From linguistic meaning to expressivity in text to speech

N. Obin[1,2], A. Lacheret[2], G. Beller[3],
I.Grichkovtsova[4], M. Morel[5]

[1, 3] IRCAM, Paris,
[2] MODYCO, Nanterre
[2] Institut universitaire de France, Paris,
[4, 5] CRISCO, Caen

Situation & Hypotheses (1)
Researchers usually define two types of prosody

Linguistic prosody

- Demarcative function
  - Rhythmical level
  - Syntactical level
  - Communicative level
- Lexical function
- Discourse function

Ectolinguistic prosody

Vocal signature

- Phonostyles

Paralinguistic prosody

Expressive function

- Attitudes & emotions

Situation & Hypotheses (2)

Situation

Researchers explore either area (linguistic prosody vs expressive prosody) without investigating into how the two levels of processing interact

In this context of lack of interfacing in standard research, **our Hypotheses are the following**

1° Rules used to generate linguistic prosody in the speech synthesis impact on the paralinguistic prosody

2° In order to generate correct expressive prosody, precise linguistic gating points must be used

3° Coding of different emotions in "neutral" utterances like *"I am going home now"* is not associated with phonological differences of syllabic localisation (the same phonological domaine, i.e. the same syllables are stressed, whatever the emotion is), but with acoustic differences: types of activated features for the realisation of emotions, level of activation, various combinations of features

Hypotheses: from linguistic to expressive prosody, features and cues for accent placement (1)

Demarcative function

- whatever the emotion is, boundary tones can be calculated from syntactical and rhythmic constraints
- Whatever the emotion is, stress position will also be linked to the information structure: discourse segmentation in topic & focus (cf. *obligatory prosodic boundary* at the end of the topic, focal stress)

Hypotheses: from linguistic to expressive prosody, features and cues (2)

Lexical function

- Words carrying an expressive stress also carry a rich informational content. In order to fulfill those two functions (informational & expressive), they should have certain lexical characteristics

→Strong connections between the informational content and lexical fields of a specific emotion: words which convey an emotion in the discourse are also those which allow the evolution of the thematic progression and the informational structure

- Some semantic features are prosodically marked (negation, intensive semantic value)

  → Some words are prosodically very flexible (adverbial status, status of verbal constructions (support verb or not, aspectual content of the unit : *devoir)*)

Hypotheses: from linguistic to expressive prosody, features and cues (3)

Discourse level

- Discourse articulation and cohesion
  - Prosodic marking of co-reference strings, anaphora & deixis: the anaphoric or deixis status of the unit in the discourse determine if it can receive or not an expressive stress (to be illustrated later)
- Speaker viewpoint (epistemical modality): linked to phonostylistic variations (some speakers invest more in their discourse and in the expression of their emotion)

→ 2 levels of stress (functional analysis)

- Primary stress: demarcative function
- Secondary stress: other functions

Corpus example

French version

Vous appelez ça une chambre d'hôtel ? Regardez un peu ces draps : ils sont ignobles ! Vous ne croyez quand même pas que je vais dormir ici ? C'est révoltant ! Je vais rentrer à la maison maintenant. Ce n'est pas un hôtel ici, c'est un élevage de cafards !

English version

Do you call this a hotel room? Look at these sheets! You don't think that I am going to sleep here! It is disgusting! I am going home now! It is not a hotel here, it is a cockroach farm.

Stress prediction model
Illustration

First step → primary stress (8 obligatory)
Second step → secondary stress (8 optional)

- Vous appelez ça une chambre **d'hôtel**
- Regardez un peu ces **draps**
- ils sont i**gnobles**
- Vous ne croyez quand même pas que je vais dormir **ici**
- C'est **révoltant**
- Je vais rentrer à la mai**son maint'nant**.
- Ce n'est pas un **hôtel** i**ci**
- c'est un élevage de **cafards**

Rules in details

1° identify all the syllables with primary stress

8 positions

2° identify lexical units carrying expressivity:

ignobles, révoltants, cafards

3° Semantic specific features (negation, intensive value)

pas

4° Modality

quand-même

5° Deictic/anaphoric

ça

Dormir ici: stressed deictic

Ce n'est pas un hôtel ici: unstressed anaphoric

Results

Vous appelez **ça** deictic une chambre **d'hôtel** demarcative stress carrying expressive modality

Regardez **un peu** adverbial ces **draps** demarcative stress carrying expressive modality

ils sont **ignobles** demarcative + lexical + focal stress carrying expressive modality

Vous ne croyez **quand même** modality **pas** adverbial que je vais dormir **ici** demarcative stress carrying expressive modality

C'est **révoltant** demarcative + lexical stress carrying expressive modality

Je vais rentrer à la **maison** demarcative stress ? **maintenant** demarcative stress carrying expressive modality .

Ce n'est **pas** adverbial un **hôtel** demarcative + focal stress carrying expressive modality ici unstressed postfix

c'est un élevage de **cafards** demarcative + lexical + focal stress carrying expressive modality

Corpus design

Starting point (Grichkovtsova & al 2008)

- 14 affective states (anger, fear, sadness, joy, disgust, grief, astonishment-surprise, uncertainty-hesitation, incredulity, embarrassment-shame, politeness-respect, obviousness, directive-authority, contempt) plus a neutral statement and a neutral question.
- An affectively coloured text was written for each studied emotion and attitude. The same neutral utterance was inserted in each text : *I am going home now./Je vais rentrer à la maison maintenant.*

  The idea was that the neutral utterance would carry the affective modality acted by the speaker throughout the text.
- 22 French native speakers (11 males, 11 females)

Corpus validation

The recorded corpus was validated through a psycholinguistic perception test with 10 French native listeners.

