Speech Processing And Prosody

Denis Jouvet
TSD 2019
11-13 September 2019

LORIA – Inria – Nancy - France

Speech Processing and Prosody

- Prosody conveys various types of information over the linguistic content
  - Prosody structures the utterances
  - May be used to emphasized words
  - Speaker emotional state
  - …

- Speech prosody neglected
  - In automatic speech recognition
  - In manual transcriptions
- But critical for expressive speech synthesis

- Prosody is a suprasegmental information, and is characterized by
  - Duration of the sounds
  - Fundamental frequency
  - Energy of the sounds

Outline

- Prosodic features, computation and reliability
  - Phone duration
  - Fundamental frequency
  - Phone energy

- Prosodic features in automatic speech processing
  - Computer assisted language learning
  - Structuring speech utterances
  - Sentence modality
  - Prosodic correlates of discourse particles
  - Expressive speech

- Conclusion

Phone duration

- Is determined from the phone boundaries that can be set
  - Manually
  - Automatically through forced speech-text alignment

- Some boundaries are clear, some are more ambiguous, for example
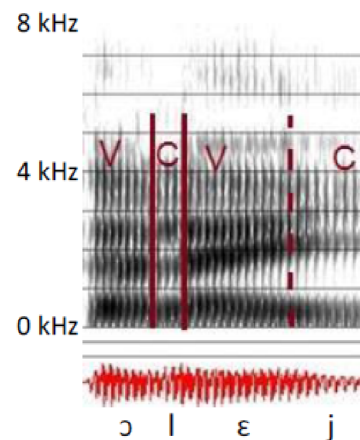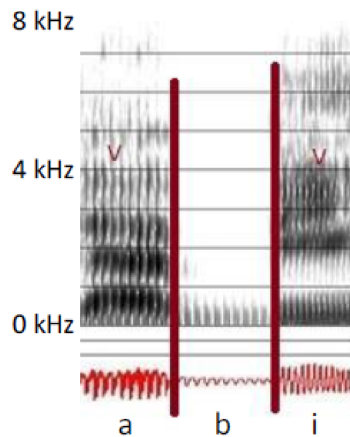
  - Clear between vowel and occlusive
  - Ambiguous between vowel and semi-vowel

Automatic speech-text alignment

- Needs only a manual transcription of the speech signal into words
  ⇨ sequence of words corresponding to the speech segment
- Uses pronunciation variants for each word (lexicon or grapheme-to-phoneme tools)
- Relies on automatic speech recognition tools
  ⇨ find the sequence of phones that best matches with the speech signal
     (and the associated word and phone boundaries)

- Works well when
  - Good quality speech data and reliable acoustic models
  - Transcription perfectly matches with the actual content
  - Pronunciation variants include the actual pronunciations

- Performance degrades
  - On noisy speech data
  - On non-native speech (difficult to predict every possible pronunciation deviations)
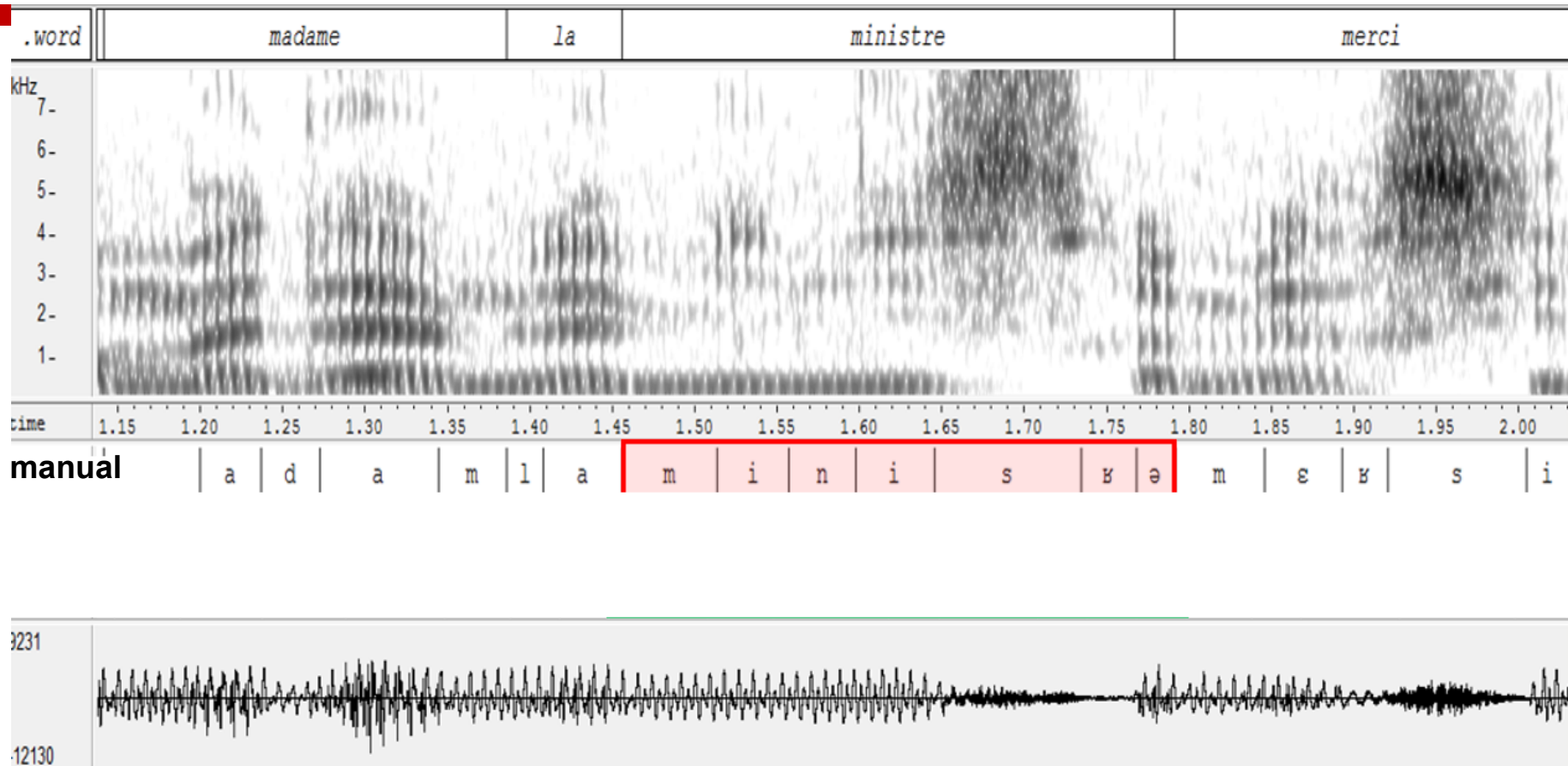
Automatic speech-text alignment

- Example for « *Madame la Ministre, merci* » (⇔ Madame Minister thanks)
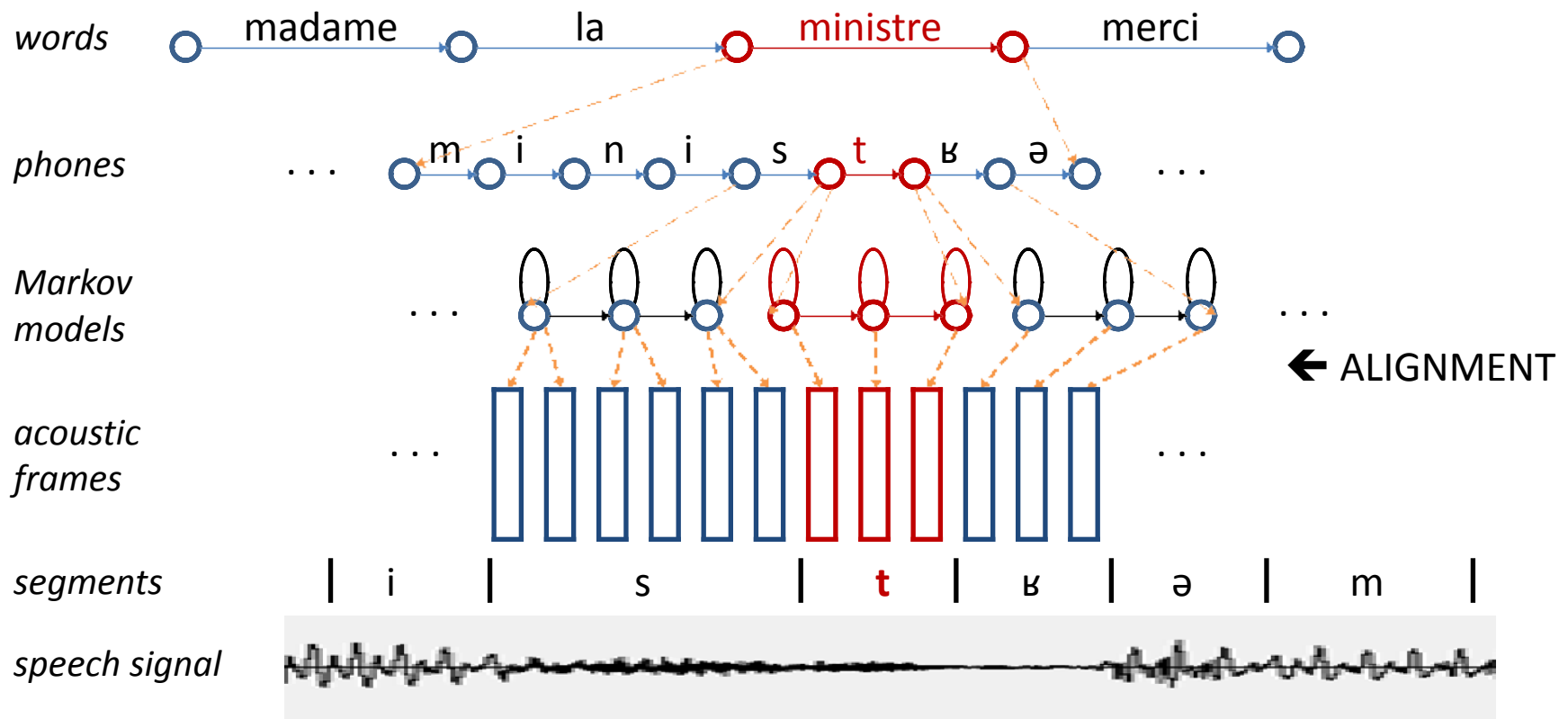
- 3 states per phone model * 10 ms per frame ➜ 30 ms minimum duration per phone

Example of speech segmentation

- absent in manual segmentation
- very short in automatic segmentation with 5 ms frame shift
- 30 ms long in automatic segmentation with 10 ms frame shift

[t]

Analysis of final consonantal clusters

- Analysis of a frequent final cluster / t ʁ / as in / m i n i s t ʁ / (*ministre*)

- Extended pronunciation lexicon where all pronunciation variants are allowed
  - Adding final schwa / ə /
  - Eliding consonants / t / or/and / ʁ /

- This leads to an extended set of pronunciation variants
  Example for *ministre*:

/ m i n i s t ʁ ə / [+t][+ʁ][+ə]	/ m i n i s ʁ ə / [-t][+ʁ][+ə]	⎤
/ m i n i s t ʁ / [+t][+ʁ][-ə]	/ m i n i s ʁ / [-t][+ʁ][-ə]	/ʁ/ pronounced
/ m i n i s t ə / [+t][-ʁ][+ə]	/ m i n i s ə / [-t][-ʁ][+ə]	⎤
/ m i n i s t / [+t][-ʁ][-ə]	/ m i n i s / [-t][-ʁ][-ə]	/ʁ/ elided

  /t/ pronounced   /t/ elided

Comparing frequency estimations

- Word *ministre*

- Comparing frequencies estimated with 5 and 10 ms frame shifts
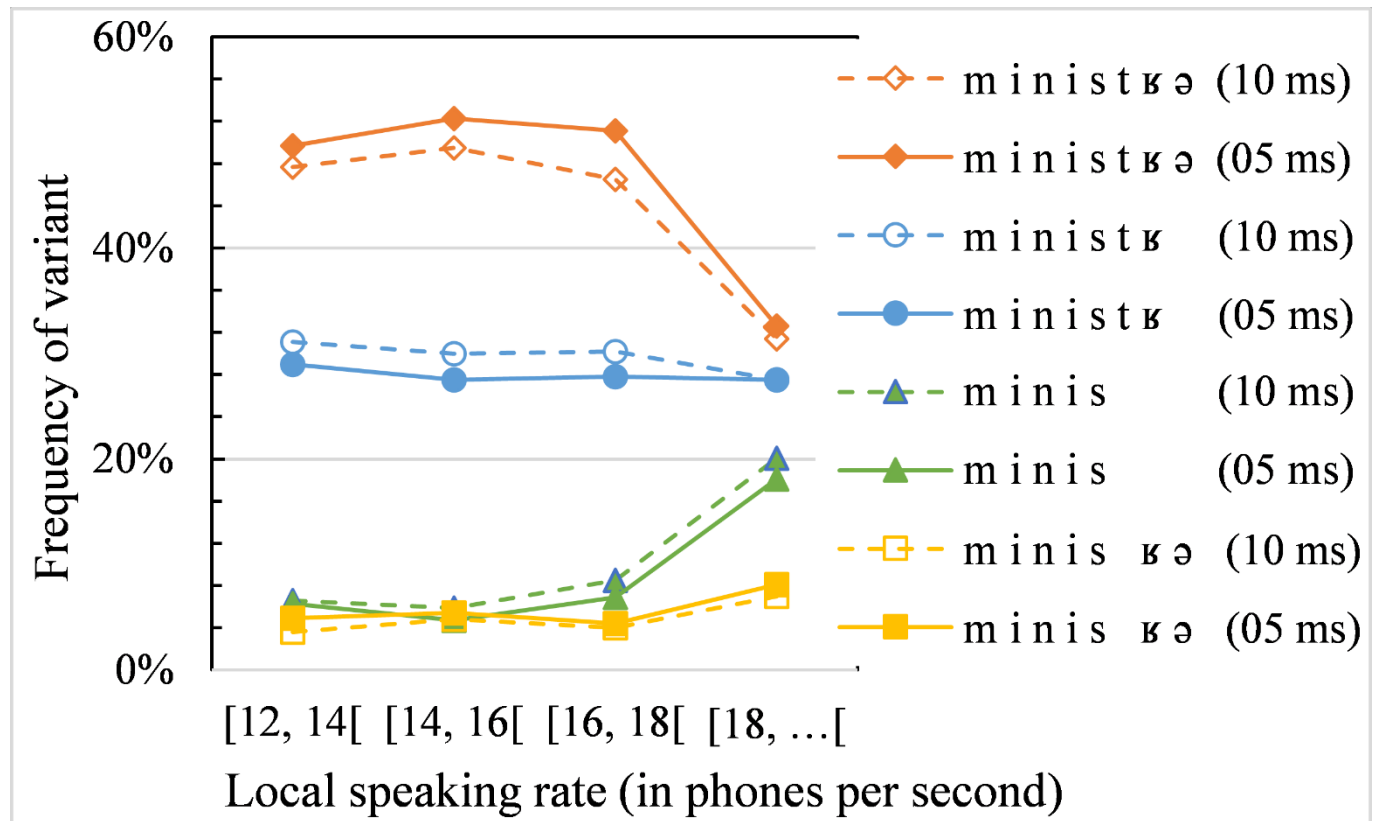
- 5 ms frame shift acoustic analysis leads to higher frequency of occurrences for longest pronunciation variant (here / m i n i s t ʁ ə /)

