Automatic Assessment of Children Speech to Support Language Learning

10th of June, 2009

Christian Hacker

Lehrstuhl für Mustererkennung

Technische Fakultät

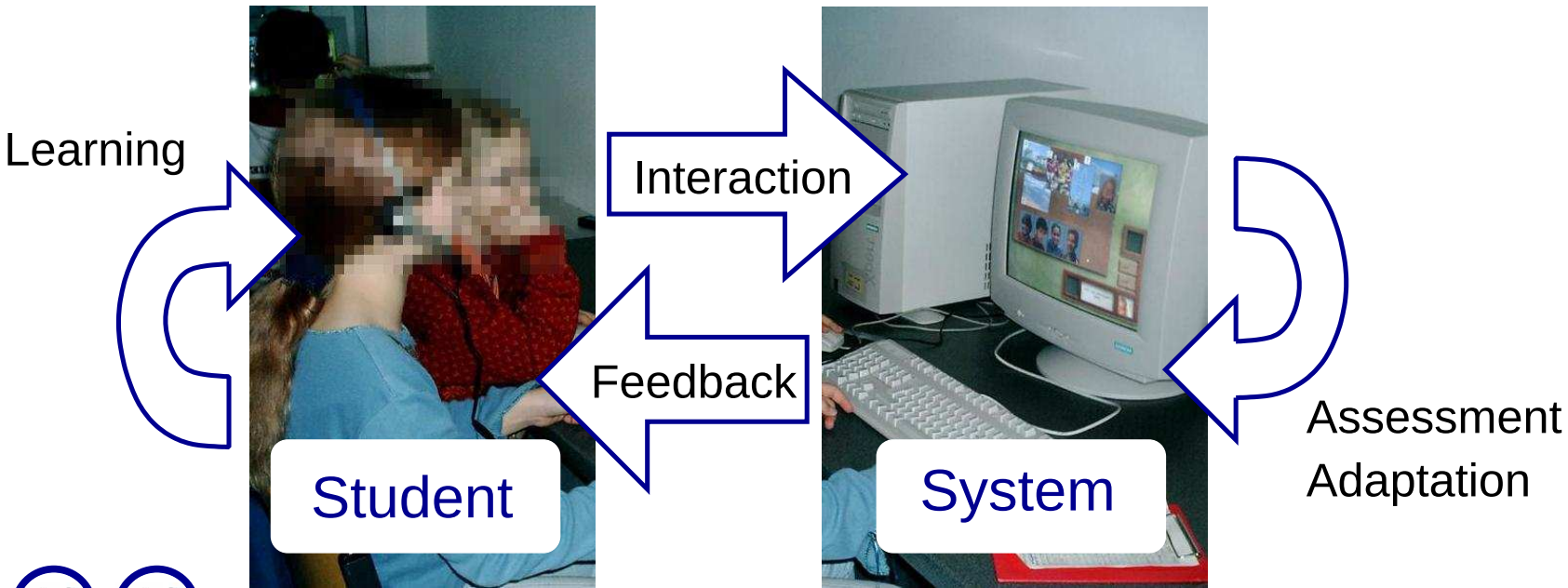
Friedrich-Alexander Universität Erlangen–Nürnberg

Introduction

- CALL: Computer aided language learning
- CAPT: Computer aided pronunciation training

Co-operation with the Ohm-Gymnasium, Erlangen

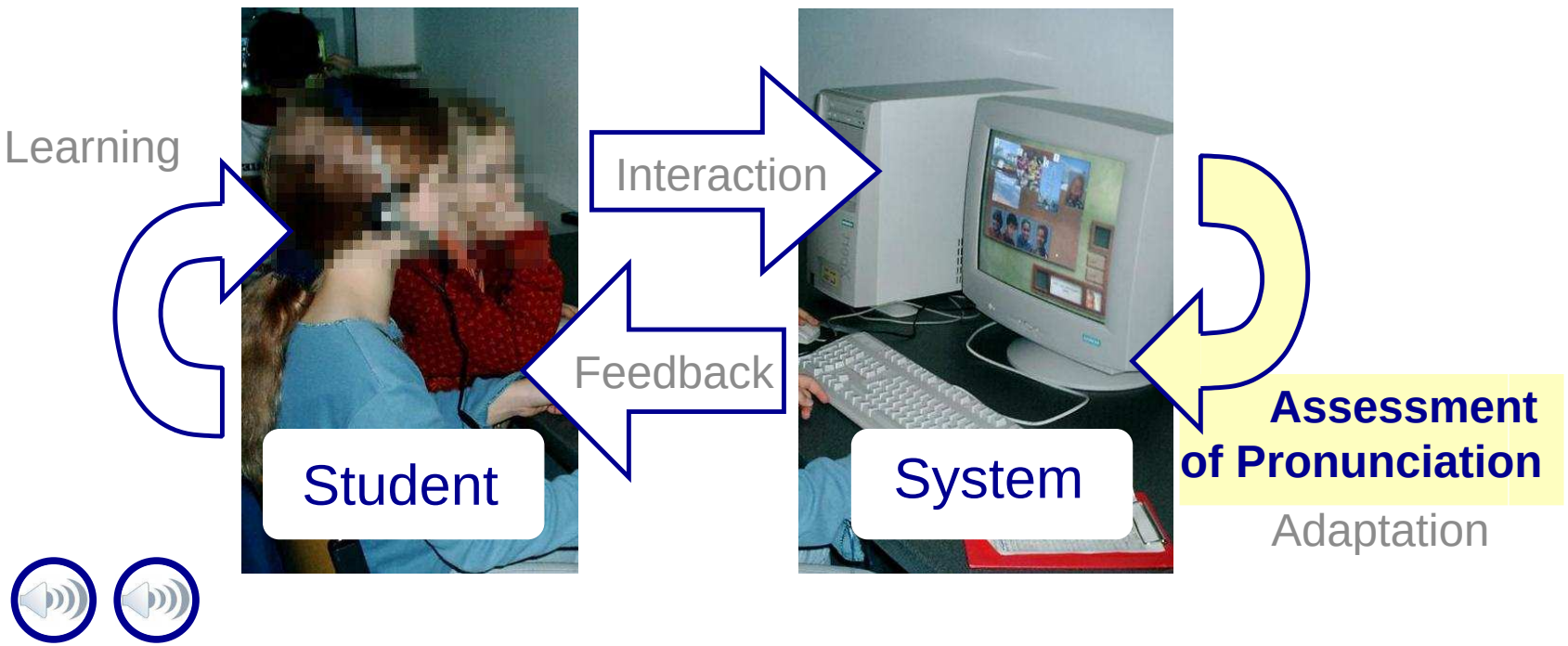
Introduction



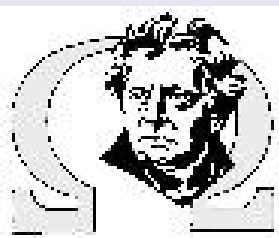
- CALL: Computer aided language learning
- CAPT: Computer aided pronunciation training

Co-operation with the Ohm-Gymnasium, Erlangen

Introduction



- CALL: Computer aided language learning
- CAPT: Computer aided pronunciation training



Co-operation with the Ohm-Gymnasium, Erlangen

Outline

1 CALL Applications

Outline

1 CALL Applications

2 Corpus and Annotations

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment

5 Experimental Results

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment

5 Experimental Results

Computer Aided Language Learning

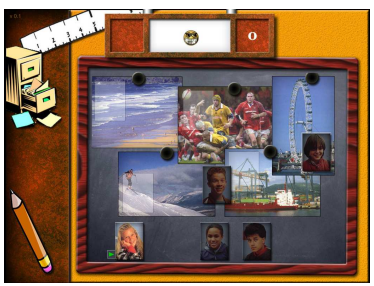
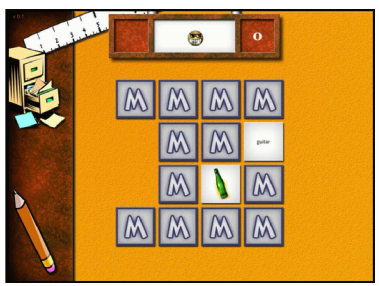
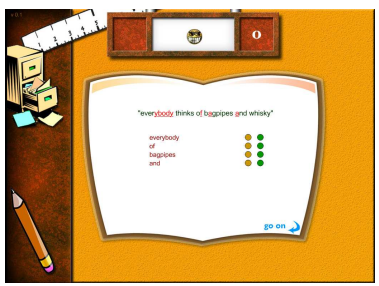
- Many commercial CALL systems exist
- Focus on reading, listening comprehension, writing
- Pronunciation training still requires
  - better automatic speech recognition (ASR) for non-natives
  - more robust pronunciation scoring algorithms

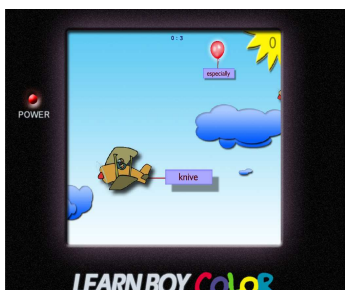
Computer Aided Language Learning

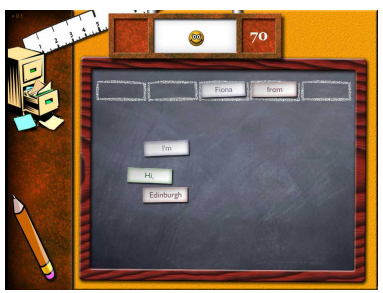
- Many commercial CALL systems exist
- Focus on reading, listening comprehension, writing
- Pronunciation training still requires
  - better automatic speech recognition (ASR) for non-natives
  - more robust pronunciation scoring algorithms

Caller

Computer **A**ssisted **L**anguage **L**earning from **Er**langen





Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment

5 Experimental Results

Corpus and Annotations

Pf-Star Non-Native Corpus (3.2 hrs. recorded in Erlangen)

- German children speaking English
- Vocabulary: 934 words
- Realistic speech containing repetitions of words, word fragments, non-verbal sound

Corpus and Annotations

Pf-Star Non-Native Corpus (3.2 hrs. recorded in Erlangen)

- German children speaking English
- Vocabulary: 934 words
- Realistic speech containing repetitions of words, word fragments, non-verbal sound

Focus on part of the data (1.2 hrs.)

- 28 children age 10 – 11 (learning English in their 1st year)
- Annotated by 14 experts
- Evaluation: leave-one-speaker-out

Corpus and Annotations

Pf-Star Non-Native Corpus (3.2 hrs. recorded in Erlangen)

- German children speaking English
- Vocabulary: 934 words
- Realistic speech containing repetitions of words, word fragments, non-verbal sound

Focus on part of the data (1.2 hrs.)

- 28 children age 10 – 11 (learning English in their 1st year)
- Annotated by 14 experts
- Evaluation: leave-one-speaker-out

Pf-Star Native Corpus (7.8 hrs. recorded in Birmingham)

- British children age 4 – 14

Ratings by 14 Experts

Experts:

S	German student of English (graduate level)
$\mathbf{T_1 - T_7}$	German teachers of English
$\mathbf{T_8 - T_{12}}$	German student teachers of English
N	Teacher, native speaker of English

Instructions:

- **S**: *„Mark all phone deviations“*
- $\mathbf{T_i}$, **N** : *„Mark words, where you would have stopped the student in class“*

Ratings by 14 Experts

Experts:

S	German student of English (graduate level)
$\mathbf{T_1} - \mathbf{T_7}$	German teachers of English
$\mathbf{T_8} - \mathbf{T_{12}}$	German student teachers of English
N	Teacher, native speaker of English

Instructions:

- **S**: *„Mark all phone deviations“*
- $\mathbf{T_i}$, **N** : *„Mark words, where you would have stopped the student in class“*

Ratings:

Word-level	**X** (wrongly), **O** (correctly pronounced)
Sentence-level	Grades **1** (best) – **5** (worst) (only **S**)
Text-level	Grades **1** (best) – **5** (worst)

Text: on average 11 sentences

Corpus and Annotations: Examples

Liz [000000000] it's [000000000] one [000000000] o'clock [000000000]

Sentence: **Grade 2**

Liz [000000000] it's [00000X000] one [X0000X000] o'clock [000000000]

Sentence: **Grade 3**

Liz [X00000000] it's [00000000X] one [XXXXXXXXX] o'clock [00000X00]

Sentence: **Grade 5**

S (graduate student), **T_1 – T_7**, **N (native teacher)**

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment

5 Experimental Results

Agreement Measures

Strictness:

- % words marked as mispronounced (X)
- 3.7 % - 7.3 %, average: 4.9 %
- Robust reference: X, if at least 3 vote with X (5.3 %)

Agreement Measures

Strictness:

- % words marked as mispronounced (X)
- 3.7 % - 7.3 %, average: 4.9 %
- Robust reference: X, if at least 3 vote with X (5.3 %)

Pearson correlation ρ:

- Sentence-/text-level
- Measures the linear relation between expert/system and reference
- Robust reference: average grade

Agreement Measures

Strictness:

- % words marked as mispronounced (X)
- 3.7 % - 7.3 %, average: 4.9 %
- Robust reference: X, if at least 3 vote with X (5.3 %)

Pearson correlation ρ:

- Sentence-/text-level
- Measures the linear relation between expert/system and reference
- Robust reference: average grade

Class-wise averaged classification rate (CL)

Classification rate with tolerance

Agreement Measures (cont.)

