Analysis and Modelling of Speech Prosody and Speaking Style

Nicolas Obin

Sound Analysis and Synthesis Department
IRCAM - CNRS - UMR 9912 - STMS

23 June 2011

ircam Centre Pompidou

Introduction

ircam Centre Pompidou

Text-To-Speech Synthesis

Current Methods

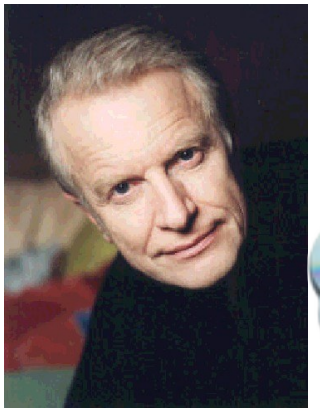
- **Unit selection** [Hunt and Black, 1996]
- **HMM-based** [Yoshimura et al., 1999]: model speech characteristics based on parametric statistical methods

Current State

- intelligible ✓

Limitations

- **natural**
- **variety**



Improvement

- modelling speech prosody

ircam Centre Pompidou

Speech Prosody

Definition

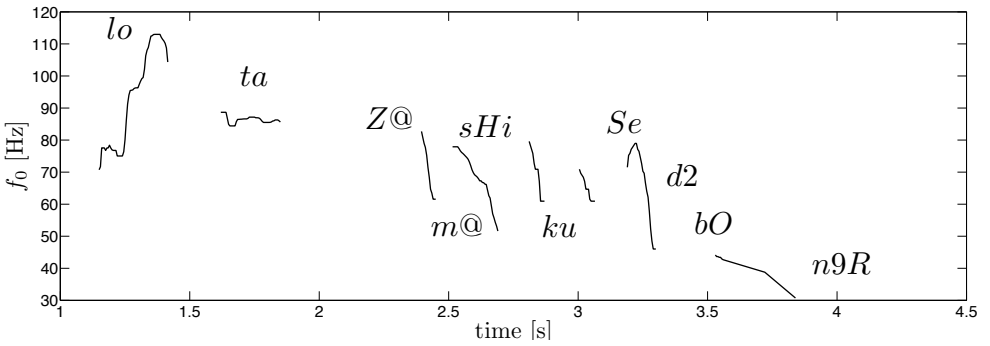
- Suprasegmental variations of speech - "*the music of speech*" (e.g., intonation, accent)
- Conveys meaning, emotions, intentions, and many information about the background of a speaker
- Vocal signature of a speaker, contribute as a part of his identity

Speech Prosody

Continuous Characteristics

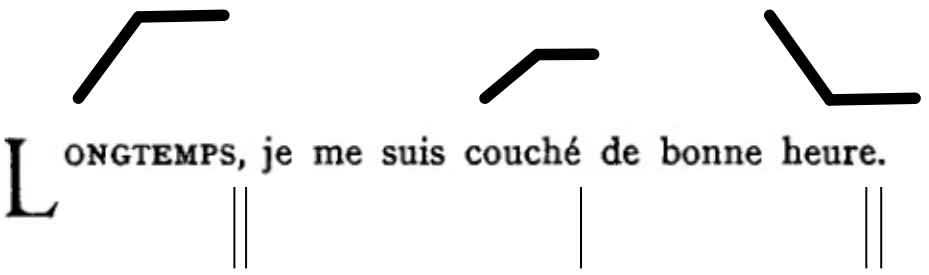
- fundamental frequency
- duration
- intensity
- voice quality
- articulation degree

L ONGTEMPS, je me suis couché de bonne heure.