Speech-text alignment

- Besides correct transcription, adequate pronunciation variants, …
  better to rely on a 2 pass process
  - First, determine the pronunciation variants actually used
    with context-dependent models
  - Then, re-align with context-independent acoustic models
    which leads to a better precision of the boundaries

- To get a better precision
  - Use 5 ms frame shift
    Note, that is what is done in parametric speech synthesis

- Other difficulties stem from
  - Non adequate noise models
  - Annotation conventions for noises, laughing, hesitations, …, that vary among corpora

Fundamental frequency (F0)

- Fundamental frequency vs. pitch
  - Pitch is linked to the perception of the frequency
  - F0 is a physical property of the sounds
- However the term 'pitch' is often used when talking about the F0

- F0 detection can be done
  - In the time domain
  - In the spectral domain
  - Using both time and spectral domains

F0 detection – time domain

$$T = \frac{1}{F0}$$
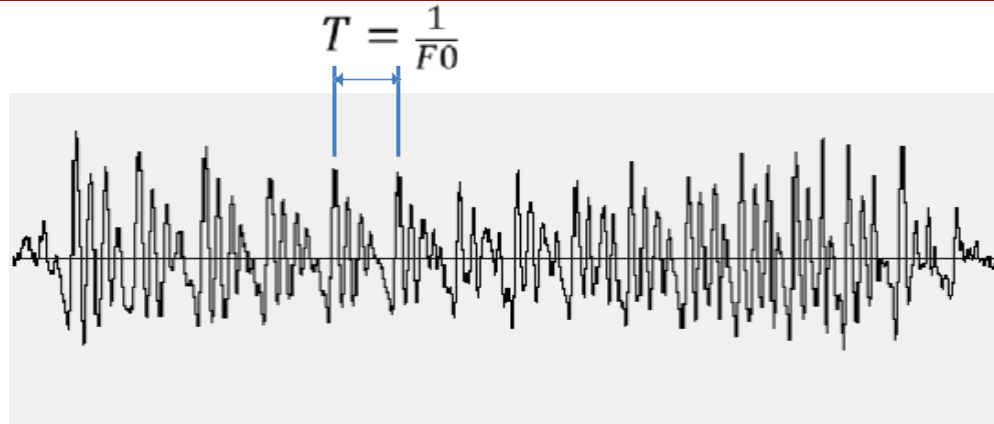
- Rely on the time shift at which the signal (almost) repeat itself (in voiced sounds)
  - ACF   (Praat)   Auto Correlation Function
  - AMDF   (snack library)   Average Magnitude Difference Function
  - CCF   (Praat)   Cross Correlation Function
  - Kaldi   (speech recognition toolkit)
  - REAPER   (REAPER)
  - RAPT   (SPTK and snack library)   Robust Algorithm for Pitch Tracking
  - SRPD   (ESTL)   Super Resolution Pitch Determinator
  - TEMPO   (STRAIGHT)
  - YIN   (YIN and JSNOORI)

F0 detection – frequency domain
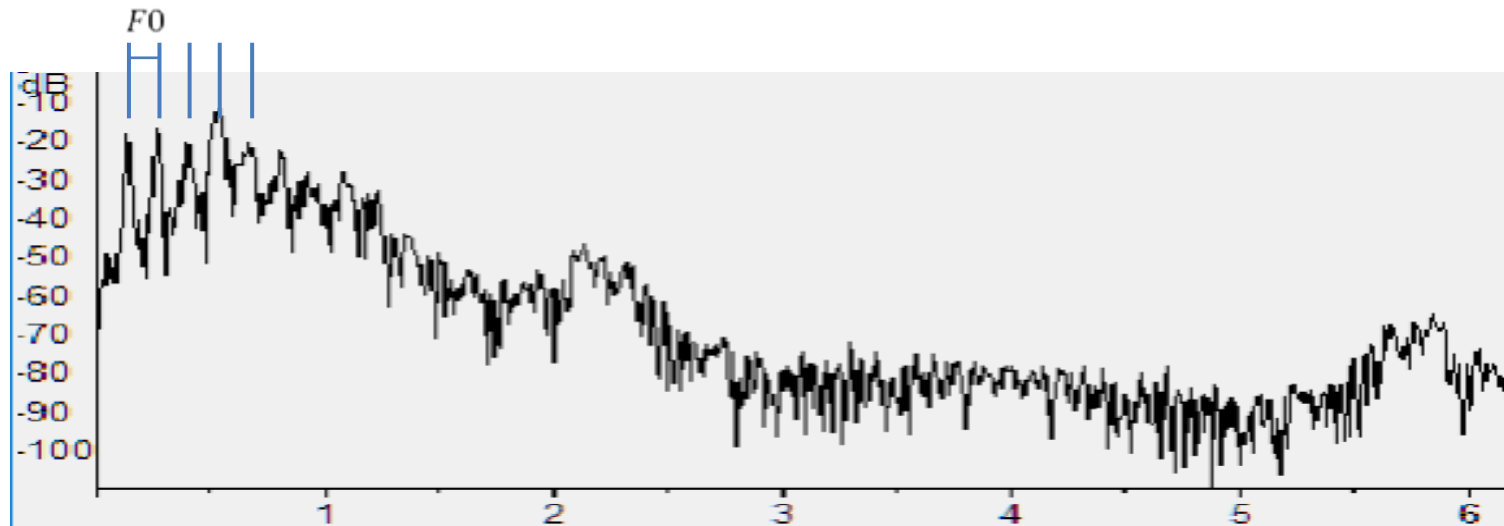
- Exploit the harmonic structure of the spectrum for voiced sounds
  - Martin   (JSNOORI)
  - SHS   (Praat)   Sub-Harmonic Summation algorithm
  - SWIPE   (SPTK and JSNOORI)   Sawtooth Waveform Inspired Pitch Estimator

F0 detection – combined approaches

- Combine time and frequency cues
    - Aurora   (ETSI)
    - NDF   (STRAIGHT)   Nearly Defect-free F0

F0 detection – comments

- Time and frequency approaches provide F0 candidates
  - Main challenge is to select the "good" candidate
    and to avoid pitch halving (F0/2) or doubling (2*F0) estimations
    which lead to the numerous variants

- Voicing decision is a critical step
  [unvoiced sounds and silence ➔ no F0 values; voiced sounds ➔ F0 values]
  - Usually carried on by applying thresholds on numerical criteria used to compute F0

- Dynamic programming-based post processing in some approaches
  - E.g., RAPT, REAPER, Martin
  - For minimizing jumps in the F0 curve (thus reducing halving and doubling errors, and to improve voicing decision)

Performance evaluation measures

- VDE: Voicing Decision Error
  - Proportion of frames for which a voicing decision error is made
  - Two types of errors
    - v->uv ⇔ voiced frame classified as unvoiced
    - uv->v ⇔ unvoiced frame classified as voiced

- FFE: F0 Frame Error
  - Provides a global error measure
  - Consider as error
    - Voicing decision error (v->uv and uv->)
    - Gross pitch error (voiced frame classified as voiced, but estimated F0 differs from the reference F0 by more than 20%)
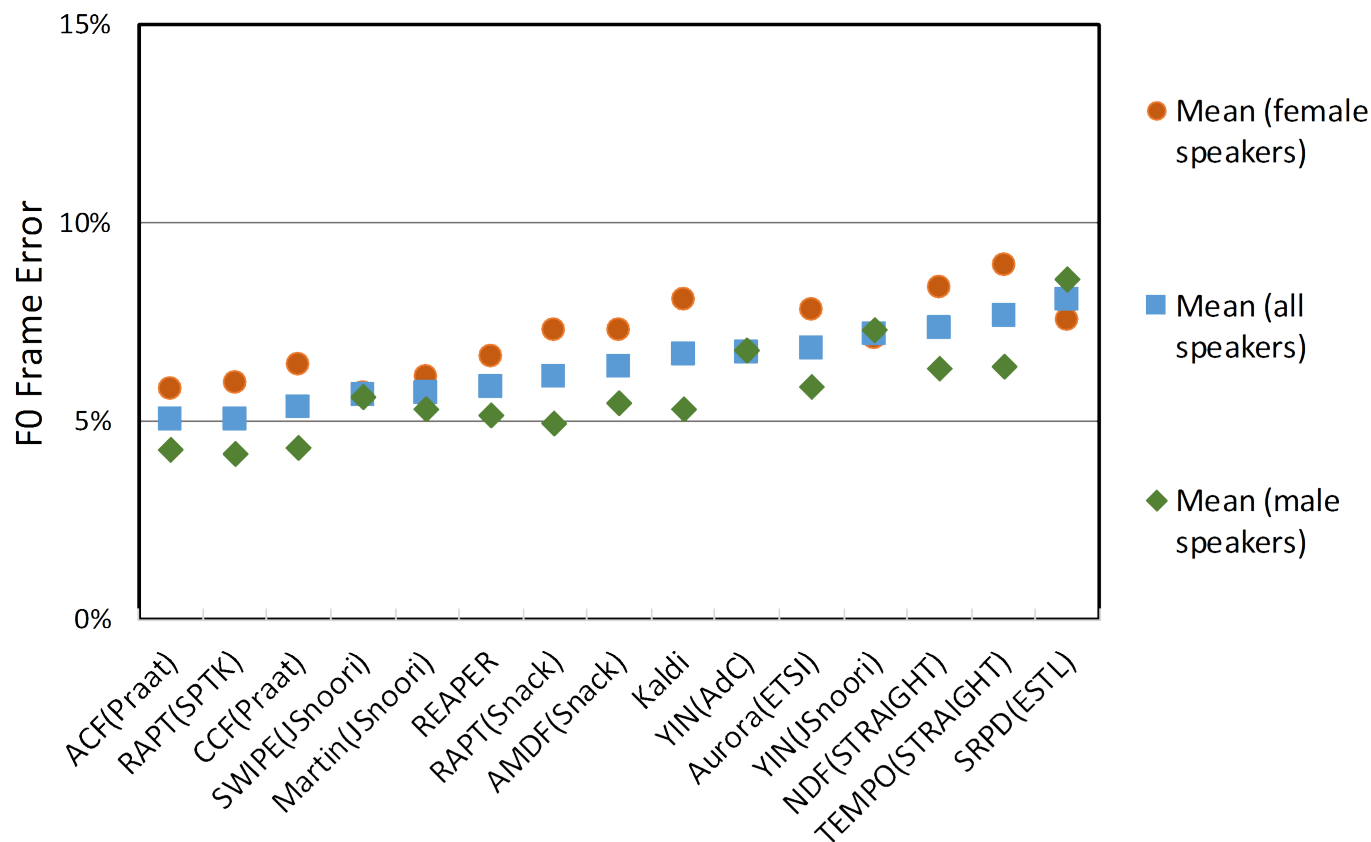
Evaluation on clean data

PTDB-TUG corpus, 20 speakers, 4720 utterances
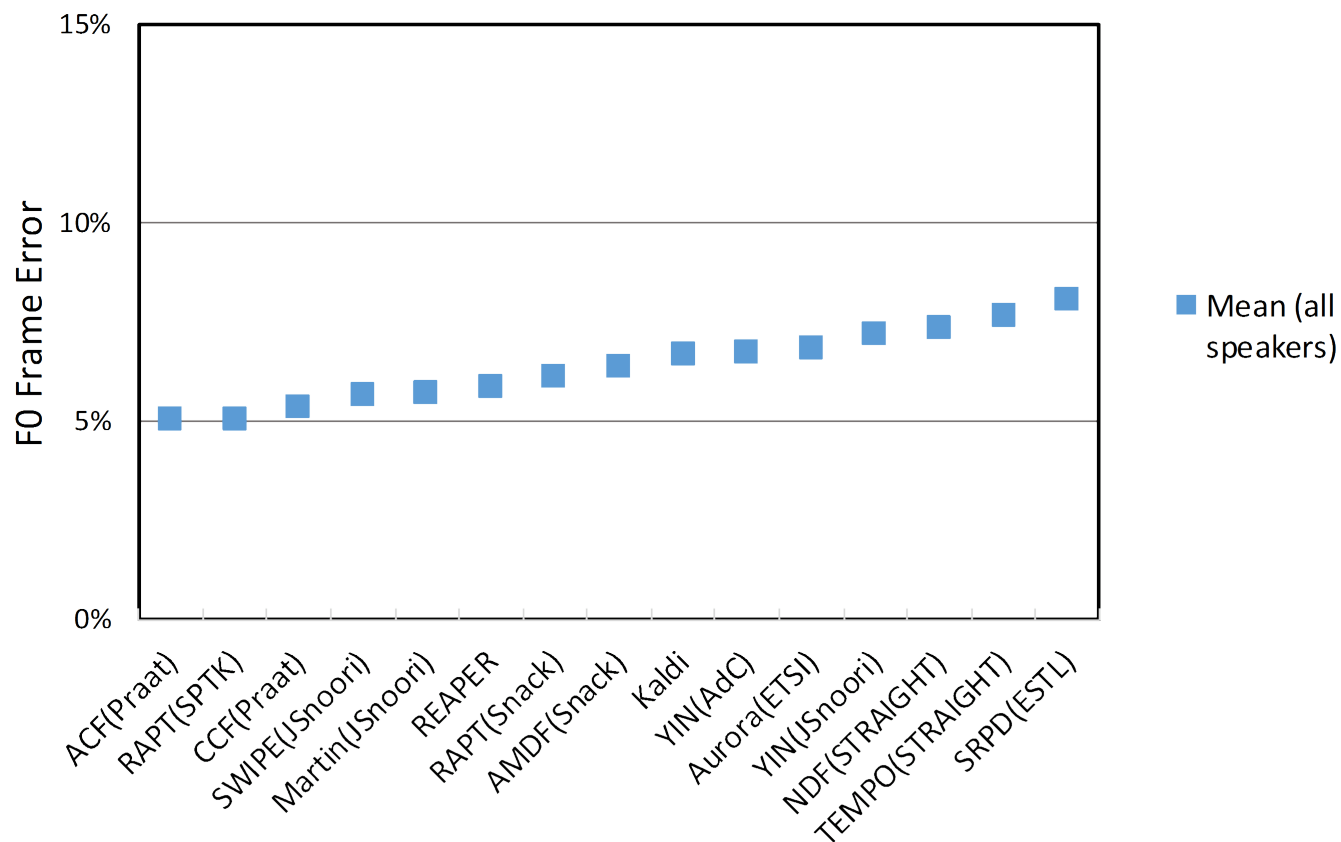
- Mean (over all speakers) ranges from 5% to 8%
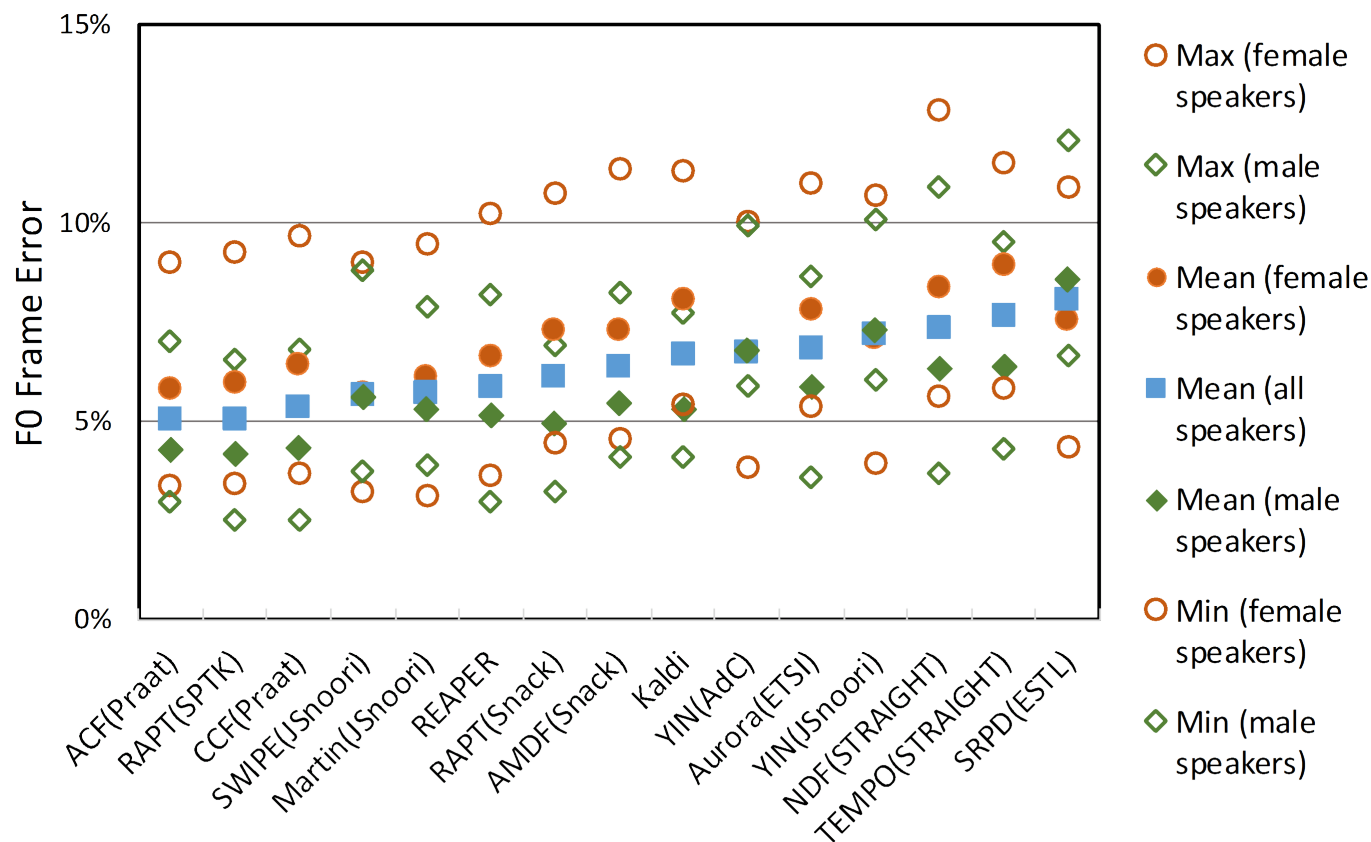
Evaluation on clean data

