Combining lexical and prosodic features for automatic detection of sentence modality in French

Luiza Orosanu, Denis Jouvet

INRIA-Loria, Nancy, France

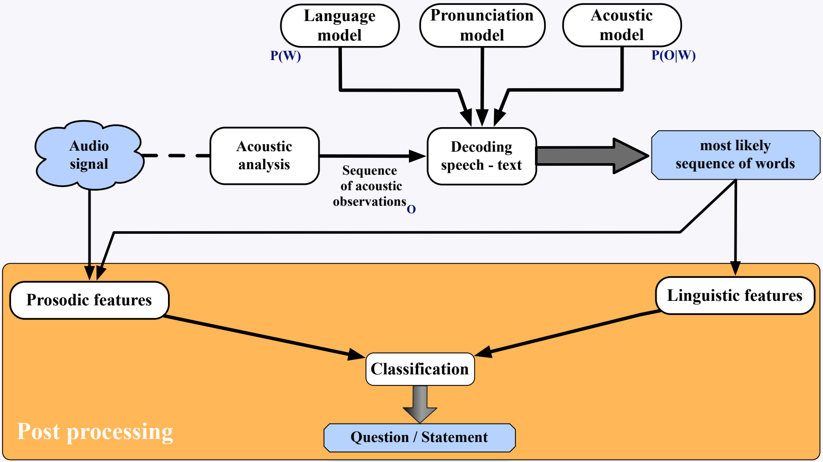
Multispeech Team

Summary

1. Context and approach
2. Prosodic and linguistic features
3. Experiments
4. Conclusions and future work

Context

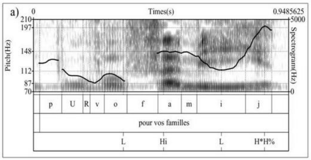
Objective : state from the automatic transcription if the sentence is a question or a statement



Approach

- **prosodic classifier** : uses the intonation

  → sentences perceived as questions through the intonation



- **linguistic classifier** : uses the linguistic information

  → sentences perceived as questions through the interrogative forms

  ∗ qu'est ce qu'on doit comprendre ?

  (→ what should we understand?)

  ∗ est ce que vous souhaitez une confrontation ?

  (→ do you want a confrontation?)

- **combined classifier** : uses both types of information

Approach

- evaluate classifier on **manual transcriptions**

  → ideal conditions - 0% word error rate

- evaluate classifier on **automatic transcriptions**

  → real conditions - 26% word error rate

Summary

1. Context and approach
2. Prosodic and linguistic features
3. Experiments
4. Conclusions and future work

Prosodic features (#10)

- generally, a question has a final rising pitch
- we compute 10 prosodic features that take into account
  - ∗ the duration
  - ∗ the energy
  - ∗ the pitch

  of the last prosodic group of the sentence

→ the F0 and energy values are computed every 10ms using the ETSI/AURORA acoustic analysis

Prosodic features (#10)

Features vector

class	{0=statement; 1=question}		
Prosodic Features	VNDurNorm	=	the duration of the last syllable (normalized)
	VNLogENorm	=	the logarithm of the energy of the last syllable (normalized)
	VNF0Delta	=	the F0 difference between the last syllable and the first syllable
	VNF0Slope	=	the F0 slope on the last syllable
	VNF0SlopeT2	=	VNF0Slope * VNDurNorm^2
	globalSlopeSlope	=	the F0 slope on the longest ending F0 slope
	globalSlopeLength	=	the length of the longest ending F0 slope
	globalSlopeDelta	=	the F0 difference between the beginning and the end of the longest ending F0 slope
	globalSlopeSlopeT2	=	globalSlopeSlope * globalSlopeLength^2
	lastF0Level	=	the last F0 level (normalized by speaker)

Linguistic features (#3)

- **iP**: the interrogative patterns

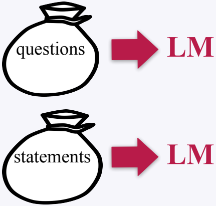
  → indicate the presence or absence
  of an interrogative pattern in a phrase

- quel *(→ which, m)*
- quelle *(→ which, f)*
- quels *(→ which, m, pl)*
- quellles *(→ which, f, pl)*
- comment *(→ how)*
- combien *(→ how much)*
- pourquoi *(→ why)*
- est ce que *(→ is/do ...)*
- est ce qu' *(→ is/do ...)*
- qu' est ce *(→ what ...)*
- qu' est ce que *(→ what ...)*
- qu' est ce qu' *(→ what ...)*

Linguistic features (#3)

