Discriminative Keyword Spotting with Limited Data

Joseph Keshet

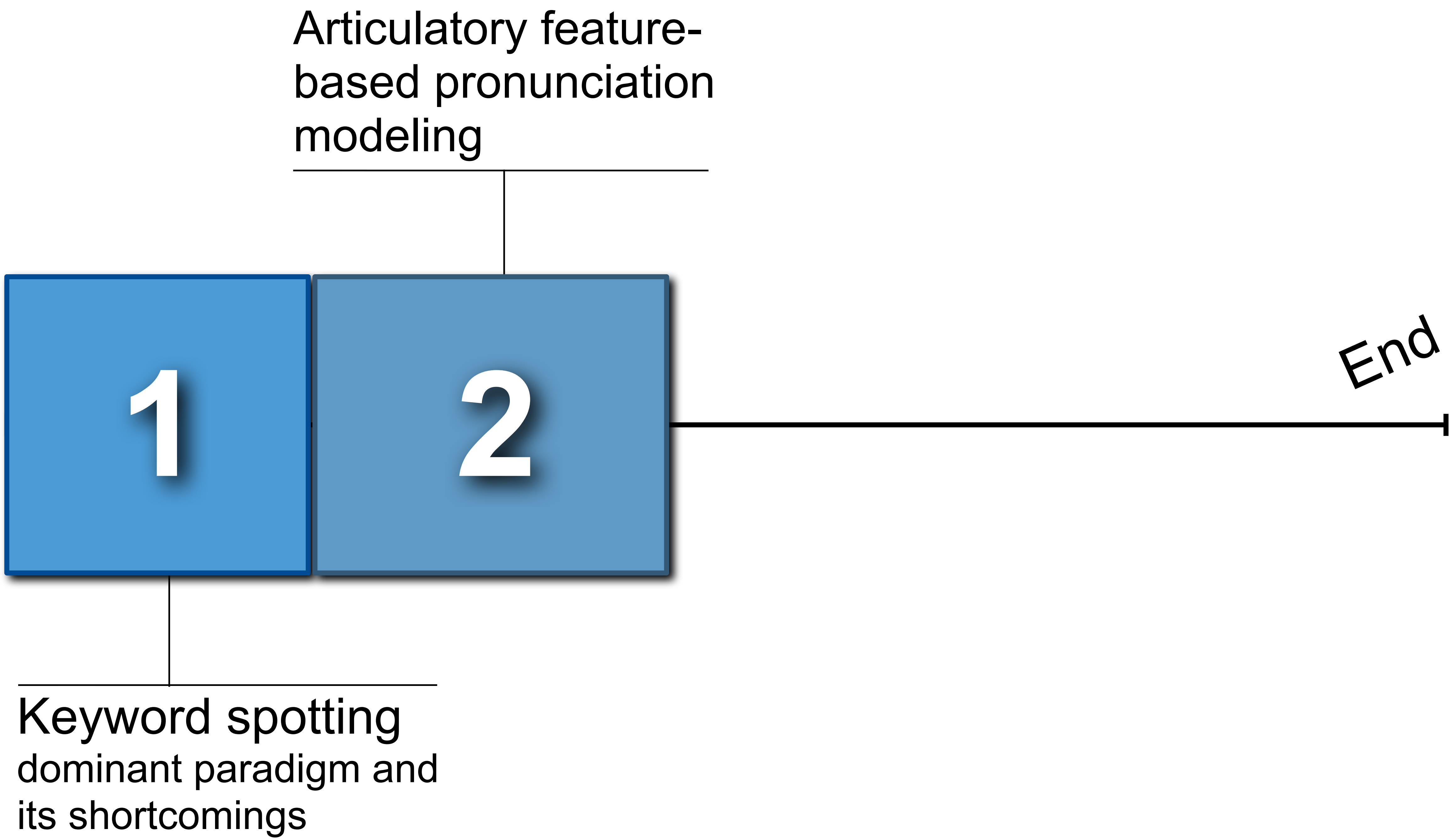
Department of Computer Science
Bar-Ilan University

Outline

Outline

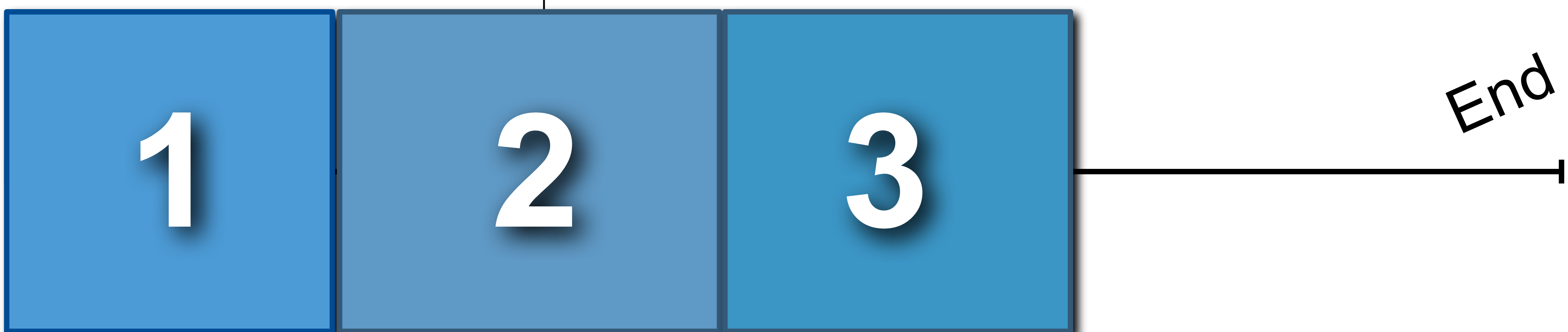
Keyword spotting
dominant paradigm and
its shortcomings

Outline


Articulatory feature-based pronunciation modeling
1
2
End
Keyword spotting
dominant paradigm and its shortcomings

Outline

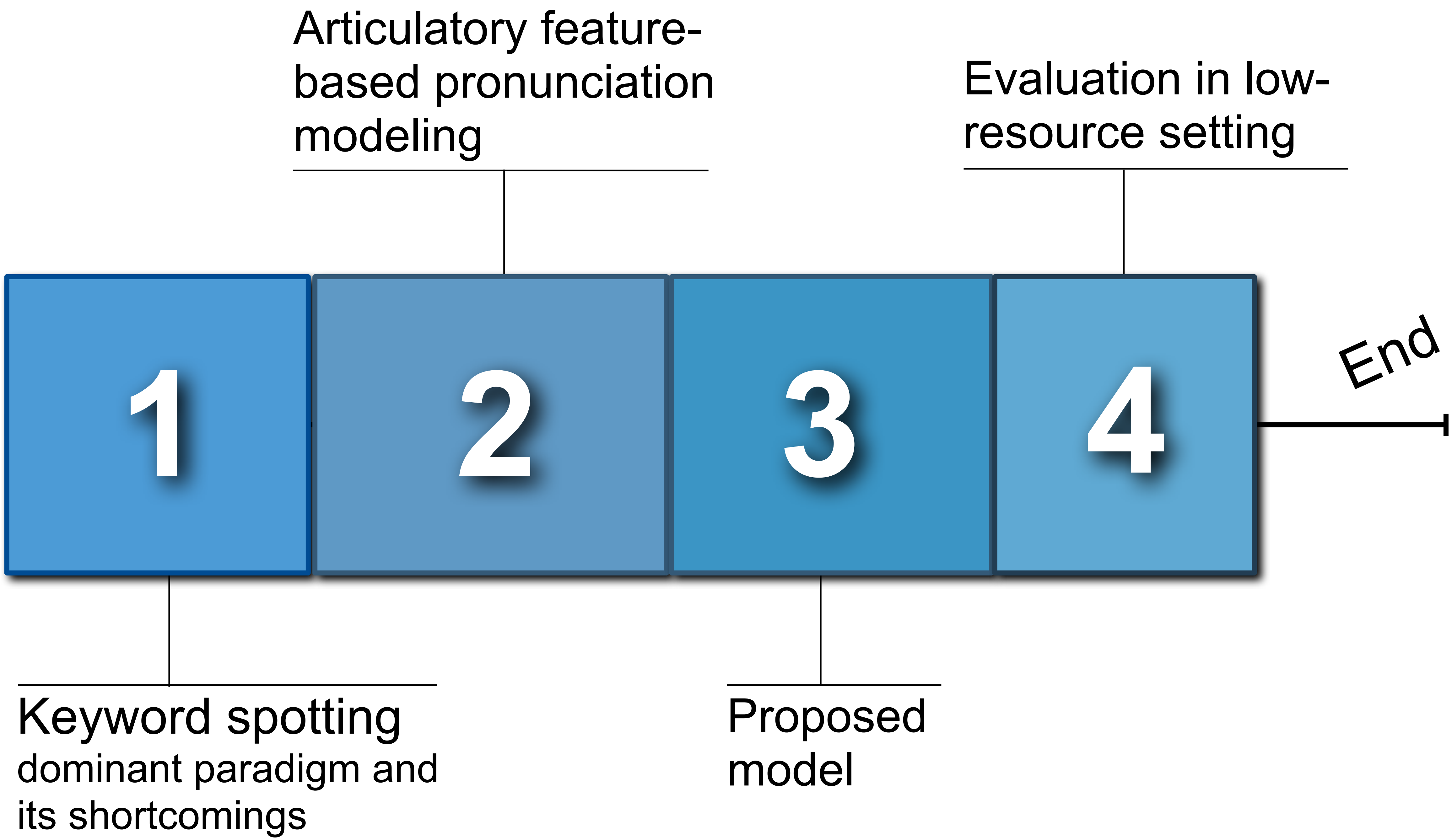
Articulatory feature-based pronunciation modeling



Keyword spotting
dominant paradigm and its shortcomings

Proposed model

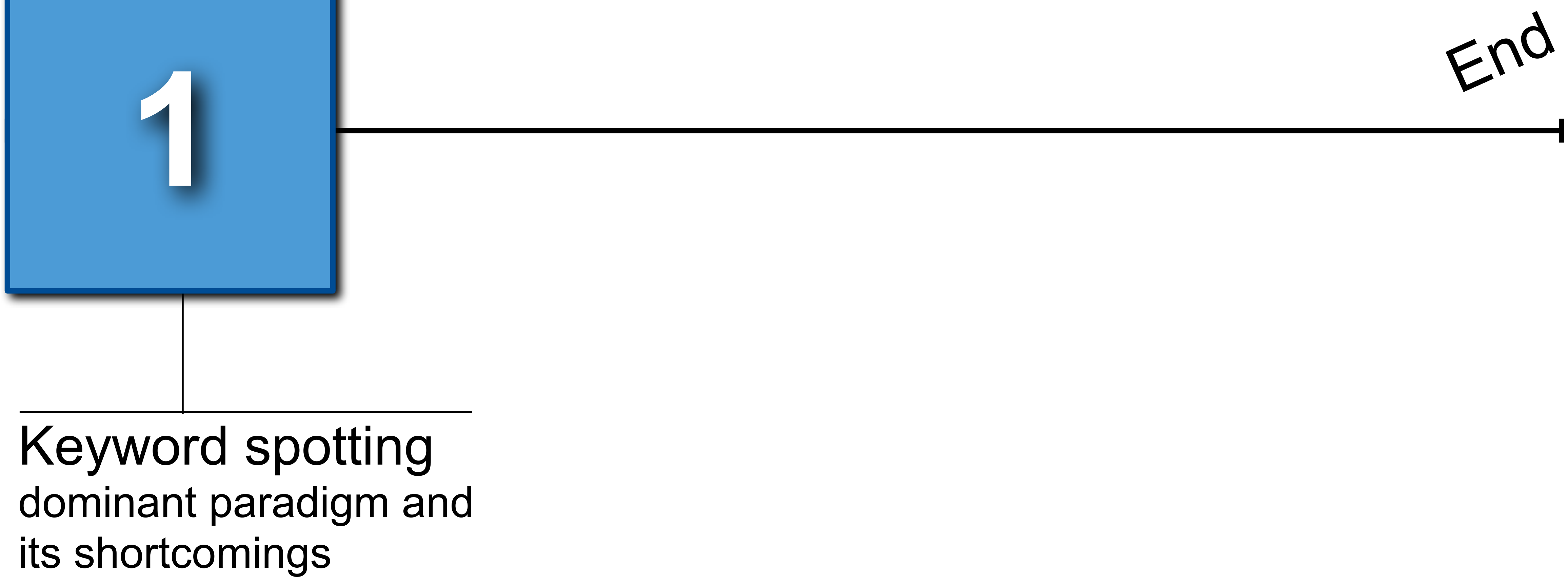
Outline


Articulatory feature-based pronunciation modeling
Evaluation in low-resource setting
1
2
3
4
End
Keyword spotting
dominant paradigm and its shortcomings
Proposed model