PTDB-TUG corpus, 20 speakers, 4720 utterances

- Mean (over all speakers) ranges from 5% to 8%
- Except SWIPE and YIN, better results on male speakers than on female speakers

Evaluation on clean data

PTDB-TUG corpus, 20 speakers, 4720 utterances

- Mean (over all speakers) ranges from 5% to 8%
- Except SWIPE and YIN, better results on male speakers than on female speakers
- Large gap in performance between best and worst speaker (for all approaches)

Evaluation on simulated noisy data

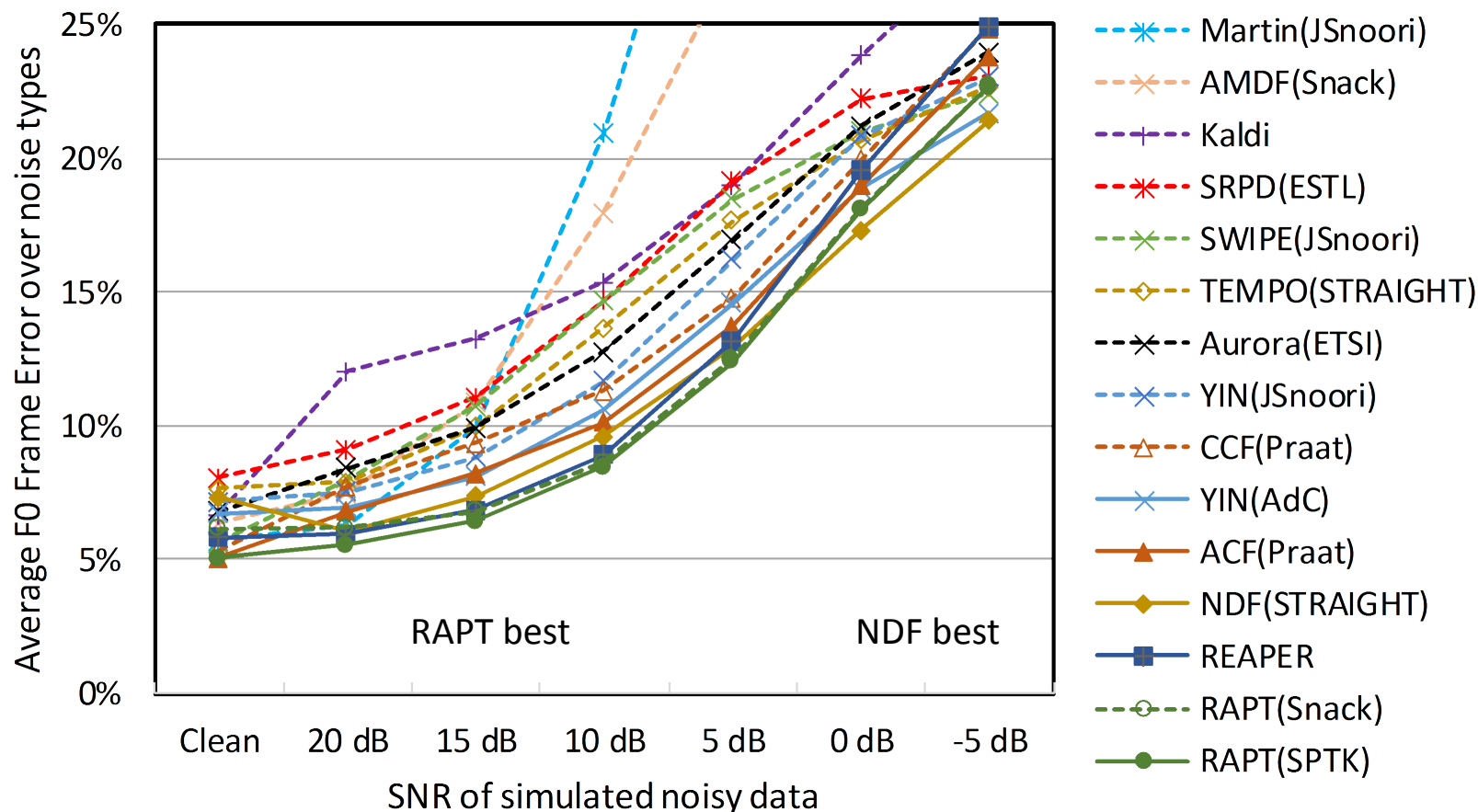## PTDB-TUG corpus, noises (babble, factory, …) added at various SNR levels
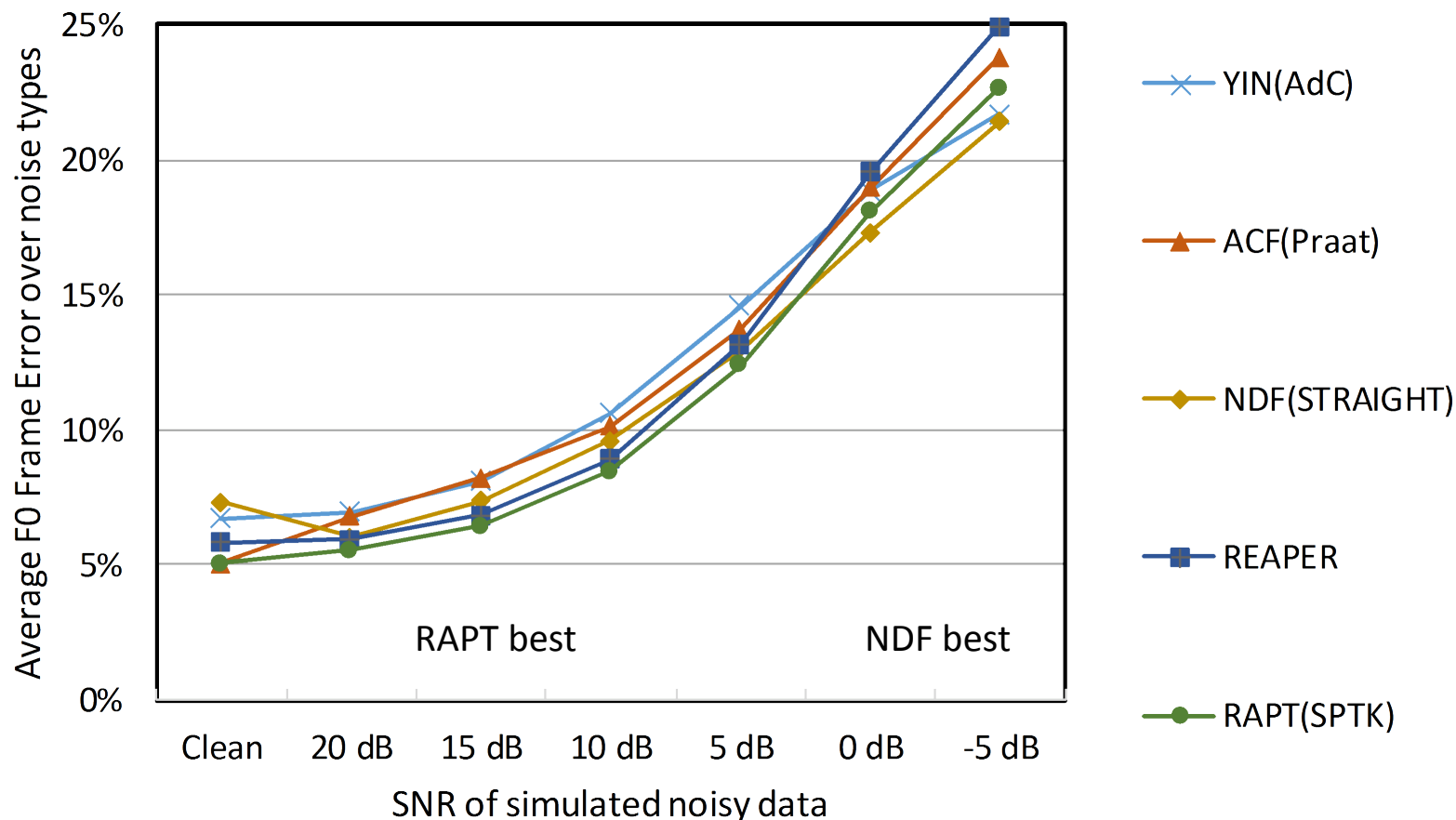
- Most approaches have the same behavior (ending at around 25% FFE for -5 dB SNR)

Evaluation on simulated noisy data

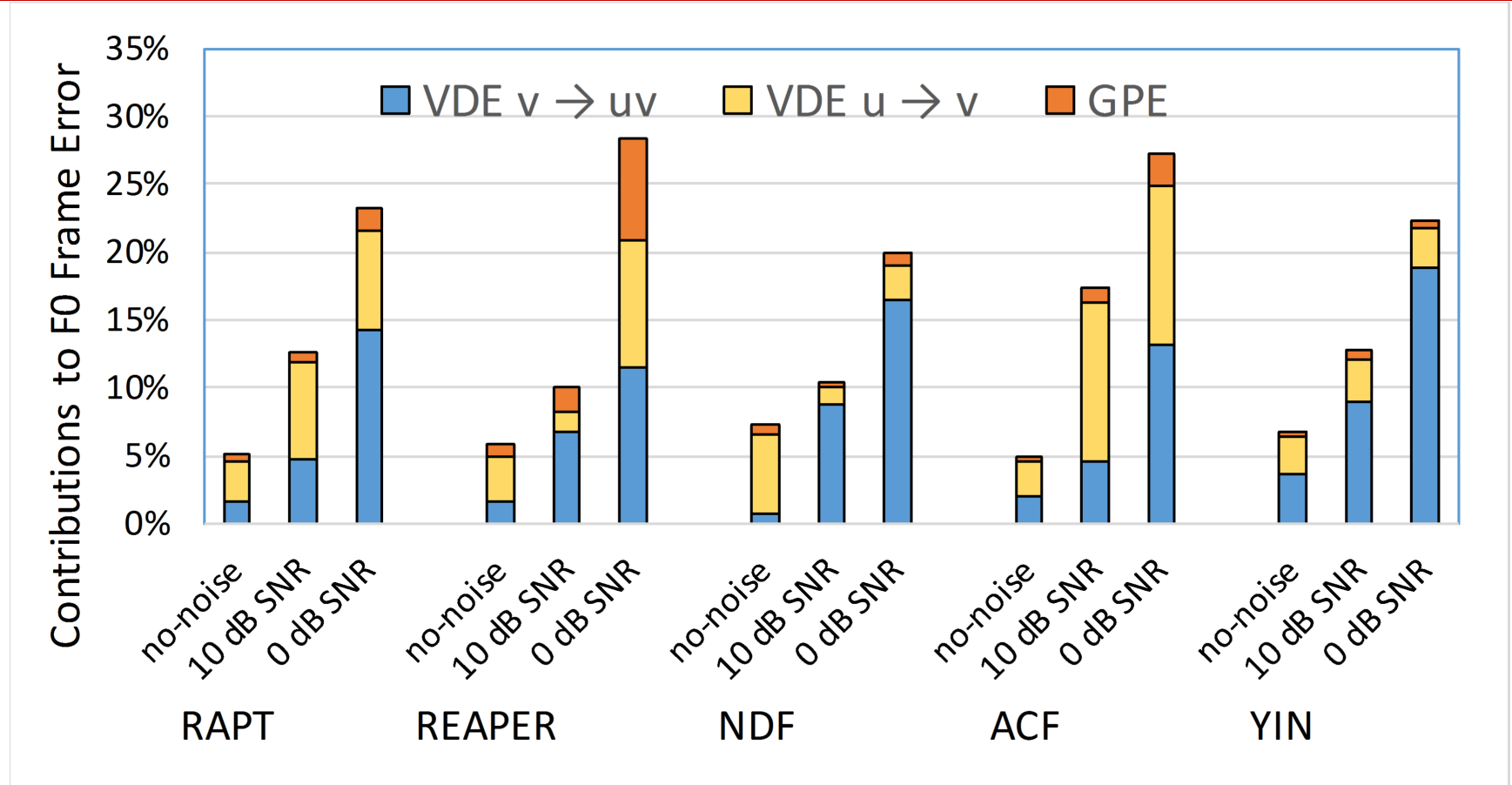## PTDB-TUG corpus, noises (babble, factory, …) added at various SNR levels