Class-wise averaged classification rate (CL)

$$\text{CL-}K := \frac{\text{HR}_1 + \ldots + \text{HR}_K}{K}, \qquad K : \text{number of classes}$$

Hit-rate HR_i: % of all i that are correctly classified

- Word: $K = 2$
- Sentence/Text: $K = 5$

Agreement Measures (cont.)

Class-wise averaged classification rate (CL)

$$\text{CL-}K := \frac{\text{HR}_1 + \ldots + \text{HR}_K}{K}, \qquad K : \text{number of classes}$$

Hit-rate HR_i: % of all i that are correctly classified

- Word: $K = 2$
- Sentence/Text: $K = 5$

Classification with tolerance (CL-10$\pm$2)

- Use average grades of 14 experts
- Map continuous grades onto 10 classes (histogram equalisation)

Evaluation of the Experts

		intra-rater
Word-level	CL-2	78 %
Text-level	CL-5	50 %
	CL-10$\pm$2	78 %
	ρ	0.71

- Intra-rater: 2nd evaluation half a year later

Evaluation of the Experts

		intra-rater	inter-rater (1 vs. rest)
Word-level	CL-2	78 %	76 %
Text-level	CL-5	50 %	56 %
	CL-10$\pm$2	78 %	80 %
	ρ	0.71	0.76

- Intra-rater: 2nd evaluation half a year later

Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

4 Approaches for the Automatic Assessment
- Robust Speech Recognition
- Approach 1: Mispronunciation Models
- Approach 2: Prosodic and Pronunciation Features
- Evaluation with Native Models

5 Experimental Results

Recognition of Children Speech

Problem:

- Robust ASR is required for automatic pronunciation scoring
- Higher word error rates (WER) for children

Solution:

- Adapt acoustic models to children speech (MAP, MLLR)
- Warp children speech to better fit to adult acoustic models (VTLN)
- Children speech recogniser with optimised feature extraction

Recognition of Children Speech

Problem:

- Robust ASR is required for automatic pronunciation scoring
- Higher word error rates (WER) for children

Solution:

- Adapt acoustic models to children speech (MAP, MLLR)
- Warp children speech to better fit to adult acoustic models (VTLN)
- Children speech recogniser with optimised feature extraction

[WER]	VM	Birm.	Non-Nat.
Training: adults (VM = Verbmobil)	35 %	85 %	73 %
Training: adults, Adaptation to Birm.		37 %	64 %
Training: children (Birmingham)		23 %	44 %

Recognition of Non-Native Speakers

Problem:

- High WER for non-native speech
- Avoid adaptation to wrongly pronounced non-native data

Solution:

- ASR trained on native speakers (Birmingham + Youth)
- Add excellent non-native speakers to the training

Recognition of Non-Native Speakers

Problem:

- High WER for non-native speech
- Avoid adaptation to wrongly pronounced non-native data

Solution:

- ASR trained on native speakers (Birmingham + Youth)
- Add excellent non-native speakers to the training

[WER]	Birm.	Non-Nat.
Training: children (Birmingham)	23 %	44 %
Training: children (Birm., Youth, Non-Nat.)	28 %	36 %

Approach 1: Mispronunciation Models

- Add acoustic models with expected wrong pronunciation
- $\rightarrow$ Wrongly pronounced phone can be found
- Lexicon (example):

```
this        /TIs/
this~e10    /sIs/
```

- Design of 44 rules
- Systematic application of rules to the vocabulary

Approach 1: Mispronunciation Models

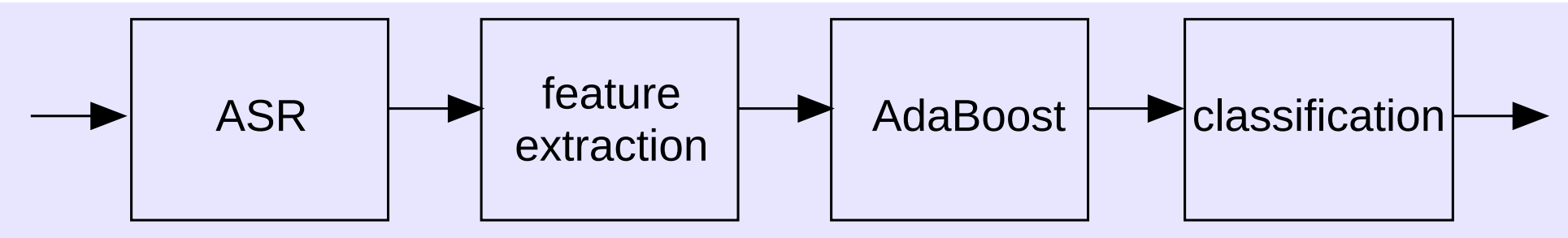
- Add acoustic models with expected wrong pronunciation
- $\longrightarrow$ Wrongly pronounced phone can be found
- Lexicon (example):

```
this        /TIs/
this~e10    /sIs/
```

- Design of 44 rules
- Systematic application of rules to the vocabulary

/..b/	$\longrightarrow$	/..p/	cab
/..d/	$\longrightarrow$	/..t/	feed
/..g/	$\longrightarrow$	/..k/	big
/Ng/	$\longrightarrow$	/N/	finger
/r/	$\longrightarrow$	/R/	right
/T/	$\longrightarrow$	/s/	think
/st/	$\longrightarrow$	/St/	stall
/..z/	$\longrightarrow$	/..s/	peas
/dZ/	$\longrightarrow$	/tS/	age
/v/	$\longrightarrow$	/f/	give
/w/	$\longrightarrow$	/v/	what
/v/	$\longrightarrow$	/w/	very
/Q/	$\longrightarrow$	/O:/	office
/aI/	$\longrightarrow$	/I/	rise
/3/	$\longrightarrow$	/e@/	early

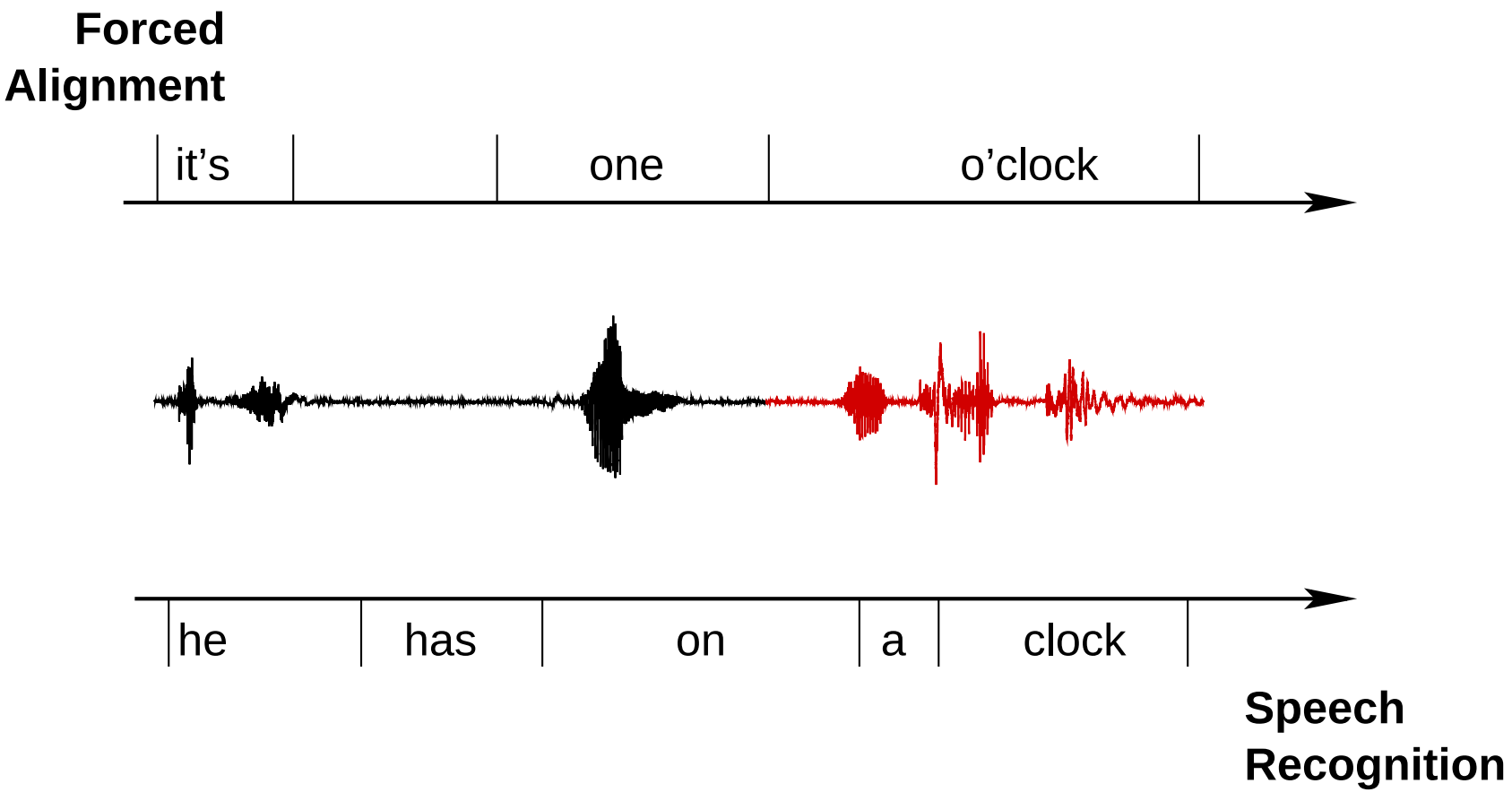
App. 2: Prosodic and Pronunciation Features



- Text and language independent approach
- Prosodic features: *how* something is said
- Pronunciation features: in particular based on ASR
- AdaBoost: Feature selection (complementary information)
- Classification: AdaBoost/LDA