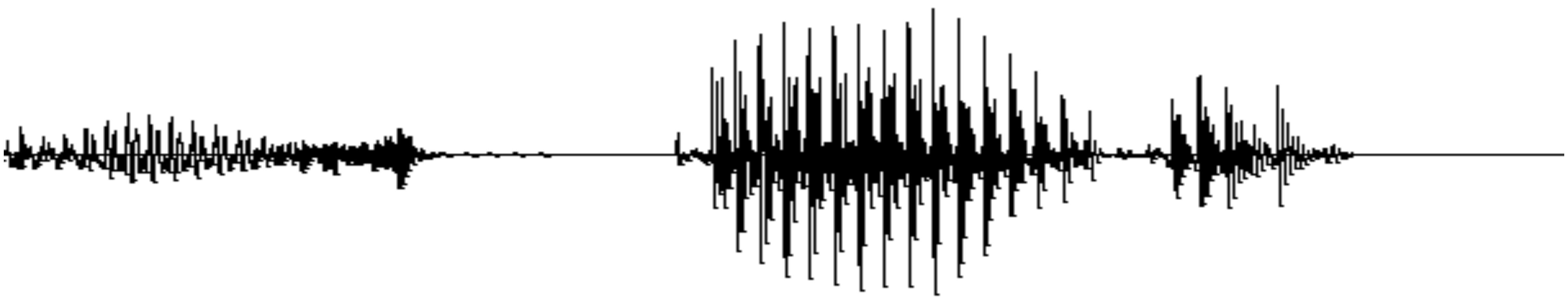
Outline

1. Keyword spotting
   dominant paradigm and its shortcomings
2. Articulatory feature-based pronunciation modeling
3. Proposed model
4. Evaluation in low-resource setting
5. Conclusions

Outline


1
End
Keyword spotting
dominant paradigm and
its shortcomings

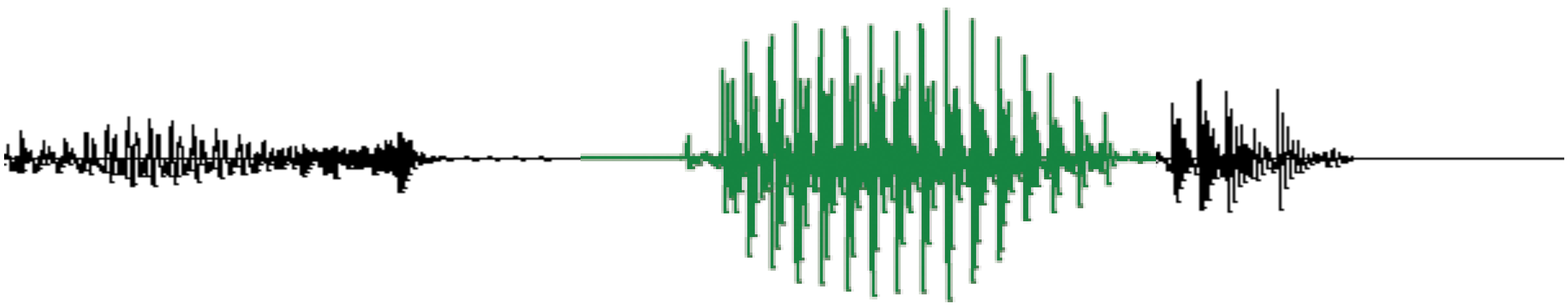
Given a speech signal



and a word **bought**

<u>Goal</u>: find if the word is uttered in the speech signal and where

Given a speech signal



and a word **bought**

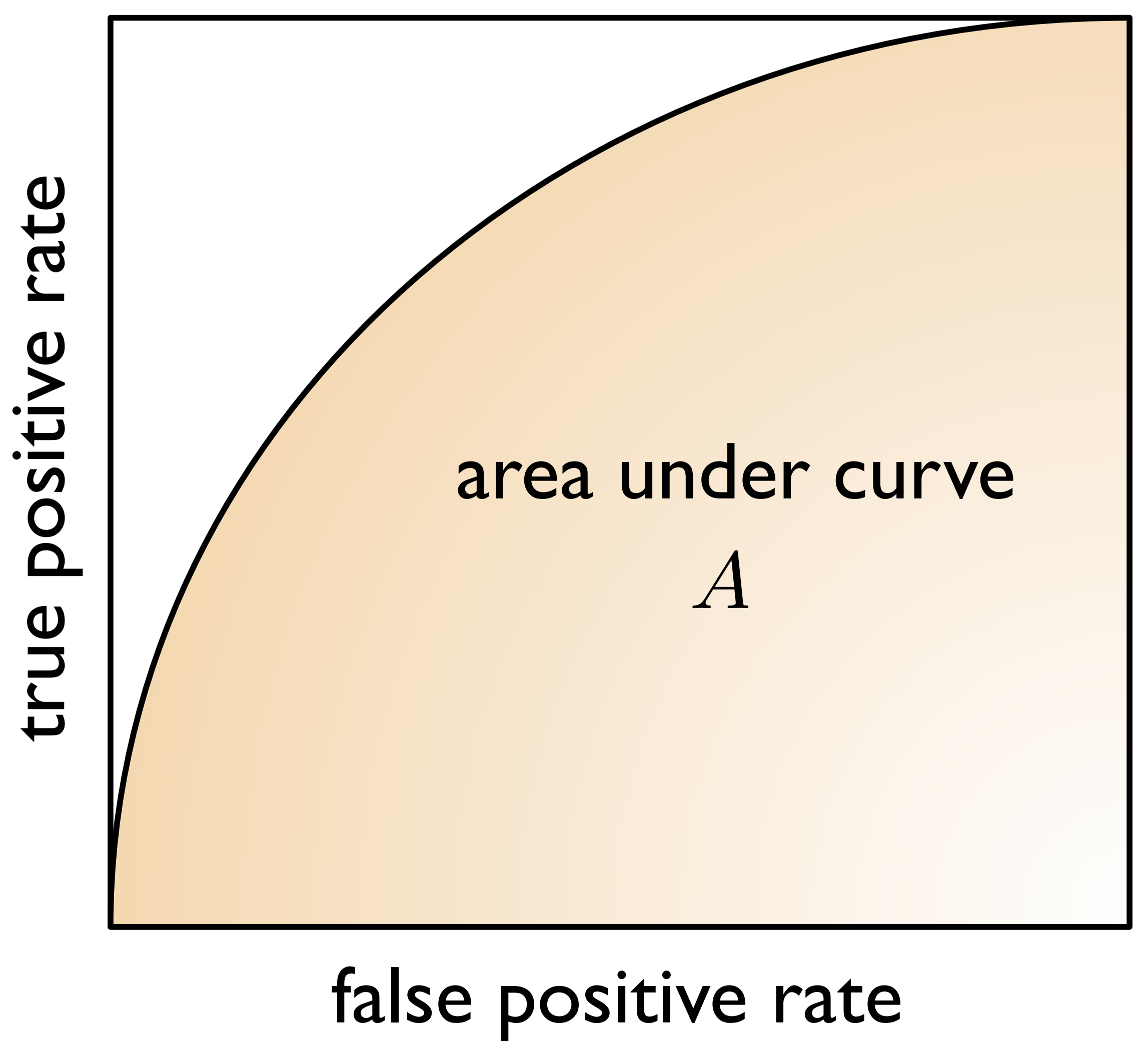
<u>Goal</u>: find if the word is uttered in the speech signal and where

The task loss

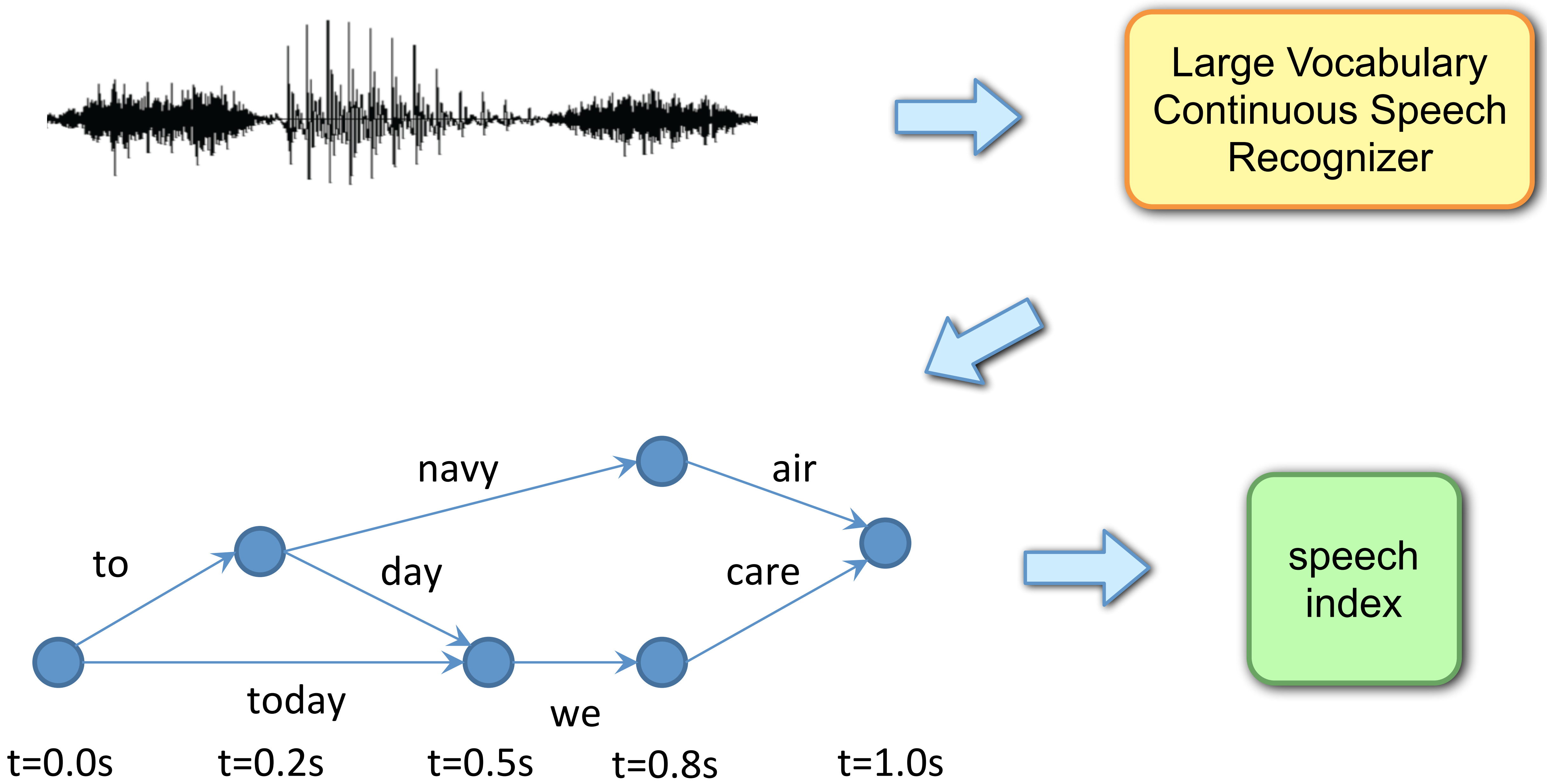
The performance of keyword spotting system is measured by <u>Receiver Operating Characteristics</u> (ROC) curve.