- the probability of the sentence being a question
  - with respect to two reference language models

$$\mathrm{LLR(sentence)} = \mathrm{Log}\left(\frac{\mathrm{P(sentence|LM\text{-}question)}}{\mathrm{P(sentence|LM\text{-}statement)}}\right)$$



- $\ast$ LLR $\geq 0 \rightarrow$ likely to be a question
- $\ast$ LLR $< 0 \rightarrow$ likely to be a statement

lexLLR	we apply the **lexical** language models on the **sequence of words**
synLLR	we apply the **syntactic** language models on the **sequence of POS tags**

Combined linguistic-prosodic features (3L-10P)

Features vector

class	{0=statement; 1=question}		
3L	lexLLR	=	the lexical log-likelihood ratio
	synLLR	=	the syntactic log-likelihood ratio
	iP	=	presence or absence of interrogative pattern
10P	VNDurNorm	=	the duration of the last syllable (normalized)
	VNLogENorm	=	the logarithm of the energy of the last syllable (normalized)
	VNF0Delta	=	the F0 difference between the last syllable and the first syllable
	VNF0Slope	=	the F0 slope on the last syllable
	VNF0SlopeT2	=	VNF0Slope * VNDurNorm^2
	globalSlopeSlope	=	the F0 slope on the longest ending F0 slope
	globalSlopeLength	=	the length of the longest ending F0 slope
	globalSlopeDelta	=	the F0 difference between the beginning and the end of the longest ending F0 slope
	globalSlopeSlopeT2	=	globalSlopeSlope * globalSlopeLength^2
	lastF0Level	=	the last F0 level (normalized by speaker)

Summary

1. Context and approach

2. Prosodic and linguistic features

3. Experiments
    - Setups for experiments
    - Results

4. Conclusions and future work

Summary

1. Context and approach
2. Prosodic and linguistic features
3. Experiments
   - Setups for experiments
   - Results
4. Conclusions and future work

Data for LM training

Textual corpus GigaWord

- extraction of **statements** : sentences ending with a '.' [#16M]
- extraction of **questions** : sentences ending with a '?' [#89K]

word sequences

question	à quel moment le raid a décidé d'intervenir?
statement	nous sommes ensemble pour 60 minutes.

⇓

the **lexical language models** of questions and statements

part-of-speech (POS) sequence

question	PRP PRO: REL NOM DET: ART NOM VER: pres VER: pper PRP VER: infi
statement	PRO: PER VER: pres ADV PRP NUM NOM

⇓

the **syntactic language models** of questions and statements

Data for training and evaluating the classifiers

- **Audio corpus**: Ester, Etape, Epac
  - training set : 300h of speech (manually transcribed)
  - evaluation set : 22h of speech (manually transcribed)
  - Ester&Epac: French broadcast news, collected from radio channels (prepared speech, plus interviews)
  - Etape: debates collected from various French radio and TV channels (spontaneous speech)
- Data sets of **questions and statements**

  → sentences ending with a '?', respectively with a '.'

	#questions	**#affirmations**
training	10.0K	10.0K
evaluation	0.8K	7.0K

Question / Statement classification

- **Classifier:** the J48 decision tree (WEKA software)

- **Settings**
  - $*$ features extracted from manual transcriptions (0% WER)
  - $*$ features extracted from automatic transcriptions (26% WER)

- **Performance**

$$\frac{1}{\mathrm{H}} = \frac{1}{2} * \left(\frac{1}{\mathrm{ccQuestions}} + \frac{1}{\mathrm{ccStatements}} \right)$$

ccQuestions = percentage of correctly classified questions
ccStatements = percentage of correctly classified statements

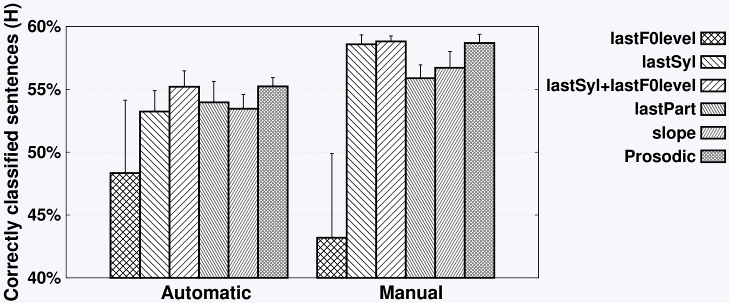
Summary

1. Context and approach
2. Prosodic and linguistic features
3. Experiments
    - Setups for experiments
    - Results
4. Conclusions and future work