- Most approaches have the same behavior (ending at around 25% FFE for -5 dB SNR)
- A large part of the errors are due to voicing decision errors
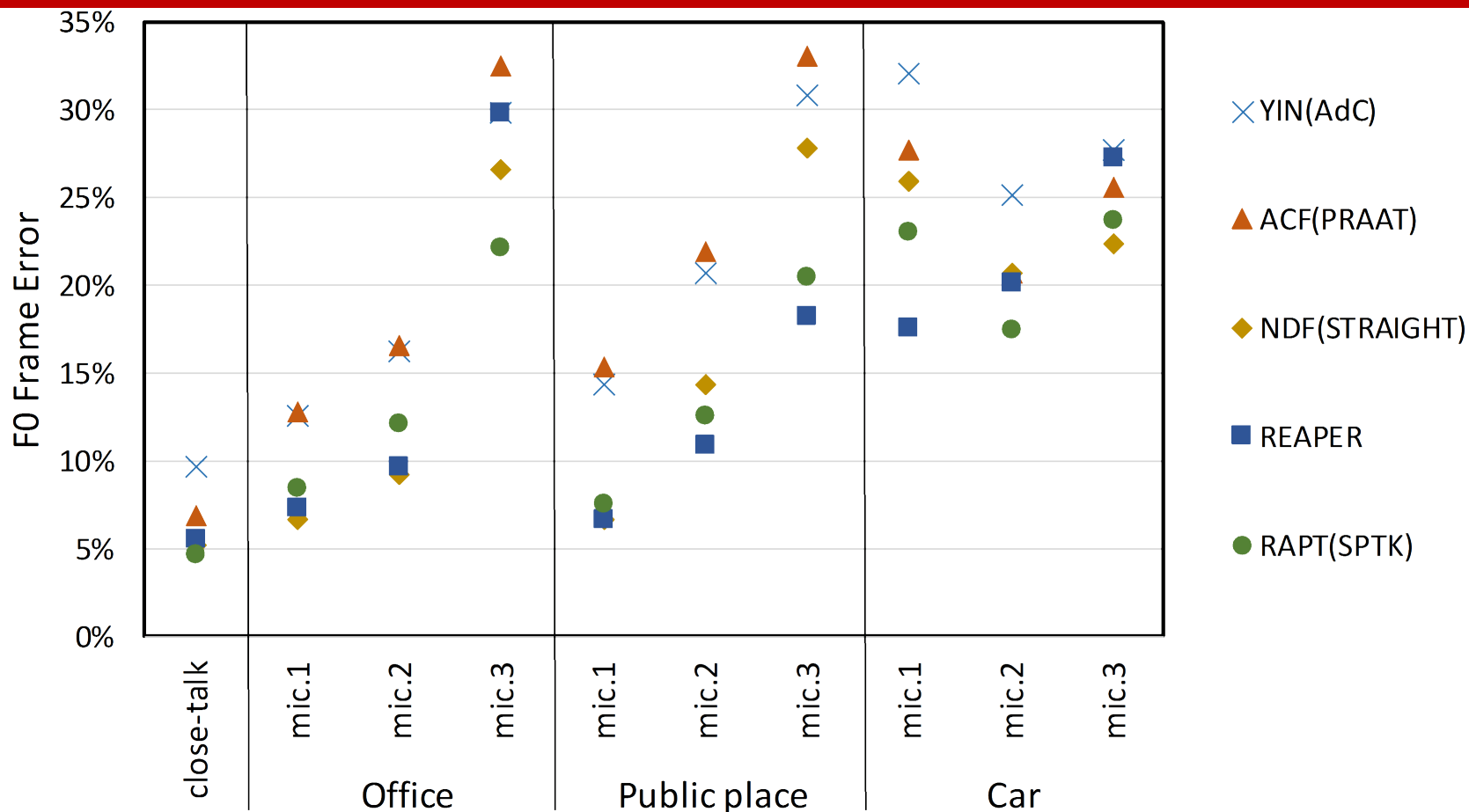
Voicing decision errors

PTDB-TUG corpus, noises (babble, factory, …) added at various SNR levels

- When noise increases, the largest part of the errors comes from v → uv decision errors
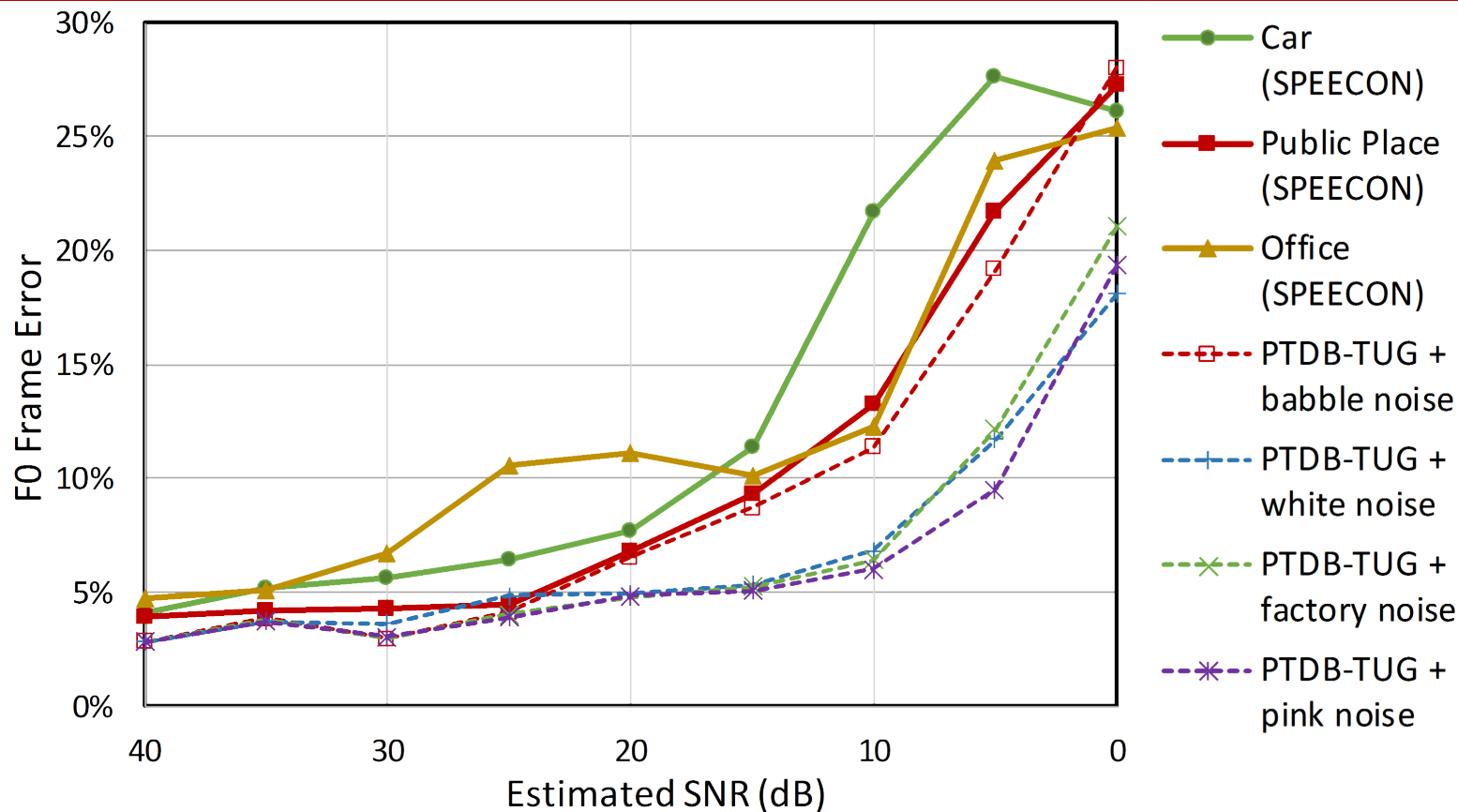
Evaluation on real noisy data

SPEECON corpus, 60 speakers, car, office and public places, close and distant microphones


- Degradation with noise (distance to speaker)
- Best algorithm vary depending on condition

Comparing performance
on real and simulated noisy data

- Degradation with respect to noise level
- For babble noise (simulated or real public places), results are very similar between simulated noisy data and real noisy data

Comparing performance
on real and simulated noisy data
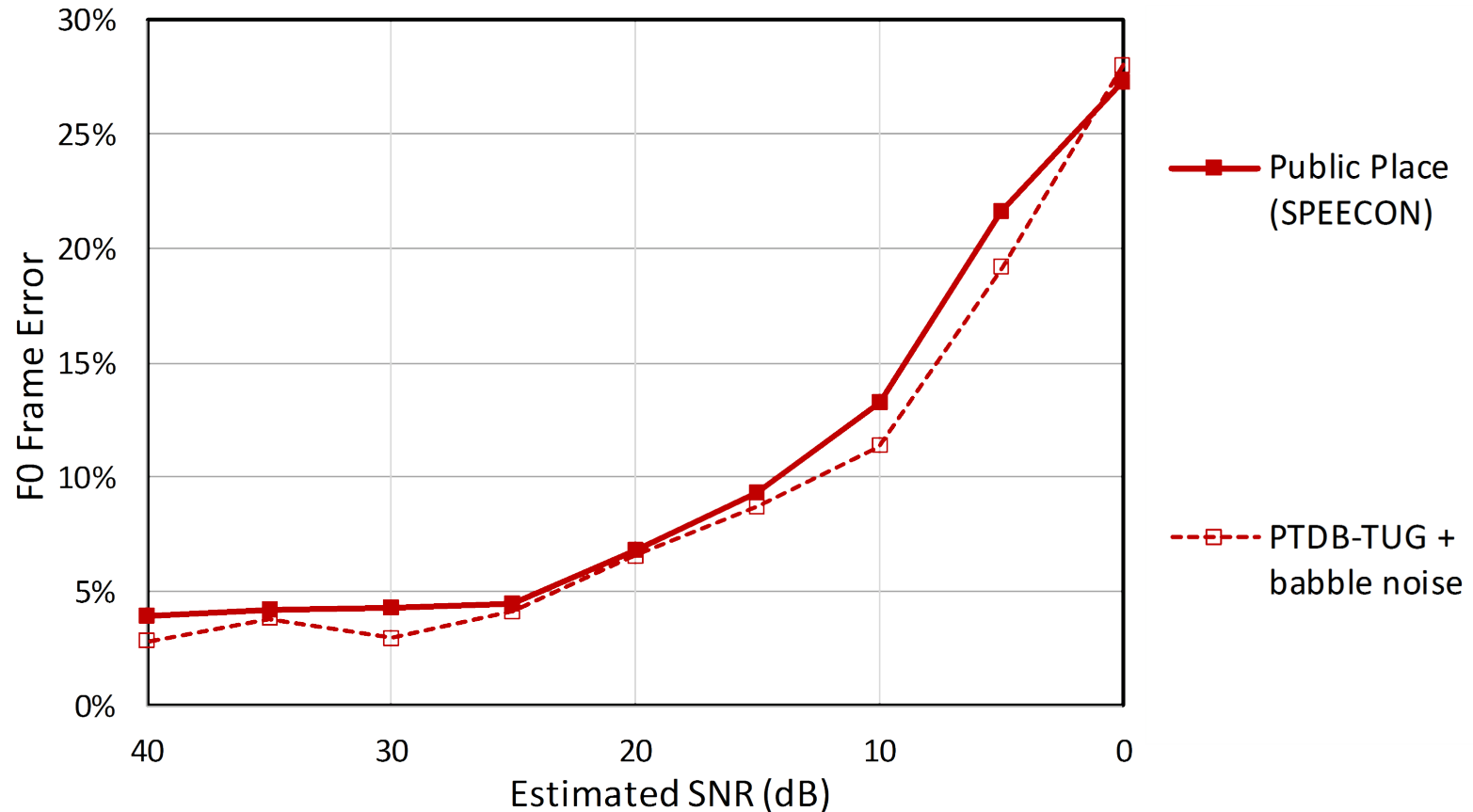
- Degradation with respect to noise level
- For babble noise (simulated or real public places), results are very similar between simulated noisy data and real noisy data

F0 detection

- Most of the algorithms provide good results on clean data (from 5% to 8% FFE)
- But large performance variation across speakers

- Performance degradation when noise is present
- Voicing detection error is the main cause of error
  (in most of the cases, voiced frames are mis-classified as unvoiced)

- Best algorithm vary depending on noise type and level
- RAPT (SPTK), REAPER and NDF (STRAIGHT) are the best approaches
- ACF (Praat), RAPT (SPTK), TEMPO (STRAIGHT), YIN and SWIPE are the most often used
  (according to a recent survey [Strömbergsson, Interspeech 2016])

- Choosing the most adequate algorithm or combining several approaches may be a solution, as well as optimizing the voicing decision

Phone energy

- How to compute it
  - Energy in the middle of the phone segment?
  - Average energy over the whole phone segment?