true positive = $\dfrac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}$

false positive = $\dfrac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}$



Dominant Paradigm



Dominant Paradigm

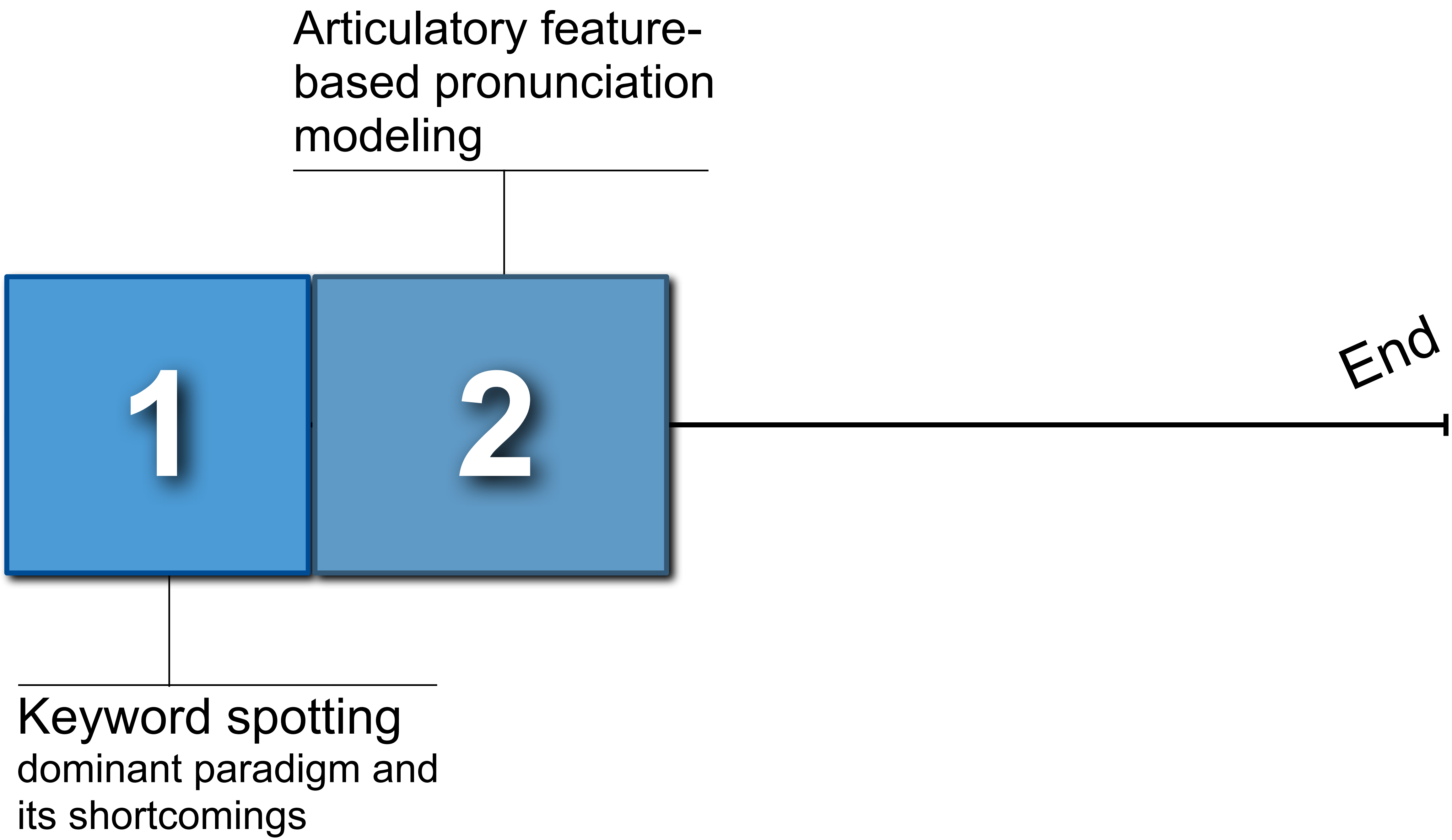
- Common for LVCSR systems to have **millions** of free parameters
  - RWTH Gale Mandarin System ≈640M (Plahl et al. 09)
- Not always appropriate to **assume availability of large amounts of training data**
  - Rapid development of systems for low-resource languages
  - Porting keyword spotting systems to new acoustic conditions or speech styles

Contributions

Articulatory feature-based pronunciation modeling

Discriminative learning by maximizing the AUC with large margin

Outline


Articulatory feature-based pronunciation modeling
1
2
End
Keyword spotting
dominant paradigm and its shortcomings

What are articulatory features?

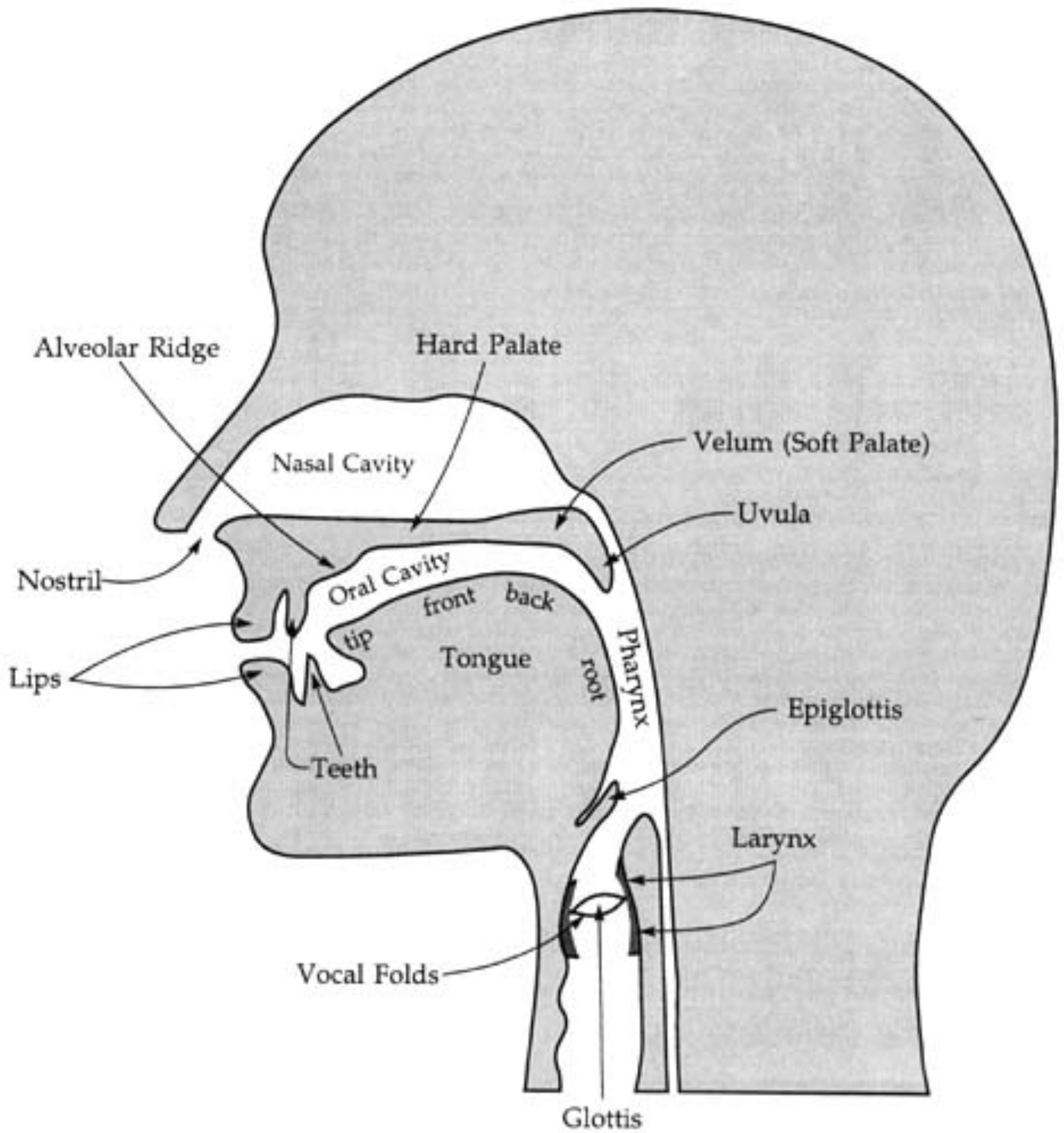
(video source: Ken Stevens, MIT)

What are articulatory features?

(video source: Ken Stevens, MIT)

Articulatory phonology

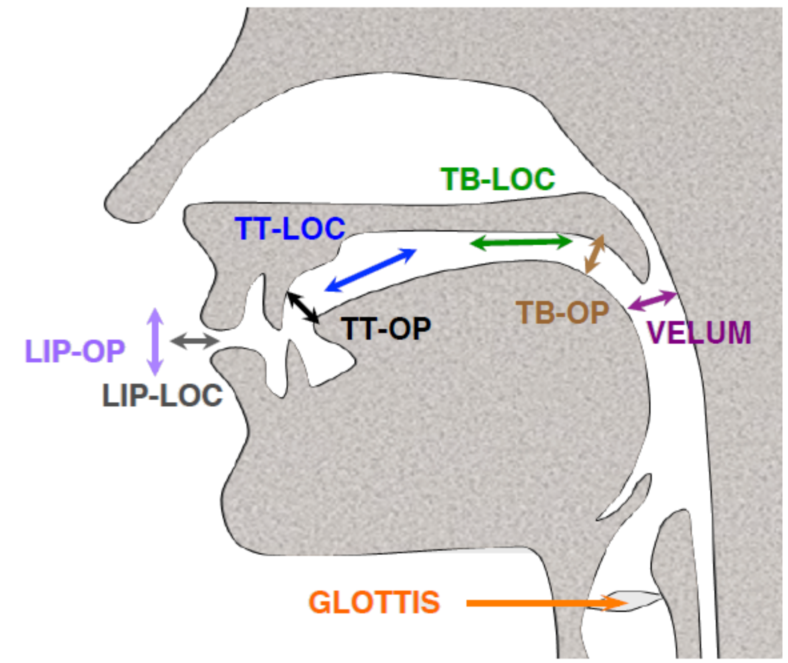
“pronunciation variations can be explained by asynchronization of the articulation”
(Browman and Goldstein, 1992)



Articulatory phonology

articulatory features (AF)

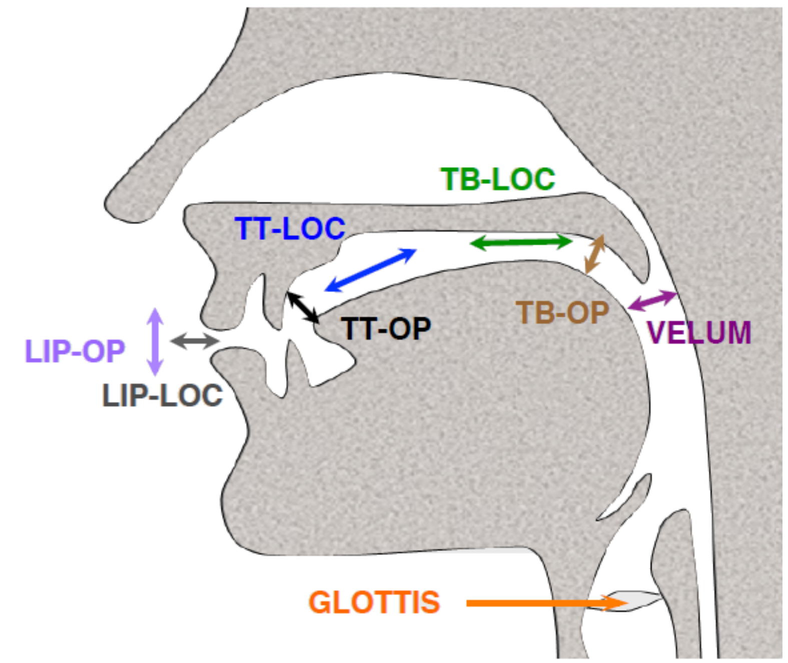
VEL	non-nasal	non-nasal	nasal	non-nasal
GLO	wide	critical	critical	wide
TB	uvular/medium	palatal/medium	uvular/medium	uvular/medium
TT	alveolar/ critical	alveolar/ medium	alveolar/closed	alveolar/critical
LIPS	wide/labial	wide/ labial	wide/labial	wide/labial
Phone	s	eh	n	s