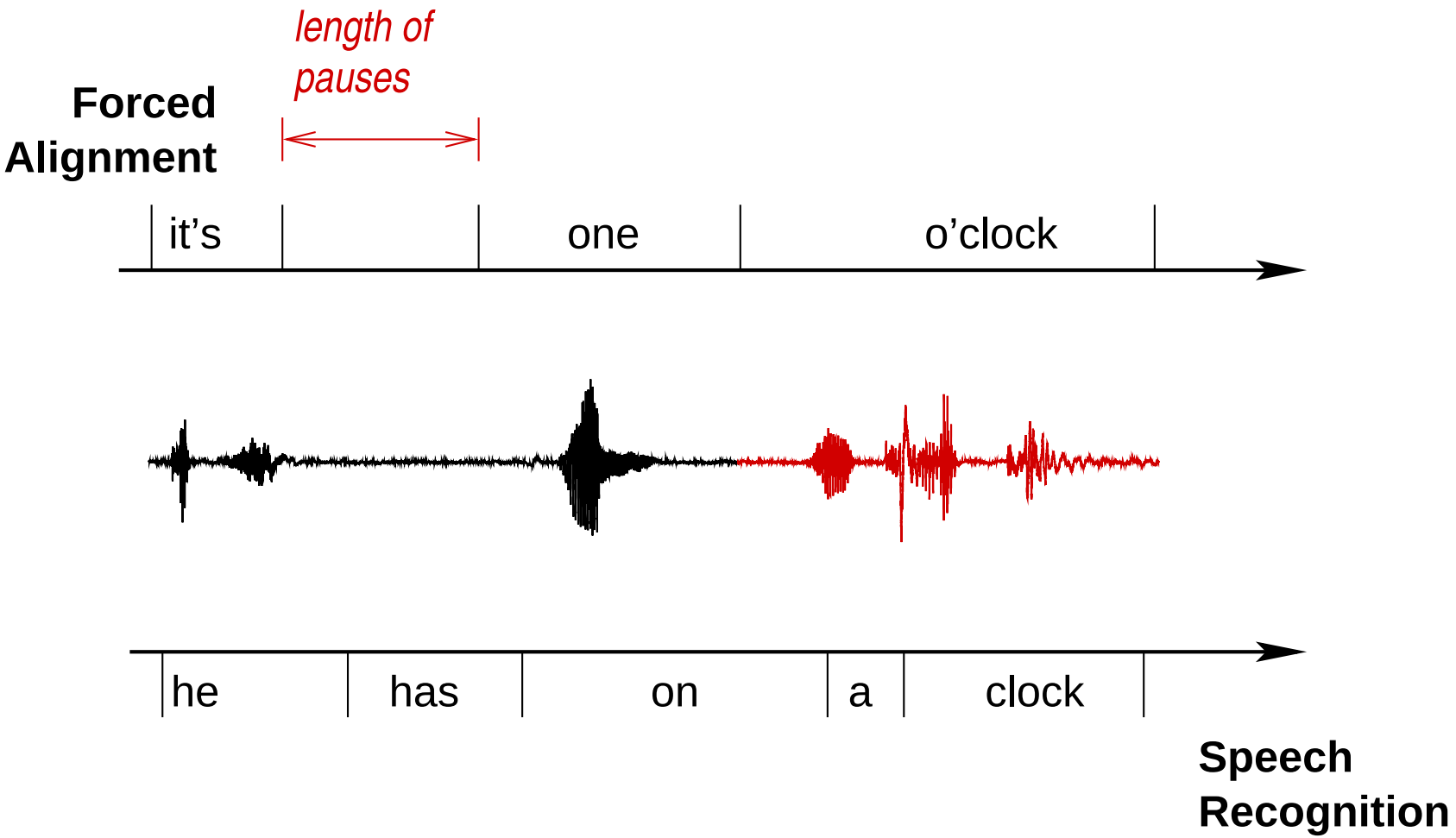
Word Based Features

Word-Level Assessment



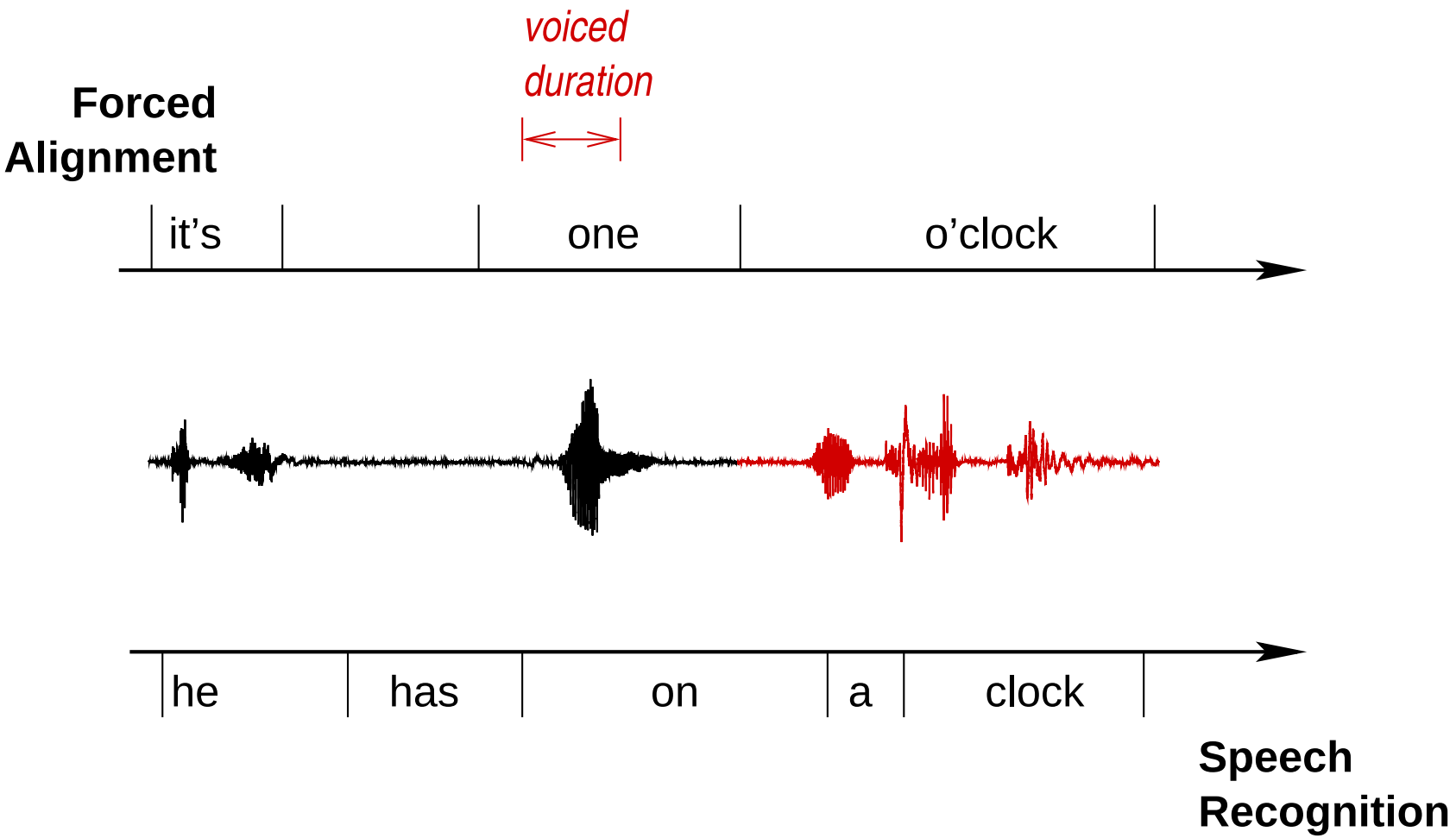
Word Based Features

Word-Level Assessment



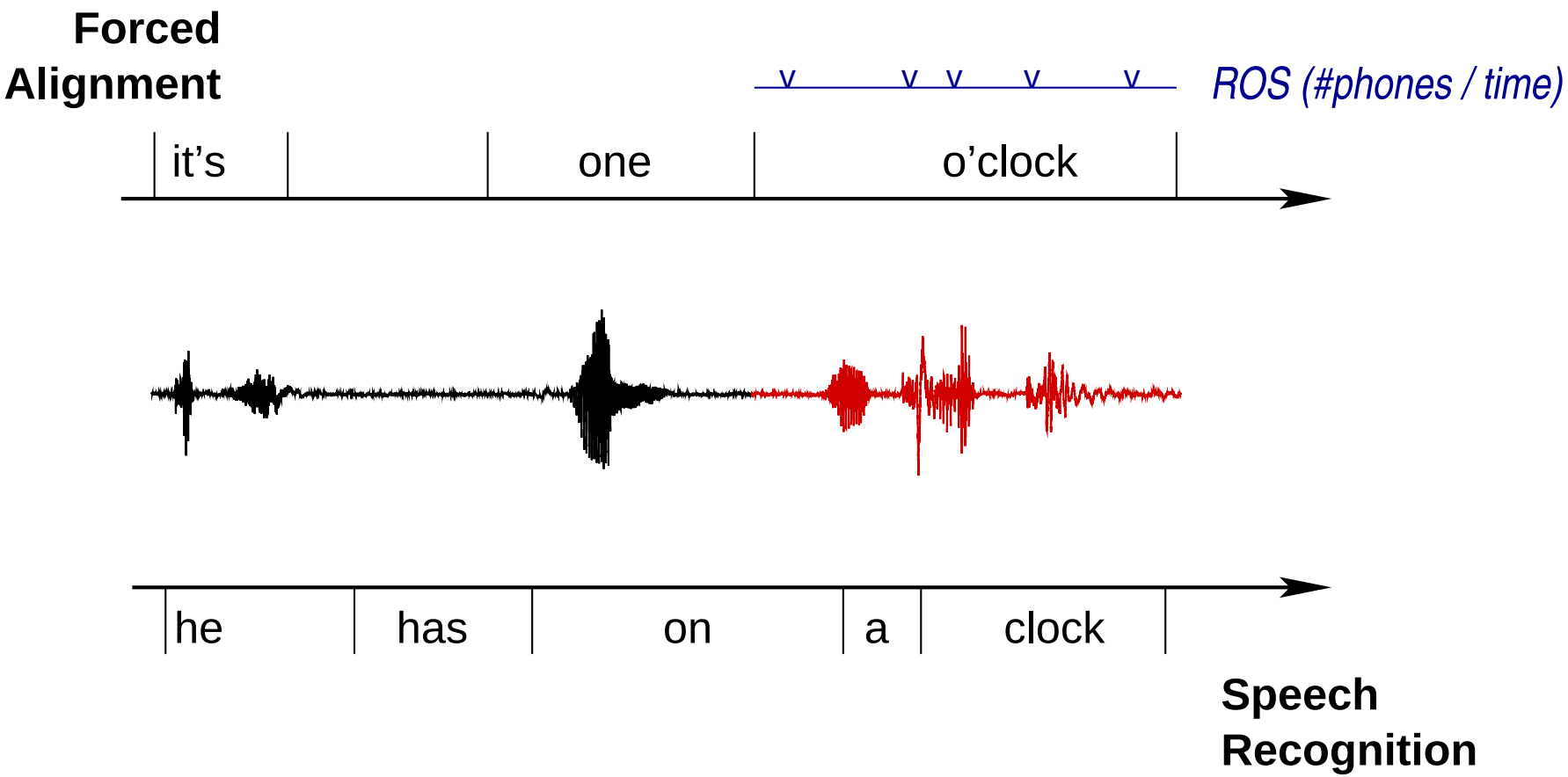
Word Based Features

Word-Level Assessment



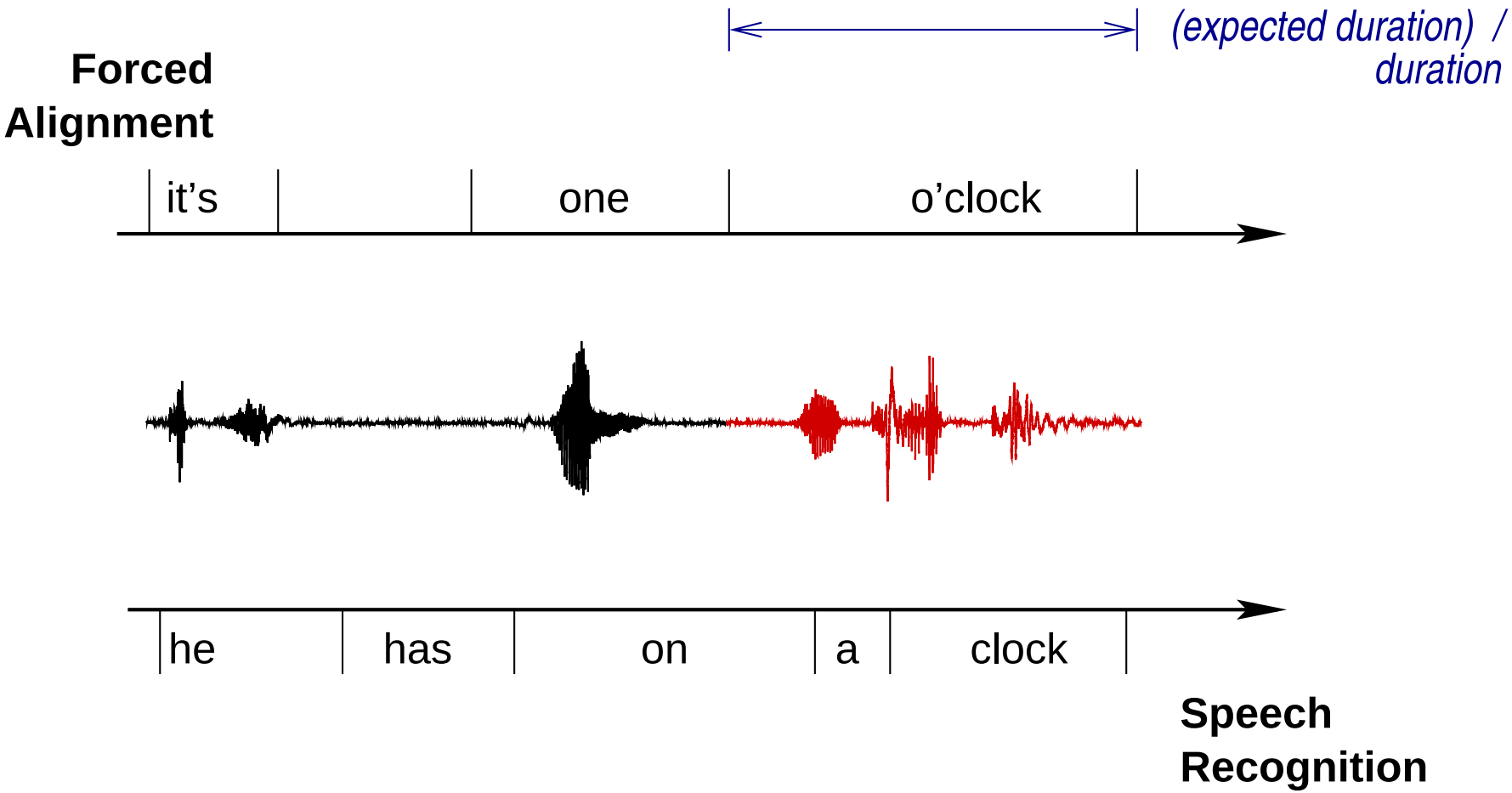
Word Based Features

Word-Level Assessment



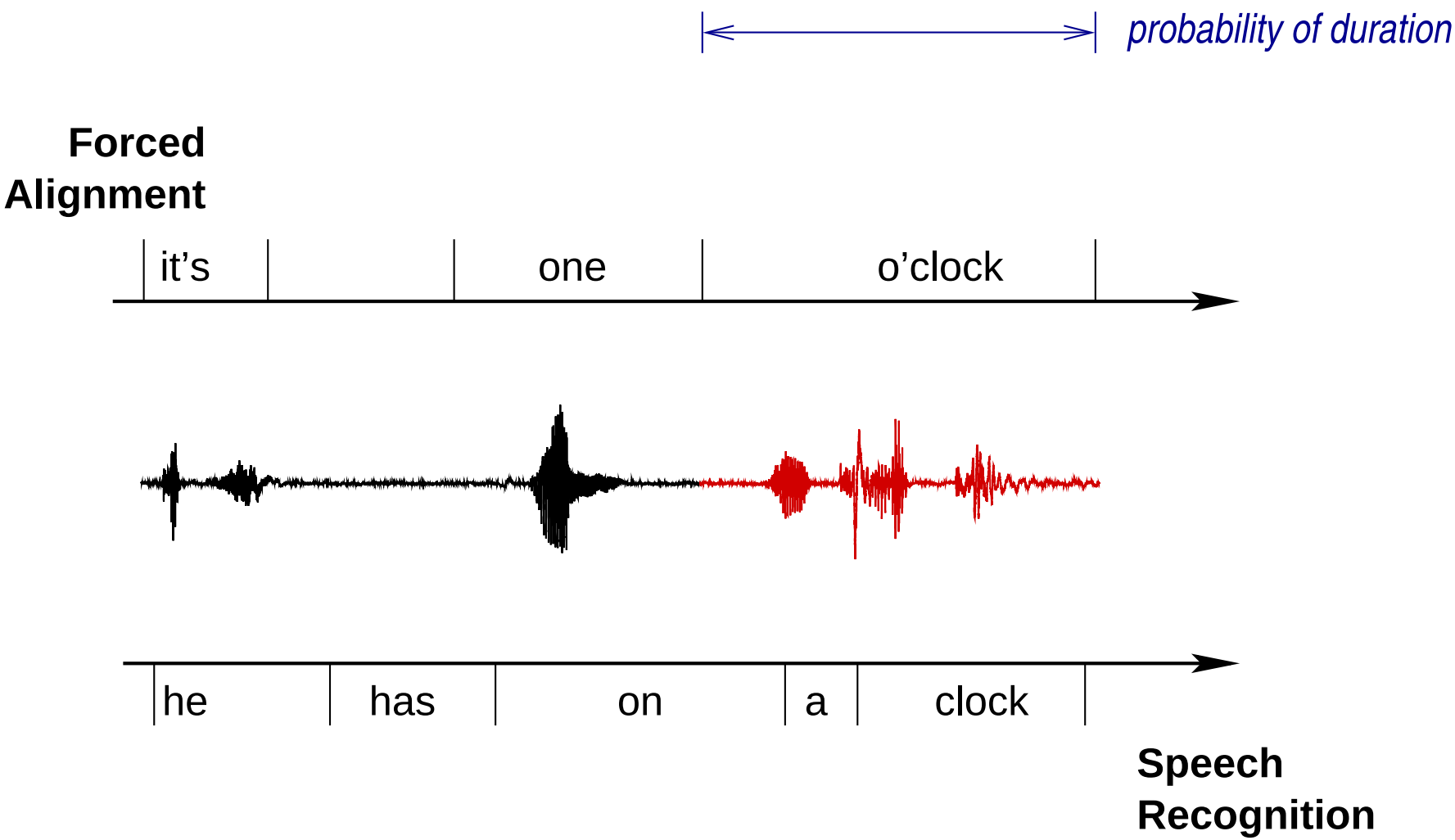
