The Challenge of Analyzing Conversational Speech

Johannes Luig

Contents

Introduction

Hidden Punctuation

- Written vs. Spoken Language
  - Written Language
    - explicit punctuation (word and sentence boundaries)
  - Spoken Language
    - stream of words
    - no obvious lexical marking
    - phrasing expressed through prosody

Hidden Punctuation

- Importance of word and sentence boundaries
  - Automatic downstream processing
    - Parsing
    - Information extraction
    - Dialog act modeling
    - Summarization
    - Translation
  - Human readability
  - (Performance of speech recognition)

Hidden Punctuation

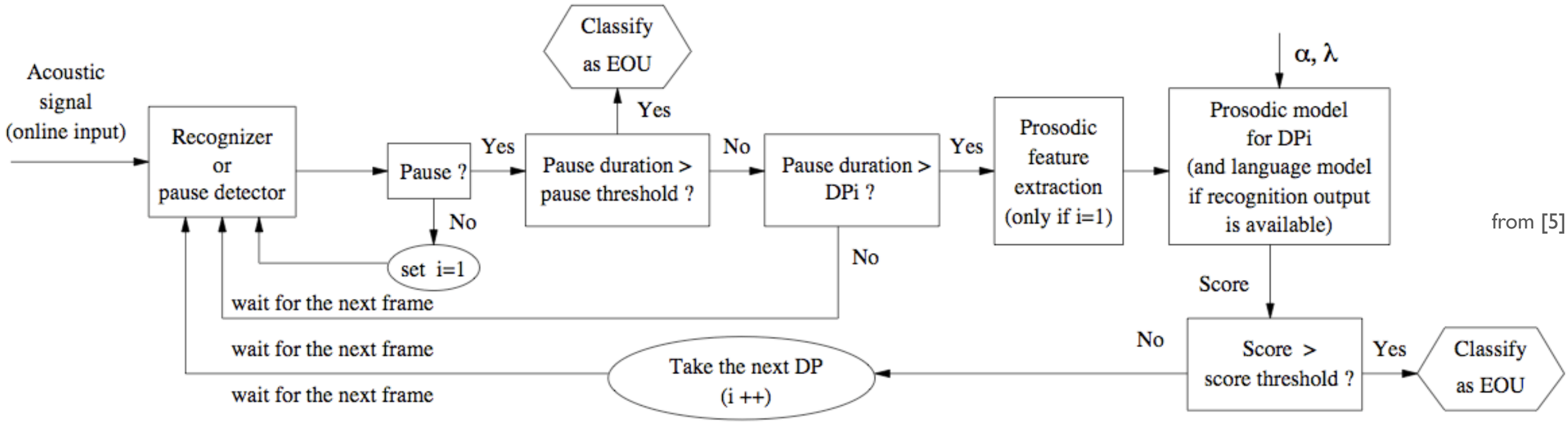
- Importance of word and sentence boundaries
  - Language Model creation (the „common“ way)
    - Recording of spoken text
    - Acoustical segmentation into sentence-like units
      - Chopping at longer pauses and speaker changes
  - Problems introduced by conversational speech
    - Pause ≠ indicator of sentence boundary:
      - Sentences strung together without pauses
      - Pauses occur at locations which are no sentence boundaries

Hidden Punctuation

- Particular Task: Endpointing
  - Determination when an utterance is „complete“
    - Application in online human-computer dialog systems
    - Online (real-time) procedure
  - Common method:
    - Waiting for a pause that is longer than pre-defined threshold

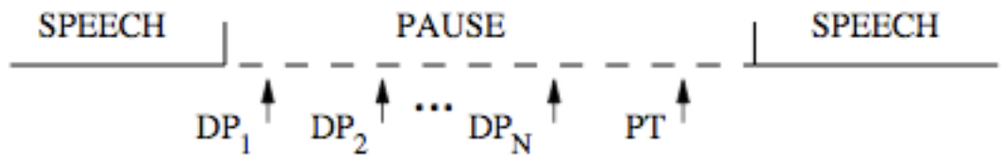
Hidden Punctuation

- Example: Prosody-based Approach



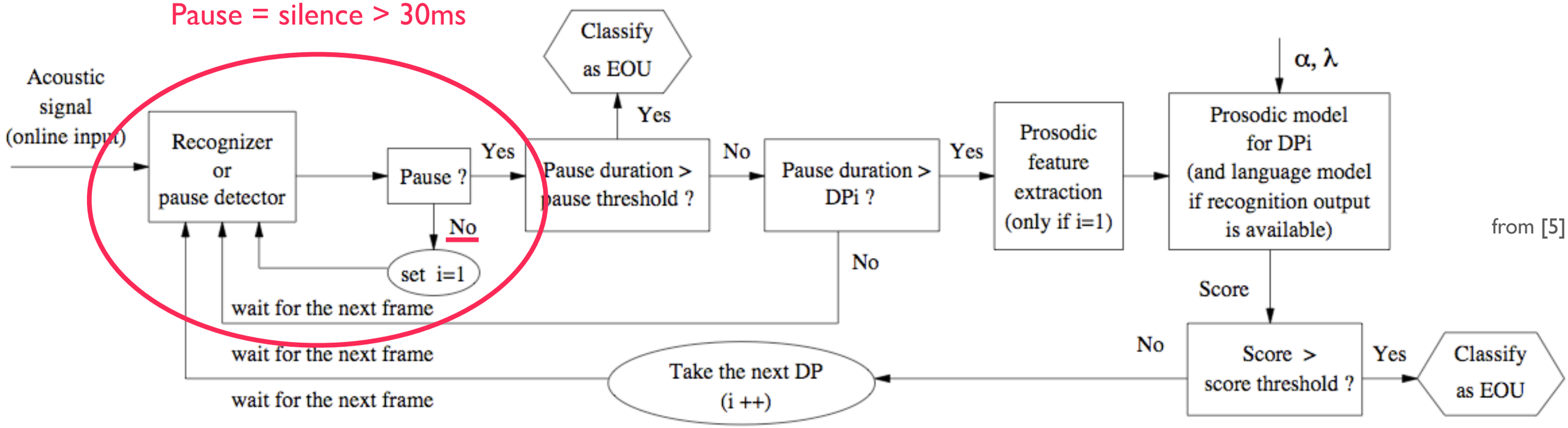
from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



