Detection of sentence modality on French automatic speech-to-text transcriptions

Luiza Orosanu, Denis Jouvet

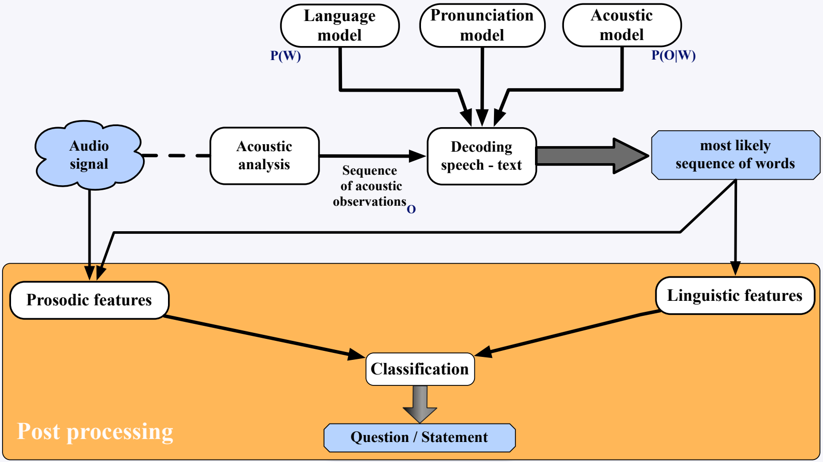
INRIA-Loria, Nancy, France

Multispeech Team

Summary

Context

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

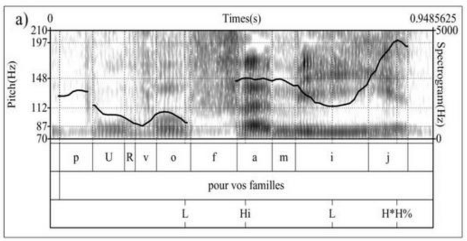


Context

Two types of questions

- expressed with interrogative forms
  - qu'est ce qu'on doit comprendre ?
  - est ce que vous souhaitez une confrontation ?
  - quelles sont les grandes annonces hein à attendre ?

- perceived as questions only through the intonation



Context

- study several approaches
  - prosodic classifier: uses intonation
  - linguistic classifier: uses the linguistic information
  - combined classifier: uses both types of information

Summary

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

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 * $\text{globalSlopeLength}^2$
	lastF0Level	= the last F0 level (normalized by speaker)

Linguistic features (#3)

- **iP**: the interrogative patterns

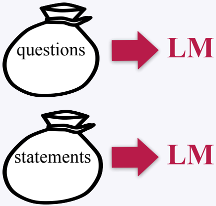
  → indicate the presence or absence
  of an interrogative pattern in a phrase

* quel
* quelle
* quels
* quellles
* comment
* combien
* pourquoi
* est ce que
* est ce qu'
* qu' est ce
* qu' est ce que
* qu' est ce qu'

Linguistic features (#3)

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

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



∗ $\mathrm{LLR} \geq 0 \rightarrow$ likely to be a question
∗ $\mathrm{LLR} < 0 \rightarrow$ 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 * $\text{globalSlopeLength}^2$
	lastF0Level	=	the last F0 level (normalized by speaker)

Summary

Summary

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.

$\Downarrow$

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

$\Downarrow$

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

- **4 classifiers**
  - LR (logistic regression)
  - J48 (decision tree)
  - JRip (decision rules)
  - MLP (multi-layer perceptron)

- evaluate classifier using
  - features extracted from **manual transcriptions**
    → ideal conditions - 0% word error rate
  - features extracted from **automatic transcriptions**
    → real conditions - 26% word error rate