- Values dependent on many parameters
  - Distance between speaker and microphone
  - Microphone and channel characteristics
  - Signal scaling

- Reasonable feature if comparisons are made inside a given utterance (assuming the speaker does not move to much during an utterance)
- Difficult to have reliable comparisons over different acquisition sessions

Normalizing prosodic features

- Phone duration depends on speaking rate
  - Phone duration ratios are often more relevant
  - Or normalization with respect to speaking rate

- F0 depends on the speaker, and large differences between males and females
  - F0 ratios (when measured in Hz) are more useful
    or delta values in semi-tones
  - Glissando threshold for perception of changing pitch (takes into account pitch variation and duration of the segment)

- Energy depends on many aspects
  - Phone energy ratios (or differences in decibels) are more relevant
  - Or normalization with respect to signal level

Confidence scoring

- Phone boundaries
  - Automatic speech-text alignment provides phone-boundaries but there are no associated confidence score
  - Just very view experiments aiming at computing the posterior probability of the boundary

- F0
  - Algorithms provide F0 values
  - A few of them provide a probability of the voicing feature
  - Some attempts at computing a confidence score on the estimated F0 values

Outline

- Prosodic features, computation and reliability
    - Phone duration
    - Fundamental frequency
    - Phone energy

- **Prosodic features in automatic speech processing**
    - Computer assisted language learning
    - Structuring speech utterances
    - Sentence modality
    - Prosodic correlates of discourse particles
   - Expressive speech

- Conclusion

Computer assisted language learning

- Providing automatic feedback to language learners, on various aspects
  - Implies detecting pronunciation defects
  - Providing reliable feedback

- Detecting pronunciations defects
  - Requires an alignment of the speech signal with the expected pronunciation
    - Pronunciation defects, such as phone insertions and deletions affect the alignment accuracy
    - If mother tongue known, some frequent pronunciation defects may be taken into account to enrich the pronunciation lexicon
  - Scoring pronunciation
    - Phoneme quality (i.e., is it the expected phoneme?) based on GOP (goodness of pronunciation) score
    - Lexical stress requires prosodic features (phone duration, fundamental frequency)
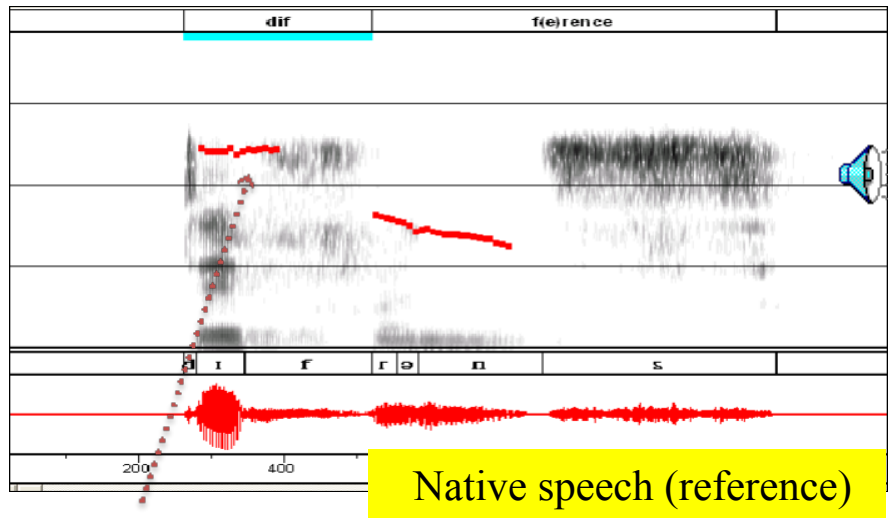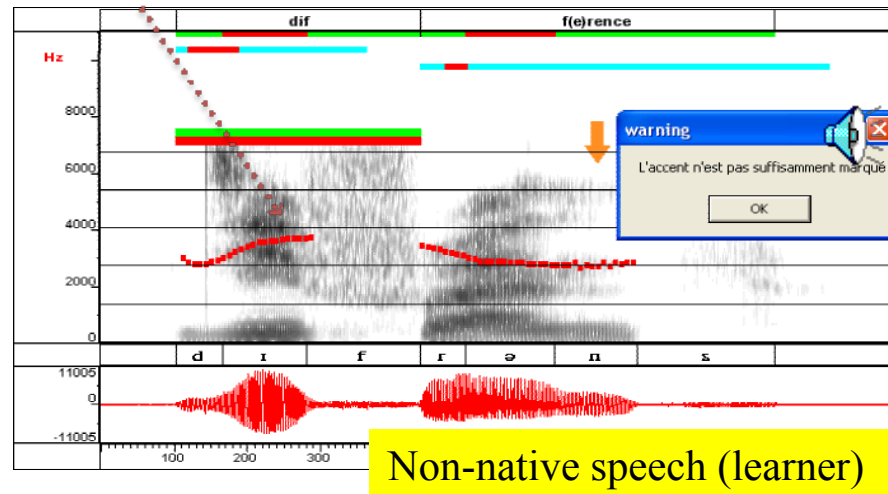
Precision of phone boundaries
on non-native speech

- Percentage of boundaries that are less than 20 ms of the reference boundary
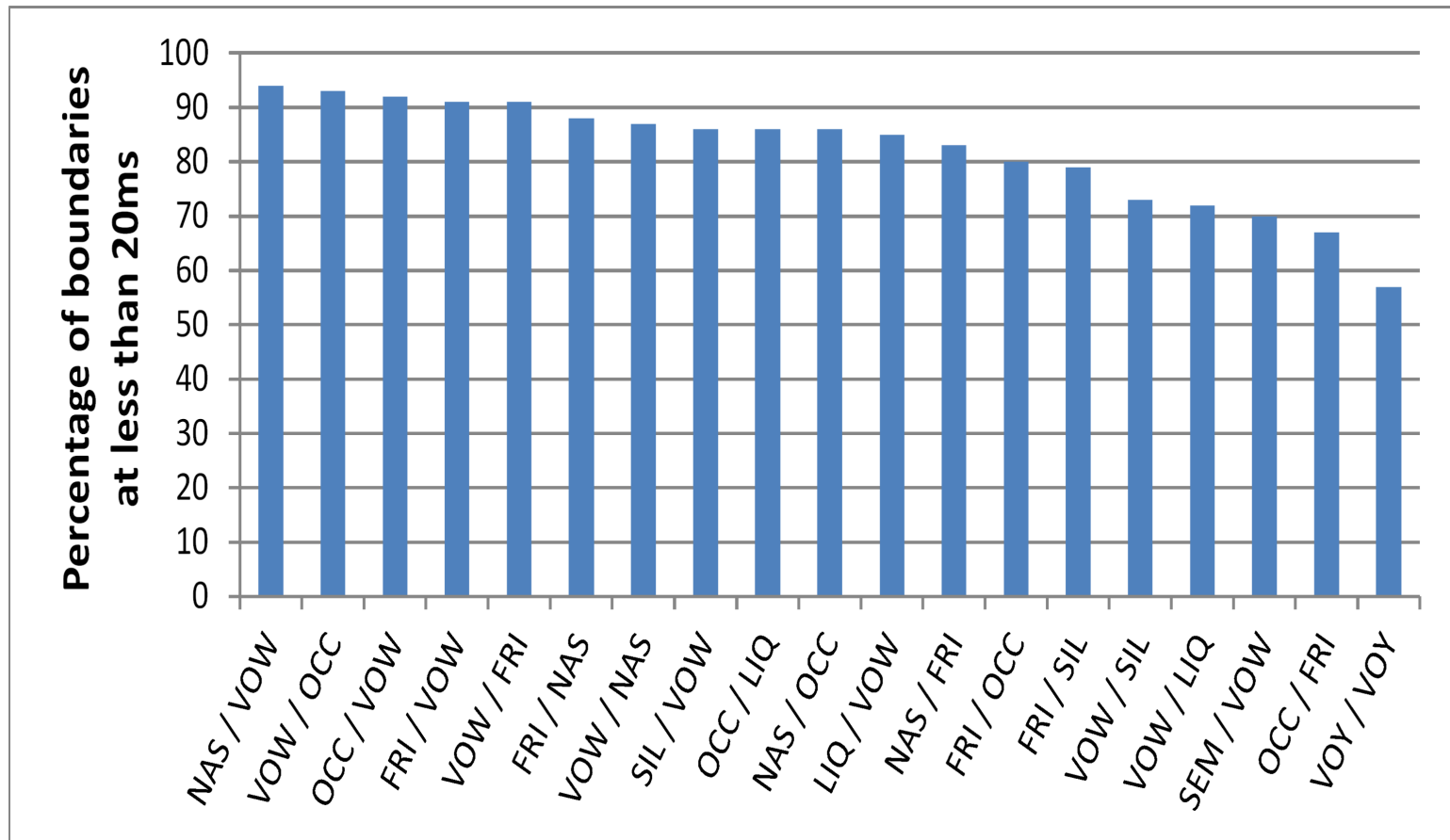
Example of audio & textual prosodic feedback

Native speech (reference)

Melodic curve (in red)

Non-native speech (learner)

Example for the word "*difference*" pronounced by a native speaker (reference) and by a learner

- Leaner: syllable S2 is too long, and syllable S1 is not stressed enough

- After analyzing the pronunciation, a textual diagnosis is provided to the learned, as well as a audio feedback

Modified learner's voice
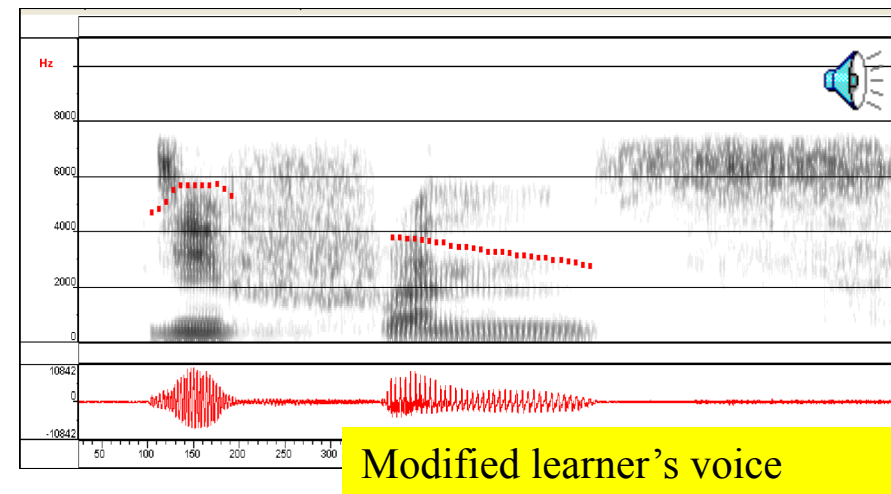
Structuring speech utterances

- Prosody structures speech utterances
  - Prosodic groups
  - Organization of prosodic groups

- Automatic approach for prosodic structure in French based on [Martin, 1987] mainly relies on
  - Amplitude of the F0 slopes
  - Inversion of F0 slopes

  at the end of the potentially stressed groups

Detection of prosodic boundaries

- Subset of ESTER and ETAPE (broadcast news) have been manually segmented in prosodic groups

- Analysis of automatic prosodic boundary detection

Speech data	Number of boundaries in reference data	Percentage		
		Found	Omitted	Inserted
ESTER subset	1405	83%	17%	20%
ETAPE subset	1167	77%	23%	13%

Examples of prosodic trees

TSD 2019 -- Speech processing and prosody
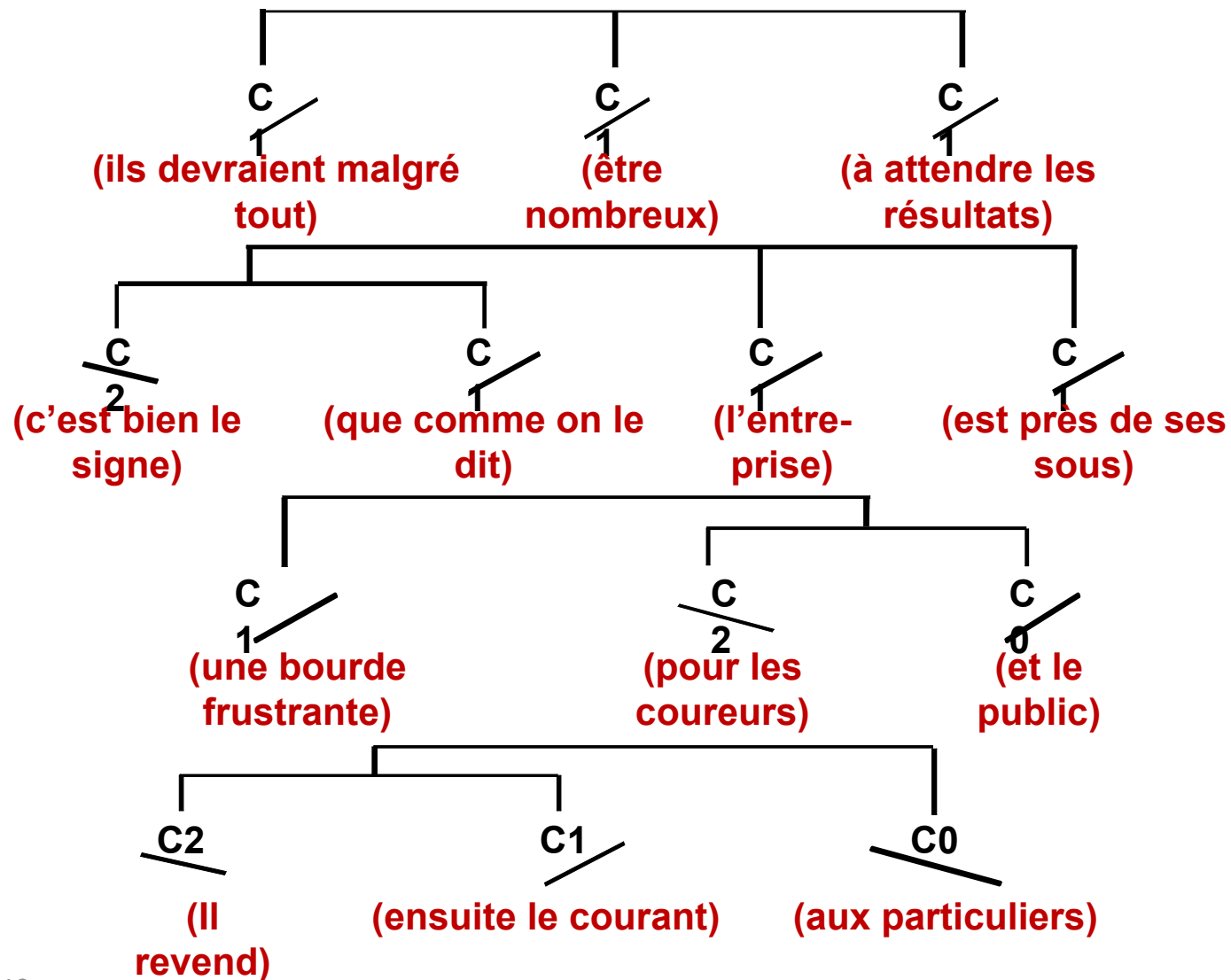
Prosodic groups and punctuation

- Using ESTER data that was manually transcribed with punctuation marks

- 96% of dots match with end of automatically detected prosodic groups

- 80% of commas match with end of automatically detected prosodic groups

Sentence modality

- Focus on statement vs. question
- Questions can be
  - Expressed with interrogative forms
  - Perceived as questions only through a rising intonation

- Classification based on
  - Linguistic features (words)
  - Prosodic features
  - Both linguistic and prosodic features

- Evaluations on speech data from ESTER and ETAPE (broadcast news) using
  - Manual transcriptions
  - Automatic speech recognition output

Detection of sentence modality

- Comparison of classification results using an MLP classifier
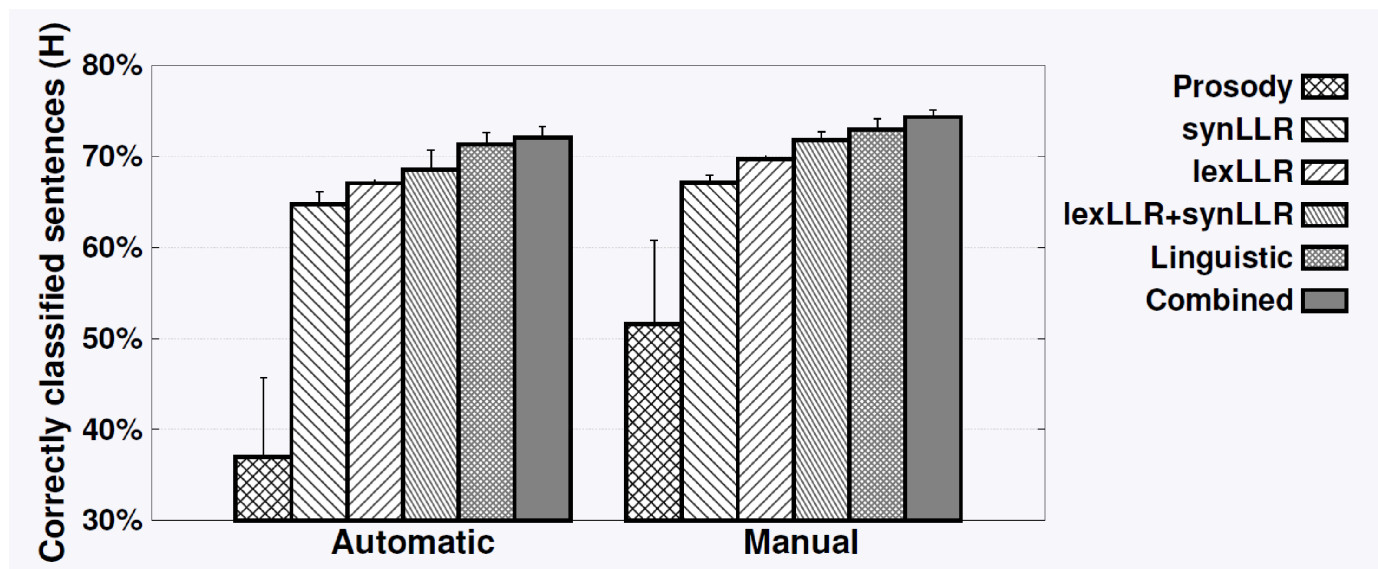
- The most important linguistic feature is the lexical log likelihood ratio (lexLLR) using two language models (one for questions, one for statements)
- The best results are obtained when combining all features