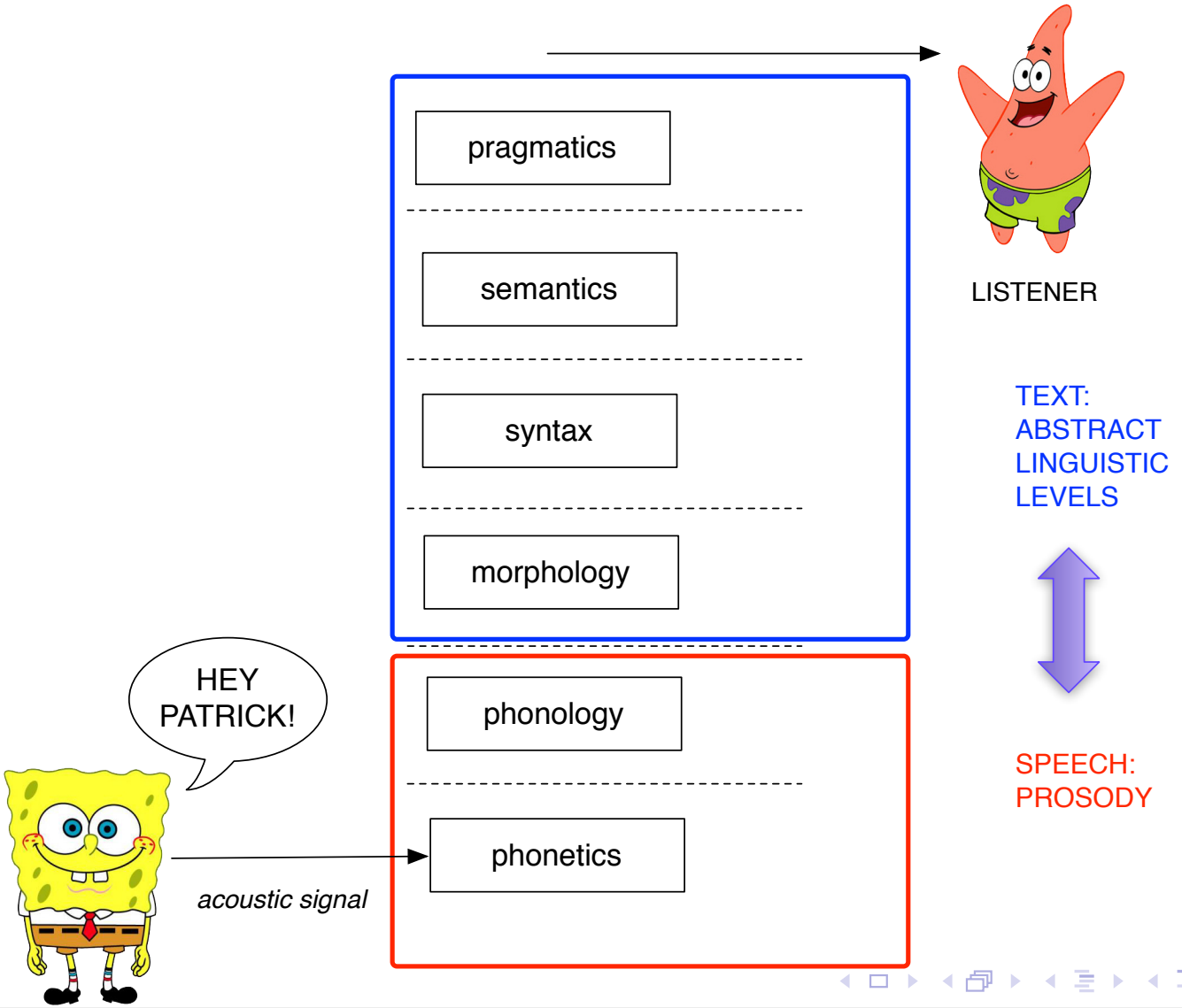
Discrete Characteristics

- Description of relevant prosodic events
  1. accent
  2. boundaries (e.g., pause)



Nicolas Obin
Sound Analysis and Synthesis Department, IRCAM - CNRS - UMR 9912 - STMS

Levels of Speech Communication



Nicolas Obin
Sound Analysis and Synthesis Department, IRCAM - CNRS - UMR 9912 - STMS

Major Contributions of this Work

Statistical Modelling of Speech Prosody

1. Statistical modelling of discrete (position of prosodic events) and continuous (F0 variations, duration) characteristics of speech prosody
2. Combination of syntactic and metric constraints to assign pauses
3. Stylization and trajectory modelling of F0 contours

Integration of a Rich Linguistic Description

4. Use of deep syntactic parsing to model speech prosody characteristics

Application to Speaking Style Modelling

5. Study on the ability of listeners to identify a speaking style
6. Reference for the evaluation of speaking style modelling
7. Ability of discrete/continuous HMMs to model the characteristics of a speaking style

Statistical Framework Used in this Work

Hidden Markov Models

- commonly used in speech recognition/synthesis

Used in Speech Synthesis

- discrete/continuous HMMs [Black and Taylor, 1994, Yoshimura et al., 1999]

Paradigms

- modelling speech characteristics
- modelling variability due to the **context** (e.g. phonemic, lexical, syntactic, or even semantic)
- context clustering: modelling contexts that are acoustically relevant

ircam Centre Pompidou

ircam Centre Pompidou

Introduction

Objective

Determine the position of prosodic events (accent, pause) from a text

Linguistic Studies

have pointed out that speech prosody is partially constrained by:

- **syntactic constraint** [Selrik, 1984]
- **metric constraint** [Liberman and Prince, 1977]

Issues

- Modelling syntactic constraint
- Modelling metric constraint
- Combining syntactic & metric constraints

Contributions of my Work

- **Integration of a deep syntactic description**
- **Use of Segmental HMMs + Dempster-Shafer Fusion**

Transcription of Speech Prosody Used in this Work[1]

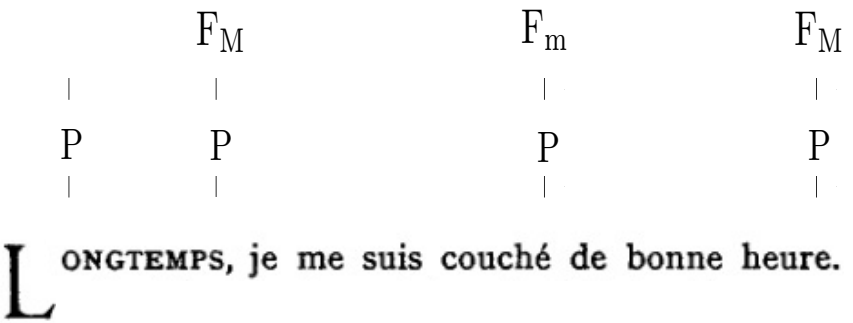
Specificities of French prosody

- French: syllable-based language (vs. English: stress-based language)
- **prosodic prominences** are primary cues used to segment speech into syntactico-semantic units
- **prosodic boundaries** have a central role in French prosody

My contribution for the transcription of French Prosody

proposed alternative to the ToBI standard [Silverman et al., 1992]

- major prosodic boundary (F_M)
- minor prosodic boundary (F_m)
- accent (P)



[1] *Rhapsodie*: reference prosody corpus of spoken French

ircam Centre Pompidou

Rich Syntactic Description [Obin et al., 2011a]

Objective

Modelling the linguistic constraint to assign prosodic events

State of the Art

- Surface description of syntactic characteristics [Black and Taylor, 1994]
- Some attempts to integrate a rich description of syntactic characteristics [Ingulfen et al., 2005]

Issue

- Integration of a rich description of syntactic characteristics

Contribution

- **Use of deep syntactic parsing to model speech prosody**

Linguistic Processing Used in this Work: ALPAGE

ALPAGE [Villemonte de La Clergerie, 2005]

- Linguistic Processing Chain developed for French
- Three modules: text pre-processing, surface and deep parsing

Text Pre-Processing

- segmentation of a text into words and sentences

Surface Parsing

- morpho-syntactic parsing (POS)

L ONGTEMPS, je me suis couché de bonne heure.

longtemps [lɔ̃tɑ̃] **1.** *adv.* long; a long time;

Linguistic Processing Used in this Work: ALPAGE

Deep Parsing

Deep parsing is used to describe the deep syntactic structure of a sentence

- Formalism: Tree Adjoining Grammar (TAG) [Joshi et al., 1975]

The syntactic structure can be described in terms of constituency or dependency.