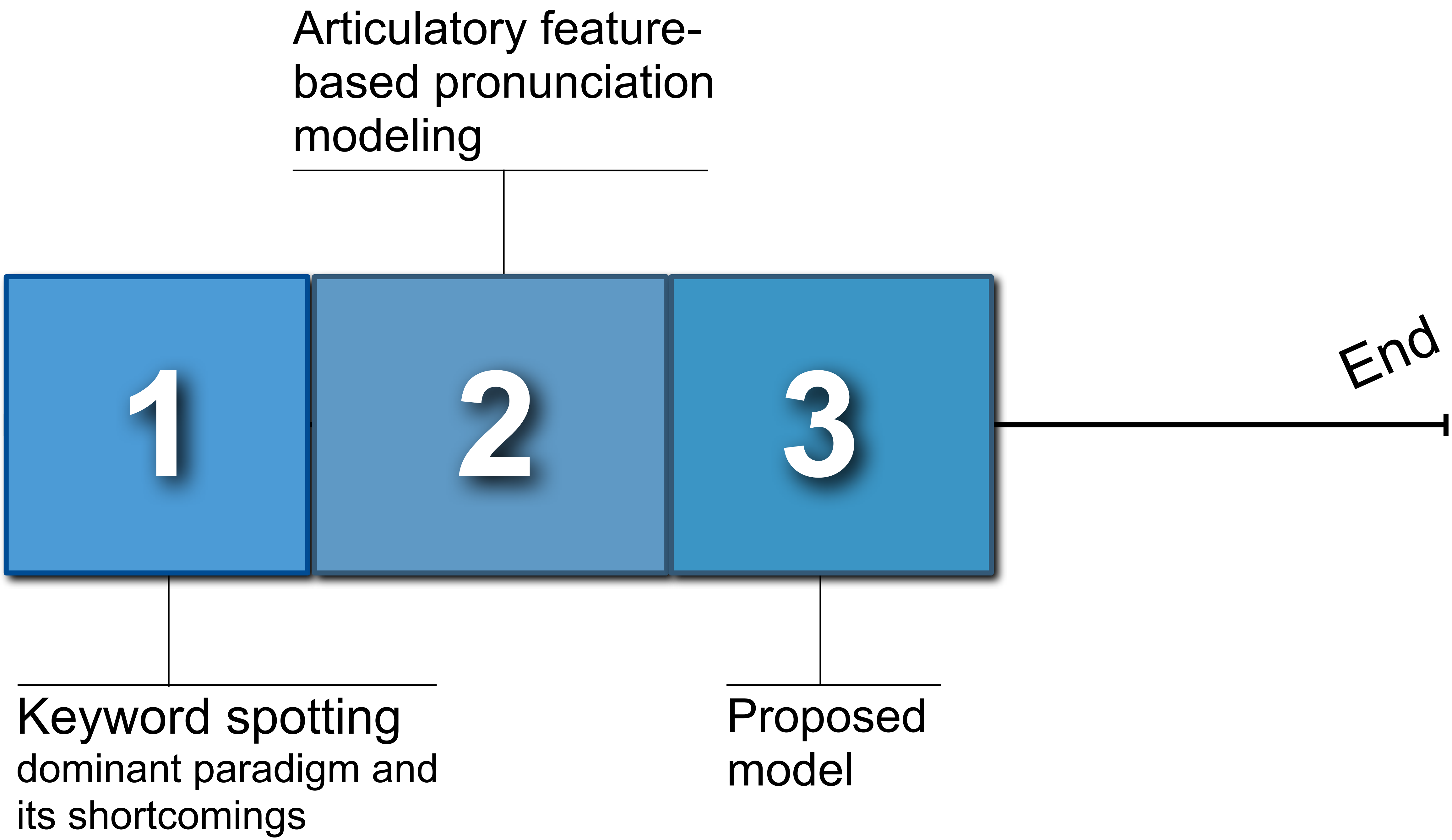
Articulatory phonology

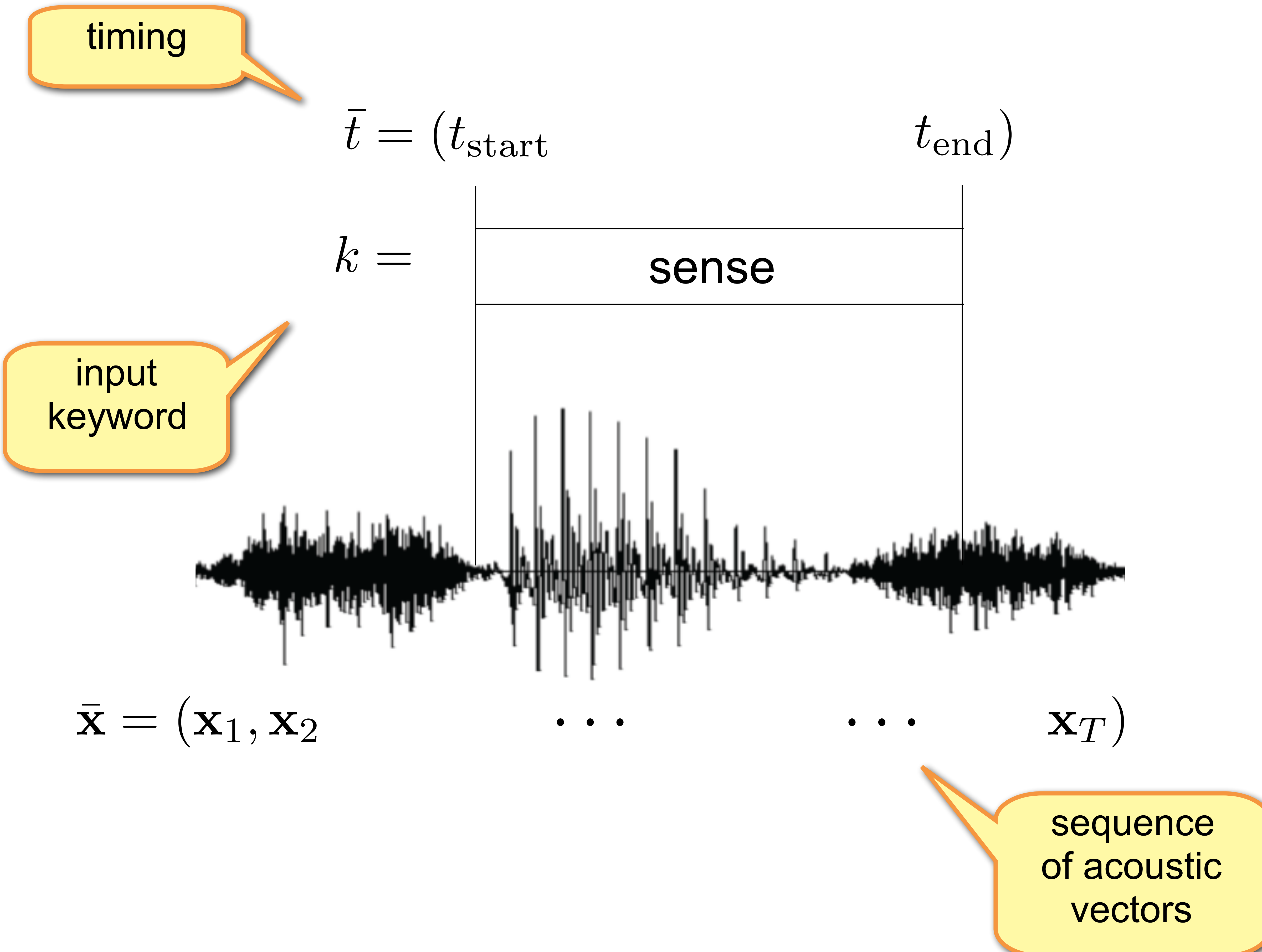
articulatory features (AF)

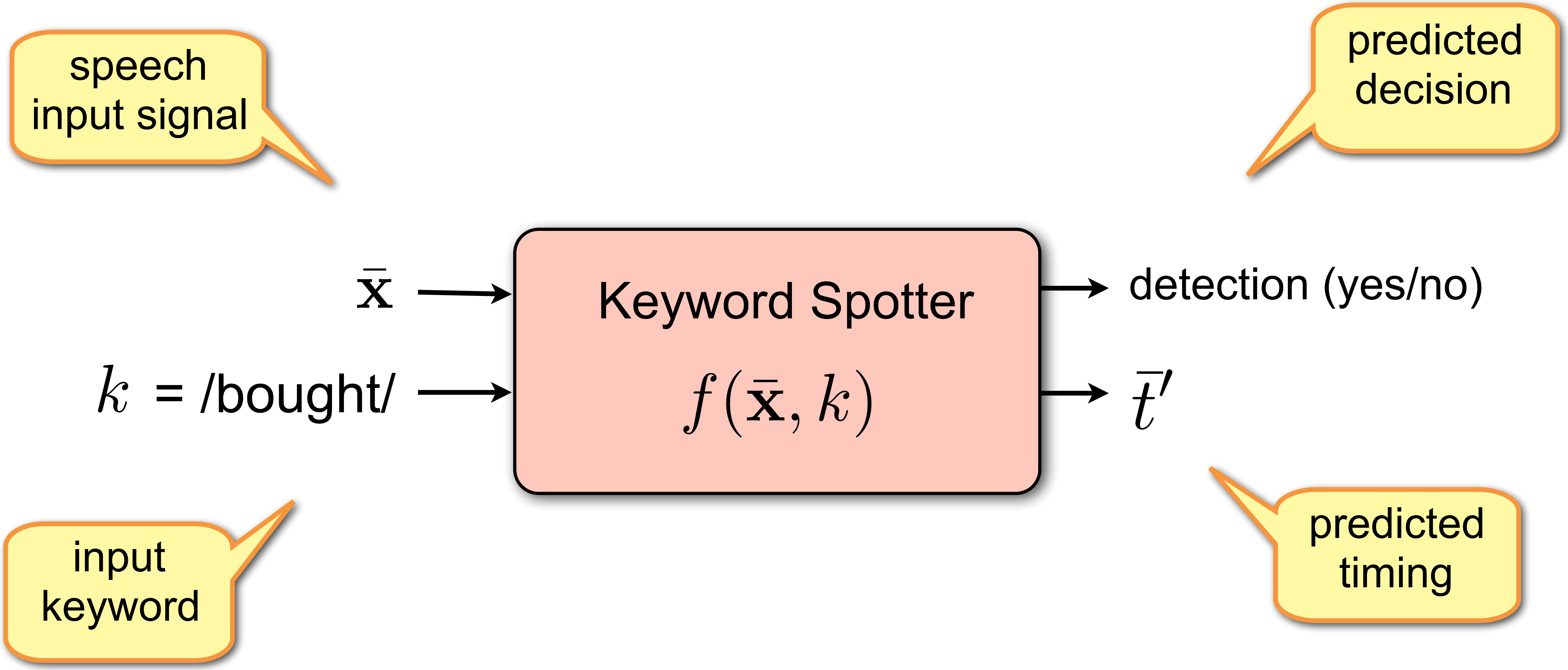
<table>
<tr><td>VEL</td><td>non-nasal</td><td>non-nasal</td><td colspan="2">nasal</td><td colspan="2">non-nasal</td></tr>
<tr><td>GLO</td><td>wide</td><td>critical</td><td colspan="2">critical</td><td colspan="2">wide</td></tr>
<tr><td>TB</td><td>uvular/medium</td><td colspan="2">palatal/medium</td><td colspan="2">uvular/medium</td><td>uvular/medium</td></tr>
<tr><td>TT</td><td>alveolar/ critical</td><td colspan="2">alveolar/ medium</td><td colspan="2">alveolar/closed</td><td>alveolar/critical</td></tr>
<tr><td>LIPS</td><td>wide/labial</td><td colspan="2">wide/ labial</td><td colspan="2">wide/labial</td><td>wide/labial</td></tr>
<tr><td>Phone</td><td>s</td><td colspan="2">eh^n</td><td>n</td><td>t</td><td>s</td></tr>
</table>



Outline


Articulatory feature-
based pronunciation
modeling
1
2
3
End
Keyword spotting
dominant paradigm and
its shortcomings
Proposed
model


timing
$\bar{t} = (t_{\text{start}} \quad t_{\text{end}})$
$k =$
sense
input keyword
$\bar{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2 \quad \cdots \quad \cdots \quad \mathbf{x}_T)$
sequence of acoustic vectors


speech input signal
predicted decision
$\bar{\mathbf{x}}$
Keyword Spotter
detection (yes/no)
k = /bought/
$f(\bar{\mathbf{x}}, k)$
$\bar{t}'$
input keyword
predicted timing

Model and inference

$$\bar{t}^* = f(\bar{\mathbf{x}}, k)$$

Model and inference

$$\bar{t}^* = f(\bar{\mathbf{x}}, k)$$

$$= \arg\max_{\bar{t}} f(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference

$$
\begin{aligned}
\bar{t}^* &= f(\bar{\mathbf{x}}, k) \\
&= \arg\max_{\bar{t}} \; f(\bar{\mathbf{x}}, k, \bar{t}) \\
&= \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})
\end{aligned}
$$