Word Based Features

Word-Level Assessment



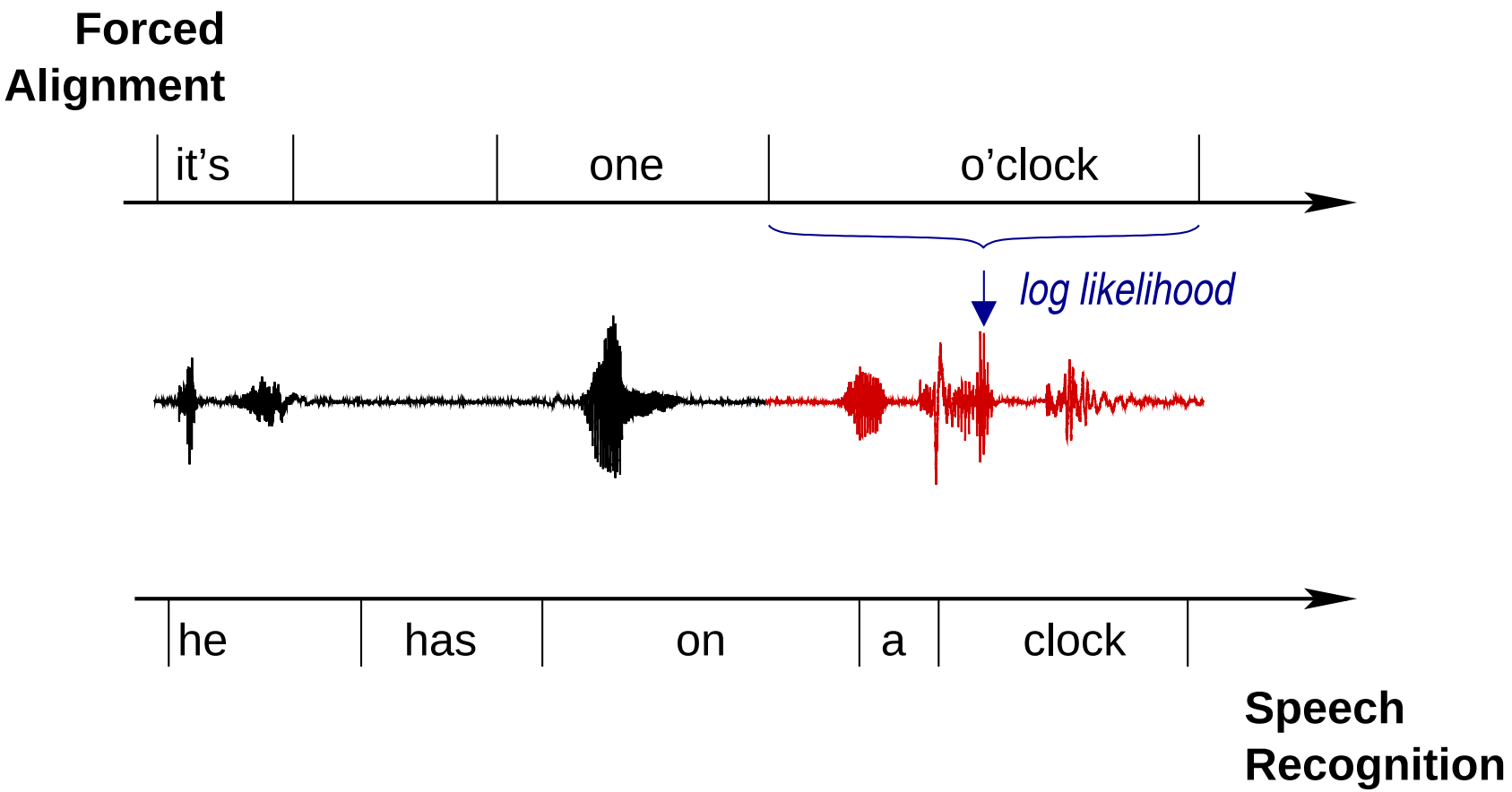
Word Based Features

Word-Level Assessment



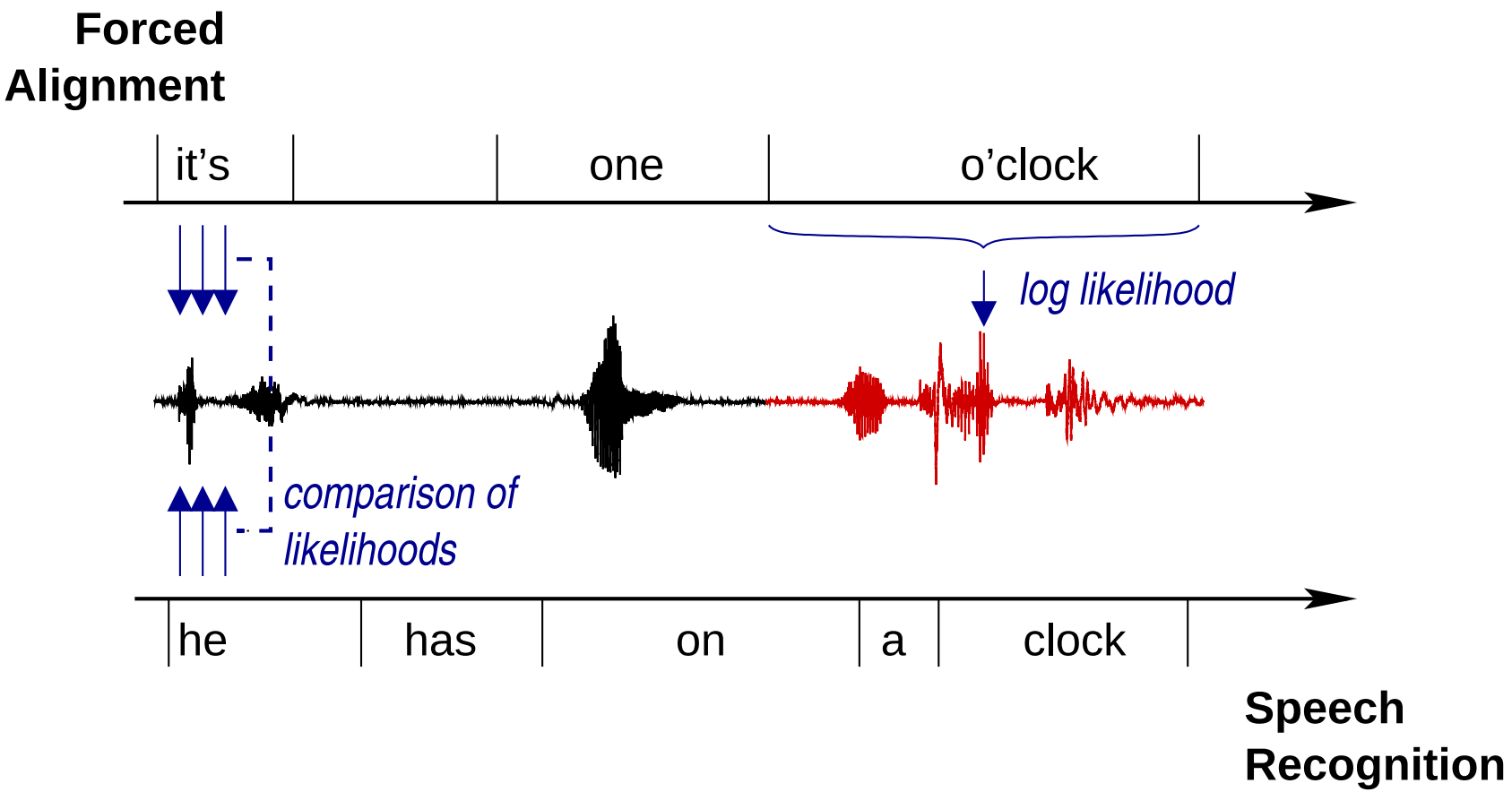
Word Based Features

Word-Level Assessment



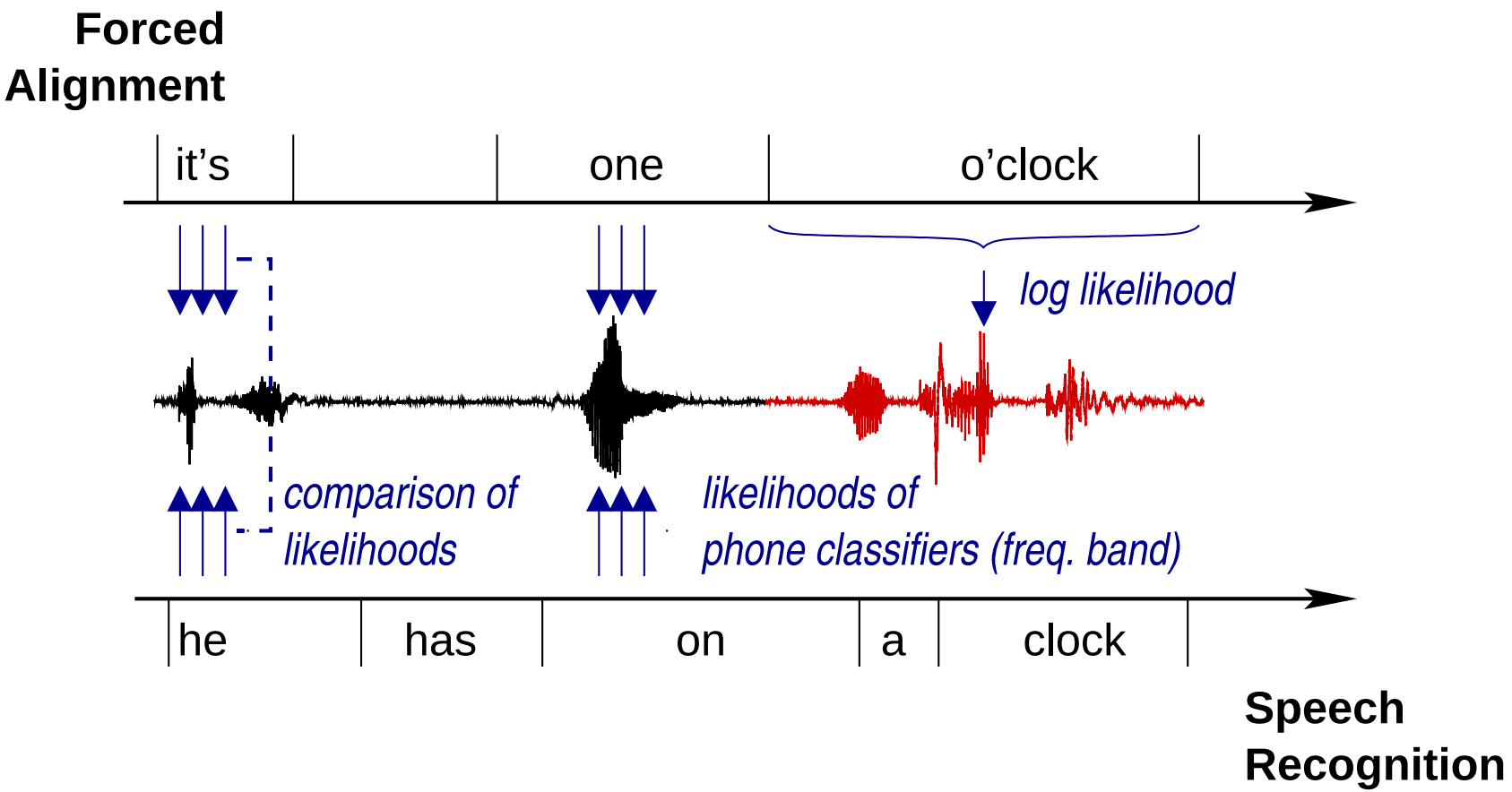
Word Based Features

Word-Level Assessment



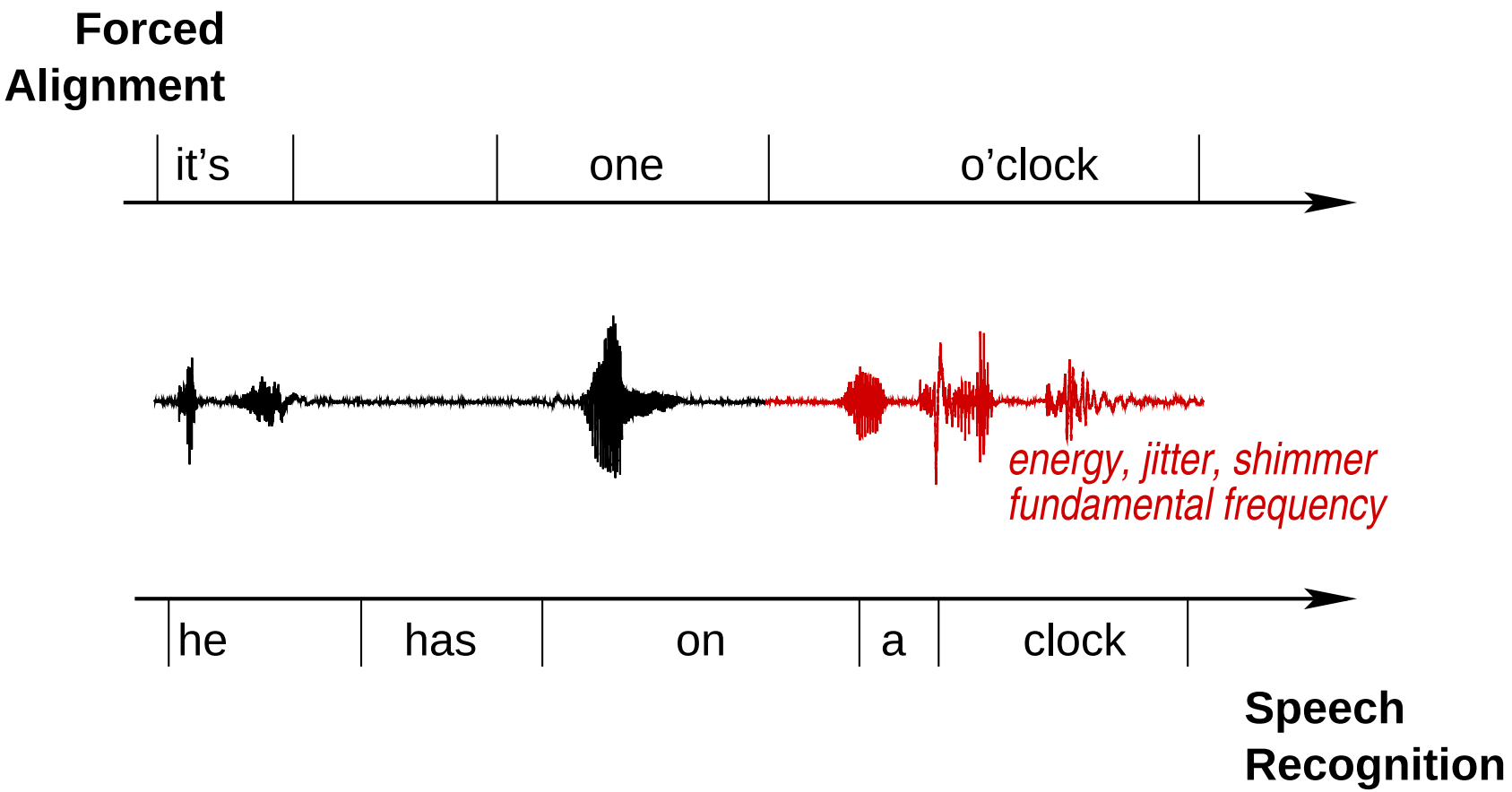
Word Based Features

Word-Level Assessment