Constituency

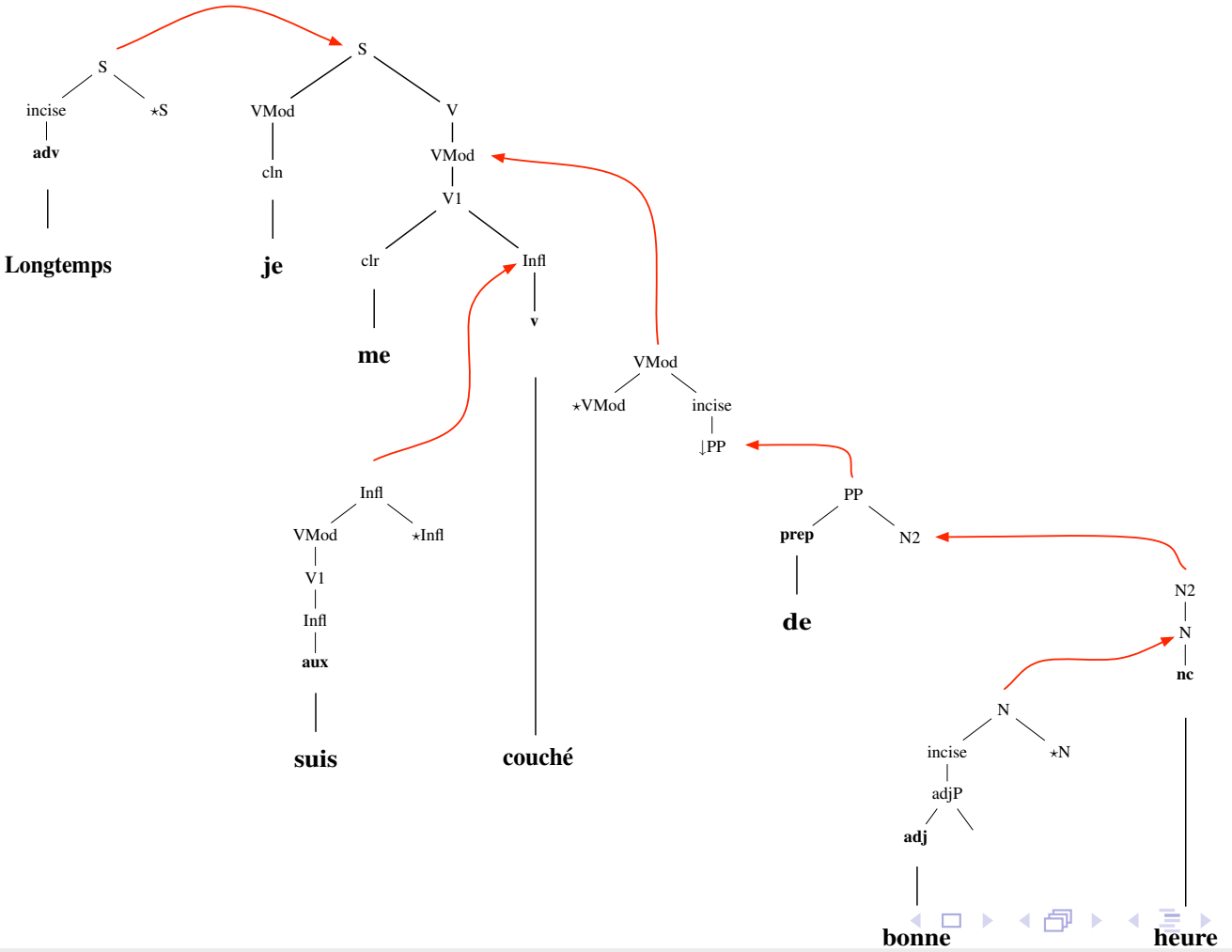
- describes the hierarchical structure of a sentence

Dependency

- describes the local dependency that relates words of a sentence

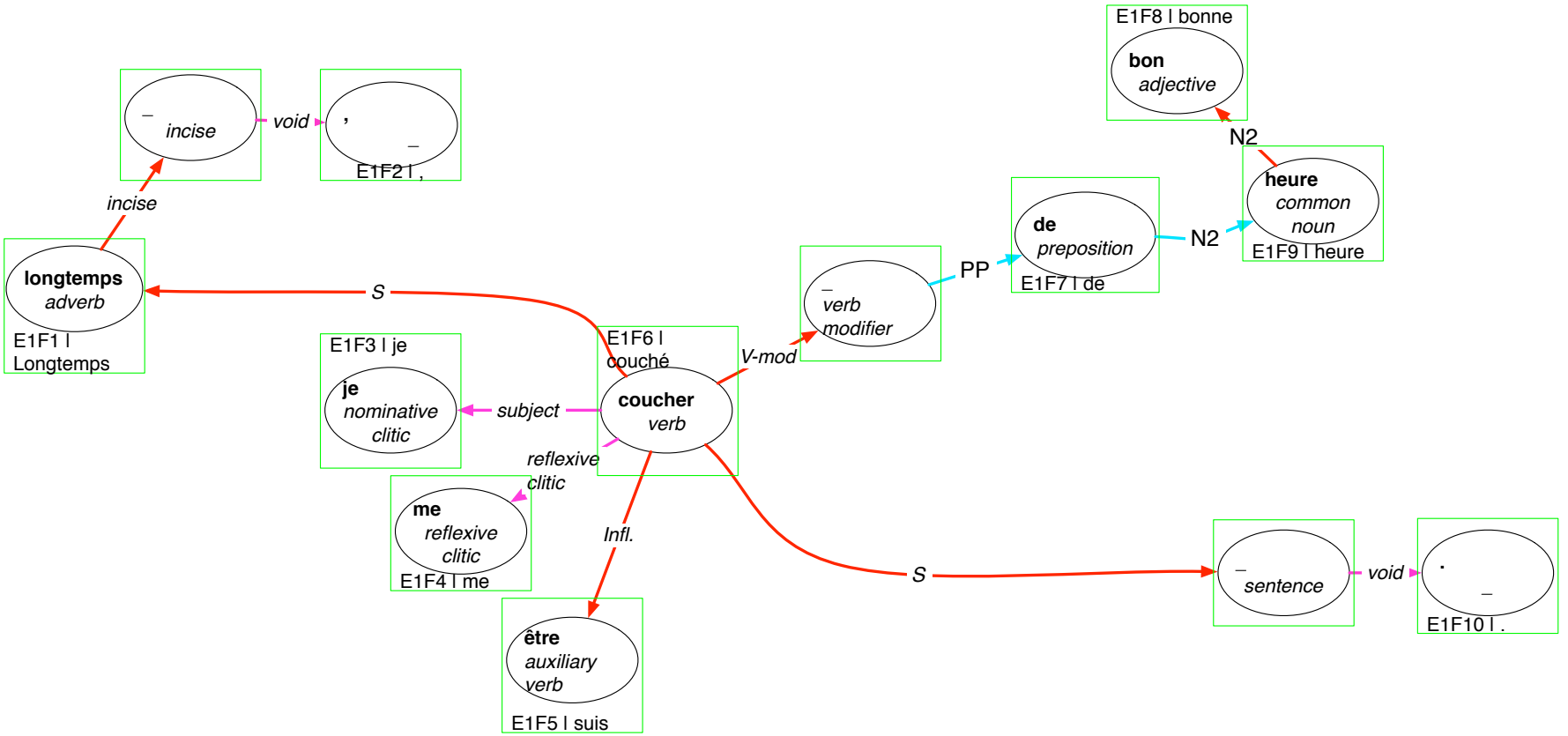
Linguistic Processing Used in this Work: ALPAGE