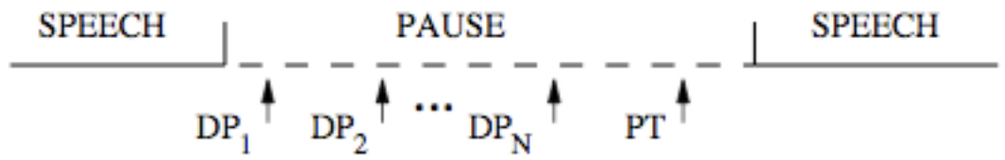
Hidden Punctuation

- Example: Prosody-based Approach



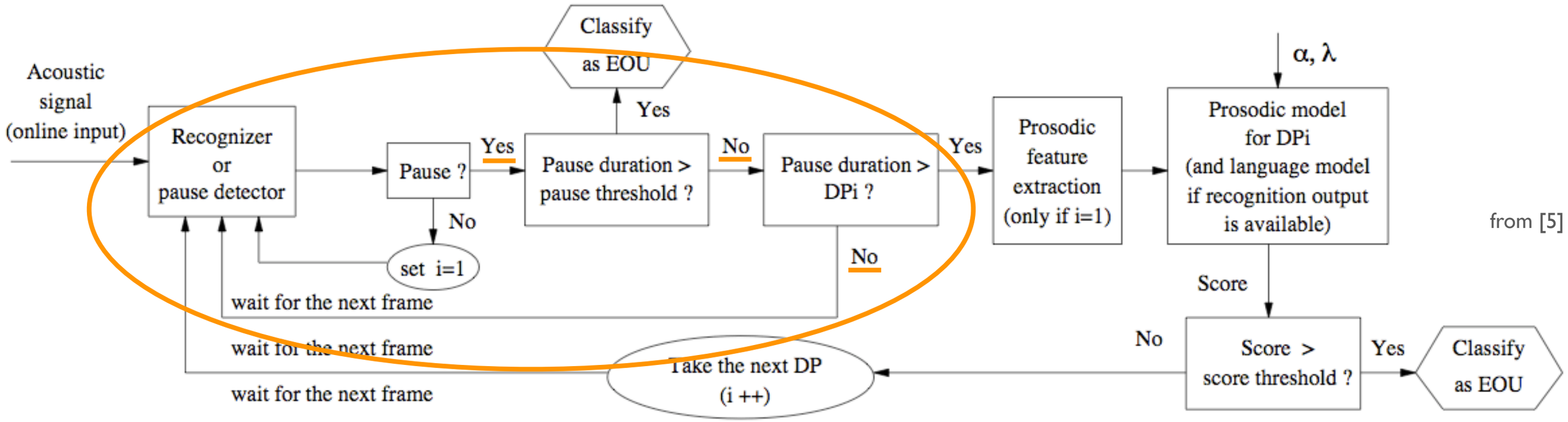
from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



