3rd International Conference on
Statistical Language and Speech
Processing, SLSP 2015

# Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features

Raheel Qader[1], Gwénolé Lecorvé[1], Damien Lolive[1], Pascale Sébillot[2]

[1] IRISA/Université de Rennes 1

[2] IRISA/INSA de Rennes

26/11/2015

# Introduction

- Many pronunciation variations occur in spontaneous speech

- Degradation in performance of  speech applications
  - Automatic Speech Recognition (ASR)
    - Low accuracy
  - Text To Speech (TTS)
    - Lack of expressivity
    - Flat style

- Example:
  - *went → [wɛnt] , [wɛn] , [wənt]*
  - *I want to go → [aɪ wɒn ɒ goʊ]*

# Introduction

**How to produce spontaneous pronunciation for TTS?**

- Adapting standard pronunciations to a spontaneous style
  - By predicting addition, deletion and substitution of phonemes
  - Using linguistic features and Conditional Random Fields (CRFs)

# Outline

- State of the art
- Corpus
- Method overview
- Feature selection
- Experiments
- Conclusion and future work

# State of the art

- Early work: phonological rules [Tajchman et al., 1995]
- Recent work [Vazirnezhad et al., 2009; Prahallad et al., 2006; Karanasou et al., 2013]
  - Machine learning: decision trees, HMMs, neural networks, random forests, CRFs
- Features types
  - Acoustic (F0, energy, duration) [Bates and Ostendorf, 2002]
  - Linguistic (syllable stress, part-of-speech, word length) [Bell et al., 2009 ; Vazirnezhad et al., 2009]

# Corpus

- Buckeye conversational English corpus (50%)
  - 20 speakers & 20 hours of recording (randomly selected)
  - Partition: 60% training set, 20% development set, 20% test set

- Existing features
  - Speech signal + orthographic transcription
  - 2 phonemic transcriptions
    - Canonical form
    - Realized form

Aligned

30% phoneme error rate

57% word error rate

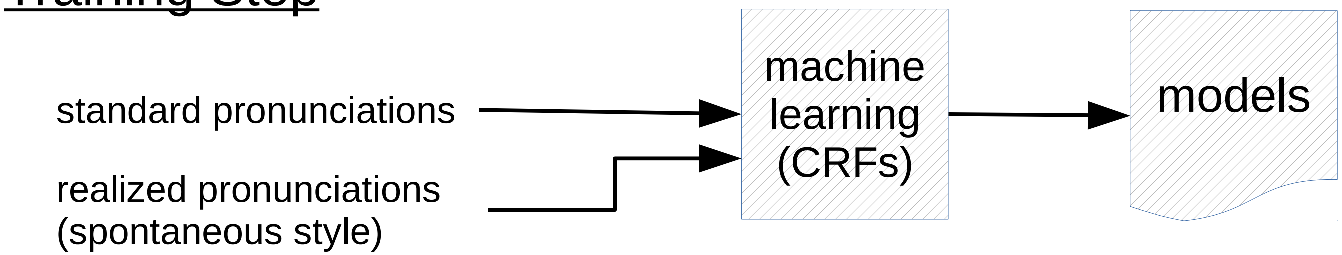
# Linguistic features

- **Utterances**
  - Utterance position
- **Words**
  - Frequency
  - Part of speech (POS)
  - Length
  - Occurrence count
  - Stems
  - Stop words
- **Syllables**
  - Syllable position
  - Syllable type
  - Syllable stress
- **Phonemes, graphemes,** etc**.**

# Method overview

- Pronunciation adaptation performed on each speaker independently

## Training Step



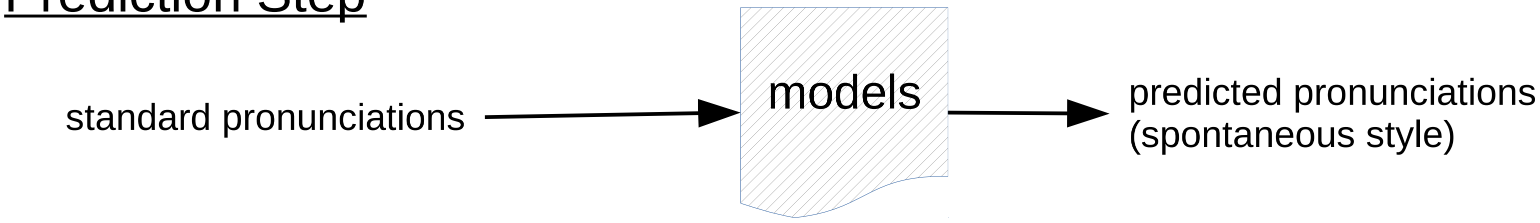
standard pronunciations

realized pronunciations
(spontaneous style)