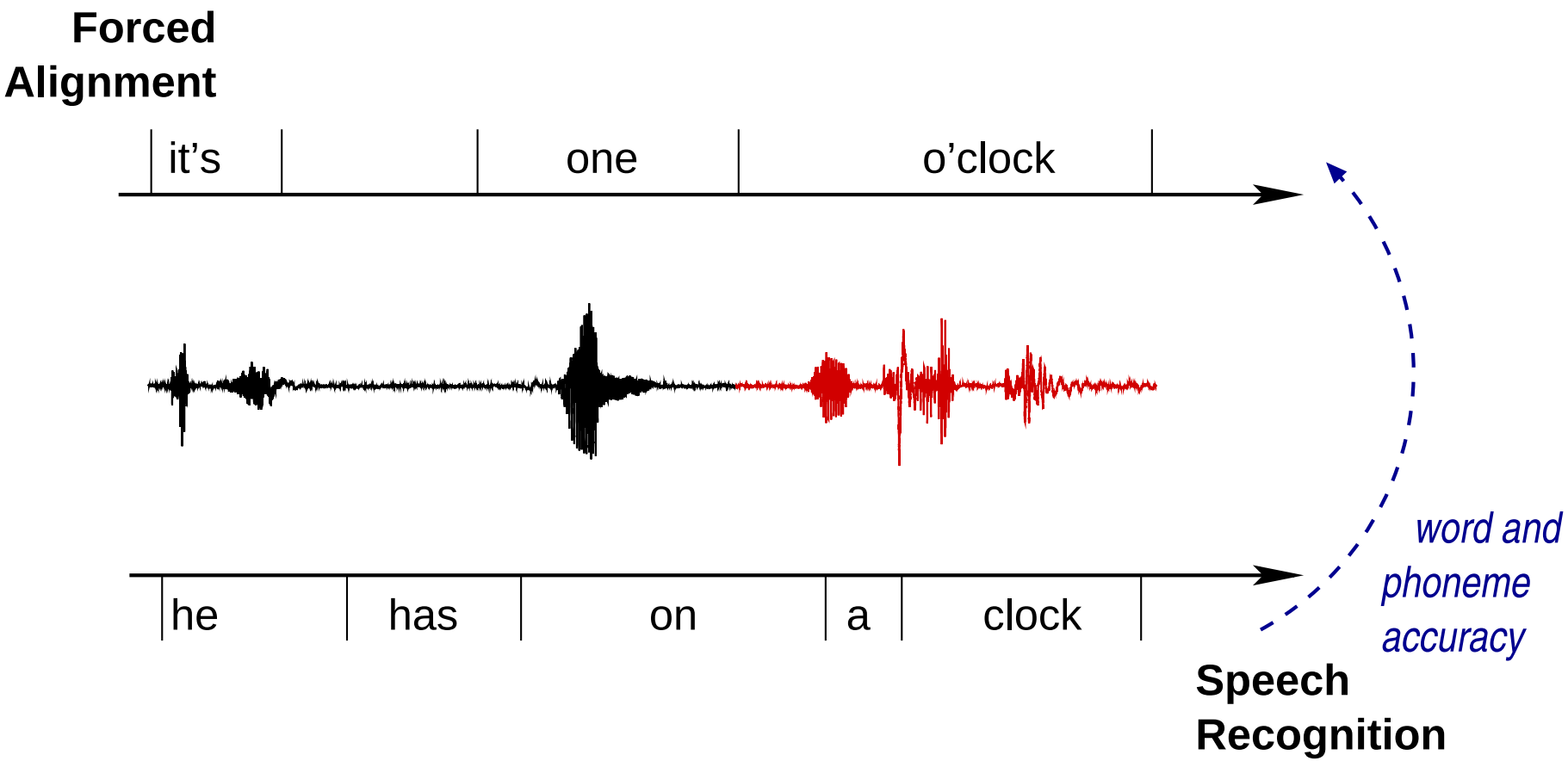
Word Based Features

Word-Level Assessment



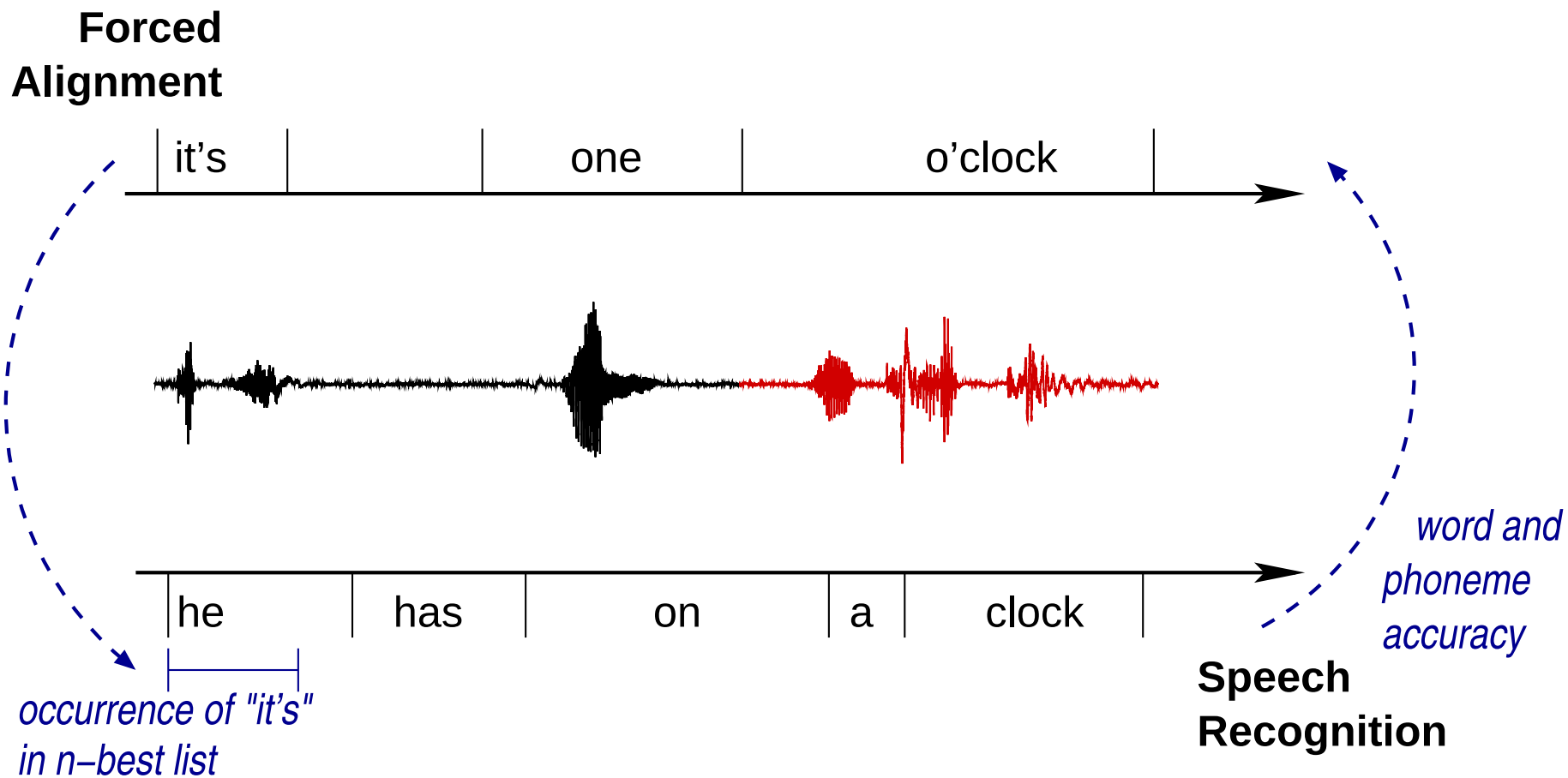
Word Based Features

Word-Level Assessment



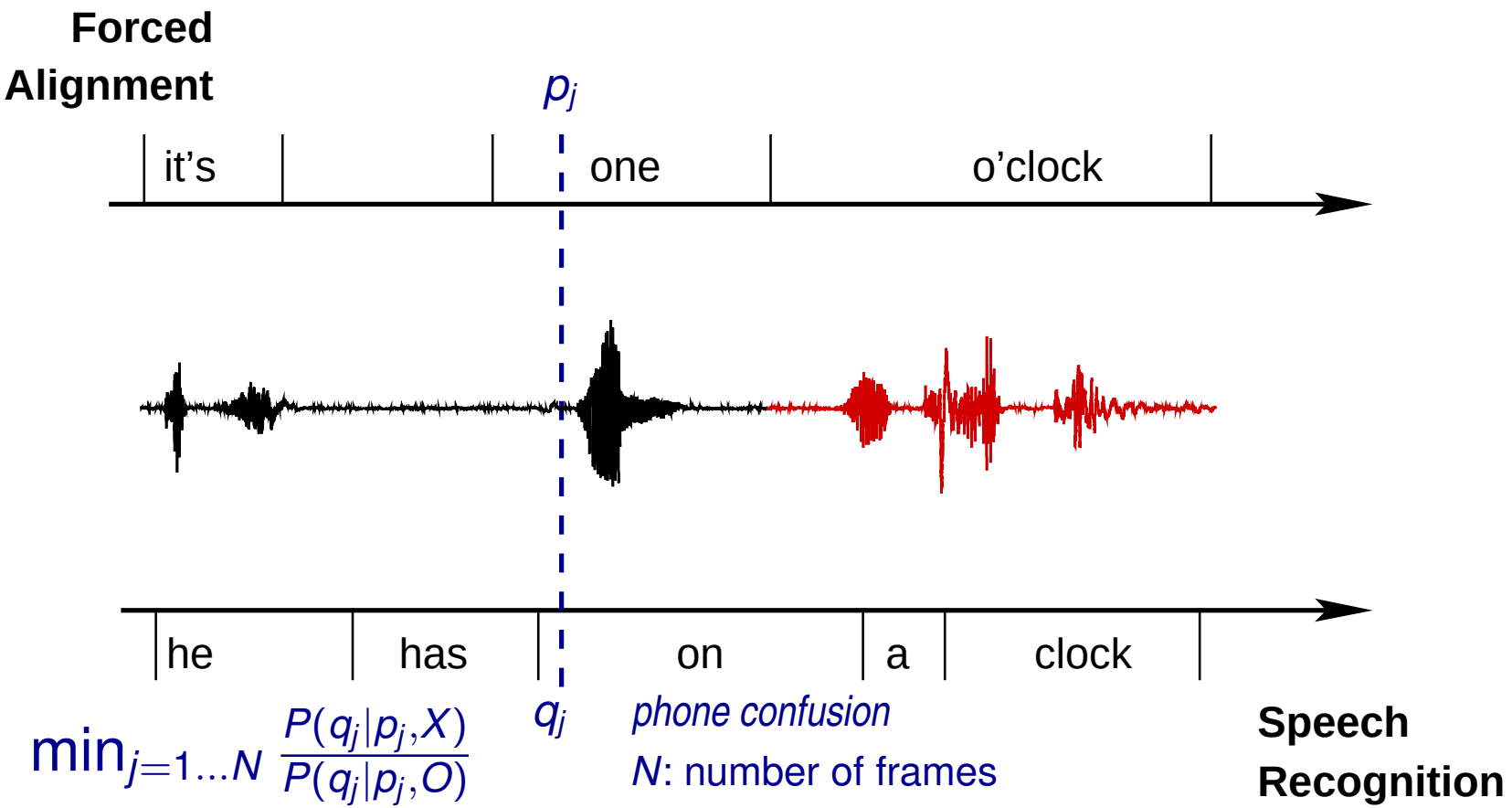
Word Based Features

Word-Level Assessment



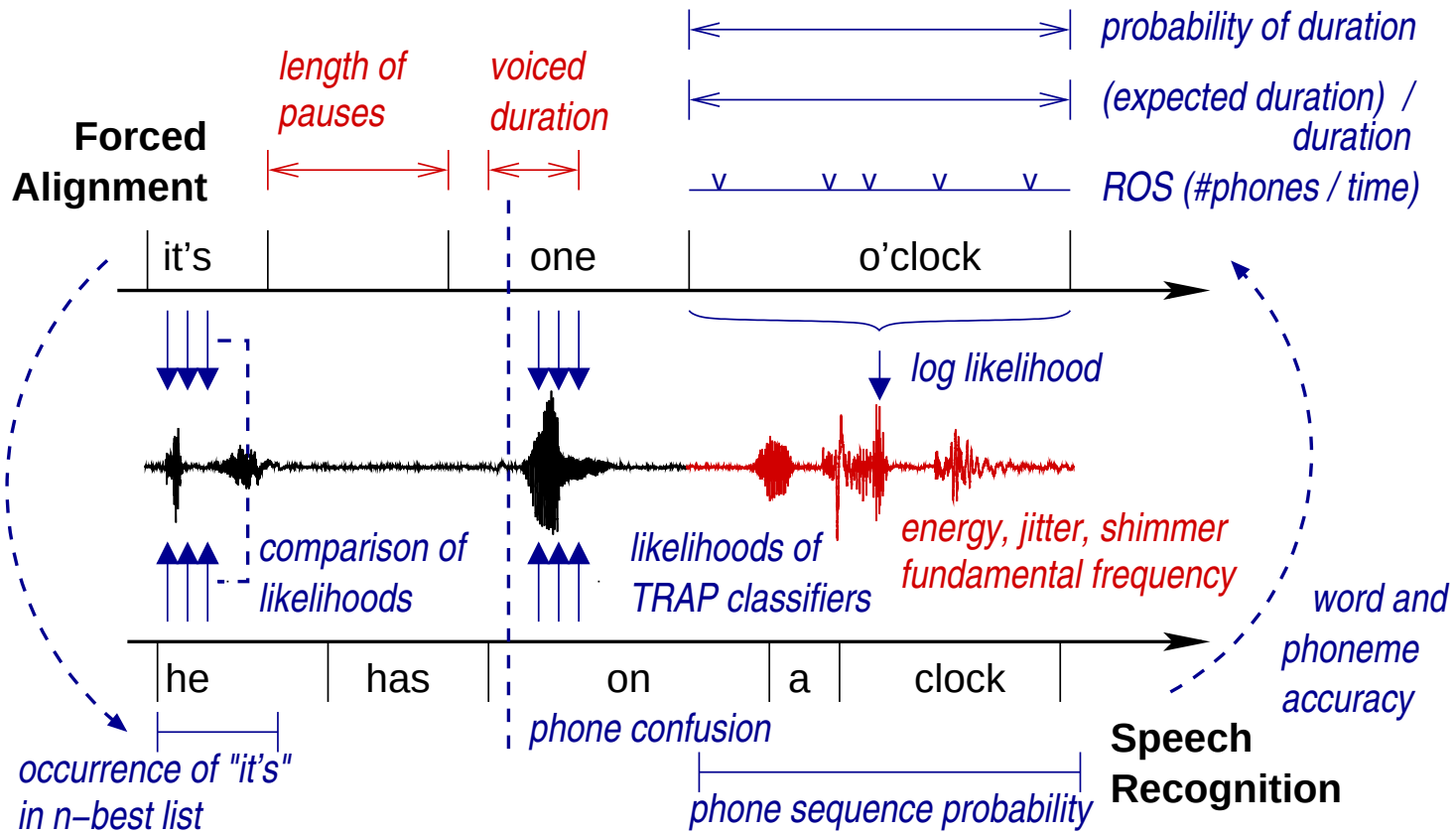
Word Based Features

Word-Level Assessment



Word Based Features

Word-Level Assessment