Constituency [Villemonte de La Clergerie, 2005]



Linguistic Processing Used in this Work: ALPAGE

Conversion into Dependency [Villemonte de La Clergerie, 2010]



Summary

Syntactic characteristics used to model speech prosody

conventional characteristics

Surface Parsing

(1) morpho-syntactic (**M**)

proposed characteristics

Deep Parsing

(2) **dependency** (**D**)

(3) **constituency** (**C**)

(4) **adjunction** (**A**): specific TAG operation

- describes a large variety of syntactic constructions (e.g., clause, incise)
- shown to be relevant to model speech prosody

Evaluation of the Automatic Assignment of Prosodic Events

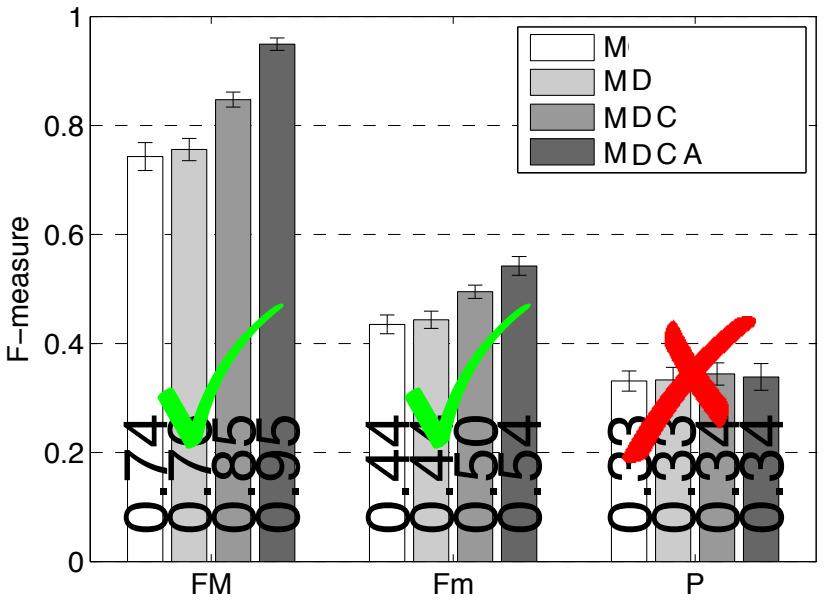
Scheme

- comparison of different sets of syntactic characteristics to model speech prosody
- speech database: neutral read speech (9hrs)
- procedure: 10-fold cross-validation
- metric: F-measure

Results

- **dramatical improvement for prosodic boundaries**
  - **F_M = 20%**
  - **F_m = 10%**
- **no improvement for prosodic prominence: probably related to semantic constraint, not syntax**

- Performance of the automatic assignment of Prosodic Events



ircam Centre Pompidou

Syntactic and Metric Constraints [Obin et al., 2011d]

Objective

Modelling syntactic and metric constraints to assign pauses

State of the Art

- Explicit integration of metric constraint in statistical modelling [Schmid and Atterer, 2004]

Issue

- Determine the adequate combination of syntactic & metric constraints

Contributions

- Use of Segmental HMMs + Dempster-Shafer Fusion

Syntactic and Metric Constraints [Obin et al., 2011d]

Segmental HMM

Segmental HMM [Ostendorf et al., 1996] addresses several limitations of conventional HMM:

- in particular, **segment duration is explicitly modelled** (metric constraint)

d= sequence that described the distance between consecutive pauses
o= sequence that describes observed syntactic characteristics

$$p(\mathbf{d}|\mathbf{o}) \propto \underbrace{p(\mathbf{o}|\mathbf{d})}_{\text{linguistic contribution}} \times \underbrace{p(\mathbf{d})}_{\text{metric contribution}} \quad (1)$$

Dempster-Shafer Fusion

- the **reliability** that can be conferred to different sources of information is explicitly formulated
- used to balance the syntactic and metric contributions to assign pauses