machine
learning
(CRFs)

models

## Prediction Step

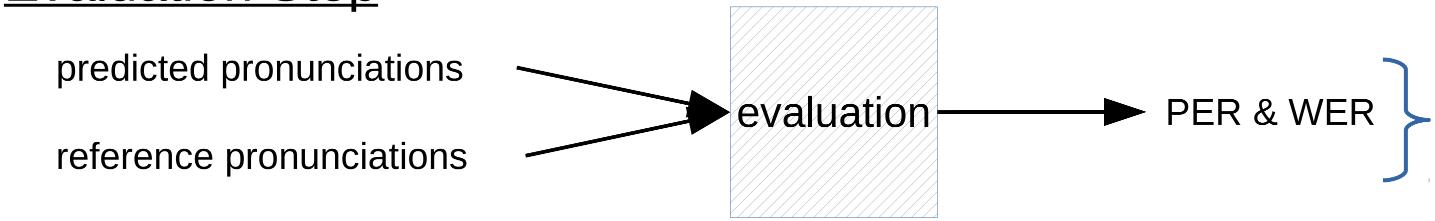


standard pronunciations

models

predicted pronunciations
(spontaneous style)

## Evaluation Step



predicted pronunciations

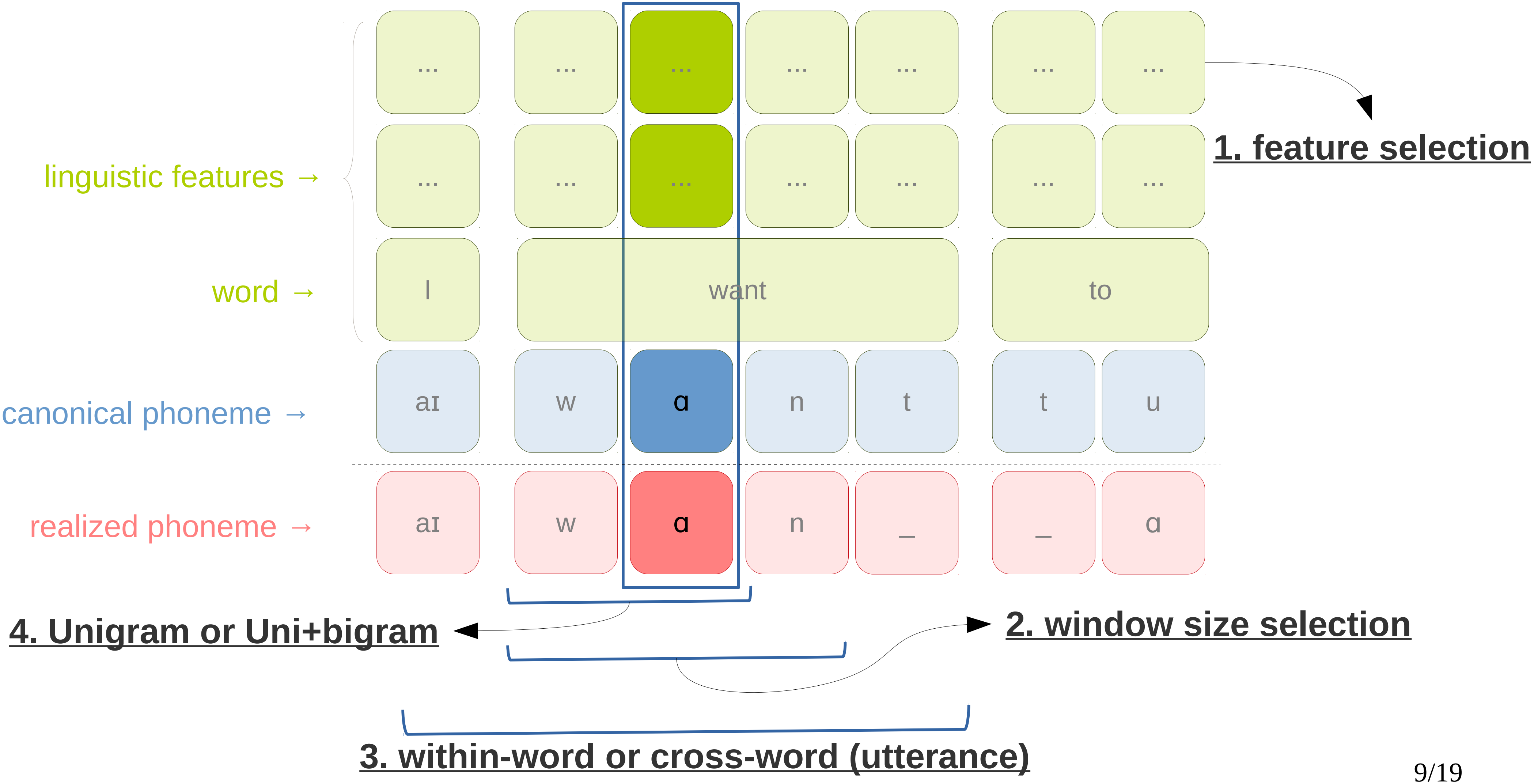
reference pronunciations

evaluation

PER & WER