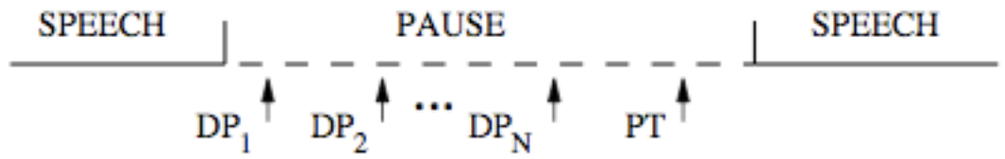
Hidden Punctuation

- Example: Prosody-based Approach



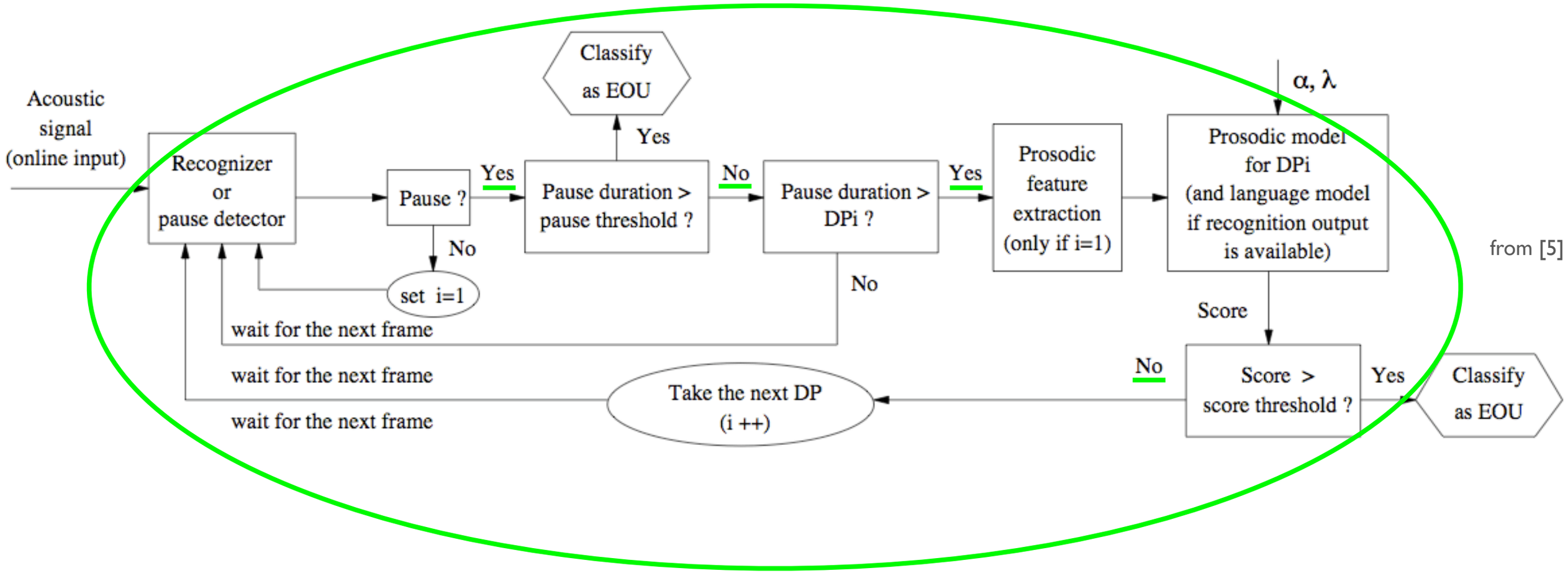
from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



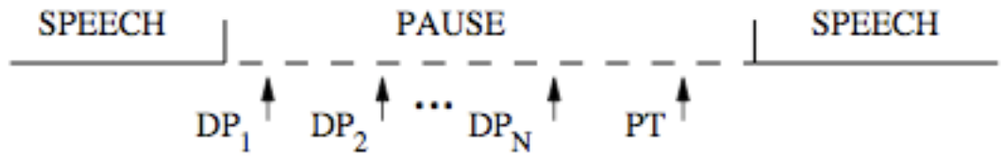
Hidden Punctuation

- Example: Prosody-based Approach



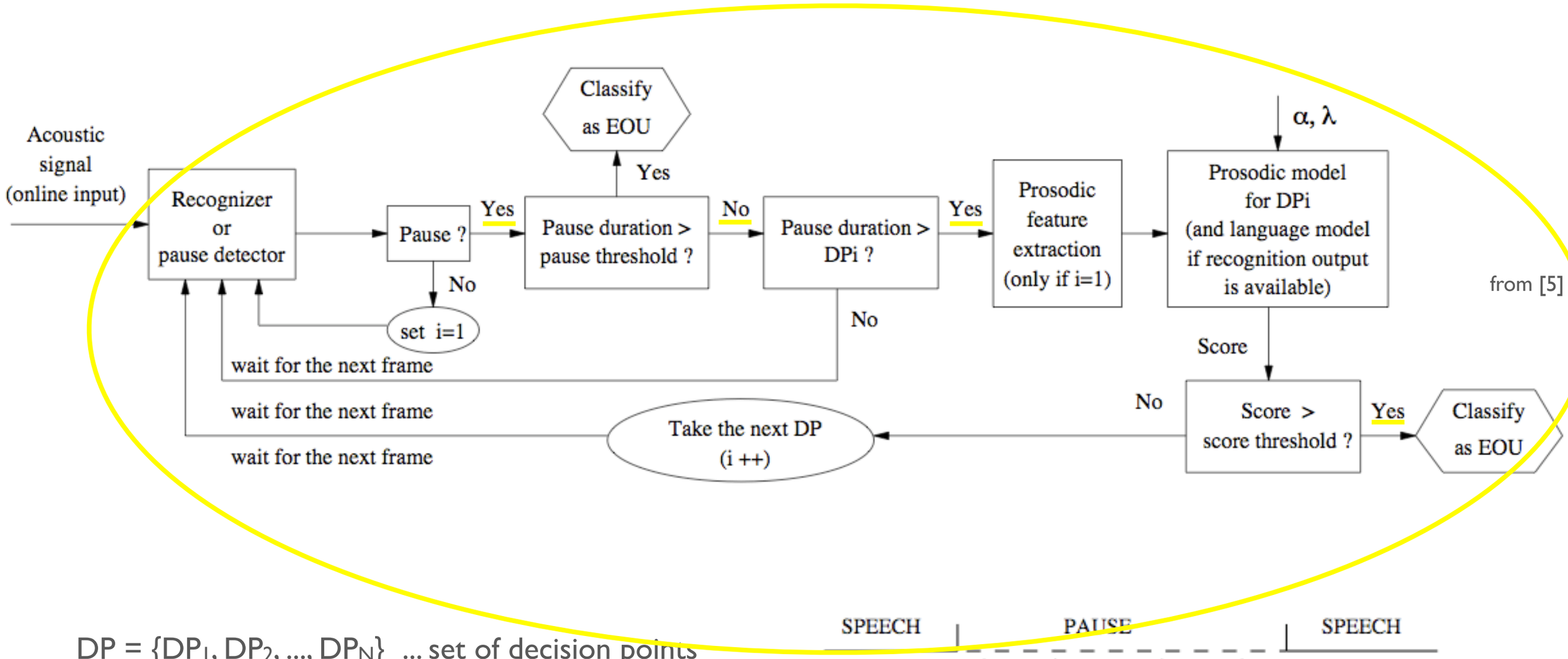
from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