Model and inference

$$\bar{t}^* = f(\bar{\mathbf{x}}, k)$$

$$= \arg\max_{\bar{t}} \; f(\bar{\mathbf{x}}, k, \bar{t})$$

$$= \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

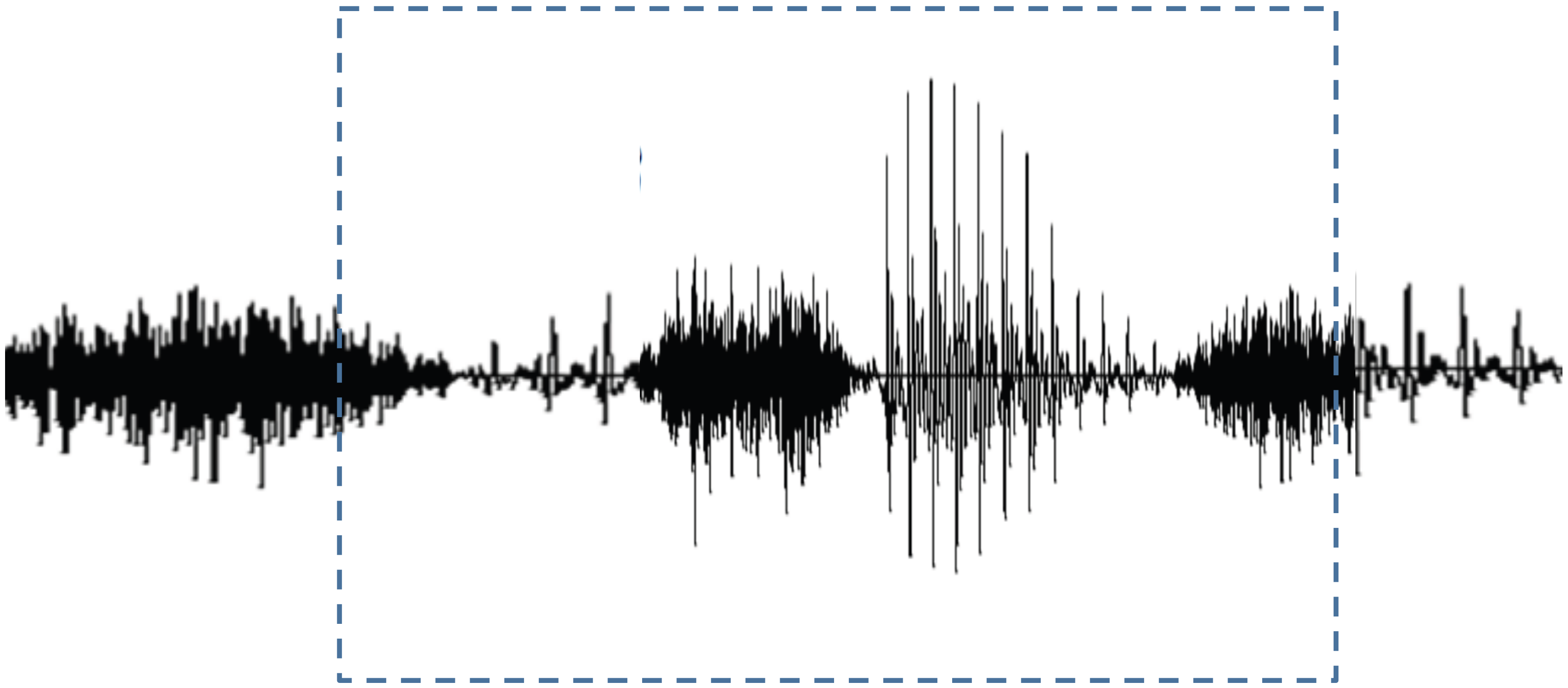
weight vector
$\mathbf{w} \in \mathbb{R}^n$

feature map

Model and inference

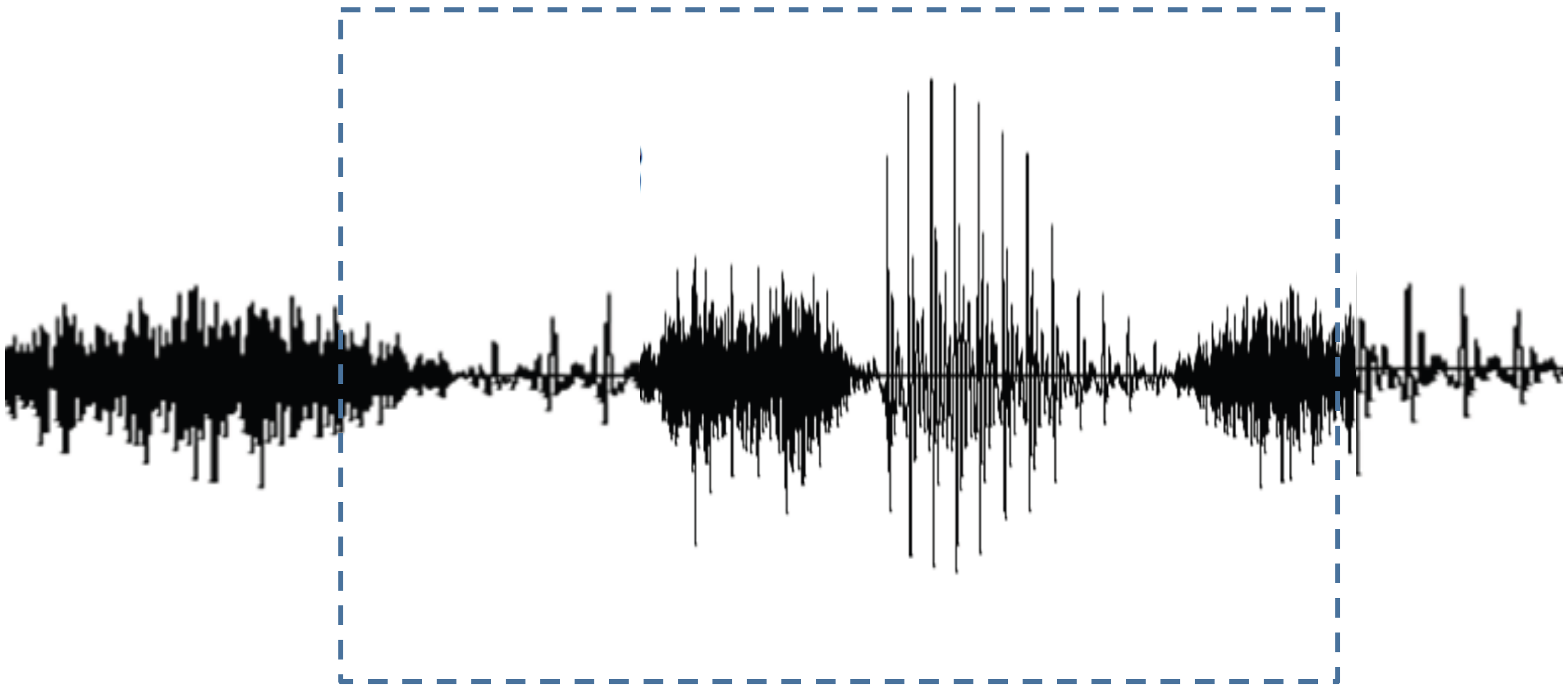
$$\begin{aligned}
\bar{t}^* &= f(\bar{\mathbf{x}}, k) \\
&= \arg\max_{\bar{t}} \; f(\bar{\mathbf{x}}, k, \bar{t}) \\
&= \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})
\end{aligned}$$

Model and inference



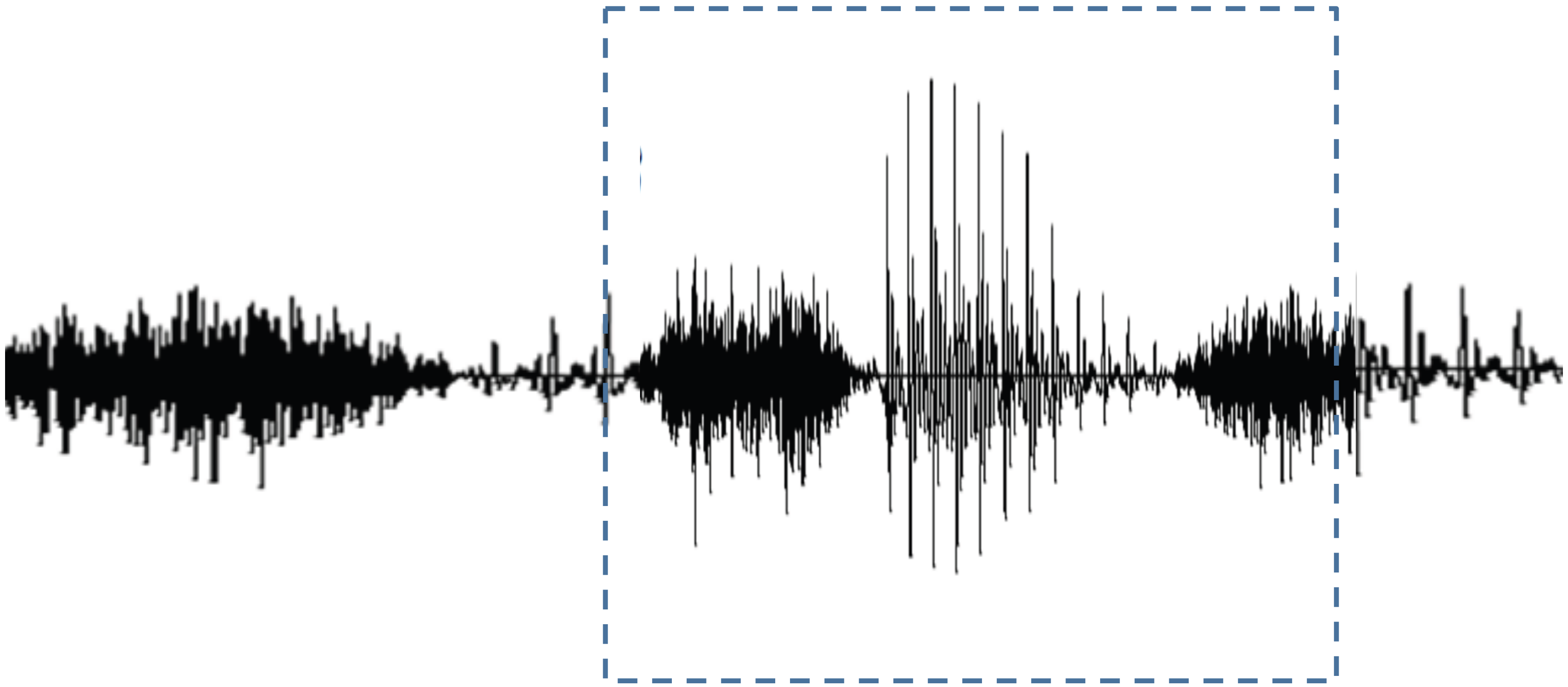
$$\bar{t}^{\,*} = \arg\max_{\bar{t}} \ \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference