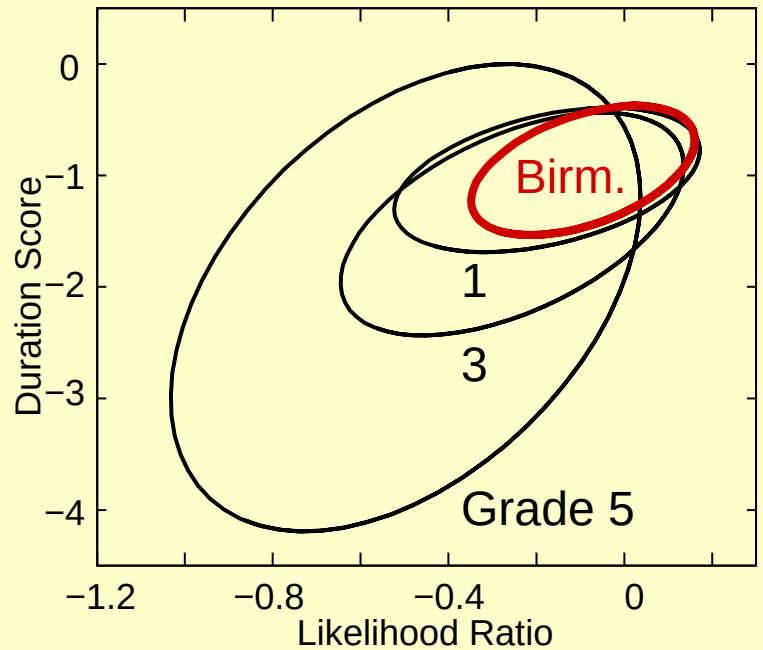


- 75 pronunciation features and **124 prosodic features** per word

Evaluation with Native Models

Sentence-Level Assessment

- Special sentence-level pronunciation and prosodic features (449)
- Mahalanobis distance from native speakers
- Convert distance values into scores; feature selection with AdaBoost
- Non-native data only required for validation