Results on prosodic features

Evaluate different **combinations of prosodic features**

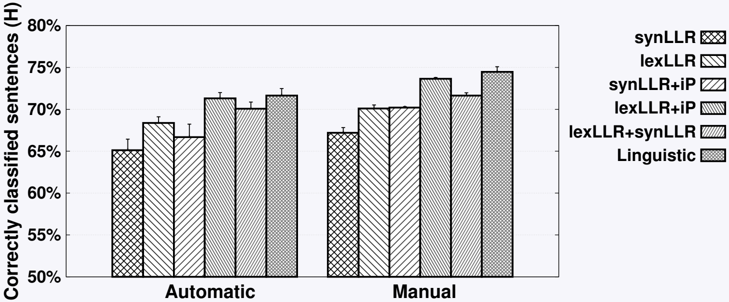
- ∗ the last F0 level (lastF0level)
- ∗ the 5 features computed over the last syllable (lastSyl)
- ∗ the 5 features computed over the last syllable + the last F0 level (lastSyl+lastF0level)
- ∗ the 5 features computed over the ending part of the utterance (lastPart)
- ∗ the 6 features related to slope measurements (slope)
- ∗ all 10 features (Prosodic)



Results on linguistic features

Evaluate different **combinations of linguistic features**

- ∗ the syntactic log-likelihood ratio (synLLR)
- ∗ the lexical log-likelihood ratio (lexLLR)
- ∗ the syntactic log-likelihood ratio + the presence of interrogative patterns (synLLR+iP)
- ∗ the lexical log-likelihood ratio + the presence of interrogative patterns (lexLLR+iP)
- ∗ the lexical log-likelihood ratio + the syntactic log-likelihood ratio (lexLLR+synLLR)
- ∗ all 3 features (Linguistic)



Results on prosodic, linguistic and combined features

Percentage of correctly classified sentences (H)

Transcripts	**Prosodic**	**Linguistic**	**Combined**
automatic	55.24%	71.64%	72.21%
manual	58.69%	74.47%	74.26%

→ linguistic classifier outperforms prosodic classifier

→ combined classifier outperforms linguistic classifier on automatic transcriptions

→ linguistic classifier: 3% alsolute difference between manual and automatic transcriptions

→ combined classifier: 2% alsolute difference between manual and automatic transcriptions

Best results with combined features

Confusion matrix between questions and statements obtained on **automatic transcriptions**

	number	classified as question	classified as statement	
question	831	**627**	204	**ccQuestions=75.45%**
statement	7005	1958	**5047**	**ccStatements=72.05%**

H=73.71%

Combine the predictions of different classifiers

- use 5 different classifiers
  - ∗ logistic regression
  - ∗ J48 decision tree
  - ∗ JRip decision rules
  - ∗ sequential minimal optimization algorithm
  - ∗ multilayer perceptron

- each classifier makes a class prediction (question / statement)

- the final decision is made by a majority vote
  - ∗ if at least 3 classifier assign the utterance to class "question"
    - → utterance assigned to class "question"

Combine the predictions of different classifiers

Average performance obtained with all 5 classifiers
and with their combination (by majority vote)

	LR	J48	JRip	SMO	MP	combination
Automatic	72.04	72.21	72.81	69.56	72.07	72.66
Manual	73.34	74.26	74.12	72.09	74.33	74.91

Summary

1. Context and approach
2. Prosodic and linguistic features
3. Experiments
4. Conclusions and future work

Conclusions and future work

- Conclusions
  - ∗ the prosodic classifier gives poor classification results
  - ∗ the linguistic classifier provides by far better results (72% on ASR transcripts, 74% on manual transcripts)
  - ∗ the combination of prosodic and linguistic features provides a slight improvement when applied on automatic transcriptions
  - ∗ all 13 features are useful in detecting questions and statements
- Investigate further
  - ∗ the use of confidence measures inside the classifier

Thank you
for your attention !

Annexe

Confusion matrix between questions and statements

	number	classified as question	classified as statement
question	831	**627**	204
statement	7005	1958	**5047**

ccQuestions=75.45%
ccStatements=72.05%
H=73.71%

- **Precision and recall on questions**

$$\begin{aligned} Qprecision &= \frac{627}{627+1958} = 24.26\% \\ Qrecall &= \frac{627}{627+204} = 75.45\% \end{aligned} \Rightarrow Qfmeasure = 36.72\%$$

- **Precision and recall on statements**

$$\begin{aligned} Sprecision &= \frac{5047}{5047+204} = 96.12\% \\ Srecall &= \frac{5047}{5047+1958} = 72.05\% \end{aligned} \Rightarrow SFmeasure = 82.36\%$$

- **weighted average F-measure** $= 77.52\%$