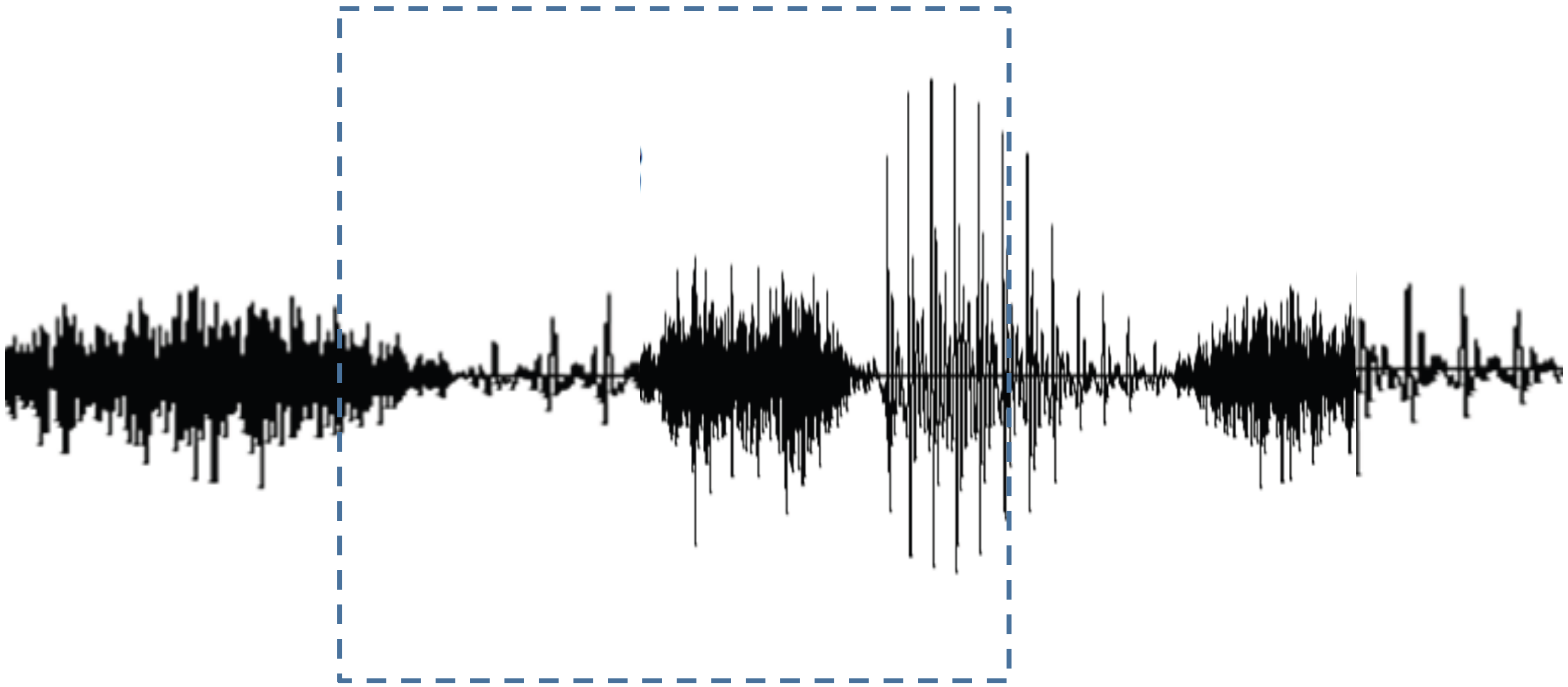
$$\bar{t}^{\,*} = \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference



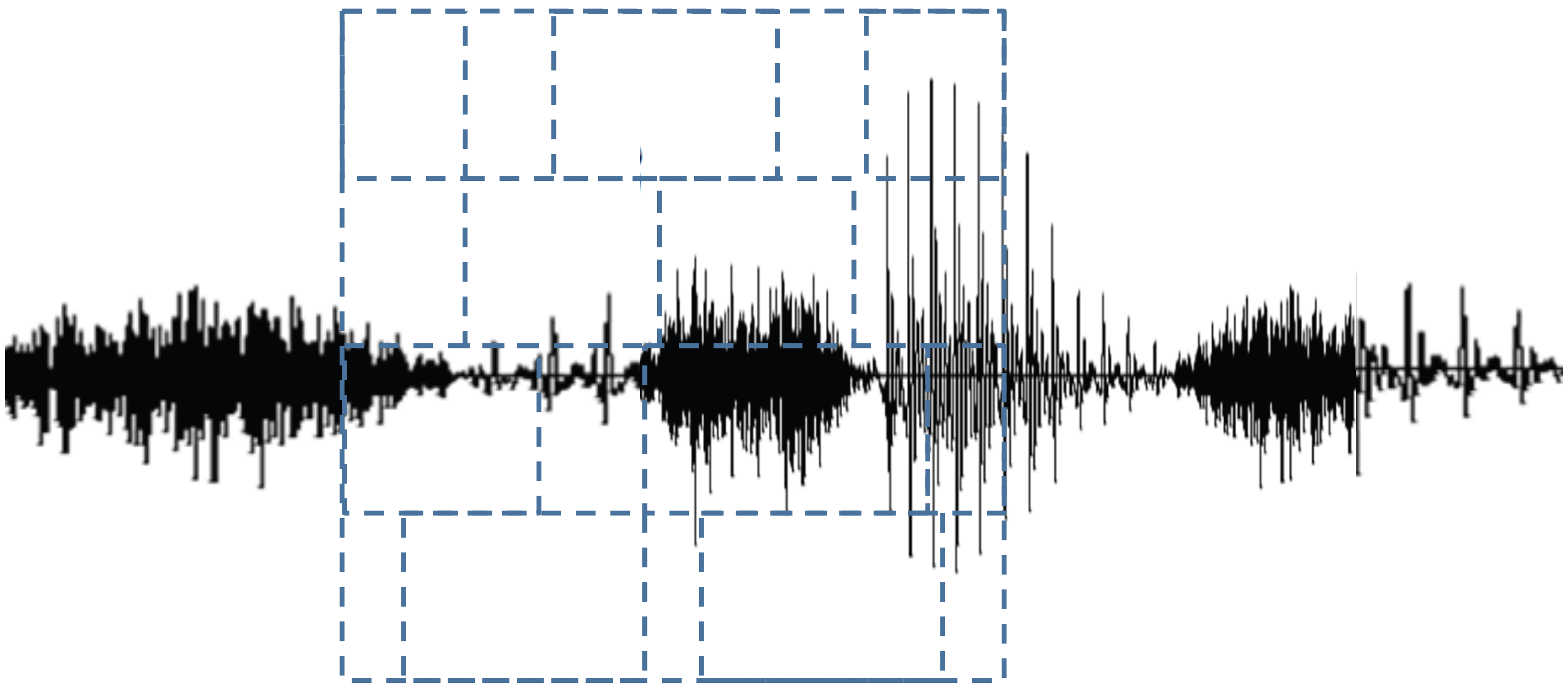
$$\bar{t}^* = \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference



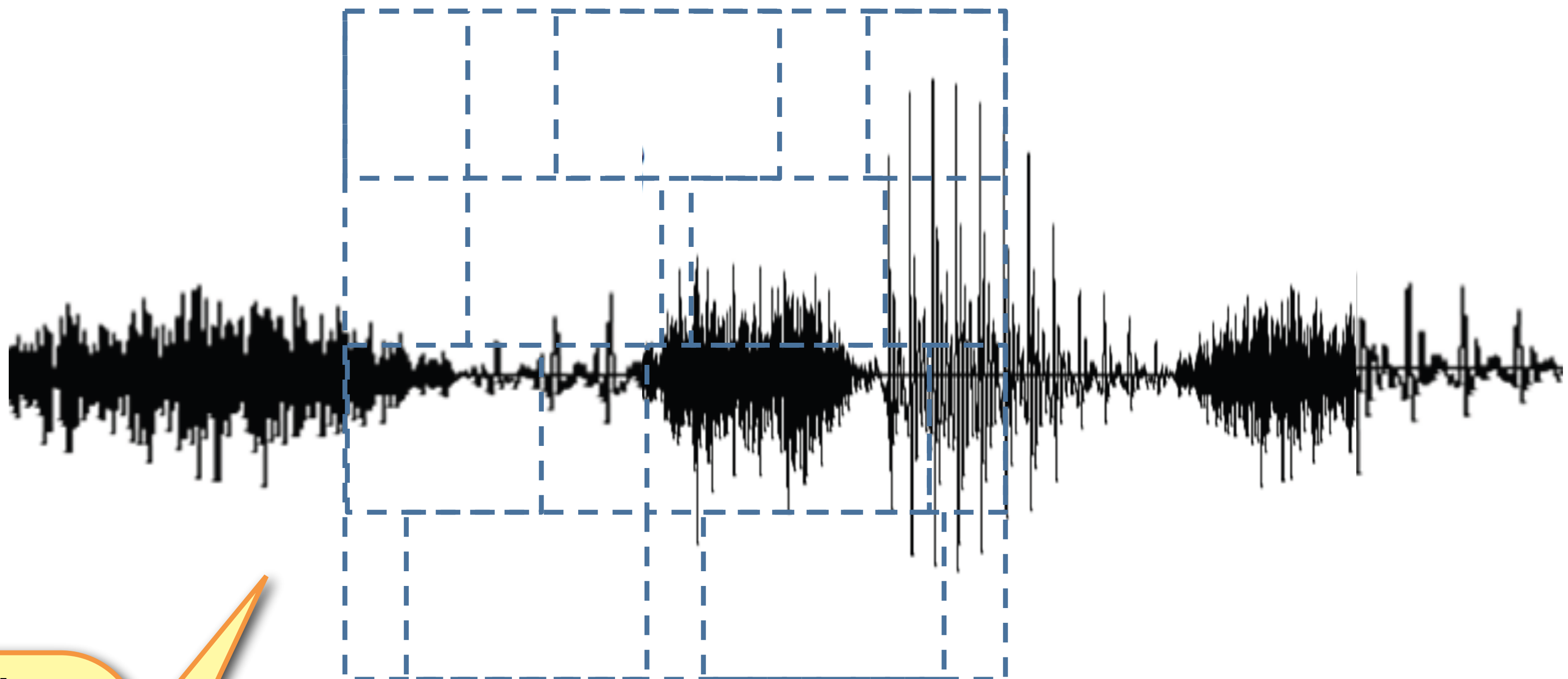
$$\bar{t}^* = \arg\max_{\bar{t}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference



$$\bar{t}^{\,*} = \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

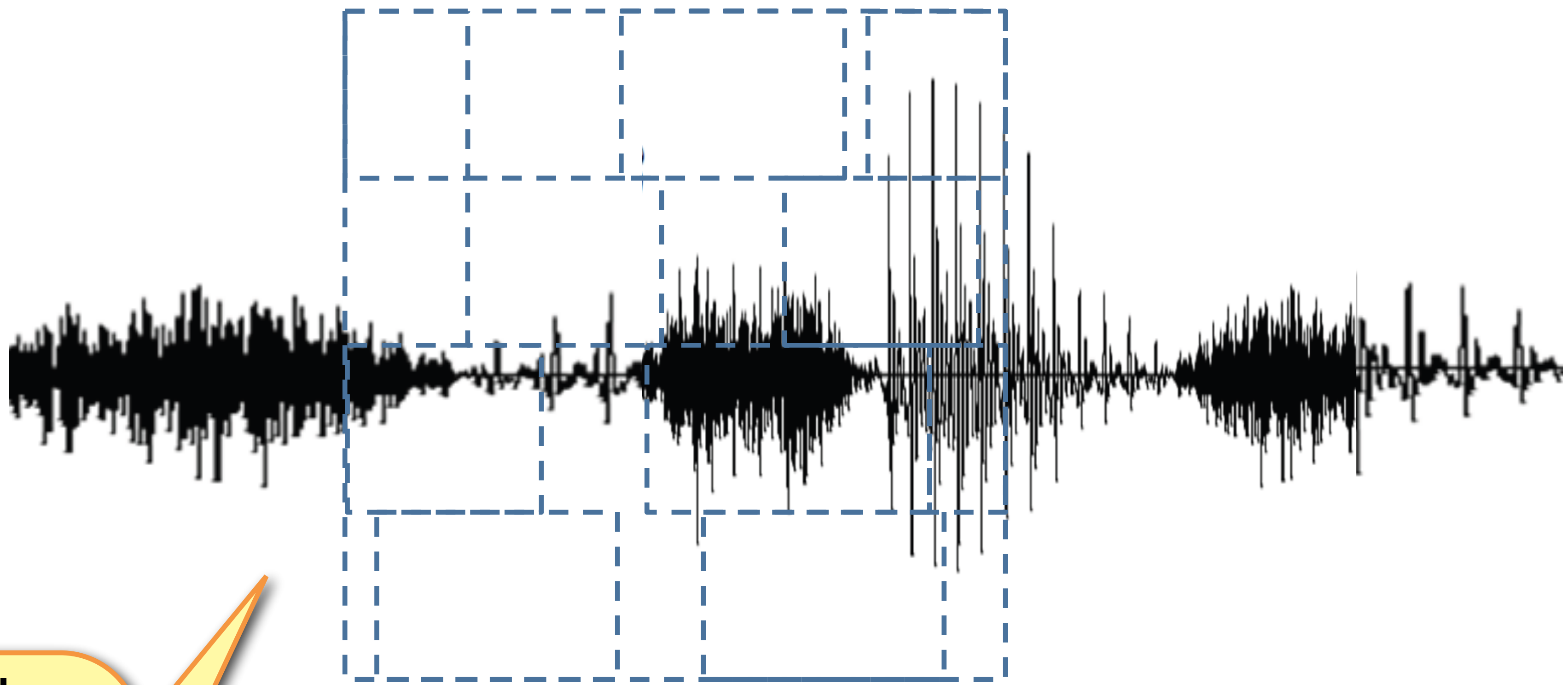
Model and inference



several streams in the case of articulatory features

$$\bar{t}^* = \arg\max_{\bar{t}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