**Average**
**PER and WER**
**over all speakers**

# Method overview

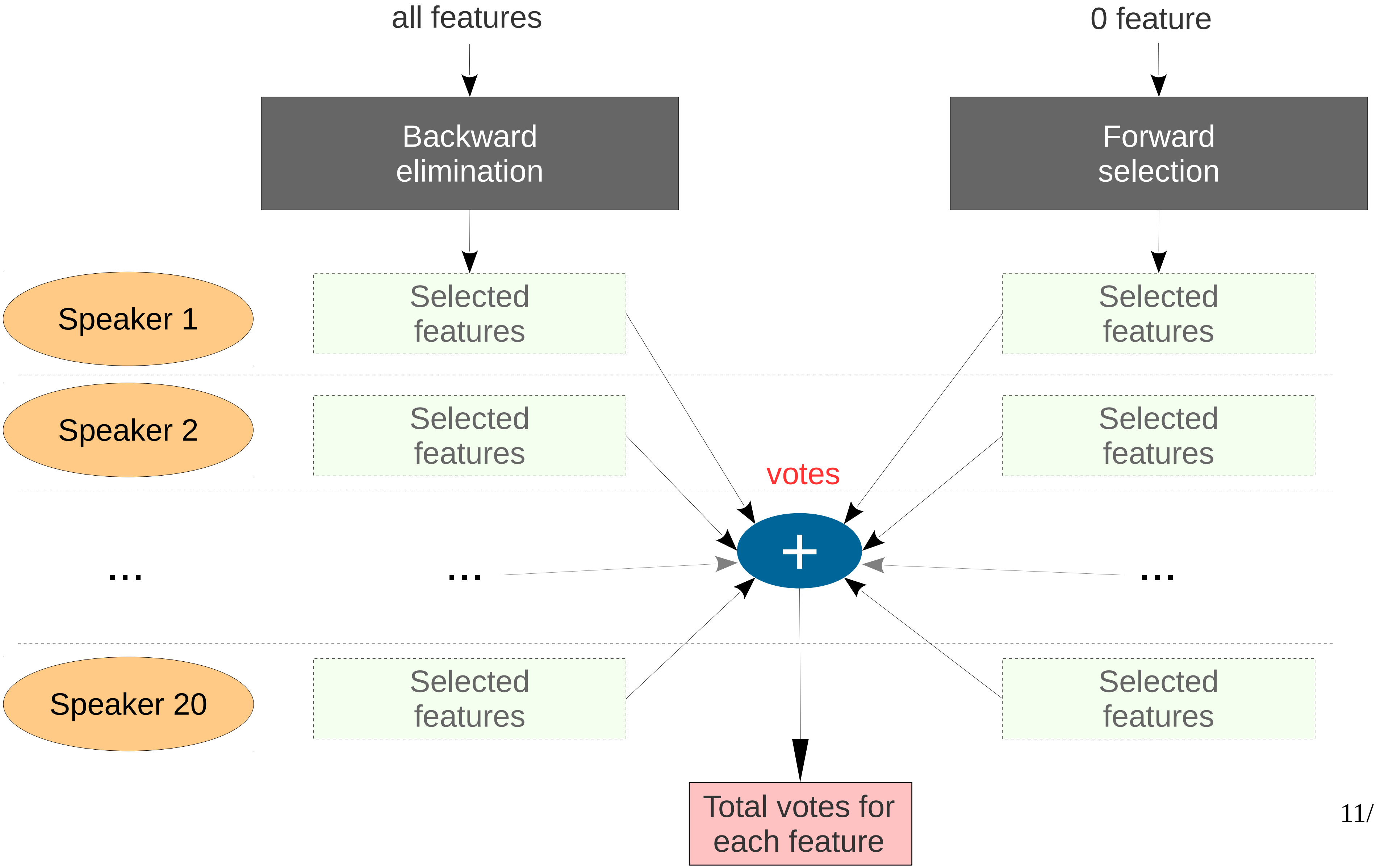


# Feature selection

- Why?
  - Having too many features
    - Results in overfitting the data
    - Increase the time needed for training process
  - Some features might be irrelevant and redundant
  - Limitations in computational resources
  - Limited training data