Evaluation of the Automatic Assignment of Pauses

Scheme

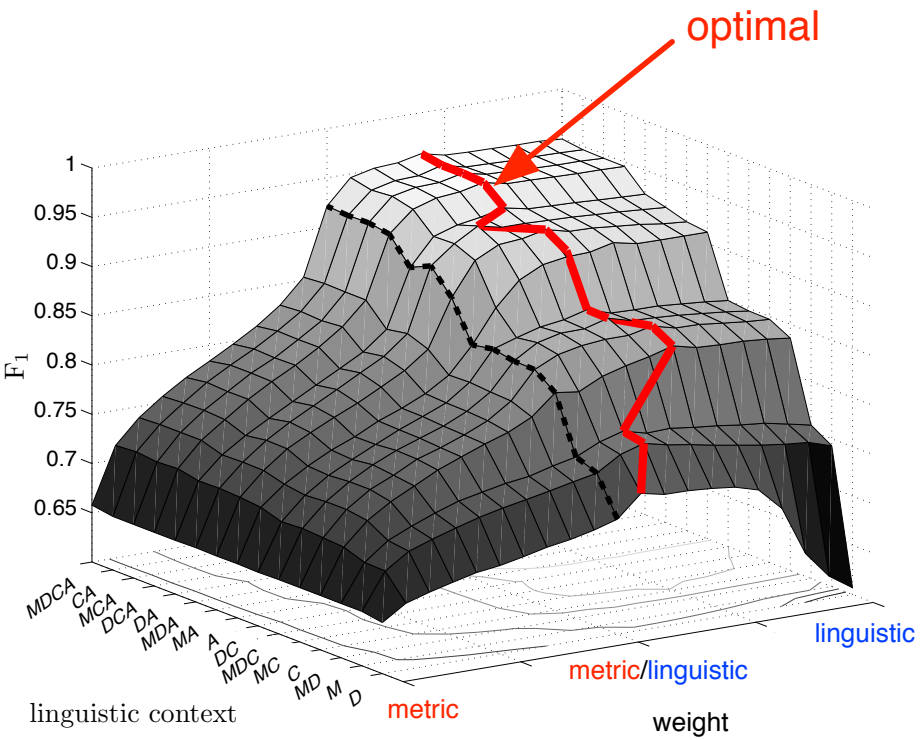
- same database, procedure, metric
- only for F_M (can be also used for F_m)

Results

- **optimal configuration of syntactic/metric constraints significantly outperforms conventional methods**
  - **optimal: 96.3%**
  - **linguistic: 95.0%**
  - **linguistic/metric: 92.1%**
- **syntactic constraint seems to have a greater influence than metric constraint**

- example of speech synthesis with automatic assignment of prosodic events

- Performance of the automatic assignment of Pauses



ircam Centre Pompidou

ircam Centre Pompidou

Objective

Modelling and Synthesizing F0 variations from a text

State of the Art

- short-term modelling + local trajectory constraint (conventional HMM) [Tokuda et al., 2003]
- stylization + no trajectory constraint [Mishra et al., 2006]
- short-term modelling + long-term trajectory constraint [Latorre and Akamine, 2008, Qian et al., 2009]

Issues

- combining stylization with trajectory modelling of F0 variations

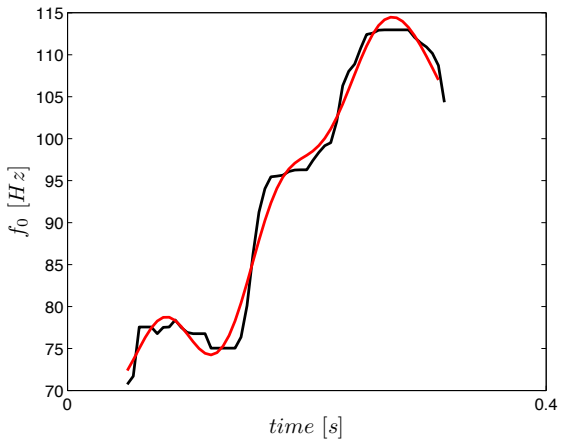
Contributions

- **trajectory model fully based on the stylization of F0 contours over various temporal domains**

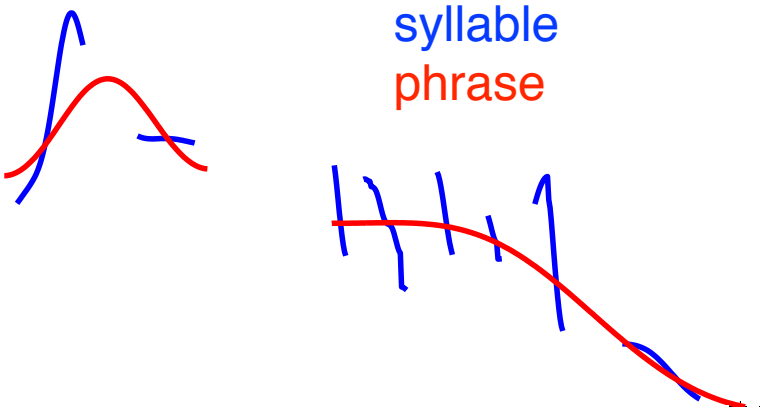
Stylization of Speech Prosody [Obin et al., 2011b]