Outline

1 CALL Applications

2 Corpus and Annotations

3 Agreement of Experts

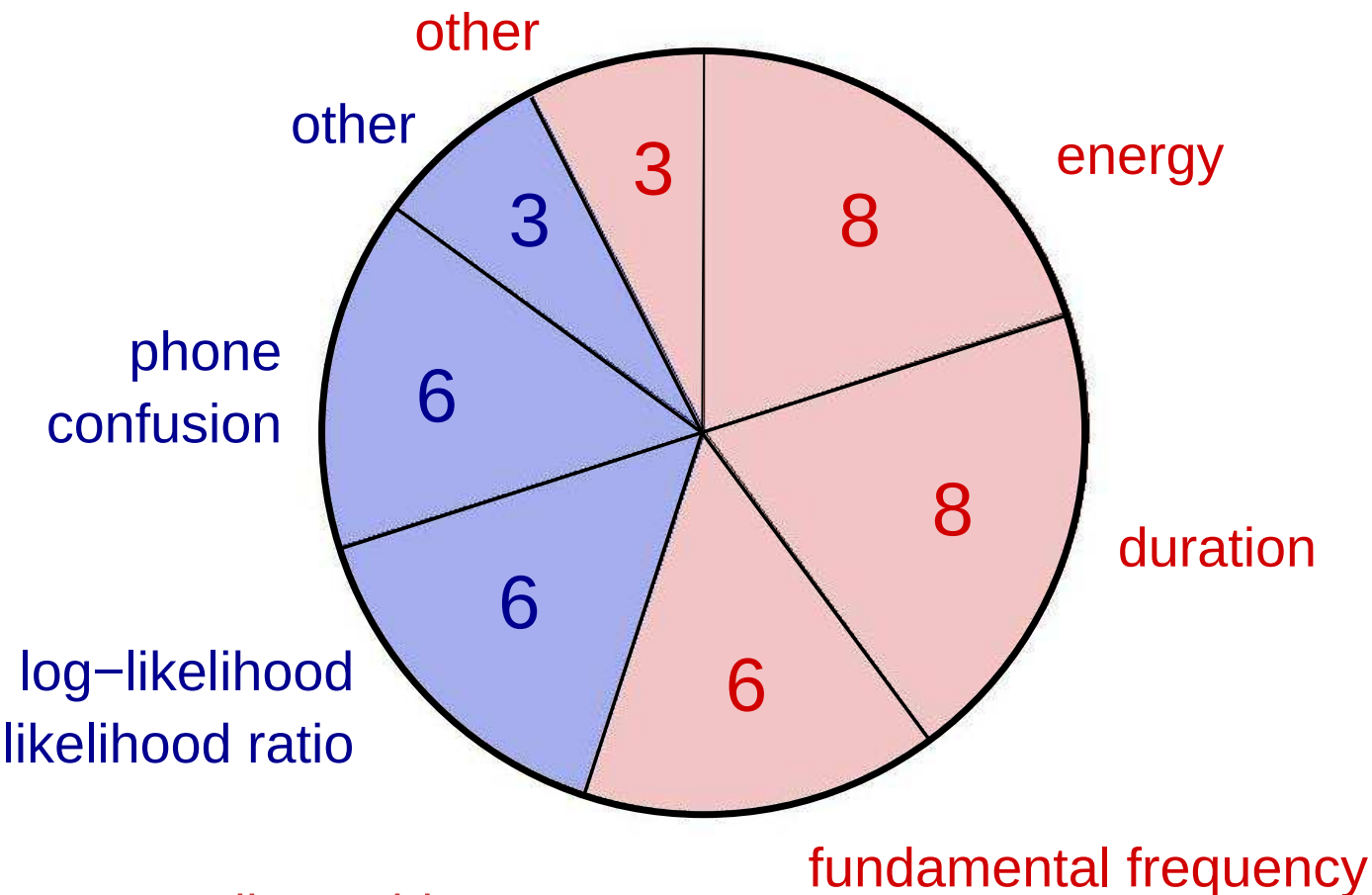
4 Approaches for the Automatic Assessment

5 Experimental Results
- Word-Level Assessment
- Text-Level Assessment

Feature Groups selected with AdaBoost

Word-Level Results

40 features, selected with AdaBoost:



- other: pauses, jitter, shimmer
- other: accuracy, confidence

Features selected with AdaBoost

Top word-level features:

1	**phone confusion**: minimum
2	**log-likelihood**: mean over phonemes
3	**duration**: expected / observed
4	**energy**: mean
5	**energy**: FFT coefficient of the en. contour

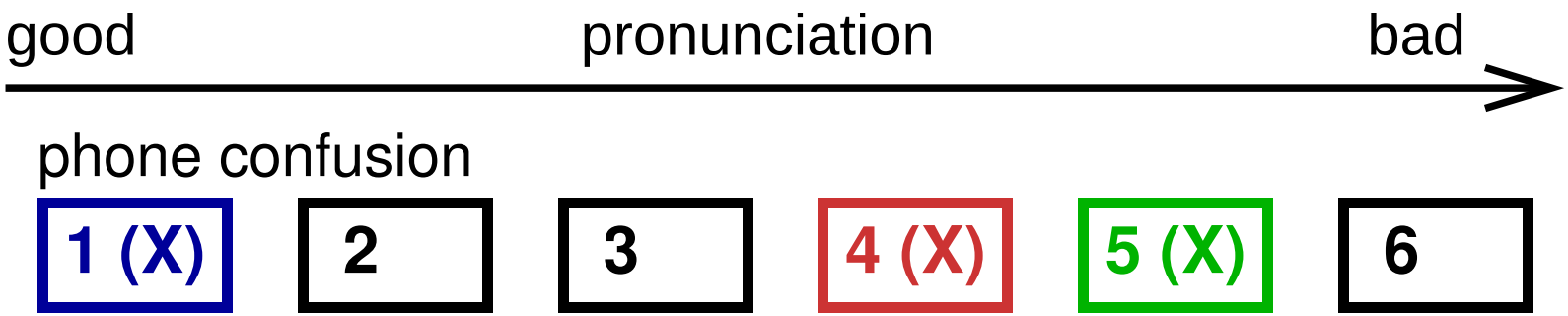
Top sentence-level features:

1	**pauses**: total duration of long pauses
2	**log-likelihood**: mean over phonemes
3	**phone confusion**: minimum
4	**fundamental frequency**: maximum of word based minima
5	**energy**: mean of normalised words

Features selected with AdaBoost (cont.)

Word-Level Results

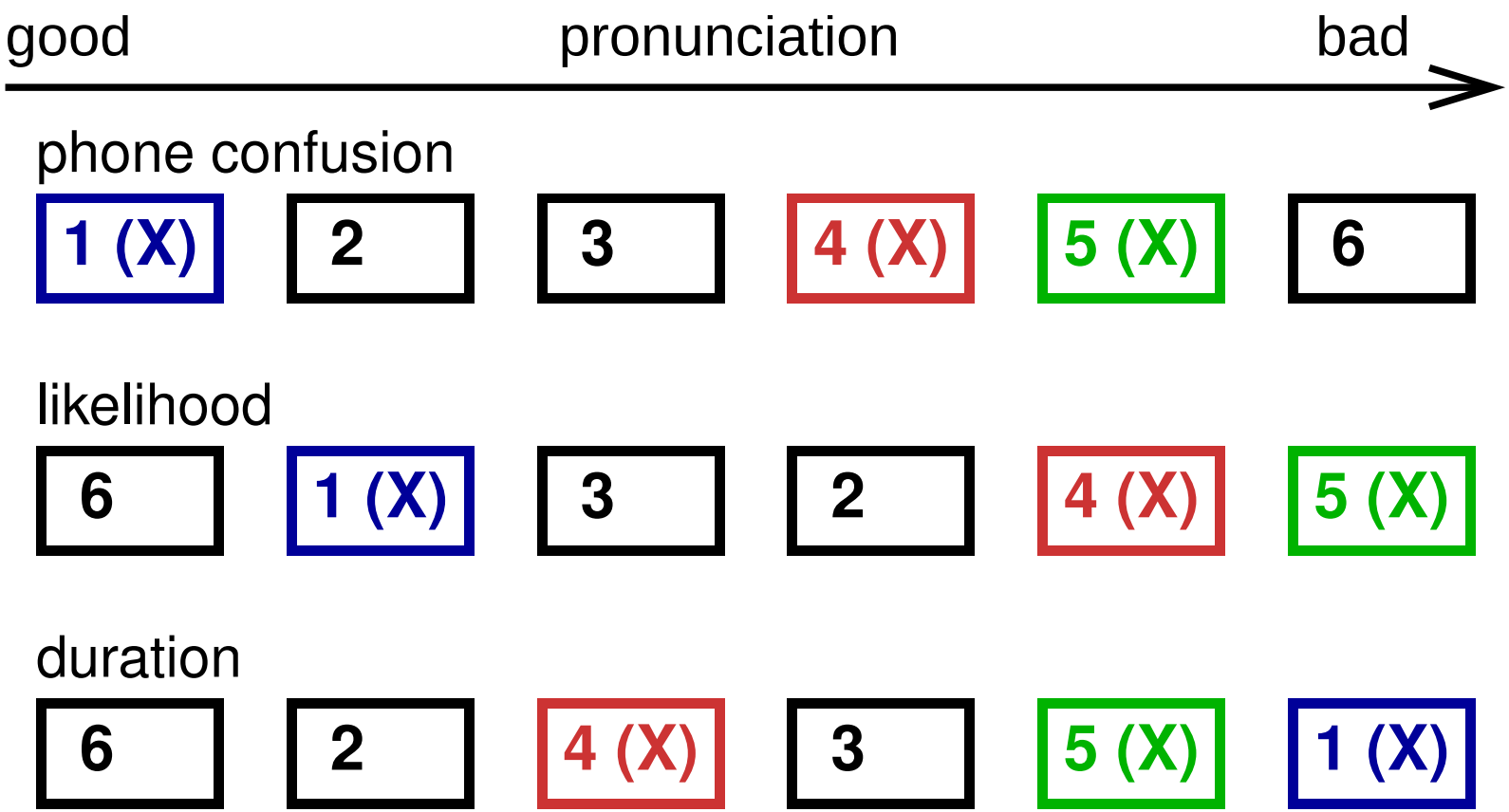
Word “thirteen”: Speakers 1–6
Wrong pronunciation (X) highlighted with colours



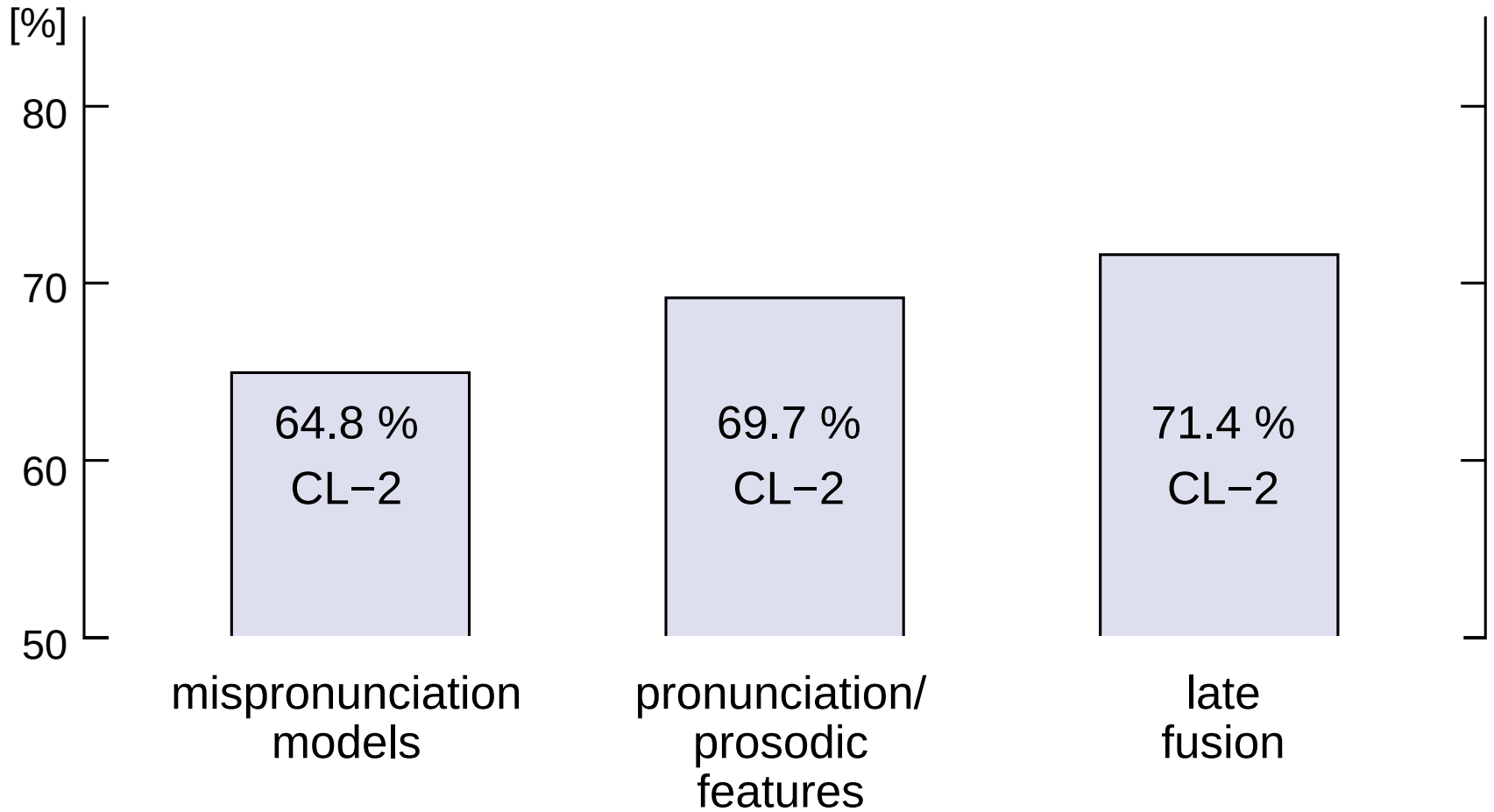
Features selected with AdaBoost (cont.)