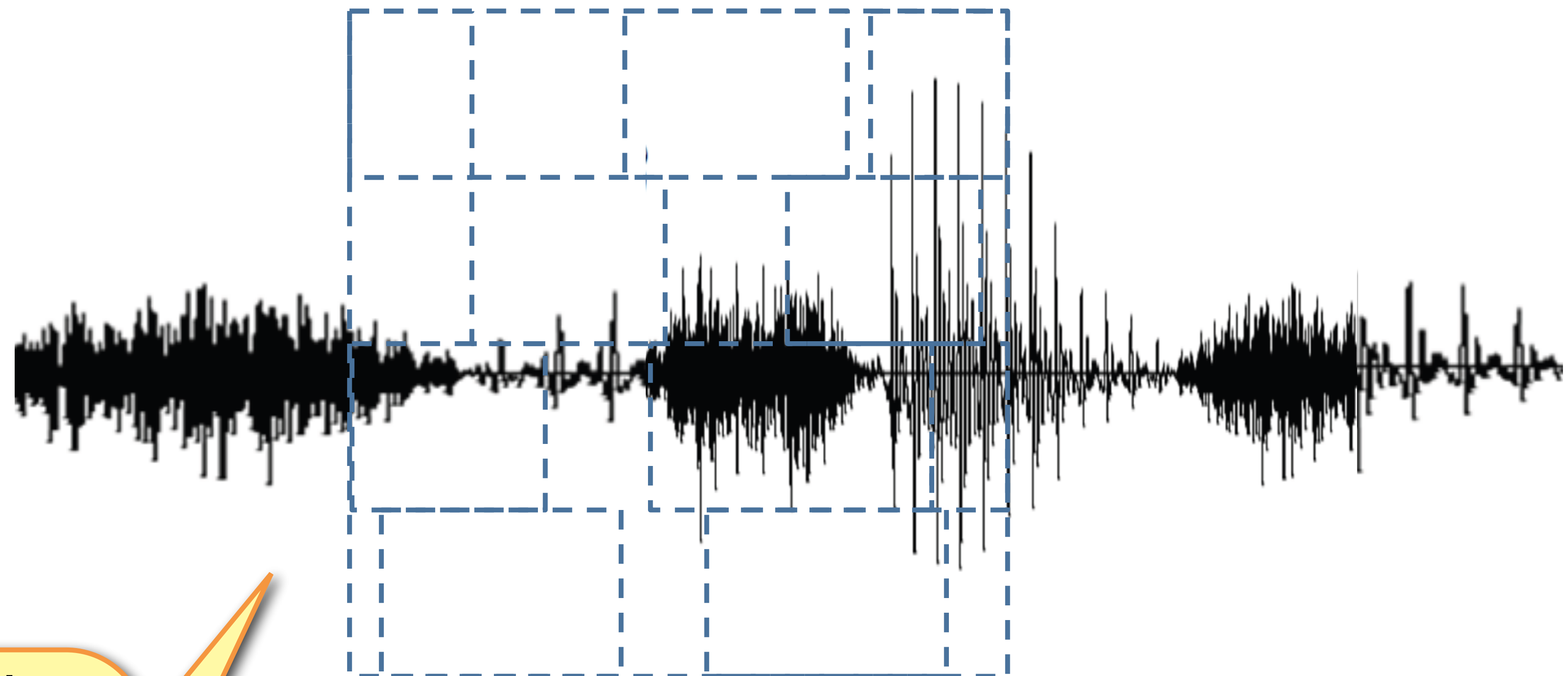
Model and inference



several streams in the case of articulatory features

$$\bar{t}^{\,*} = \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

Model and inference



several streams in the case of articulatory features

$$\bar{t}^* = \arg\max_{\bar{t}} \; \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, k, \bar{t})$$

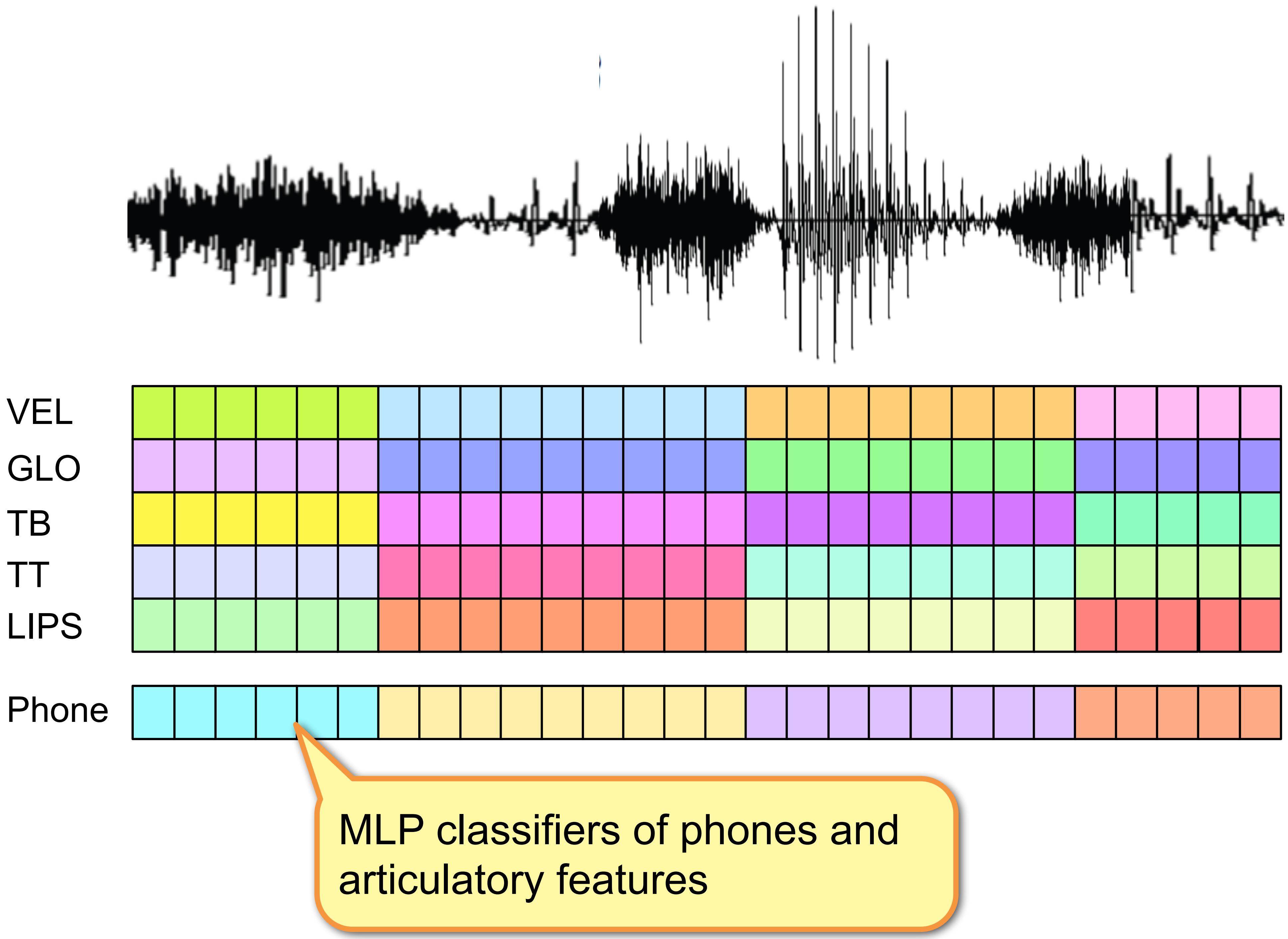
weight vector

$\mathbf{w} \in \mathbb{R}^n$

feature map

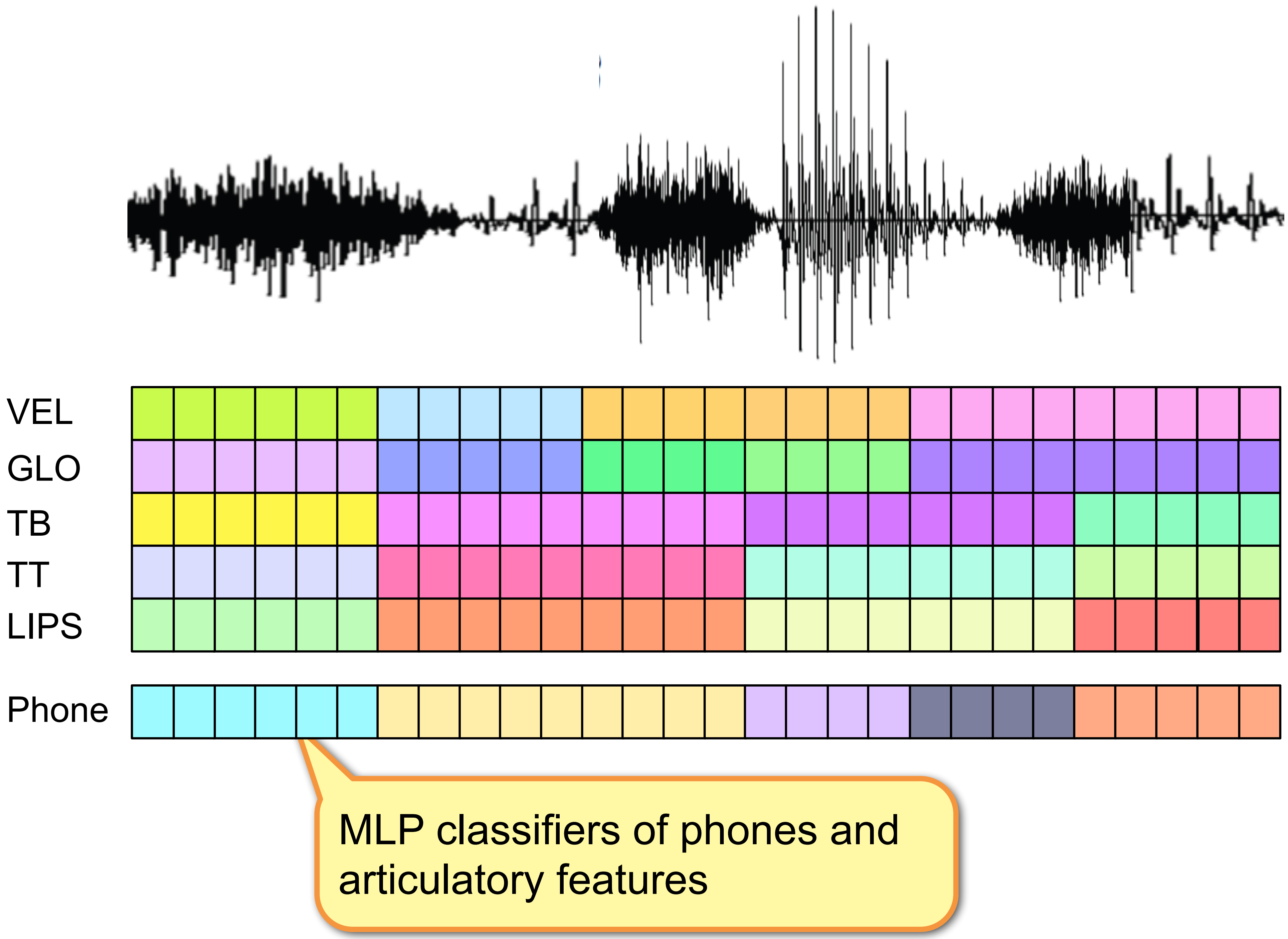
Feature map I

How likely is current frame to correspond to each of the AFs given segmentation?