Stylization of F0 contours

- Modelling relevant melodic variations



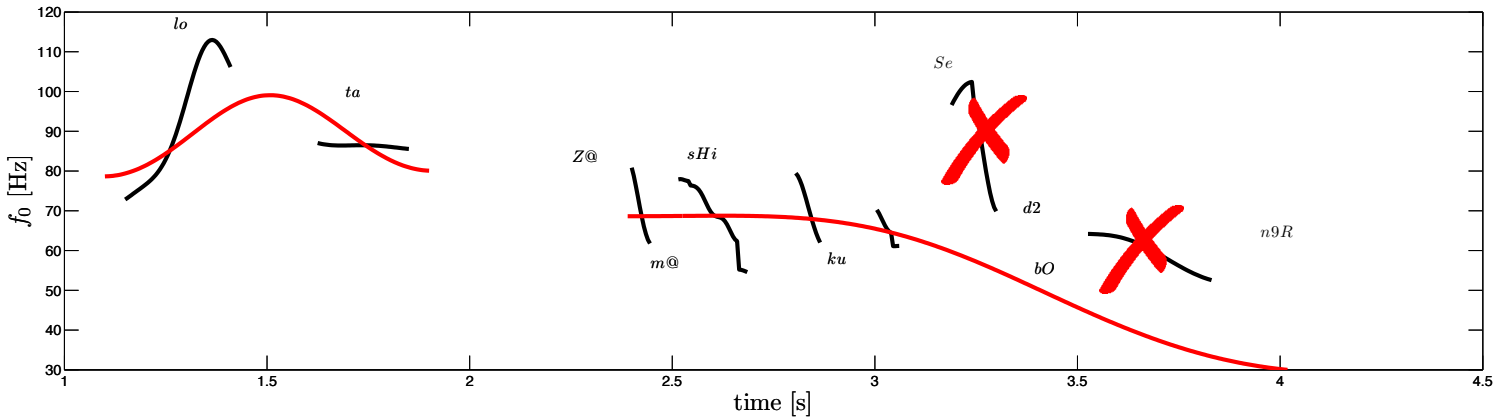
- Modelling contours over Various Temporal Domains



Trajectory Modelling

Stylization + Conventional Modelling

- During synthesis, the sequence of contours is determined as the sequence of mean contours (assumption of conditional independence)
- Example:

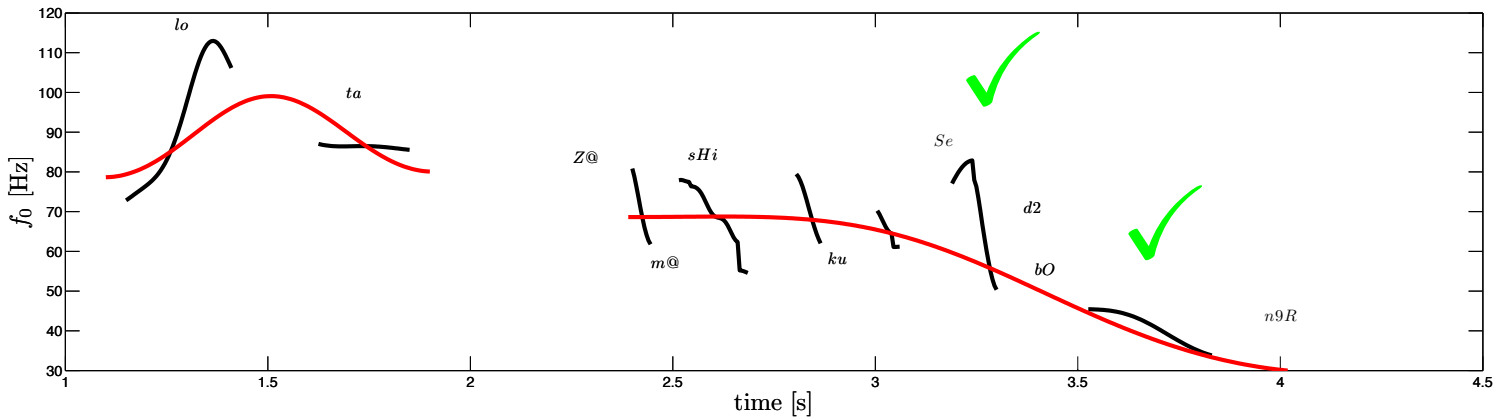


- **This may result into inadequate phrasing**

Trajectory Modelling

Stylization + Trajectory Modelling

- During synthesis, the sequence of contours is determined **under the constraint of long-term trajectories**
- Example:



- **The long-term trajectory constraints adequate phrasing**

Subjective Evaluation of Stylization/Trajectory Model

Comparison of F0 models in Speech Synthesis

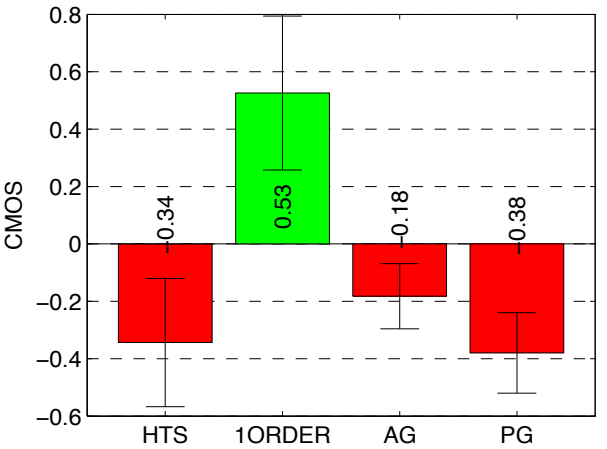
- conventional HMM (HTS) [Yoshimura et al., 1999]
- stylization + trajectory
  1. syllable + 1-order adjacent syllables (1-ORDER)
  2. syllable + minor prosodic phrase (AG)
  3. syllable + major prosodic phrase (PG)

Procedure

- CMOS preference experiment
- 8x4 synthesized speech utterances
- 20 native French speakers

- **1-order trajectory model is significantly preferred**
- **long-term trajectories (AG/PG) partially successful (due to increase in complexity)**

- CMOS



ircam Centre Pompidou

Style: a matter of Situation

Individual and Shared Identities

Each communication act instantiates a style which is composed of:

- an individual speaking style that depends on the speaker identity

SPEAKER

- a conventional speaking style conditioned by a specific situation

ircam Centre Pompidou

Objective

Modelling speaking style in speech synthesis

State of the Art

- modelling continuous characteristics of a speaking style (emotions) [Yamagishi et al., 2004]
- modelling discrete characteristics of a speaking style [Bell et al., 2006]

Issues

- modelling the discrete and continuous characteristics of a speaking style

Contributions

- **study of the ability of human listeners to identify a speaking style**
- **reference for the evaluation of speaking style modelling**
- **study of the capacity of discrete/continuous HMMs to model characteristics of a speaking style**

ircam Centre Pompidou

Identification Ability of Speaking Style [Obin et al., 2010]