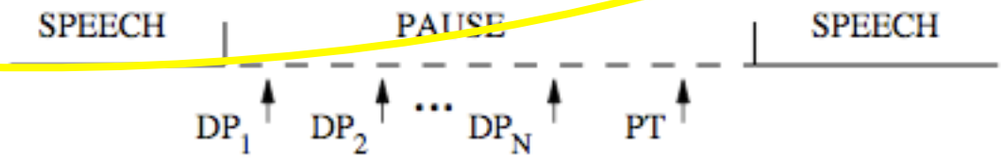
Hidden Punctuation

- Example: Prosody-based Approach



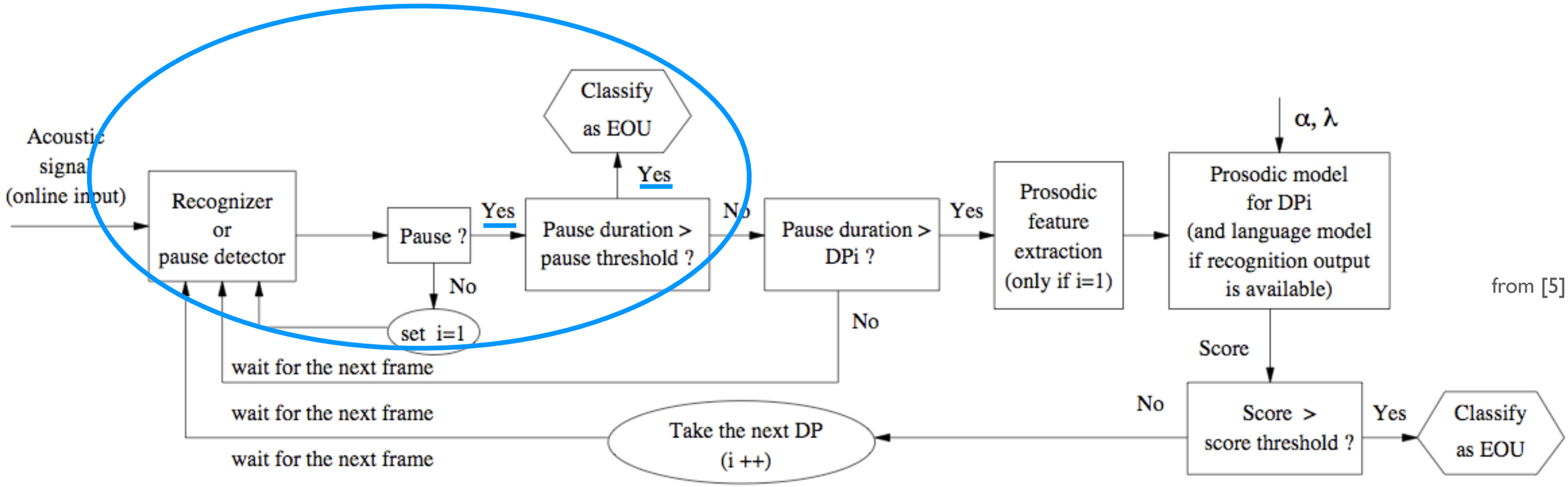
from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



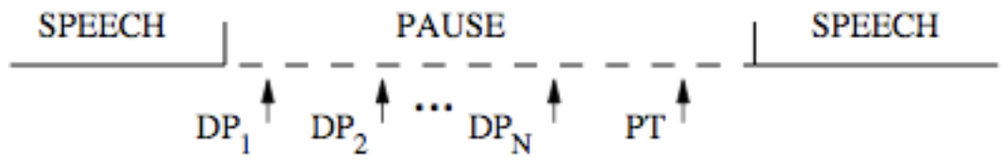
Hidden Punctuation

- Example: Prosody-based Approach



from [5]

DP = {DP_1, DP_2, ..., DP_N} ... set of decision points



Disfluencies

- Types of disfluencies
  - Repetitions
    - Speaker repeats some part of the utterance:
      I ... I like it.
  - Revisions
    - Speaker modifies some part of the utterance:
      We ... I like it.
  - Restarts
    - Speaker abandons an utterance and then starts over:
      It's also ... I like it.

Disfluencies

- Types of disfluencies
  - Filled pauses
    - Editing/correction is „commented“:
      We ... I mean, I like it.