- Proposed solution: reduce the number of features

# Feature selection



# Feature selection

| Feature | votes |
|---|---|
| Canonical phoneme | 40 |
| Word | 40 |
| Is a stop word (true/false) | 24 |
| Syllable lexical stress | 24 |
| Syllable part (onset/nucleus/coda) | 24 |
| Word frequency in English | 22 |
| Reverse phoneme position in syllable | 22 |
| Phoneme position in syllable | 20 |
| Syllable location (first/middle/last) | 20 |
| Stem frequency in the interview | 19 |
| Word frequency in the interview | 18 |
| Syllable type (open/close) | 18 |
| POS | 17 |
| Number of syllables of the word | 17 |
| Stem frequency in English | 16 |
| Grapheme | 16 |
| Word length | 13 |
| Reverse utterance position | 4 |
| Utterance position | 3 |
| Word position | 2 |
| Reverse word position | 0 |
| Word occurrence count in interview | 0 |

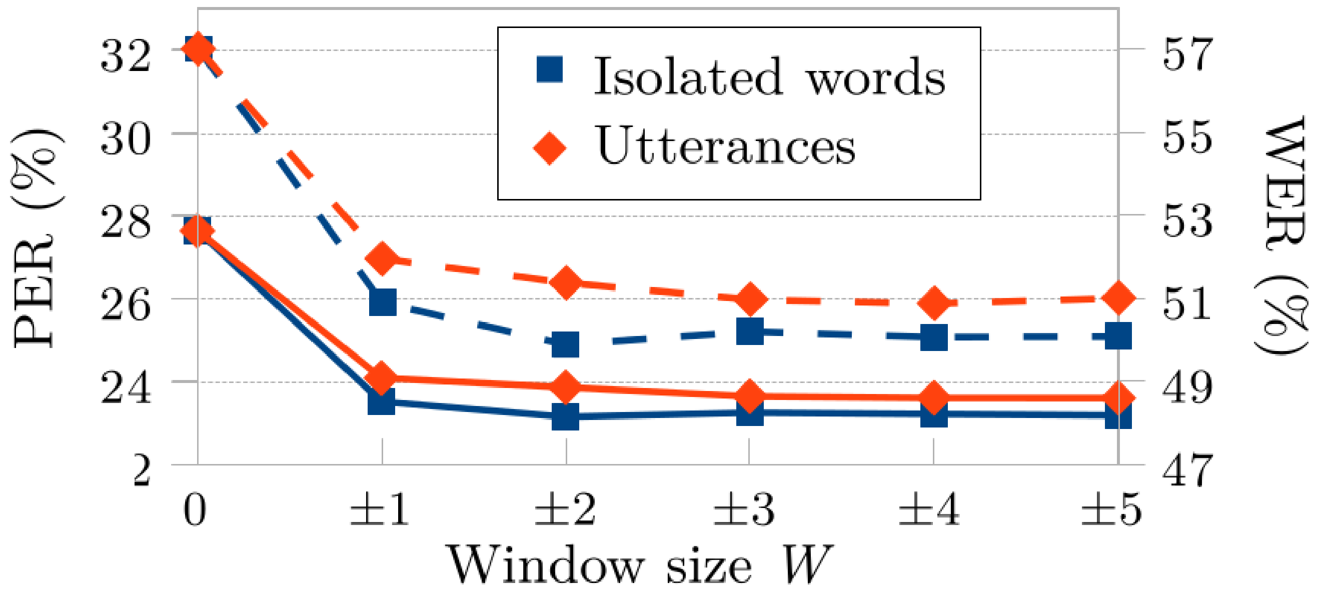
Selected features

Removed features

+ word boundary feature for utterances

# Window size selection

- PER and WER according to window size (neighborhood)



# Backend experiments

- Parameters
  - Features: Canonical phoneme, best features
  - Window size W=0 (no window), W=±2
  - Unit size: word, utterance
  - Feature configuration: unigram, uni+bigram
- Final experiments - on test sets
  - Separate parameters
  - Combined parameters

# Backend experiments

Isolated word

| | | PER (%) | |
|---|---|---|---|
| Baseline | | 30.5 | |
| Canonical phoneme | | 30.4 | [-0.1] |
| | + window | 23.8 | [-6.7] |
| + linguistic features | + window | 23.6 | [-6.9] |

