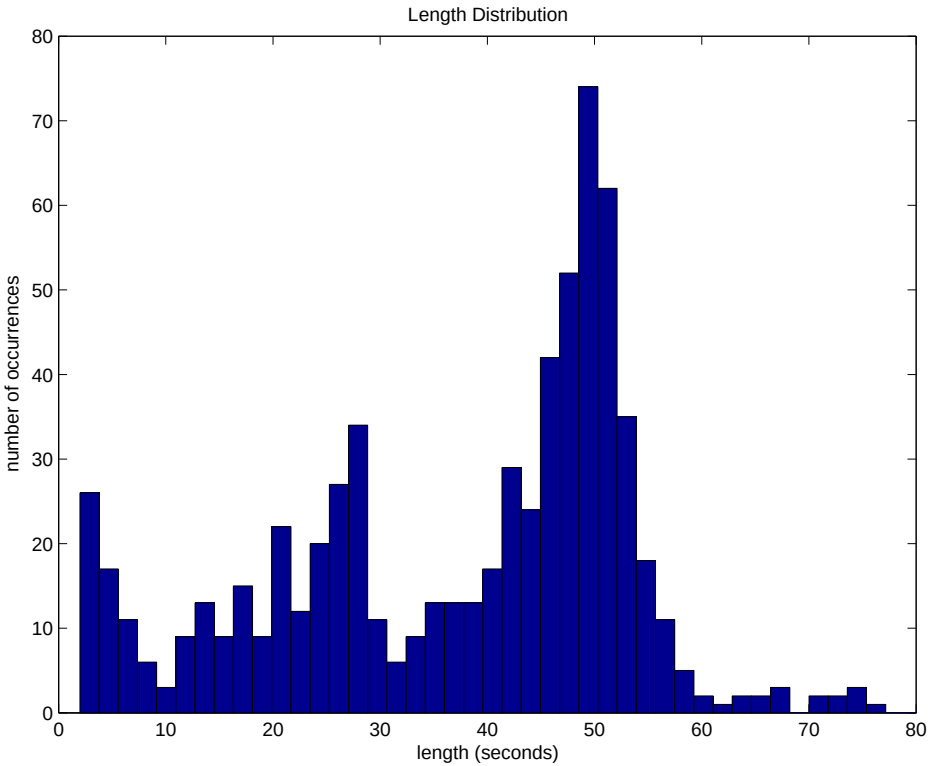
Classification

- A Gaussian $\mathcal{N}(\vec{x}|\vec{\mu}_s, \Sigma_s)$ is obtained for each of the two classes and, given an unseen feature vector, the classification is performed as follows:

$$s(\vec{x}) = \arg \max_{s \in \{0,1\}} \mathcal{N}(\vec{x}|\vec{\mu}_s, \Sigma_s) p(s) \tag{9}$$

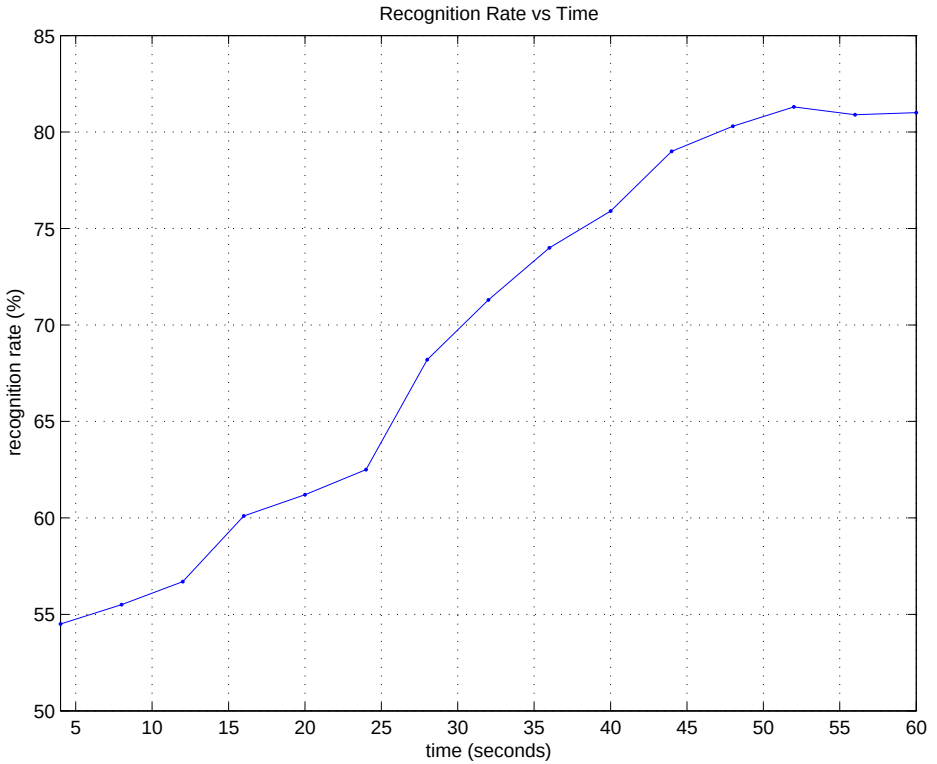
  where $s(\vec{x})$ is the class assigned to $\vec{x}$, and $p(s)$ is the a-priori probability of class s. The Gaussian parameters are estimated by maximizing the likelihood.
- The data set is split into two parts that are used alternatively as a training and test set.

The data



The data set consists of 686 single speaker segments, 313 journalists and 373 non-journalists. The number of individuals is 330 (234 non-journalists and 96 journalists).

The results



The plot shows the recognition rate as a function of the time extracted from the test set segments.

What About Humans?

- A pool of 16 human assessors has listened to 30 randomly selected clips, the total number of judgments is 480
- The clips are in French, but the mother tongues of the assessors are English (2 persons), Hindi (5 persons), Chinese (6 persons), Farsi (1 person), Serbian (1 person) and Arab (1 person).

total	women	men
82.3%	88.0%	79.0%

The performance of the automatic system over the same clips is 73.3%, but the test set is too small to conclude that the difference is statistically significant.

Conclusions

- Social Cognition seems to be a reasonable source of inspiration for multimedia indexing algorithms.
- Broadcast material offers a good compromise between spontaneous interactions and reasonable constraints.
- In a comparison with speech recognition, Social Signals seem to play the role of the acoustic features and the Social Networks seem to play the role of the language model.
- So far we used separately Social Signals and Social Networks, but in the future it can be worth to combine the two.
- We will address more ambitious problems like finding who supports whom in discussions, who are the best communicators, when there are conflicts and when there is cooperation, etc.

Thank You!