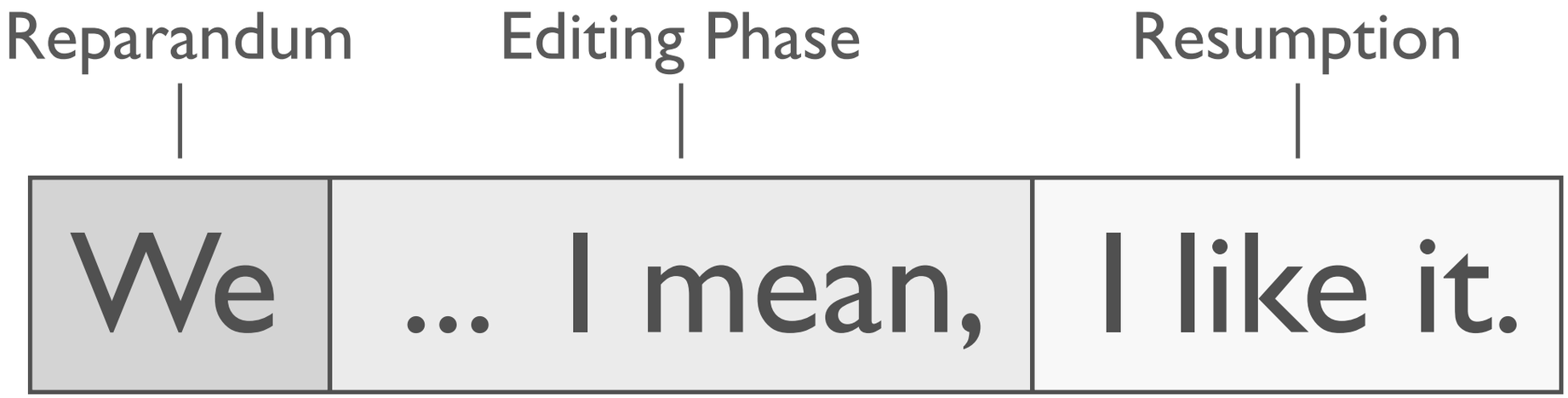
Disfluencies

- Structure of disfluencies

We	... I mean,	I like it.

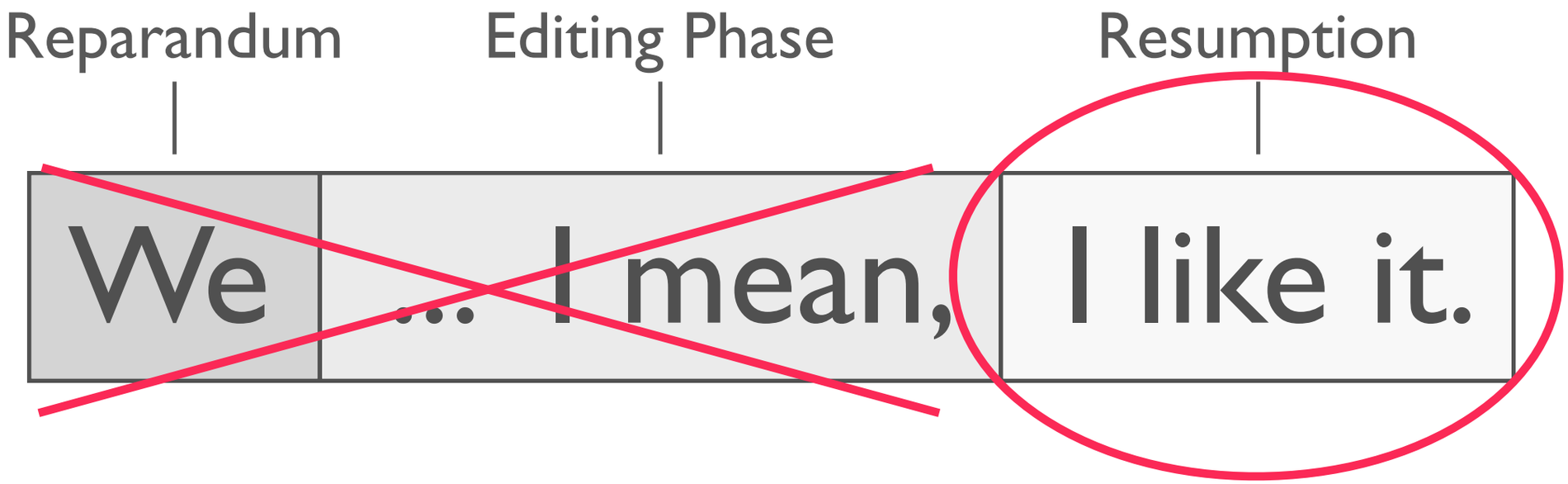
Disfluencies

- Structure of disfluencies



Disfluencies

- Structure of disfluencies



Speech Recognition

Disfluencies

- Occurrence of disfluencies
  - Frequency ~ every 20 words
  - up to 1/3 of all utterances affected
- Coping with disfluencies
  - Modeling rather than viewing them as „errors“
  - Detection of *interruption points*:
    We ... I like it.
  - For filled pauses, also the extent of the „cutaway“ region has to be determined