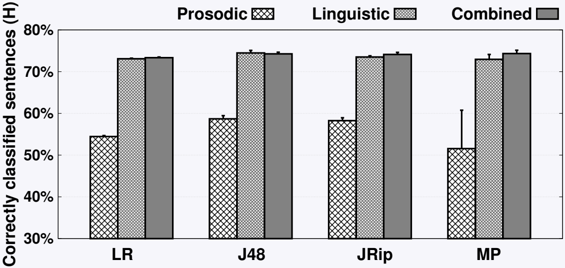
- performance

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

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

Summary

Results on manual transcriptions



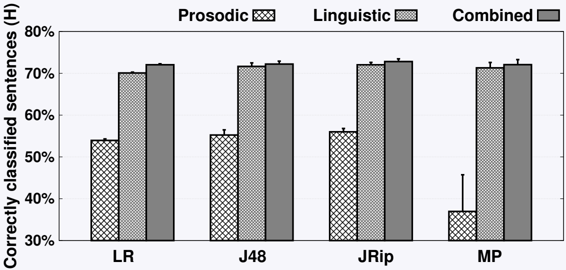
Analysis of the average classifier's performance when applied on

manual transcriptions

⇒ the linguistic classifiers outperform the prosodic classifiers

⇒ the combination of linguistic and prosodic features does not provide any significant improvement on manual transcripts

Results on automatic transcriptions



Analysis of the average classifier's performance when applied on **automatic transcriptions**

⇒ the linguistic classifiers outperform the prosodic classifiers

⇒ 3% performance loss between the manual and the automatic transcriptions

⇒ the combination of linguistic and prosodic features provides a slight improvement on automatic transcription

Best results on manual and automatic transcriptions

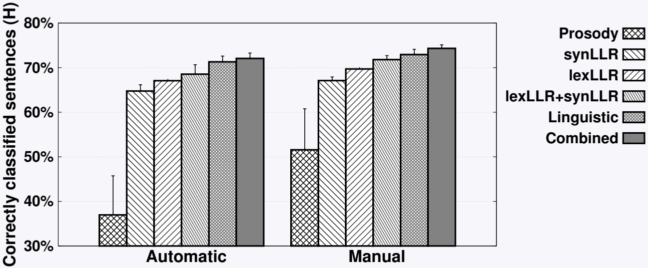
- Confusion matrix between questions and statements obtained on **manual transcriptions (MLP, H=75.05%)**

	number	classified as question	classified as statement	
question	831	**603**	228	**ccQuestions=72.56%**
statement	7005	1559	**5446**	**ccStatements=77.74%**

- Confusion matrix between questions and statements obtained on **automatic transcriptions (MLP, H=73.50%)**

	number	classified as question	classified as statement	
question	831	**611**	220	**ccQuestions=73.60%**
statement	7005	1863	**5142**	**ccStatements=73.41%**

Impact of different feature combinations



Analysis of the average performance obtained with the MLP classifier when using **different feature combinations** on automatic and manual transcriptions

⇒ the most important linguistic feature is the lexical log-likelihood ratio (lexLLR)

⇒ the best results are obtained when combining all features

Impact of sentence boundaries

Assess the performance loss when the **sentence boundaries** are not perfect

$\rightarrow$ change the predefined sentence boundaries

- $*$ by shifting each boundary (left and right) with a random value of {-300, -200, -100, +100, +200, +300}ms
- $*$ by shifting each boundary (left and right) with a random value of {-1000, -800, -600, -400, -200, +200, +400, +600, +800, +1000}ms
- $*$ by finding the longest silence-enclosed sentence

Impact of sentence boundaries

Reference sentence:

"que fallait -il faire" [947090,948370]

946230	946290	++micro++	60 [ms]
946300	946350	à	50 [ms]
946360	946660	travers	300 [ms]
946670	946760	le	90 [ms]
946770	947020	monde	250 [ms]
947030	947160	**<sil>**	130 [ms]
947230	947450	**++resp++**	220 [ms]
947460	947650	**que**	190 [ms]
947660	947920	**fallait**	260 [ms]
947930	948080	**-il**	150 [ms]
948090	**948350**	**faire**	260 [ms]
948360	948390	eh	30 [ms]
948400	948530	bien	130 [ms]
948540	948670	il	130 [ms]
948680	948870	fallait	190 [ms]
948880	949310	choisir	430 [ms]
949320	949400	++rire++	80 [ms]