Discourse particles

- Words of expressions such as « well », « then », « you see », « you know », …
- That lose their usual lexical meaning
- But have a function at the discourse level
  - For utterance interpretation
  - For the management of the interaction
  - …

- Focus on a few French words that are frequently used as discourse particles (DP)
  - *alors* (so)
  - *bon* (well)
  - *donc* (thus, therefore)
  - *enfin* (finally, anyway)
  - *quoi* (what)
  - *voilà* (there you go)

Examples

Label	Example
Non-DP	*… la question que tout le monde se posait **alors** était les ventes de ces nains de jardin refléteraient elles …* … the question that everyone was asking **then** was would the sales of these garden dwarves reflect …
DP	*… la les forces régulières les forces loyalistes vont mettre le paquet sur bouaké [pause] **alors** la question qui qui se pose à la mi journée c'est de savoir qui …* … the regular forces the loyalist forces will provide full backing on bouaké [pause] **then** the question arising at midday is to know …
DP	*… en achetant tout simplement des produits vous savez étiquetés satisfait ou remboursé **alors** c'est une gestion mais ça marche il l'a prouvé il a rempli son frigo …* … by simply buying products you know labeled satisfied or refunded **then** it is a management but it works he proved it he has filled its fridge …

Speech corpora

- Large set of speech corpora (13 subsets)
  - that were manually transcribed (by respective corpora developers)
  - And text-speech aligned (in house, or in the ORFEO project)

- French language

- Variety of speaking styles with various degrees of speech spontaneity
  - Storytelling   [0.14 million words]
  - Prepared speech   [1.82 million words]
    - Broadcast news
  - Spontaneous speech
    - Conversations, interviews, … [1.84 million words]
    - Interactions   [1.52 million words]

- About 1000 occurrences randomly selected for each word

Data annotation

- Annotation of speech data
  - Speech segments with about 15 words before and 15 words after the selected word
  - Using praat
    - ❑ Speech signal available (for listening)
    - ❑ Speech transcription also available
  - Annotation as DP or non-DP
  - If DP, further annotation with pragmatic function

- Pragmatic functions depend on discourse particles
- Examples of pragmatic functions are
 - Introduction
  - Conclusion
  - Addition
  - Confirmation
  - …

Examples

Label	Example
Non-DP	*… la question que tout le monde se posait **alors** était les ventes de ces nains de jardin refléteraient elles …* … the question that everyone was asking **then** was would the sales of these garden dwarves reflect …
DP – **introduction**	*… la les forces régulières les forces loyalistes vont mettre le paquet sur bouaké [pause] **alors** la question qui qui se pose à la mi journée c'est de savoir qui …* … the regular forces the loyalist forces will provide full backing on bouaké [pause] **then** the question arising at midday is to know …
DP – **conclusion**	*… en achetant tout simplement des produits vous savez étiquetés satisfait ou remboursé **alors** c'est une gestion mais ça marche il l'a prouvé il a rempli son frigo …* … by simply buying products you know labeled satisfied or refunded **then** it is a management but it works he proved it he has filled its fridge …

DP / non-DP analysis for word "*alors*"
## with respect to spontaneity of speech data
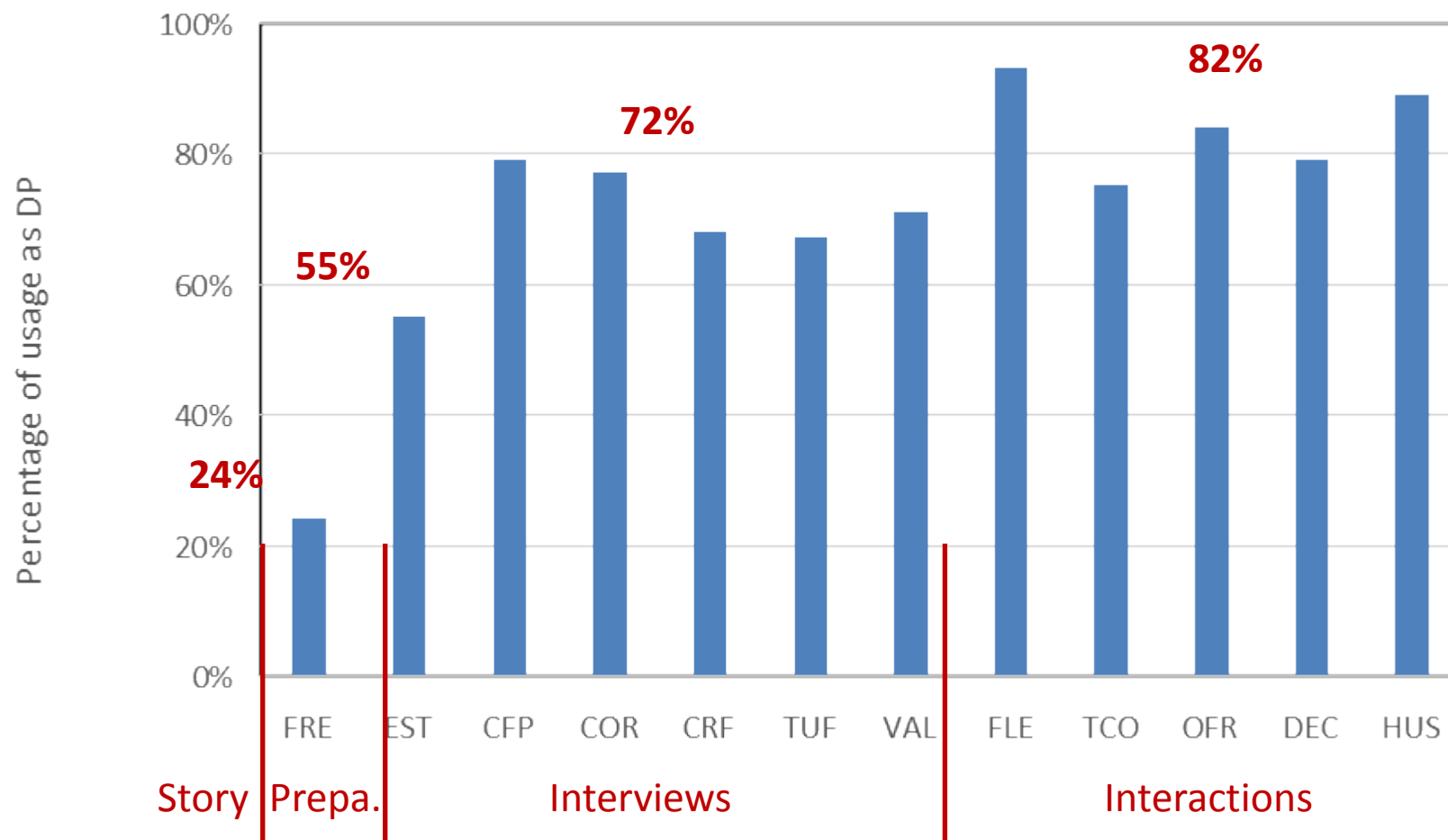
# DP / non-DP with respect to speech type
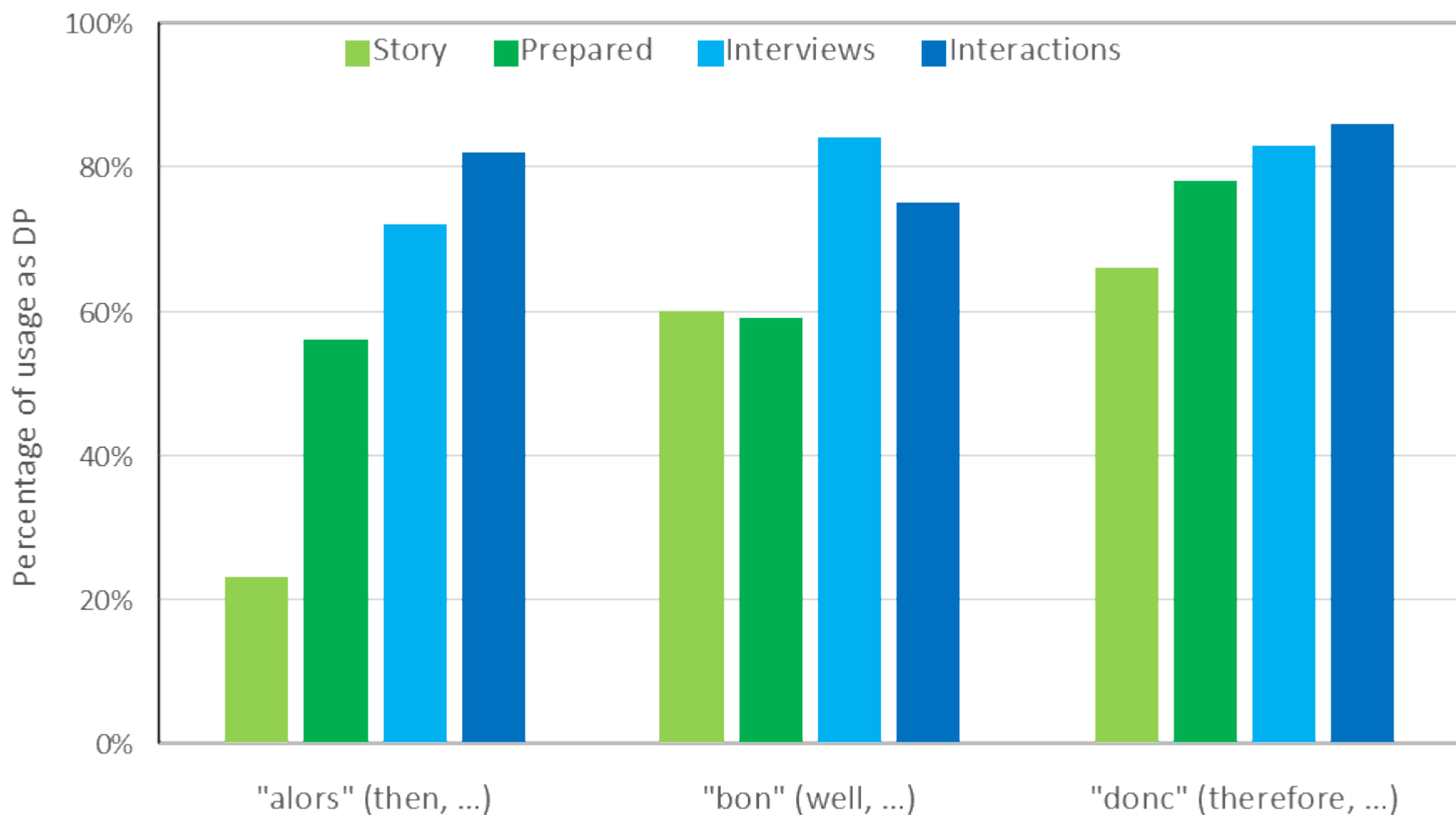
Analysis of a few prosodic correlates

- Different prosodic correlates have been analyzed
  - Pauses before and after the word
  - Position in intonation group
    (segmentation in intonation groups relies on F0 slope inversion, pitch level and vowel duration)
  - Pitch level and slope at end of words
  - Vowel duration, and lengthening
  - …

- Here, analysis is focused on
  - Pauses before and after the word
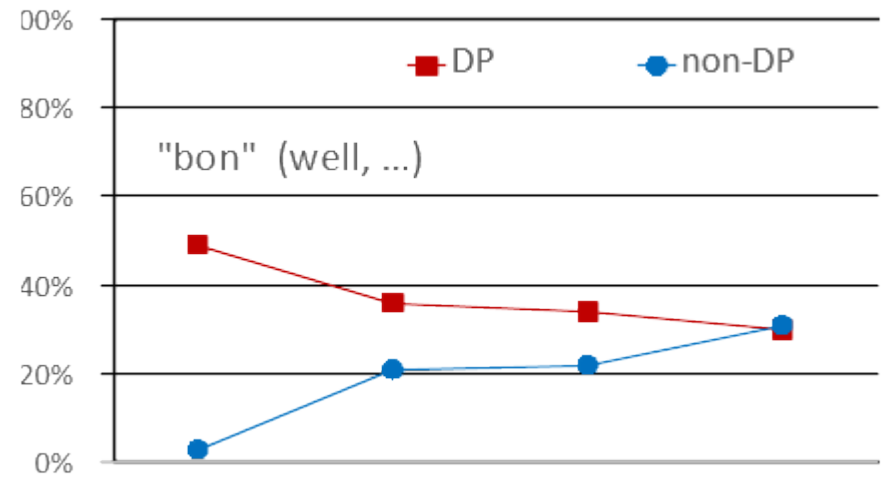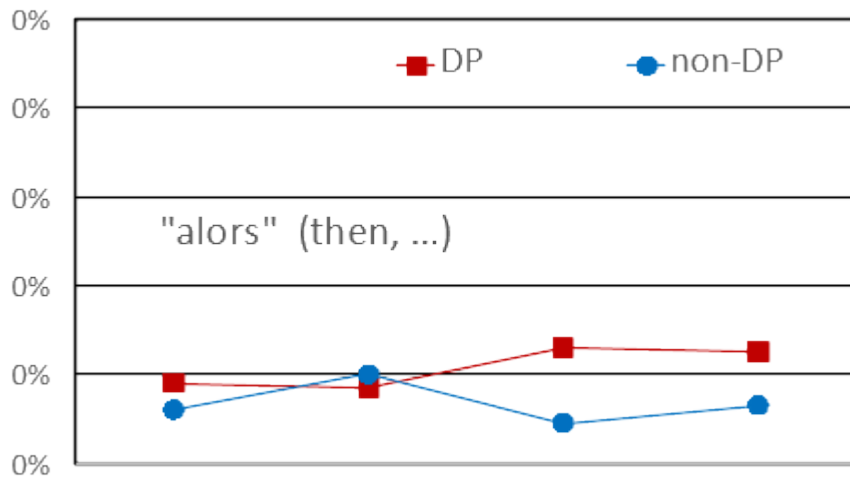 - Position in intonation group