Disfluencies

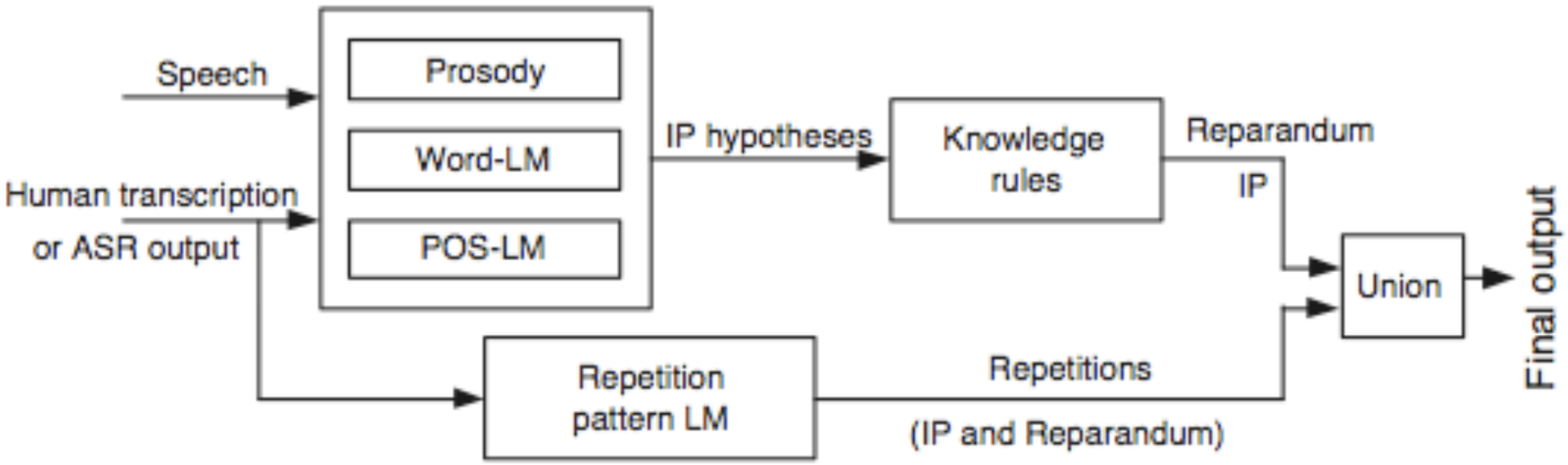
- Occurrence of disfluencies
  - Frequency ~ every 20 words
  - up to 1/3 of all utterances affected
- Coping with disfluencies
  - Modeling rather than viewing them as „errors“
  - Detection of *interruption points*:
    We ..| I like it.
  - For filled pauses, also the extent of the „cutaway“ region has to be determined

Disfluencies

- Occurrence of disfluencies
  - Frequency ~ every 20 words
  - up to 1/3 of all utterances affected
- Coping with disfluencies
  - Modeling rather than viewing them as „errors“
  - Detection of *interruption points*:
    We ... I like it.
  - For filled pauses, also the extent of the „cutaway“ region has to be determined

Disfluencies

- Example: Prosody- and LM-based Approach

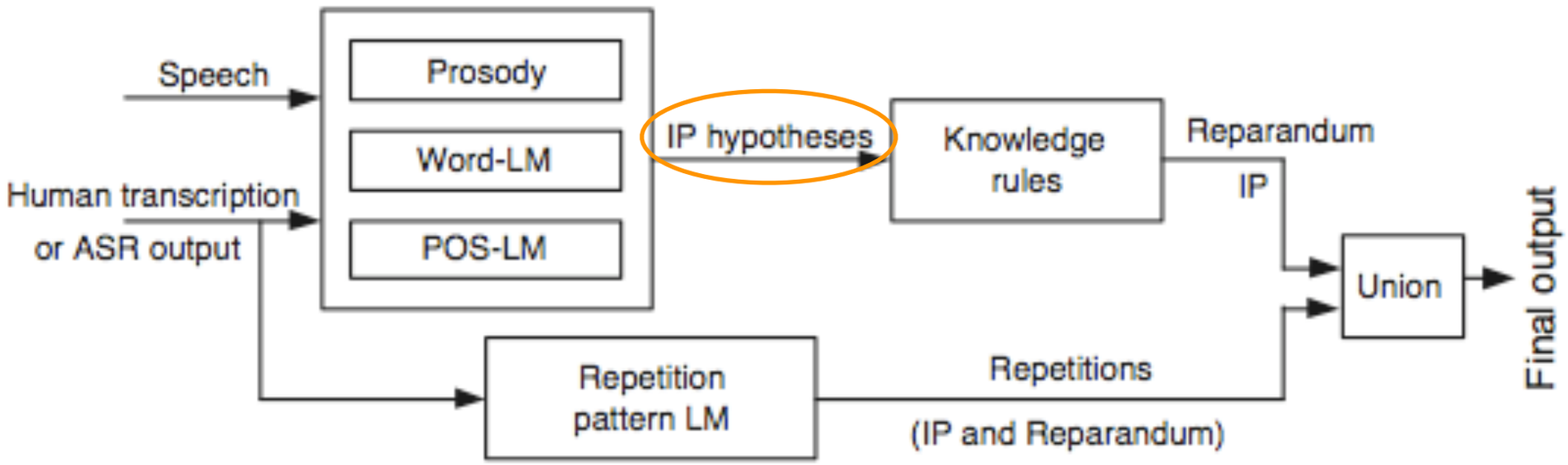


from [2]

We ... I like it.