Emotions and attitudes were evaluated separately in the following three subtests:

•emotions (anger, fear, sadness, joy, disgust, grief and neutral),

•attitudes (uncertainty-hesitation, embarrassment-shame, politeness-respect, obviousness, directive-authority, contempt and neutral),

•attitudes with an interrogative contour (surprise, incredulity, neutral question and neutral statement).

Only utterances identified by at least 50% of listeners were selected for the corpus.

Data used for the experiment

→28 utterances were taken from the validated corpus for the present study (between 1 to 3 speakers by utterance)

3 productions for anger, fear, sadness, surprise, hesitation, incredulity, obviousness)

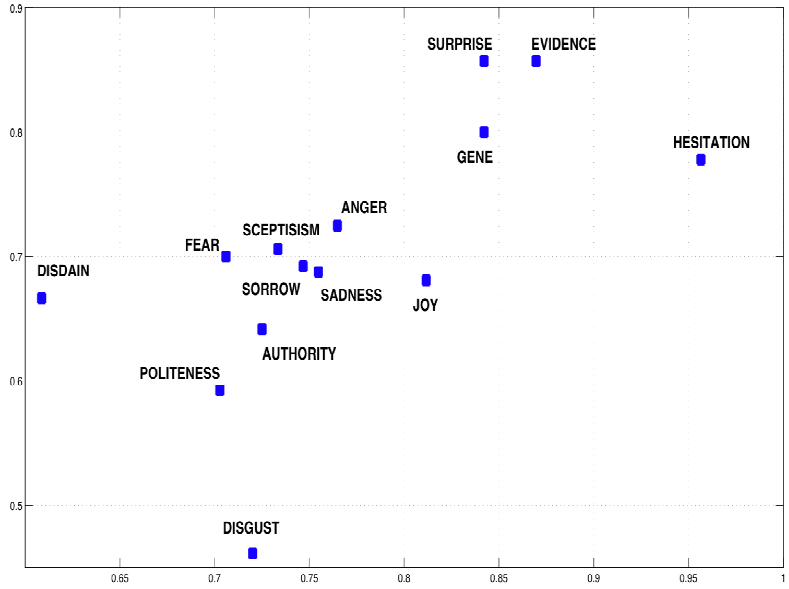
Study : stress labelling

- Prominence labelling *syllable-based*
- Prominence labelling of text exclusively based on linguistic knowledge (NP, PP, PS)
- Prominence labelling based on acoustic: two annotators
- Inter acoustic annotators agreement :
- 76% f-measure on prominence.

ANO1/ANO2	NP	P	TOTAL
NP	885	161	1046
P	12	282	294
TOTAL	897	443	

Comparison : Text prediction PP vs. annotators forced consensus

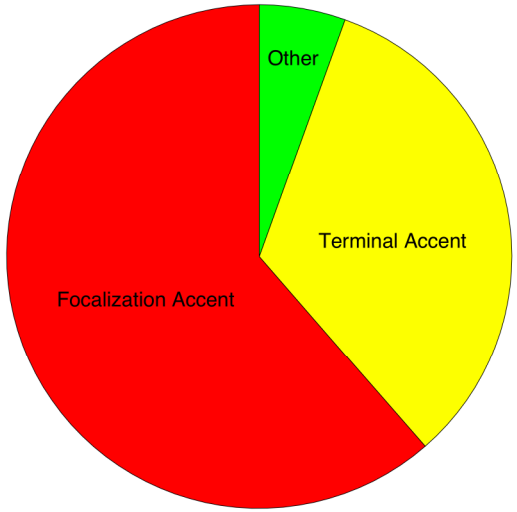
- Prominence predicted after text is precise (precision = 84%) but a lot of observed prominences are missed (accuracy = 61%)
  - Vous appelez ça une chambre d'hôtel ?
  - Vu **za**$_{\textbf{focalization}}$ p@ le sa yn Sa br@ do tEl
- Agreement is expressivity-dependent



Studying the text label PS distribution

Related to :

- Accentuation strategy within prosodic group (semantic-dependent)
  - Regardez un **peu**$_{PS}$ ces **draps**$_{PP}$
- Within-word stress distribution
  - Ils sont **i**$_{PS}$**gno**$_{PP}$bles
- These labels are related to focalization accent (80%) and pausing strategy (15%).
- This secondary stress function highly depends on expressivity: evidence (quasi null) / anger (the most)



Conclusion (1)
strategies for modeling expressive prosody in text-to-speech synthesis

- We studied the relation between linguistic meaning and prominence in expressive speech
- We put in light that *expert linguistic knowledge* should be used in a first step to infer expressivity-dependent prominence location
- Manual diagnostic and linguistic stress feature should be used for further analysis
- Then this knowledge could be used in an expressivity-dependent phonological structure learning.

Conclusion (2)
Some points to be discussed

- Pb of syllabic prediction for stress distribution within a word: gap between the phonological stress & the phonetic holistic realisation of the stress *(ils sont **ignobles**)* → is syllable a relevant psycho-acoustic anchored point ?

- Stress perception linked to consonantic articulation is not predicted by the model; actually it plays a great role in paralinguistic prosody independantly of linguistic factors

- While our hypotheses (one & only one phonological system, different phonetic variations) seems to be valide when the emotion is lexically marked, it is problematic regarding neutral utterance *je vais rentrer à la maison maintenant*

- Question : why ?
- Answer : principle of economy vs compensation principle :
  - in such a context, prosody is the only tool to mark contrasts between differrent emotions or attitudes→ all cues which may fulfill this function are used (syllabic duration, pauses, consonantic gestures, tempo, etc)

  → one apply the compensation principle