# Frequency of occurrence of pauses **before** the word
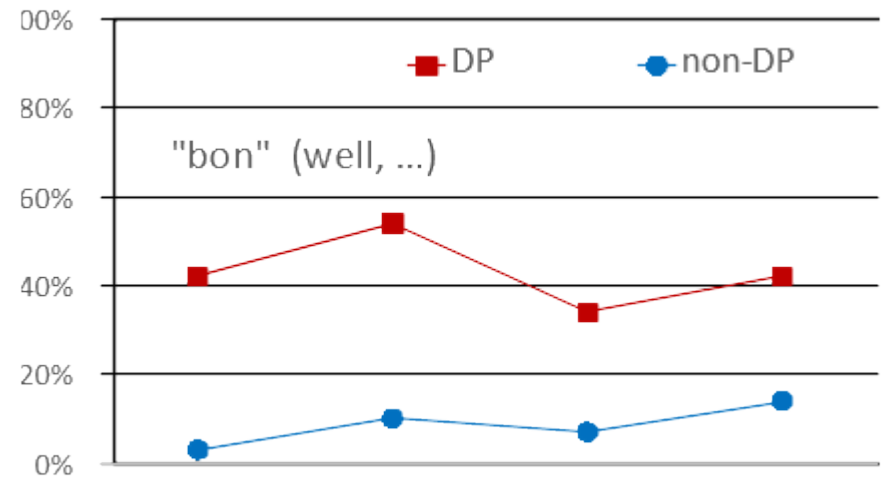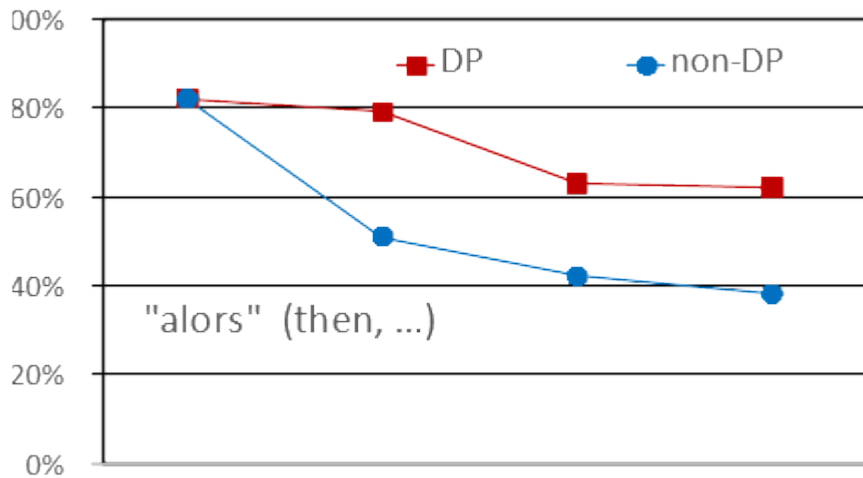
- Word "*bon*"
  - Very few pauses before when non-DP
  - Pause before much more frequent when DP
- Words "*alors*" and "*donc*"
  - More pauses before when DP than when non-DP, in spontaneous styles
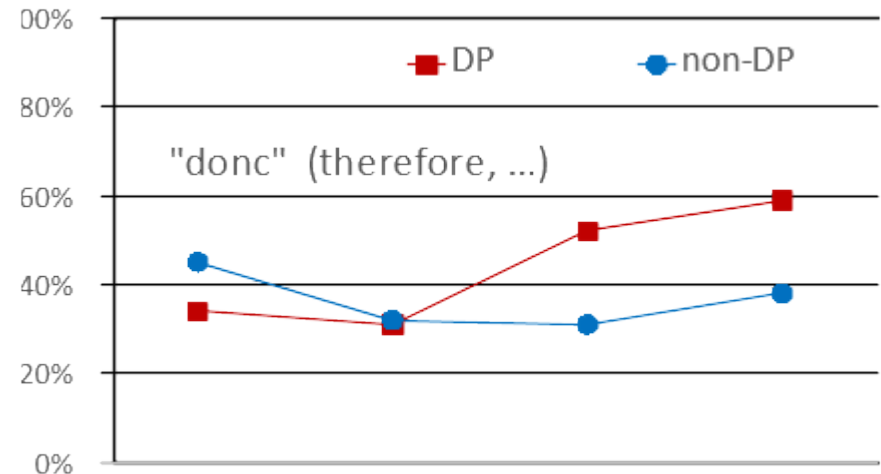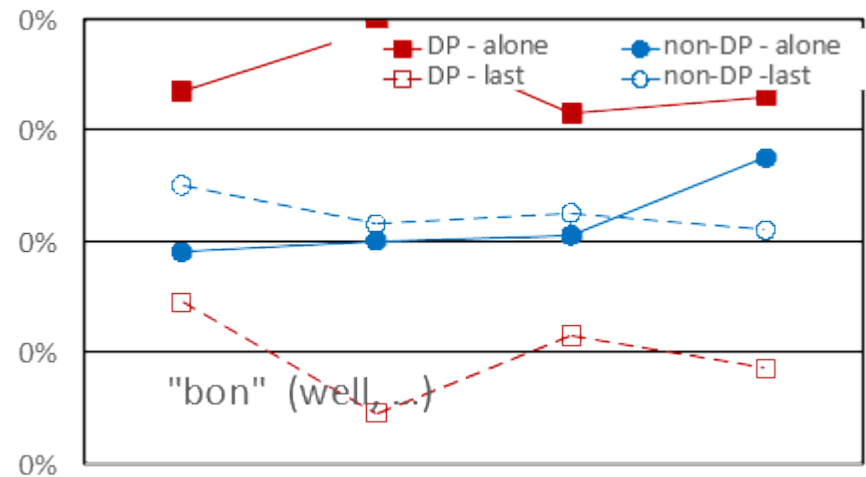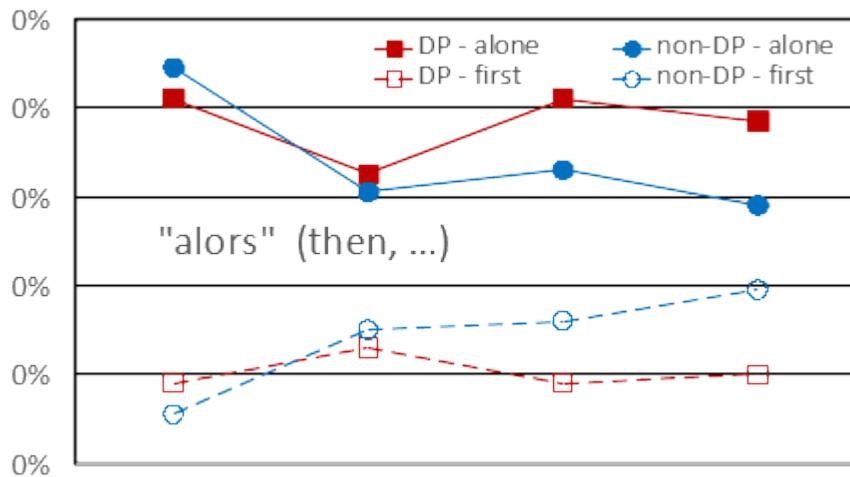
# Frequency of occurrence of pauses **after** the word

- No large differences between DP and non-DP functions, except for "*bon*"
- Word "*bon*" (well, …)
  - Largest difference for storytelling
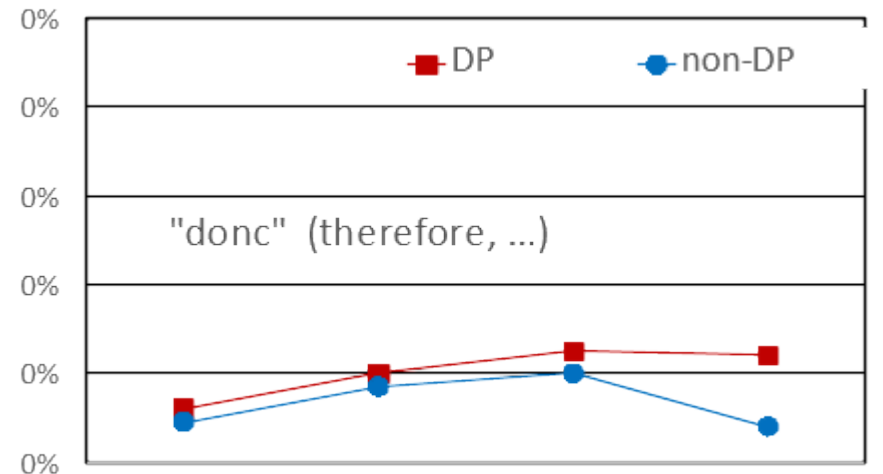
Position of the word in the intonation group

"alors" (then, …)

"bon" (well, …)

"donc" (therefore, …)

- Alone in intonation group
  - More often when DP than when non-DP
  - Largest difference for "*bon*"
- *"alors"* non-DP
  - Is getting more frequent in first position when spontaneous speech
- *"bon"* non-DP
 - More frequent in last position than when DP
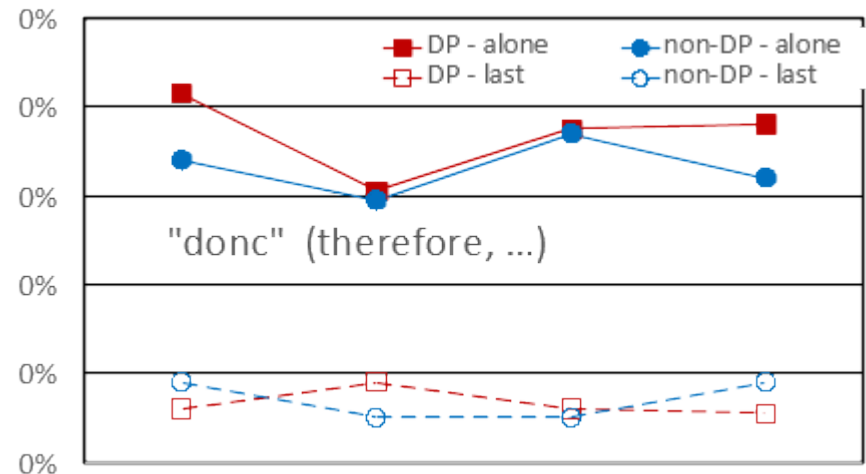
Automatic classification and detection experiments

- Data subsets
  - 60% for training, 10% for validation, 30% for performance evaluation

- Classifiers
  - Word dependent classifier
  - Neural network approach (Keras toolkit)

- Two sets of features
  - Prosodic features over a few word window
    - duration and energy of last vowel of the word
    - absolute F0 value at end of the word, and its slope
    - pause before and/or after the word
    - …
  - Fundamental frequency values over a few second window
    - F0 values computed every 10 ms