Disfluencies

- Example: Prosody- and LM-based Approach

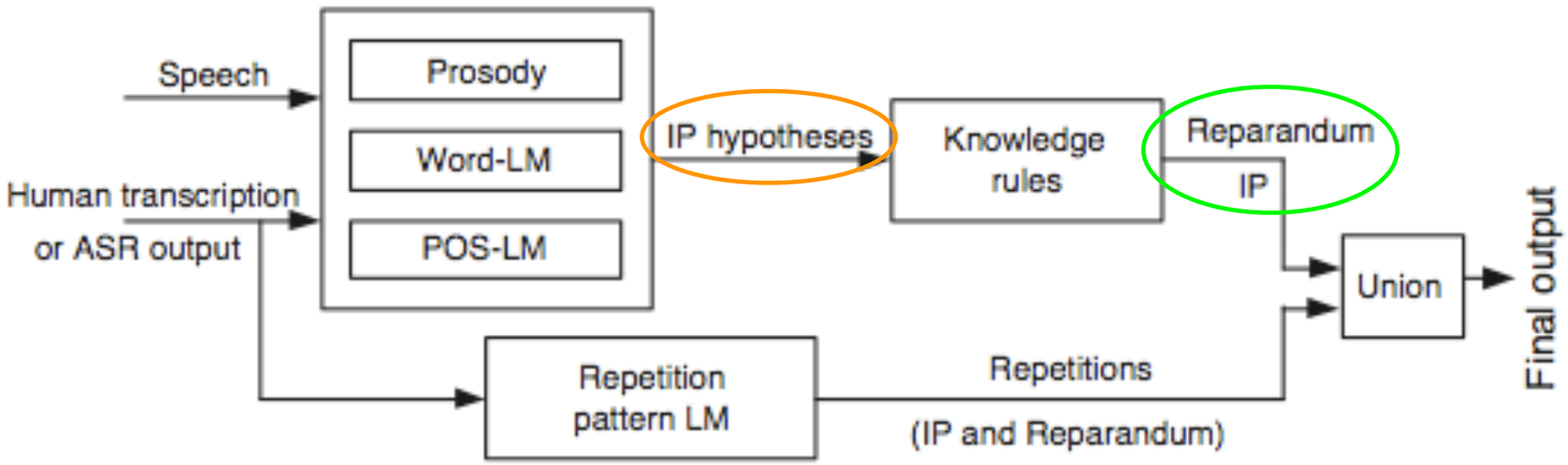


from [2]

We ... I like it.

Disfluencies

- Example: Prosody- and LM-based Approach

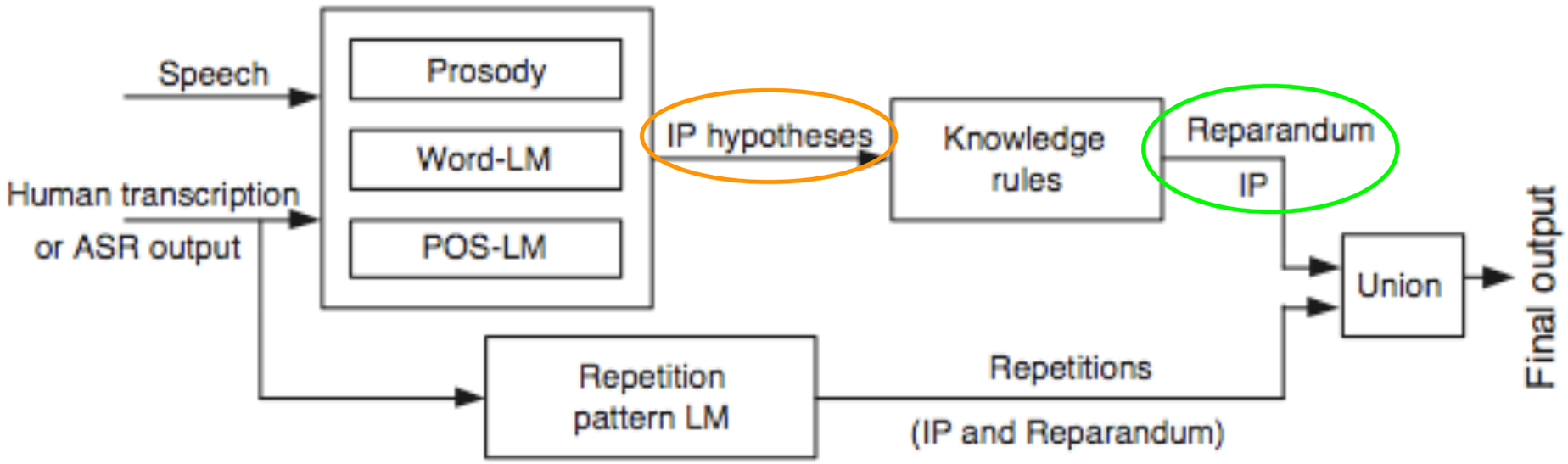


from [2]

We ... I like it.

Disfluencies

- Example: Prosody- and LM-based Approach



from [2]

Language Models: „Hidden-Event“ LM

- word-based LM
- part-of-speech based LM
- repetition pattern LM

Turn-Taking

- Speaker alternation in conversations
  - Speakers do not alternate sequentially
  - Listeners anticipate the end of a speaker‘s contribution by analyzing
    - Syntax
    - Semantics
    - Pragmatics
    - Prosody