Word-Level Results

Word “thirteen”: Speakers 1–6
Wrong pronunciation (X) highlighted with colours



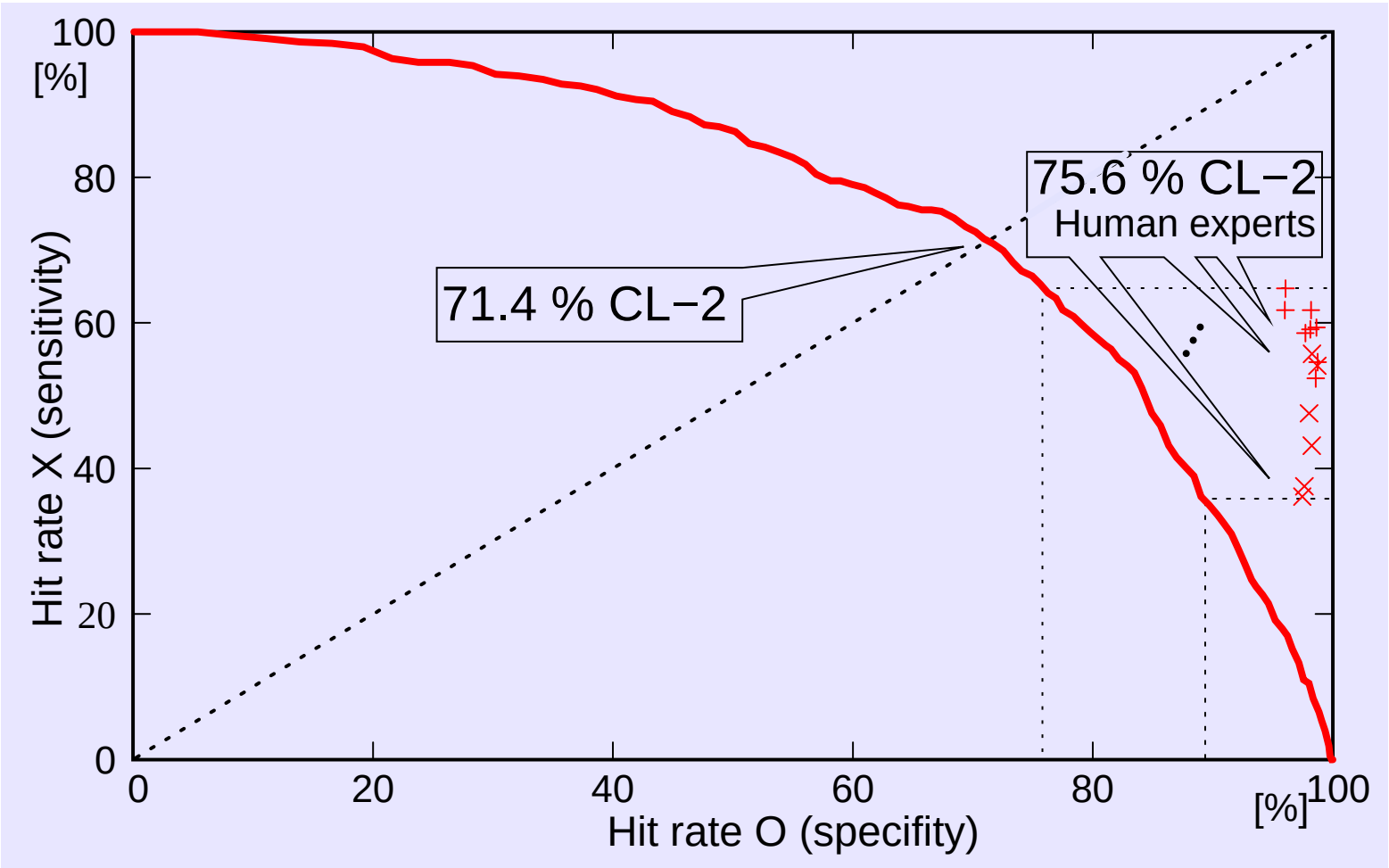
Word-Level Results



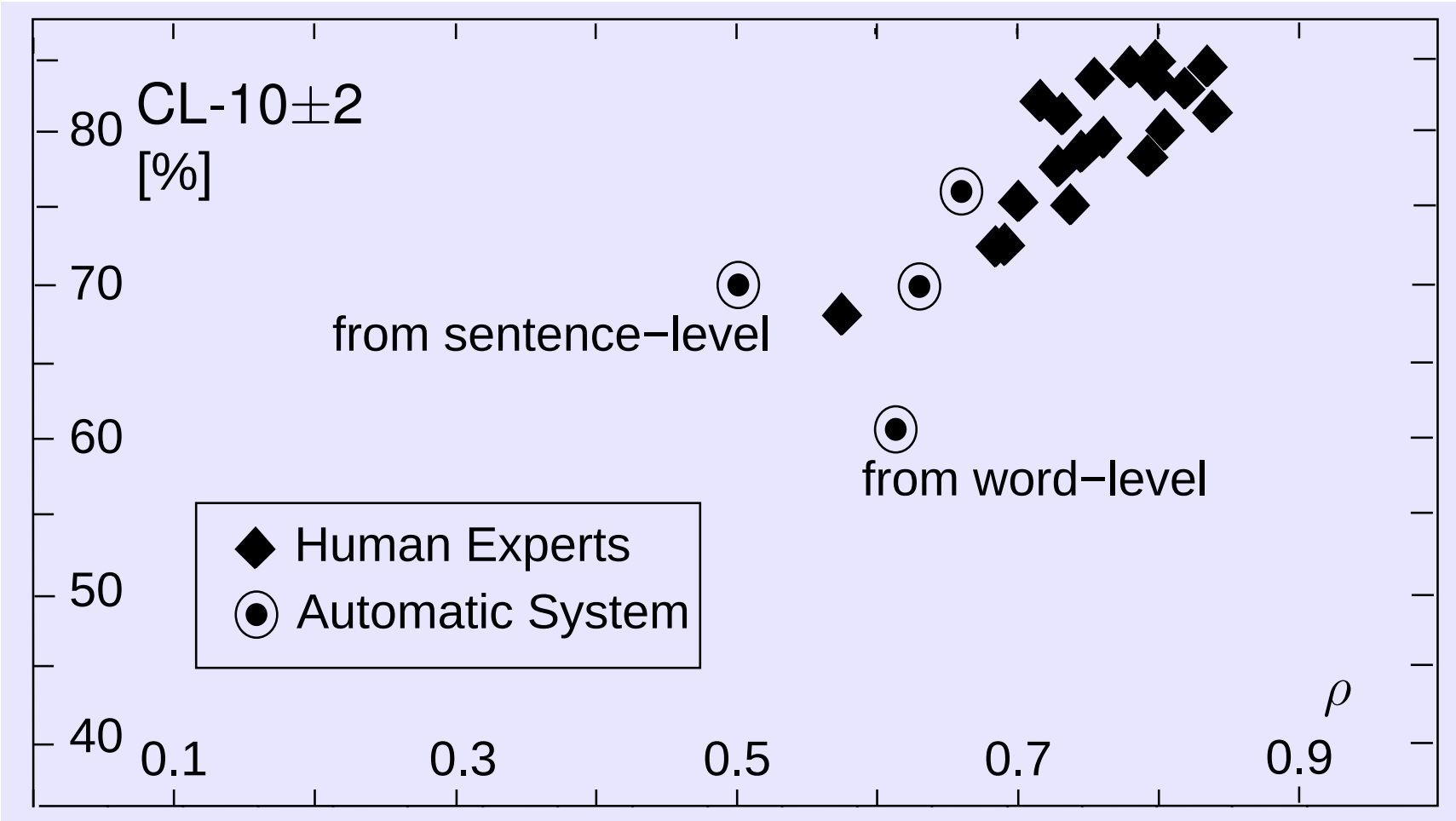
- Mispronunciation models: low WER important.
- 69.7 % → 71.4 %: significance level 0.05

ROC-Evaluation

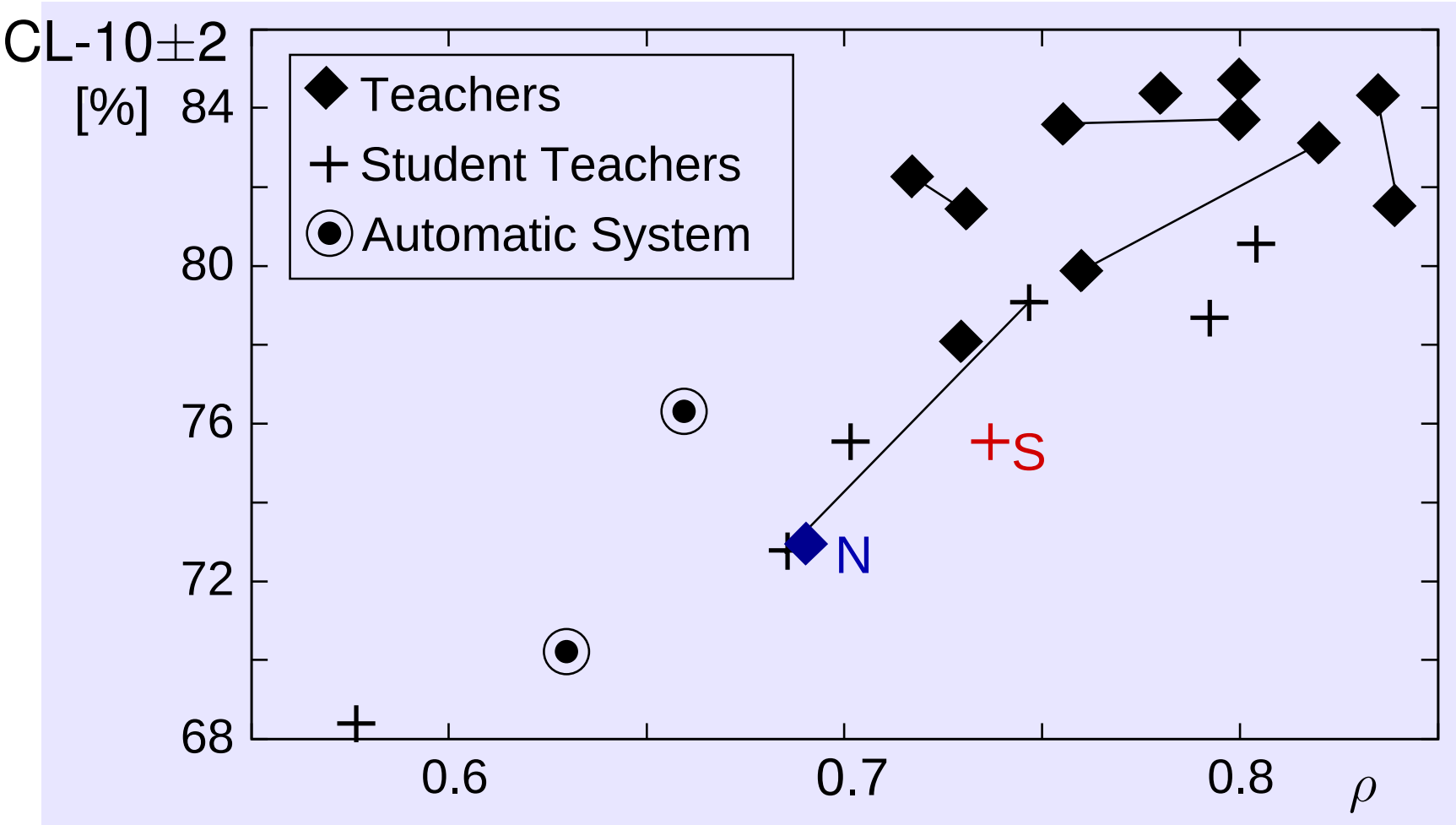
Word-Level Results



Text-Level Results



Text-Level Results (cont.)



Summary

- Pf-Star non-native corpus, annotations by 14 experts
- Speech recognition for non-native children
- Algorithms for automatic assessment:
  - Mispronunciation models
  - Pronunciation and prosodic features
  - Distance from native data

Summary

- Pf-Star non-native corpus, annotations by 14 experts
- Speech recognition for non-native children
- Algorithms for automatic assessment:
  - Mispronunciation models
  - Pronunciation and prosodic features
  - Distance from native data
- Still room for improvement
- Text-Level: Closed to human performance
- Word-Level: Too many false alarms $\rightarrow$ concentrate on important words

Thank you for your attention.