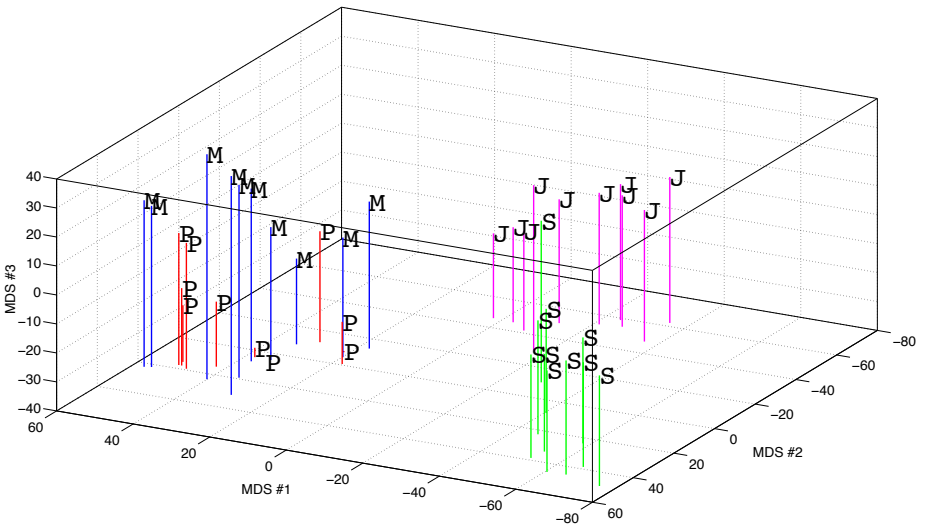
Objective

- Study of the ability of human listeners to identify the speaking style of natural speech

Experiment

- 40 speech utterances with 4 speaking styles (church office, political speech, journalistic speech, sport commentary)
- delexicalized to remove lexical access
- 72 participants (various language background)
- multiple choice identification of speaking style by human listeners

- Confusion of natural speaking styles by human listeners



ircam Centre Pompidou

Modelling Speaking Style [Obin et al., 2011c]

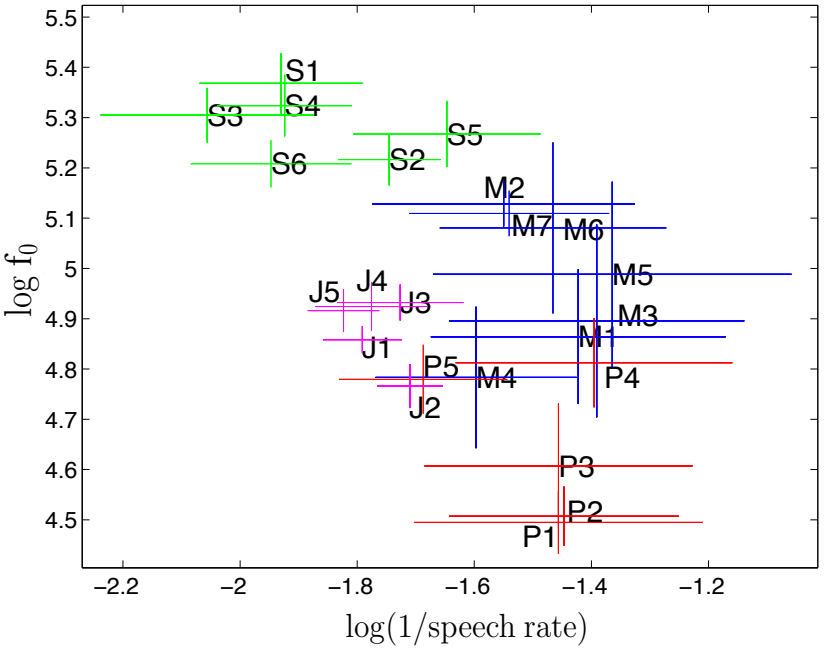
Objectives

- application of discrete/continuous model to speaking style
- capacity of HMM-based speech synthesis to model speaking style

Principle

- average modelling of discrete and continuous characteristics of a speaking style (multiple speakers)

- Description of speakers characteristics



Evaluation of Ability of Human Listeners to Identify a Synthetic Speaking Style

Experiment

- 40 synthesized speech utterances + delexicalized (same as previously)
- 50 participants (various language background)
- multiple choice identification of speaking style by human listeners

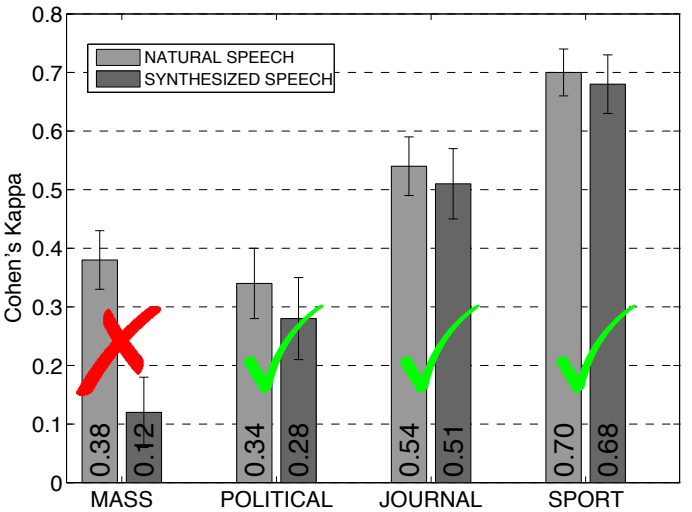
Results

- **discrete/continuous HMMs consistently model the characteristics of a speaking style** (with exception of church office)