Impact of sentence boundaries

Modified borders ± 300ms: [+200,+300ms]

"que fallait -il faire eh bien il" [947290,948670]

946230	946290	++micro++	60 [ms]
946300	946350	à	50 [ms]
946360	946660	travers	300 [ms]
946670	946760	le	90 [ms]
946770	947020	monde	250 [ms]
947030	947160	<sil>	130 [ms]
947230	947450	**++resp++**	220 [ms]
947460	947650	**que**	190 [ms]
947660	947920	**fallait**	260 [ms]
947930	948080	**-il**	150 [ms]
948090	948350	**faire**	260 [ms]
948360	948390	**eh**	30 [ms]
948400	948530	**bien**	130 [ms]
948540	**948670**	**il**	130 [ms]
948680	948870	fallait	190 [ms]
948880	949310	choisir	430 [ms]
949320	949400	++rire++	80 [ms]

Impact of sentence boundaries

Modified borders ± 1000ms: [-400ms,-600ms]

"le monde que fallait" [946690,947770]

946230	946290	++micro++	60 [ms]
946300	946350	à	50 [ms]
946360	946660	travers	300 [ms]
946670	946760	**le**	90 [ms]
946770	947020	**monde**	250 [ms]
947030	947160	**<sil>**	130 [ms]
947230	947450	**++resp++**	220 [ms]
947460	947650	**que**	190 [ms]
947660	**947920**	**fallait**	260 [ms]
947930	948080	-il	150 [ms]
948090	948350	faire	260 [ms]
948360	948390	eh	30 [ms]
948400	948530	bien	130 [ms]
948540	948670	il	130 [ms]
948680	948870	fallait	190 [ms]
948880	949310	choisir	430 [ms]
949320	949400	++rire++	80 [ms]

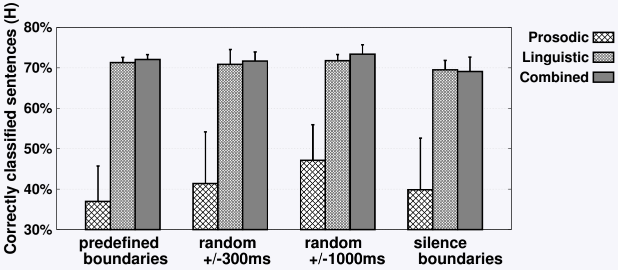
Impact of sentence boundaries

Modified borders: the longest silence-enclosed sentence

"que fallait -il faire eh bien il fallait choisir" [947460,949310]

946230	946290	++micro++	60 [ms]
946300	946350	à	50 [ms]
946360	946660	travers	300 [ms]
946670	946760	le	90 [ms]
946770	947020	monde	250 [ms]
947030	947160	<sil>	130 [ms]
947230	947450	++resp++	220 [ms]
947460	947650	**que**	190 [ms]
947660	947920	**fallait**	260 [ms]
947930	948080	**-il**	150 [ms]
948090	948350	**faire**	260 [ms]
948360	948390	**eh**	30 [ms]
948400	948530	**bien**	130 [ms]
948540	948670	**il**	130 [ms]
948680	948870	**fallait**	190 [ms]
948880	**949310**	**choisir**	430 [ms]
949320	949400	++rire++	80 [ms]

Impact of sentence boundaries



Analysis of the average performance obtained with the MLP classifier on automatic transcriptions when modifying the predefined boundaries

⇒ even if an automatic segmentation module wrongly assigns the sentence boundaries, our classifier still manages to correctly classify the question/statements entries between 69% and 72%

Summary

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
  - * even if an automatic segmentation module wrongly assigns the sentence boundaries, our classifier still manages to correctly classify the question/statements entries between 69% and 72%
- Investigate further
  - * the use of confidence measures inside the classifier

Thank you
for your attention !