Automatic classification and detection
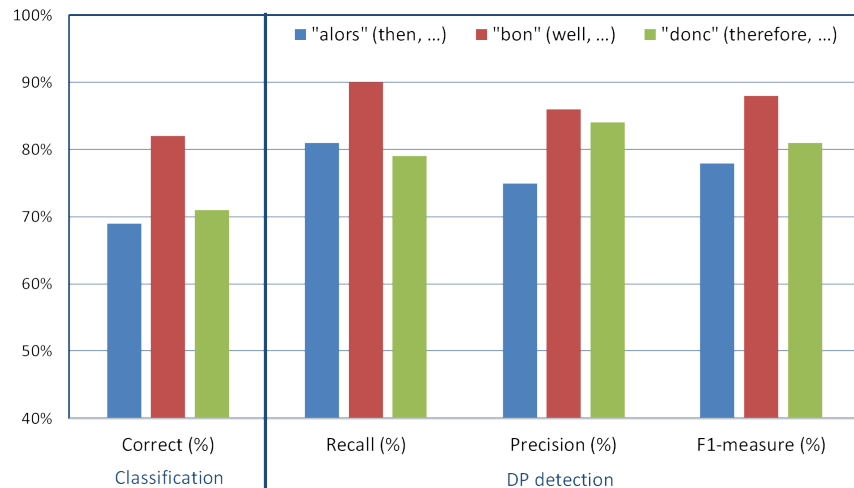# **using prosodic features**
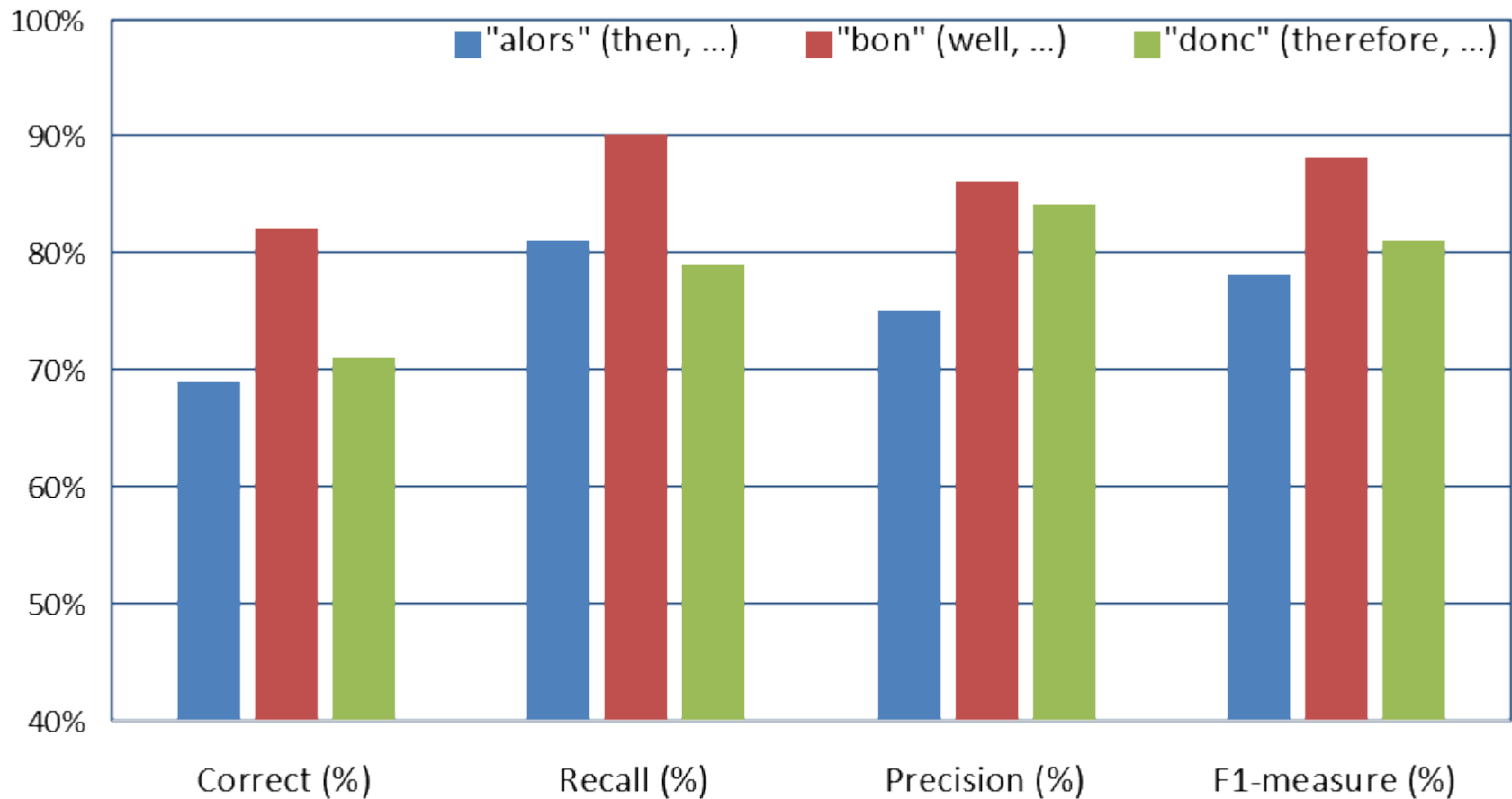
Automatic classification and detection
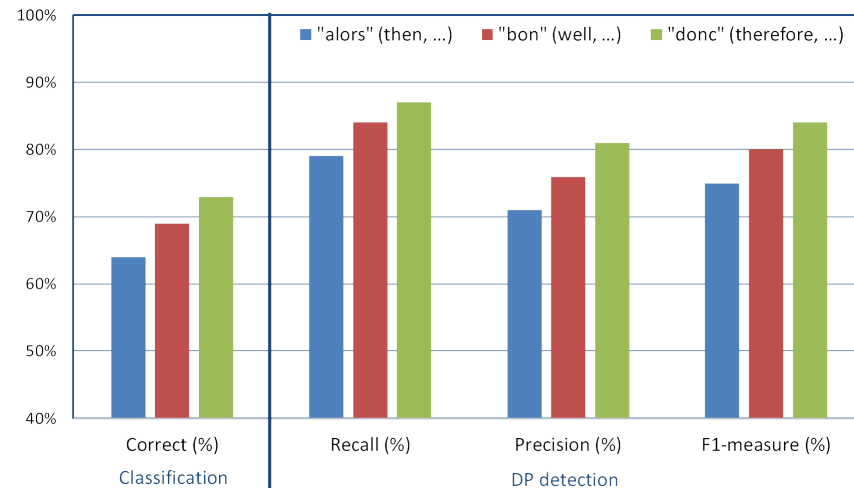## using fundamental frequency values
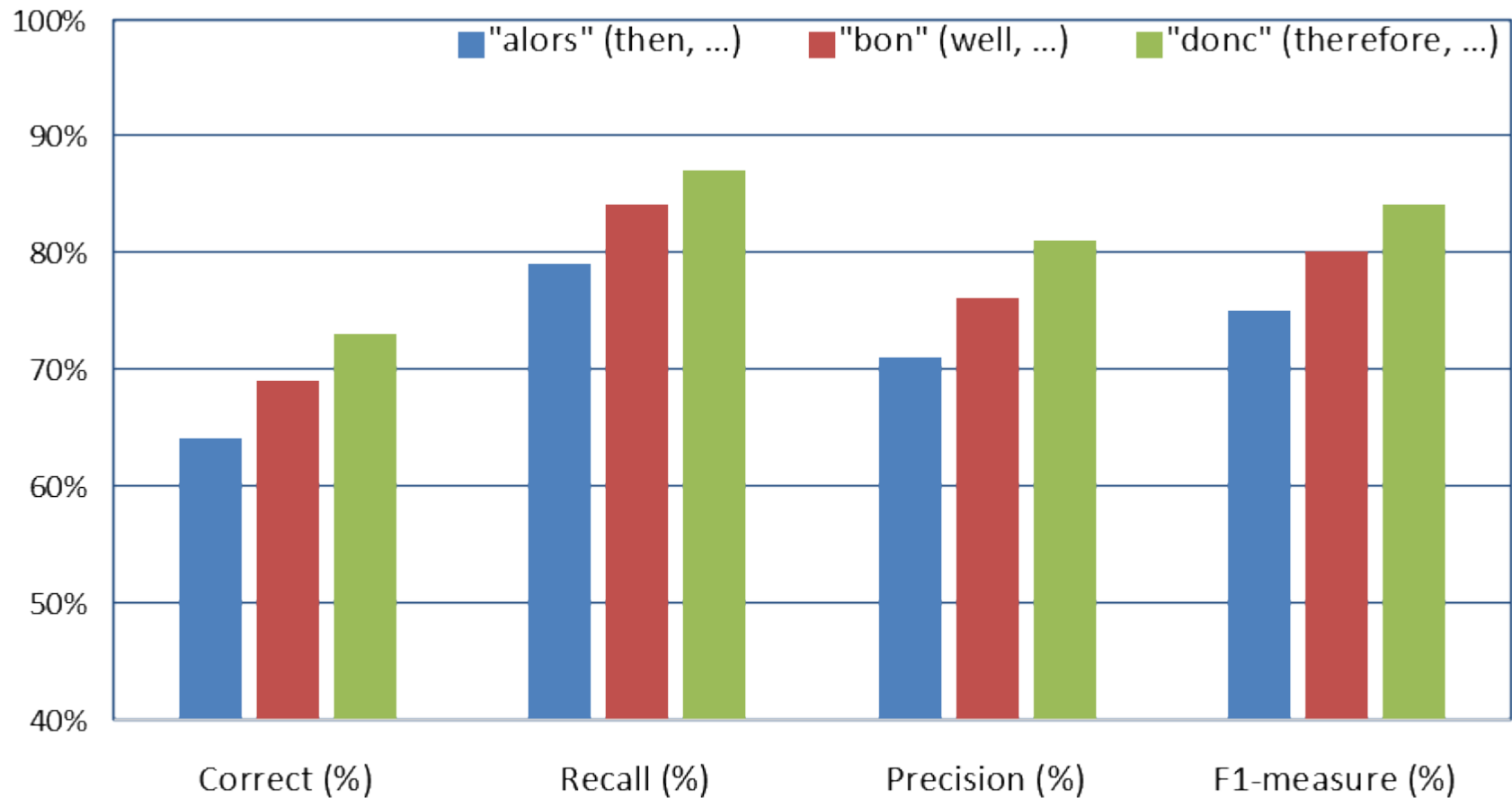
Automatic classification and detection

Prosodic features

Fundamental frequency

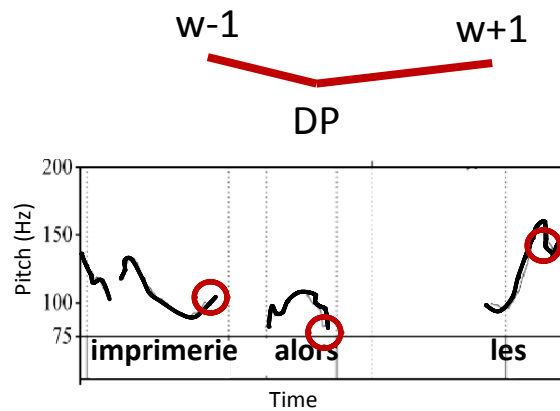- *"alors"* (then, …) & *"bon"* (well, …) ➔ Prosodic features more relevant than F0

- *"donc"* (therefore, …) ➔ F0 slightly more relevant than prosodic features

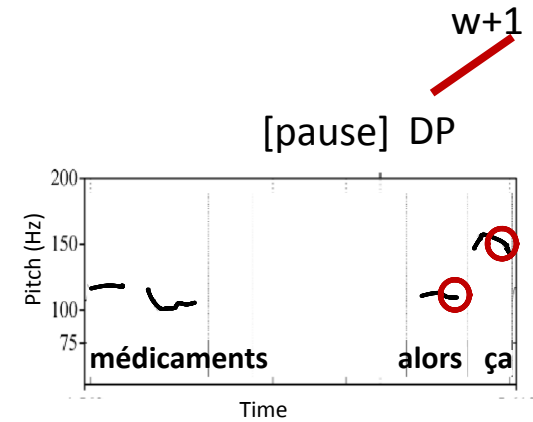- It might be interesting to combine these two sets of features

F0 patterns

- F0 movements with respect to
    - Last syllable of previous word
    - First syllable of next word
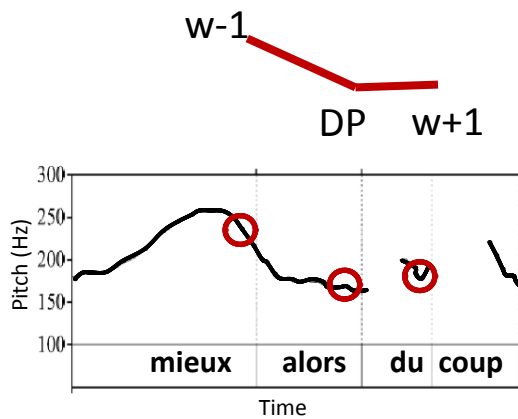
falling_plateau

falling_rising

rising_w+1

F0 patterns

Most frequent F0 patterns with respect to discourse particle and pragmatic function

Discourse Particle	Pragmatic function	F0 patterns	
alors	conclusion	falling-rising	falling-plateau
	introduction	rising	rising-plateau
	reintroduction	falling-plateau	plateau
donc	conclusion	falling-plateau	plateau
	reintroduction	rising-plateau	plateau
	addition	falling-plateau	plateau
bon	conclusion	falling-rising	falling-plateau
	interruption	plateau	
	confirmation	falling-rising	plateau
	incident	falling-plateau	

F0 patterns

addition and incident ➜ add an information or a comment

Discourse Particle	Pragmatic function	F0 patterns	
alors	conclusion	falling-rising	falling-plateau
	introduction	rising	rising-plateau
	reintroduction	falling-plateau	plateau
donc	conclusion	falling-plateau	plateau
	reintroduction	rising-plateau	plateau
	addition	**falling-plateau**	plateau
bon	conclusion	falling-rising	falling-plateau
	interruption	plateau	
	confirmation	falling-rising	plateau
	incident	**falling-plateau**	

F0 patterns

Conclusion and confirmation ➔ expression of look-back; semantic action of finality

Falling-rising and falling-plateau highlight a strong semantic break

Discourse Particle	Pragmatic function	F0 patterns	
alors	**conclusion**	**falling-rising**	**falling-plateau**
	introduction	rising	rising-plateau
	reintroduction	falling-plateau	plateau
donc	**conclusion**	**falling-plateau**	plateau
	reintroduction	rising-plateau	plateau
	addition	falling-plateau	plateau
bon	**conclusion**	**falling-rising**	**falling-plateau**
	interruption	plateau	
	confirmation	falling-rising	plateau
	incident	falling-plateau	

Expressive speech

- Expressive speech is now attracting a lot of interest
  - Expressive text-to-speech synthesis
  - Recognition of emotions

- Emotional speech can be collected
  - Recording of spontaneous speech – then annotation of the emotion
  - Recording through induced situations
  - Recording of acted speech from professional actors

Prosody of emotional speech

- Considering for example the F0 range, in comparison with neutral speech
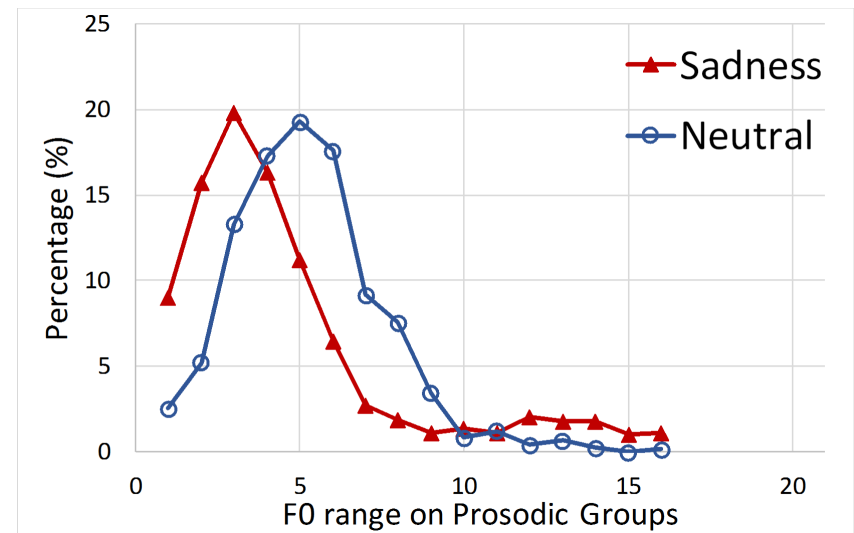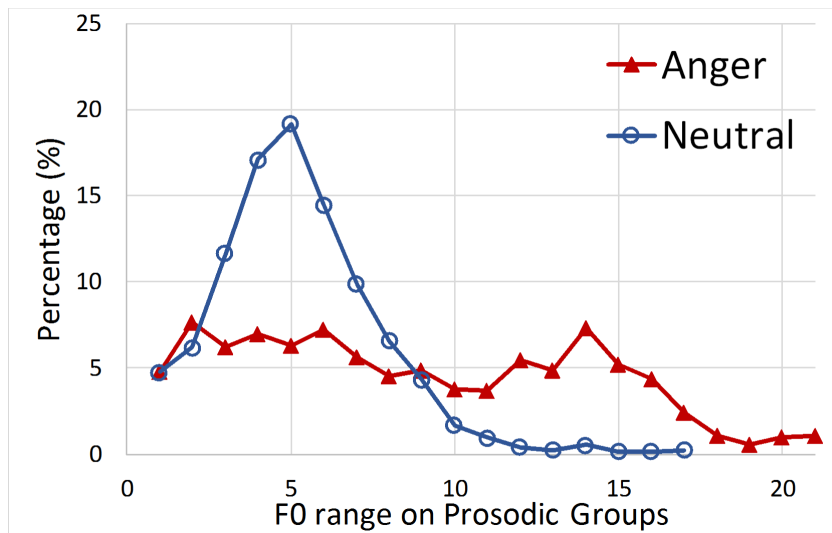
- Larger F0 ranges are much more frequent for anger
- And, slightly more frequent for fear, surprise and joy
- Smaller F0 ranges are more frequently observed for sadness.

Segmental level analysis

- Compared to neutral speech, pronunciation of emotional speech is often modified

- Many omissions of the schwa like vowel
- Omissions are more frequently observed In the first and last breathing groups
- Slightly vary with emotions – highest percentage was observed for disgust, fear and joy

- There exist also some other modifications, as for example
the omission of liquid consonants in consonantal clusters

Expressive speech synthesis

- Currently relies on an expressive speech synthesis corpus

- Recent approaches are based on deep learning approaches

- This opens research tracts for
  - Adjusting the level of the emotions
  - Investigating mixing of emotions
  - Investigating transfer learning approaches
  - …

Outline

- Prosodic features, computation and reliability
    - Phone duration
    - Fundamental frequency
    - Phone energy

- Prosodic features in automatic speech processing
    - Computer assisted language learning
    - Structuring speech utterances
    - Sentence modality
    - Prosodic correlates of discourse particles
    - Expressive speech

- **Conclusion**

Conclusion

- Computation of prosodic features
  - Forced speech-text alignment is used for phone duration
  - Many algorithms exists for fundamental frequency

- Approaches work well on clean and good quality speech
- However performance degrades on noisy speech

- Missing of reliable confidence estimators

- Prosody features are involved in many speech processing tasks