- ➔ Increasing window size leads to significant improvement

Utterance

| | | PER (%) | |
|---|---|---|---|
| Baseline | | 30.3 | |
| + linguistic features | + window | 23.4 | [-6.9] |

- ➔ Including cross-word information provides minimal improvement

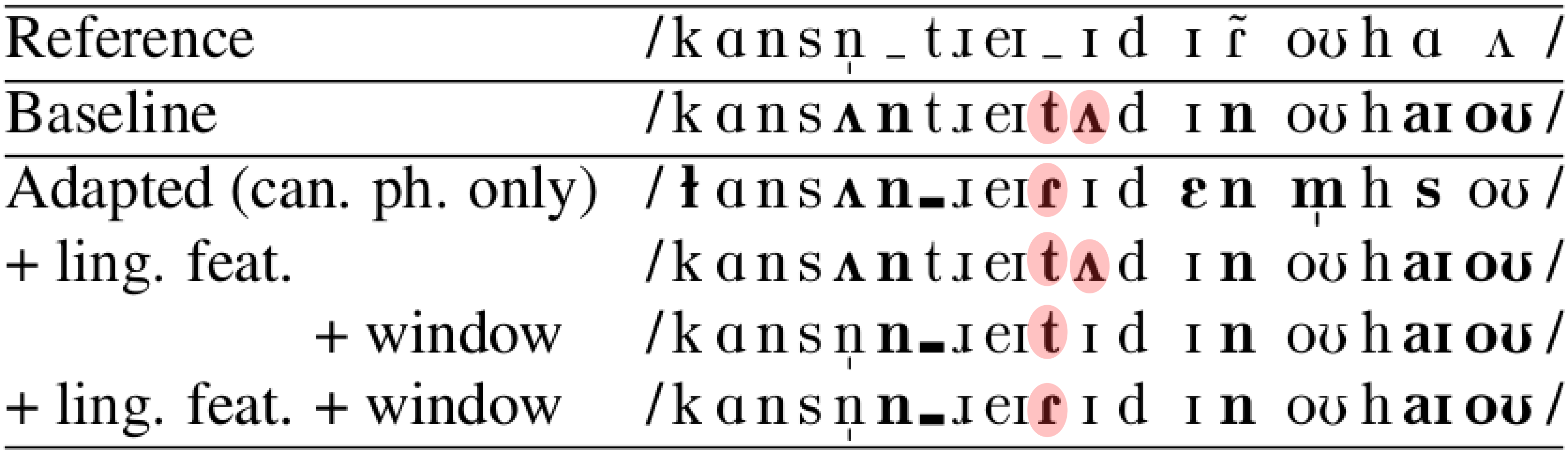
Unigram vs Uni+bigram (using linguistic features + window)

| | | |
|---|---|---|
| Isolated word | Unigram | 23.6 |
| | Uni+bigram | 24.2 |
| Utterance | Unigram | 23.4 |
| | Uni+bigram | 24.4 |

- ➔ Uni+bigram configuration Increases the error rate

# Example

Pronunciation samples predicted by different configurations for the phrase ”**concentrated in Ohio**”



| | | |
|---|---|---|
| Reference | /k ɑ n s n̩ _ t ɹ eɪ _ ɪ d ɪ ɾ̃ oʊ h ɑ ʌ / | |
| Baseline | /k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d ɪ **n** oʊ h **aɪ oʊ**/ | [7 errors] |
| Adapted (can. ph. only) | /**ɫ** ɑ n s **ʌ n _** ɹ eɪ **ɾ** ɪ d **ɛ n** **m̩** h **s** oʊ/ | [10 errors] |
| + ling. feat. | /k ɑ n s **ʌ n** t ɹ eɪ **t ʌ** d ɪ **n** oʊ h **aɪ oʊ**/ | [7 errors] |
| + window | /k ɑ n s n̩ **n _** ɹ eɪ **t** ɪ d ɪ **n** oʊ h **aɪ oʊ**/ | [6 errors] |
| + ling. feat. + window | /k ɑ n s n̩ **n _** ɹ eɪ **ɾ** ɪ d ɪ **n** oʊ h **aɪ oʊ**/ | [6 errors] |

- *Evaluation of spontaneous pronunciations is a difficult task!*

# Conclusion

- Pronunciation adaptation:
  - Probabilistic approach
  - Speaker independent
  - Linguistic features
- Considerable improvement:
  - When adding context information
- Extra improvement:
  - Adding linguistic features
  - Using Utterances
- Feature selection process is necessary

# Future work

- Articulatory and signal features
- N-best hypotheses
- Perceptual tests

# Q & A