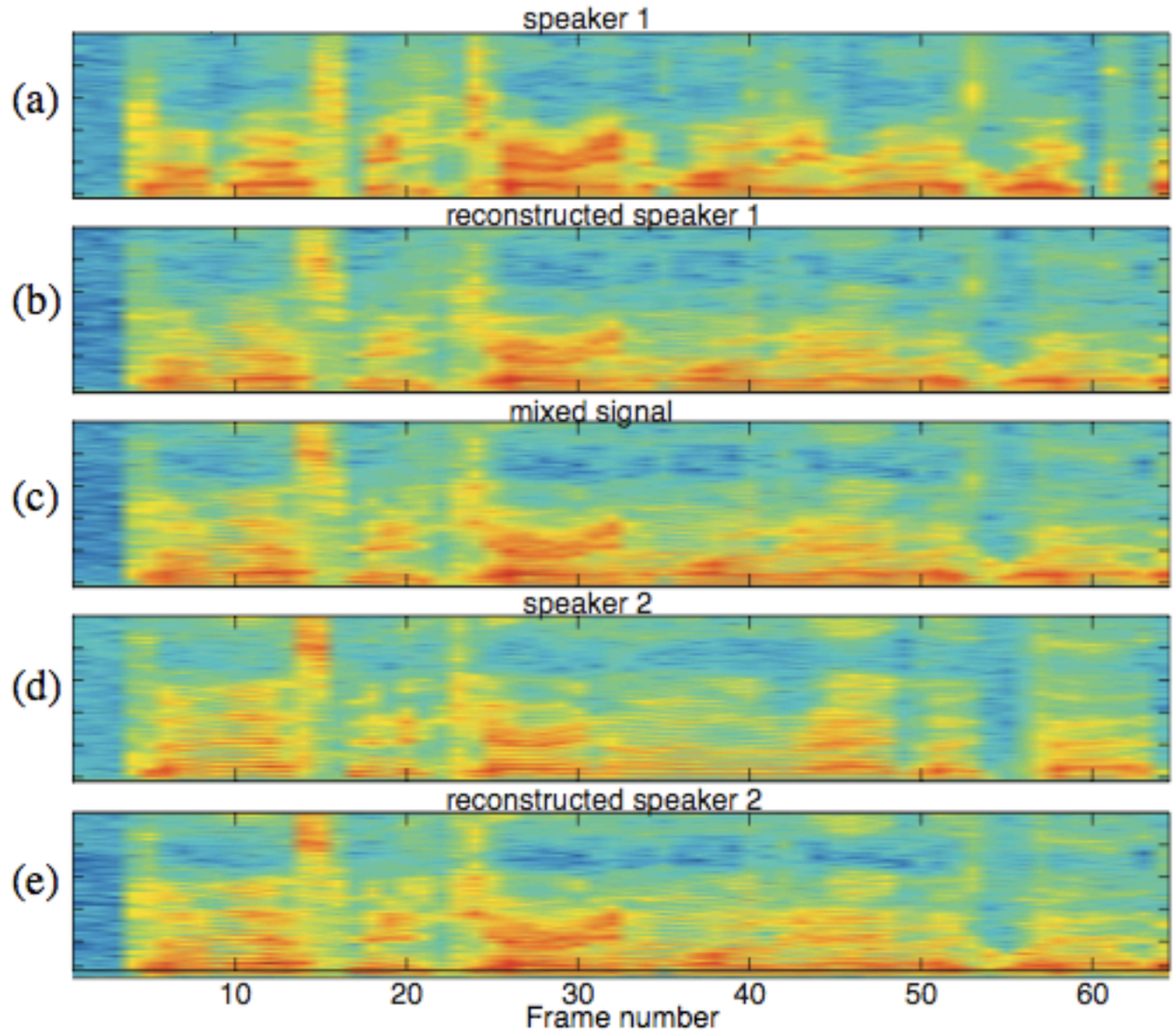
  and start talking before current speaker is finished

Turn-Taking

- Example: Spectrogram Decomposition
  - Assumption:
    - STFT vectors are outcomes of a discrete random process generating frequency bin indices
  - Modeling:
    - Random process: mixture multinomial distribution
    - Speaker distribution learned from training data (supervised approach)
    - separation → maximum likelihood estimates

Turn-Taking

- Example: Spectrogram Decomposition



from [6]

Emotions

- Hearing „more than words“
  - Inherently difficult or a machine, since
    - Rhetorical devices like irony require interpretation and contextual relation
    - „Natural“ emotion data is very rare
    - Labeling of gathered data is difficult even for humans
  - Intractable Problems:
    - Finding the classification reference („neutral“ state)
    - Setting the „emotion analysis time window“

Emotions

- Example: Features used for Emotion Analysis
  - „Classic“ Features
    - Time domain: duration / speaking rate / pause
    - Frequency domain: pitch / energy / spectral tilt
  - Features introduced by recent Approaches:
    - Loudness (RMS) across critical bands → Bark scale
    - Glottal-excitation-derived features
    - Voice quality features (Jitter, ...)
    - Harmonicity features

Conclusion / Outlook

- Four Main Challenges:
  - Hidden Punctuation
  - Disfluencies
  - Turn-Taking
  - Emotions
- Approaches mainly incorporate:
  - Prosody
  - Language Models

Conclusion / Outlook

- Perspective
  - Modeling and synthesis of conversational speech remains an interesting and challenging task
- Improvement Opportunities
  - Improved basic features
  - Incorporation of „longer-range“ information (greater time windows)
  - Speaker-dependent modeling not only in frame-level acoustics, but also in lexical / prosodic patterns

References

[1] Spontaneous Speech: How People Really Talk and Why Engineers Should Care
Elizabeth Shriberg
Proc. EUROSPEECH, Lisbon (Portugal), 2005

[2] Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources
Yang Liu, Elizabeth Shriberg, Andreas Stolcke
Proc. EUROSPEECH, Geneva (Switzerland), 2003

[3] How to Find Trouble in Communication
Anton Batliner et al.
Speech Communication, 40, 2003

[4] Conversational Speech Synthesis and the Need for Some Laughter
Nick Campbell
Journal of LaTeX Class files, Vol. 1, November 2002

References

[5] A Prosody-based Approach to End-of-Utterance Detection that does not require Speech Recognition
Luciana Ferrer, Elizabeth Shriberg, Andreas Stolcke
Proc. ICASSP, 2003

[6] Latent Variable Decomposition of Spectrograms for Single-Channel Speaker Separation
Bhiksha Raj, Paris Smaragdis
IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2005

Thank you.