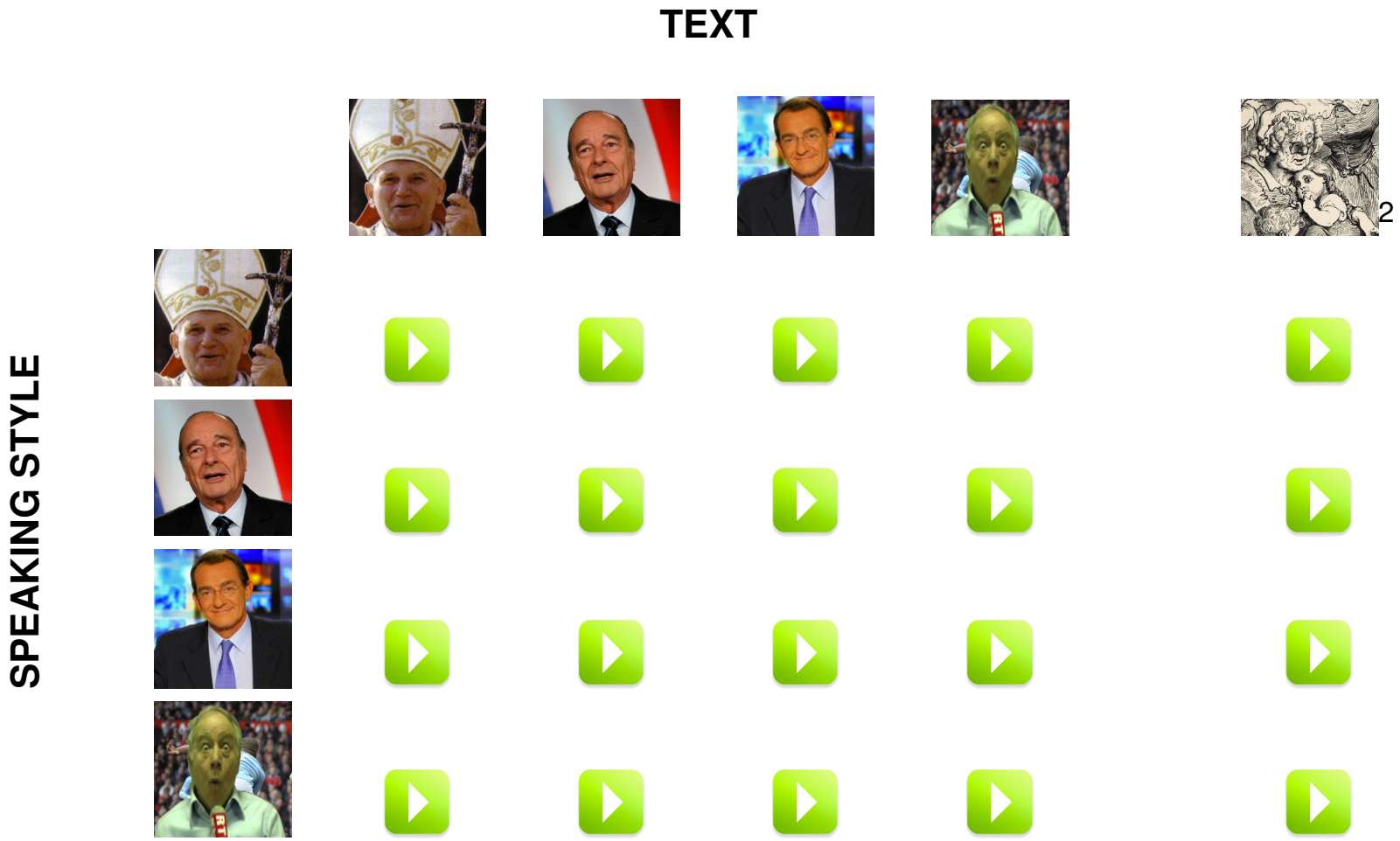
Question

- **what is the contribution of discrete/continuous characteristics?**

- Comparison of identification obtained for natural and synthetic speech



Synthesis of Speaking Styles



[2] Neutral, Fairy Tale, "*Litlle Tom Thumb*"

ircam Centre Pompidou

ircam Centre Pompidou

Conclusion

Contributions to the Statistical Modelling of Speech Prosody

1. Statistical modelling of discrete and continuous characteristics of speech prosody
2. Combination of linguistic and metric constraints to assign pauses
3. Stylization and trajectory modelling of F0 contours

Contribution to the Integration of a Rich Linguistic Description

4. Use of deep syntactic parsing to model speech prosody characteristics

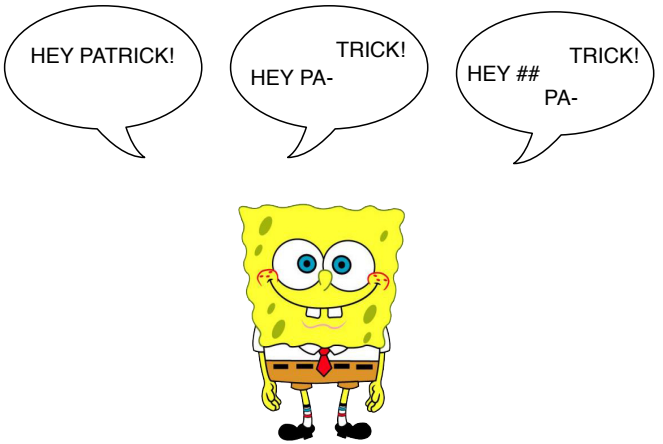
Contributions to the Modelling of Speaking Style

5. Study of the ability of listeners to identify a speaking style
6. Reference identification performance for the evaluation of speaking style modelling
7. Ability of discrete/continuous HMMs to model the characteristics of a speaking style

Further Directions

Modelling the Variety of Speech Prosody

- a speaker has various alternatives to pronounce a same utterance (intra-speaker variability)
- varying speech prosody will certainly improve the naturalness of synthetic speech
- examples of various speech prosody determined for the same sentence



Unified Modelling

- joint modelling of discrete/continuous characteristics
- that also accounts for alternatives
- objective: improving the coherence and variety of speech prosody

ircam Centre Pompidou