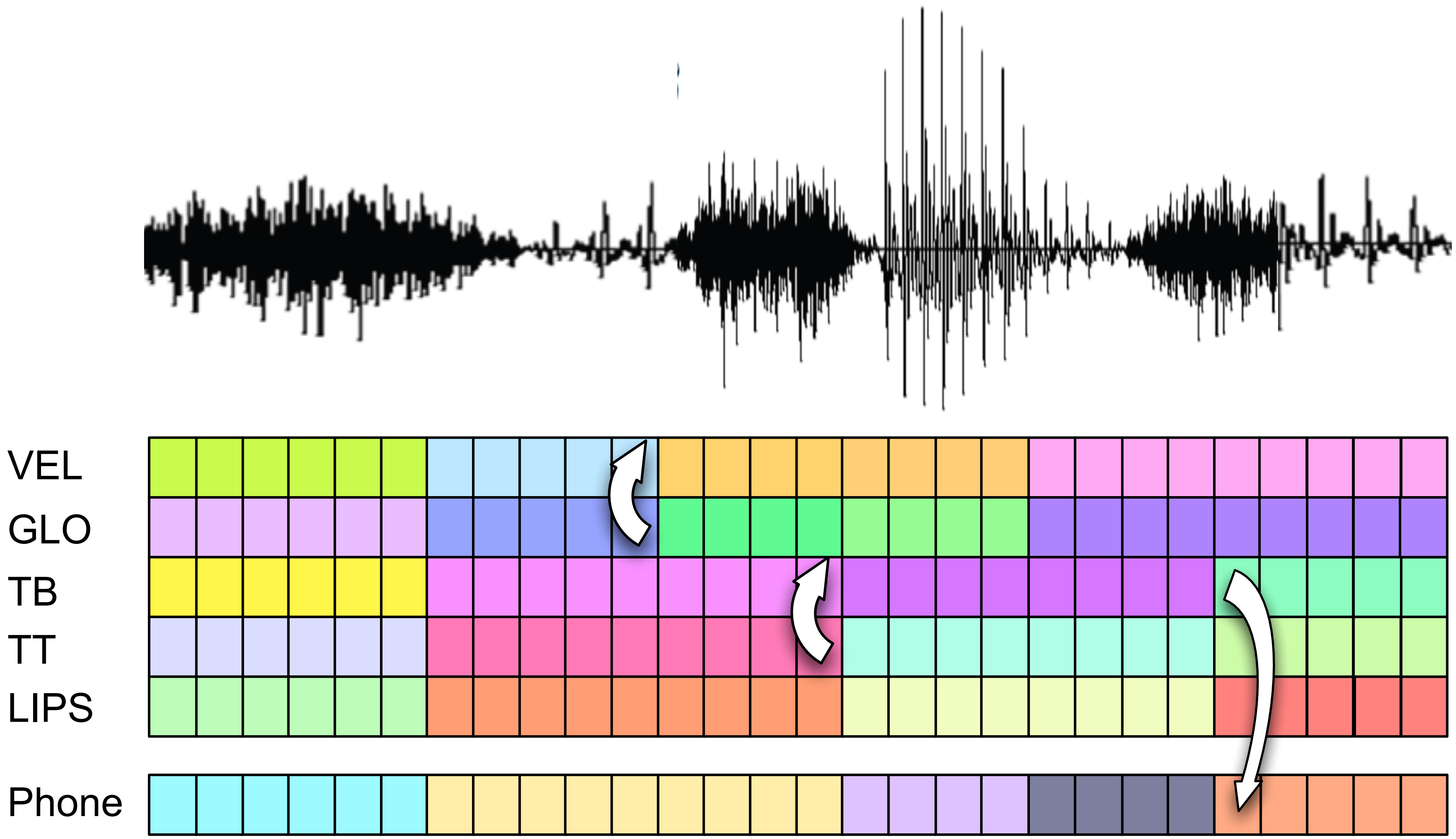
Feature map I

How likely is current frame to correspond to each of the AFs given segmentation?



Feature map II

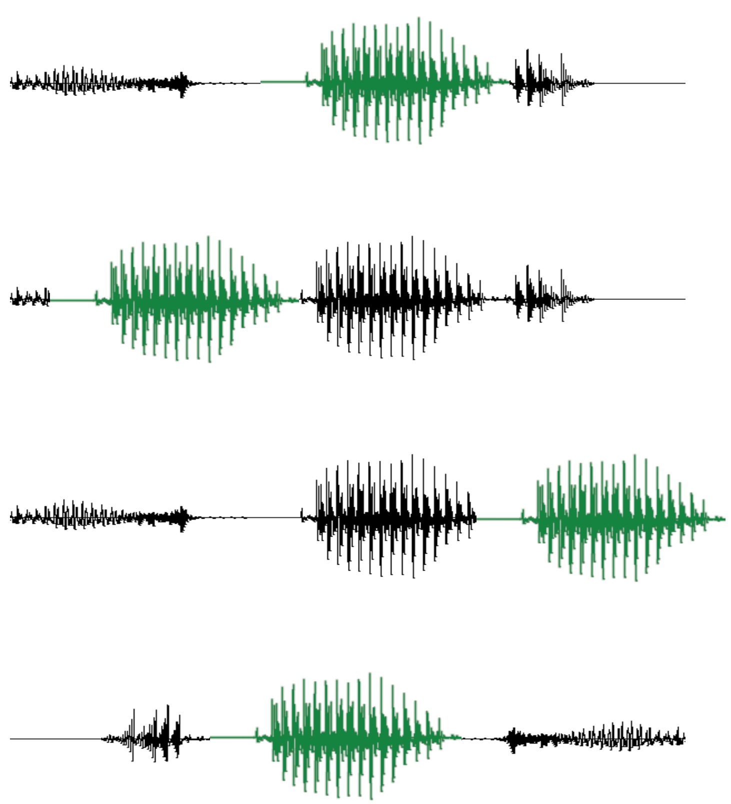
How likely is AF in stream i at previous frame corresponds to AF stream j at current frame

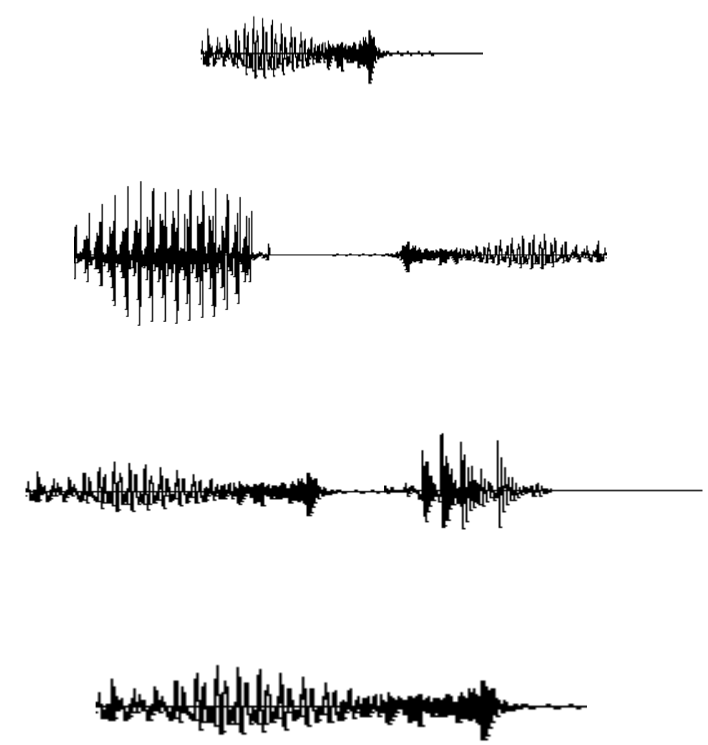


Maximizing area under ROC (AUC)

For every event (keyword) k define two sets of input signals (speech utterances):

$\mathcal{X}_k^+$ $\mathcal{X}_k^-$





Maximizing area under ROC (AUC)

By definition of the area under the ROC:

$$A = \mathbb{P}\left[\max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{+}, k, \bar{t}) > \max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{-}, k, \bar{t})\right]$$

(**Keshet**, Grangier and Bengio, *2009)*

Maximizing area under ROC (AUC)

By definition of the area under the ROC:

$$A = \mathbb{P}\left[\max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{+}, k, \bar{t}) > \max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{-}, k, \bar{t})\right]$$

(**Keshet**, Grangier and Bengio, *2009)*

Maximizing area under ROC (AUC)

By definition of the area under the ROC:

$$A = \mathbb{P}\left[\max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{+}, k, \bar{t}) > \max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}^{-}, k, \bar{t})\right]$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \left[1 - \max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}_i^{+}, k_i, \bar{t}) + \max_{\bar{t}} f_{\mathbf{w}}(\bar{\mathbf{x}}_i^{-}, k_i, \bar{t})\right]_+ + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

(**Keshet**, Grangier and Bengio, *2009)*

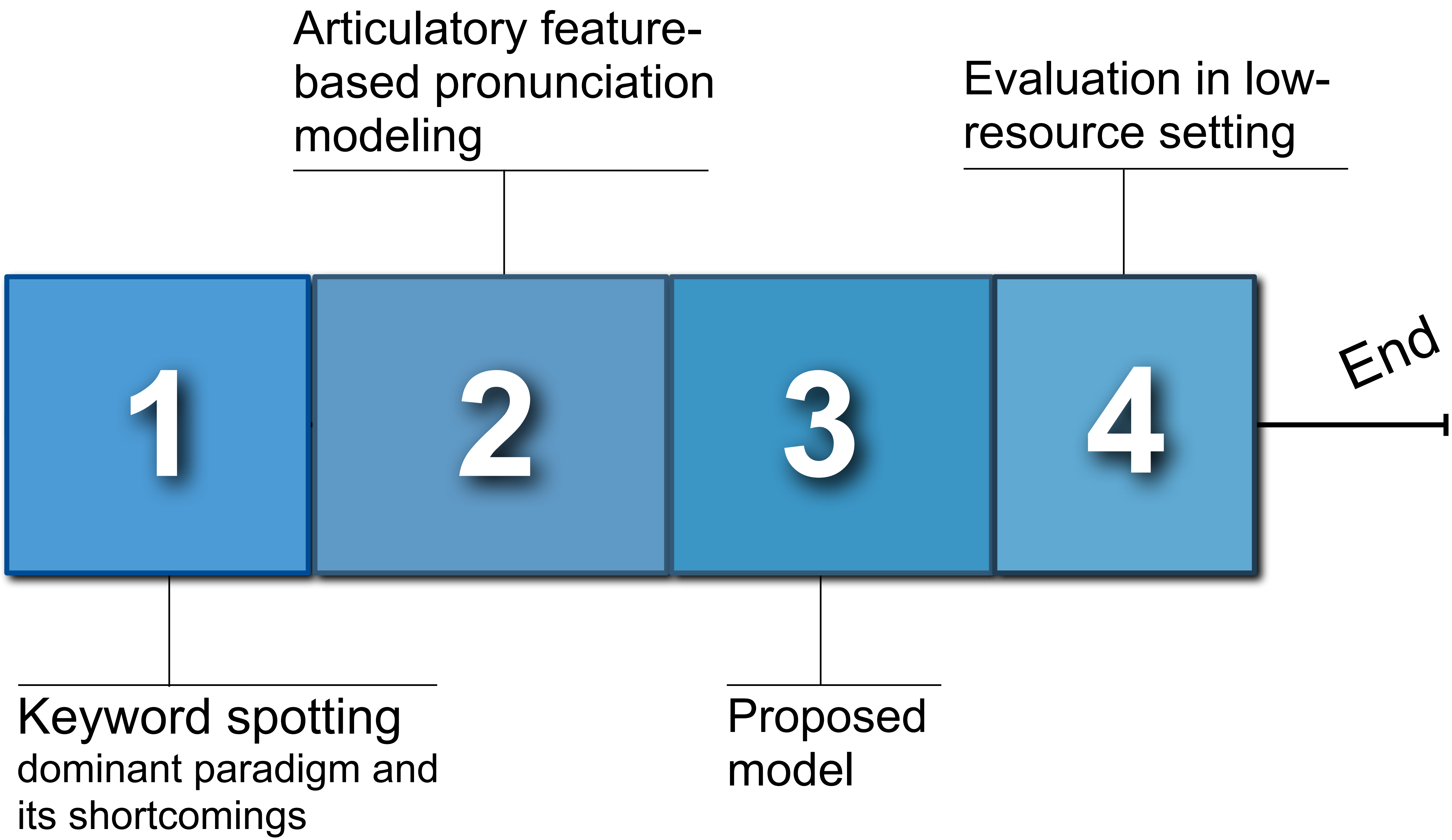
Implementation

- Iterative algorithm to solve the optimization problem efficiency on huge data (millions of examples)

- Theorems support the maximization of AUC

(**Keshet**, Grangier and Bengio, *2009;* Prabhavalkar, **Keshet**, Livescu and Fosler-Lussier*, 2012)*

Outline


Articulatory feature-based pronunciation modeling
Evaluation in low-resource setting
1
2
3
4
End
Keyword spotting
dominant paradigm and its shortcomings
Proposed model

Experiments

- Constructed four corpora containing 500-5000 utterances respectively by randomly selecting utterances from Switchboard
- Development set (40 terms) and Test set (60 terms)
  - 20 positive and negative sentences each

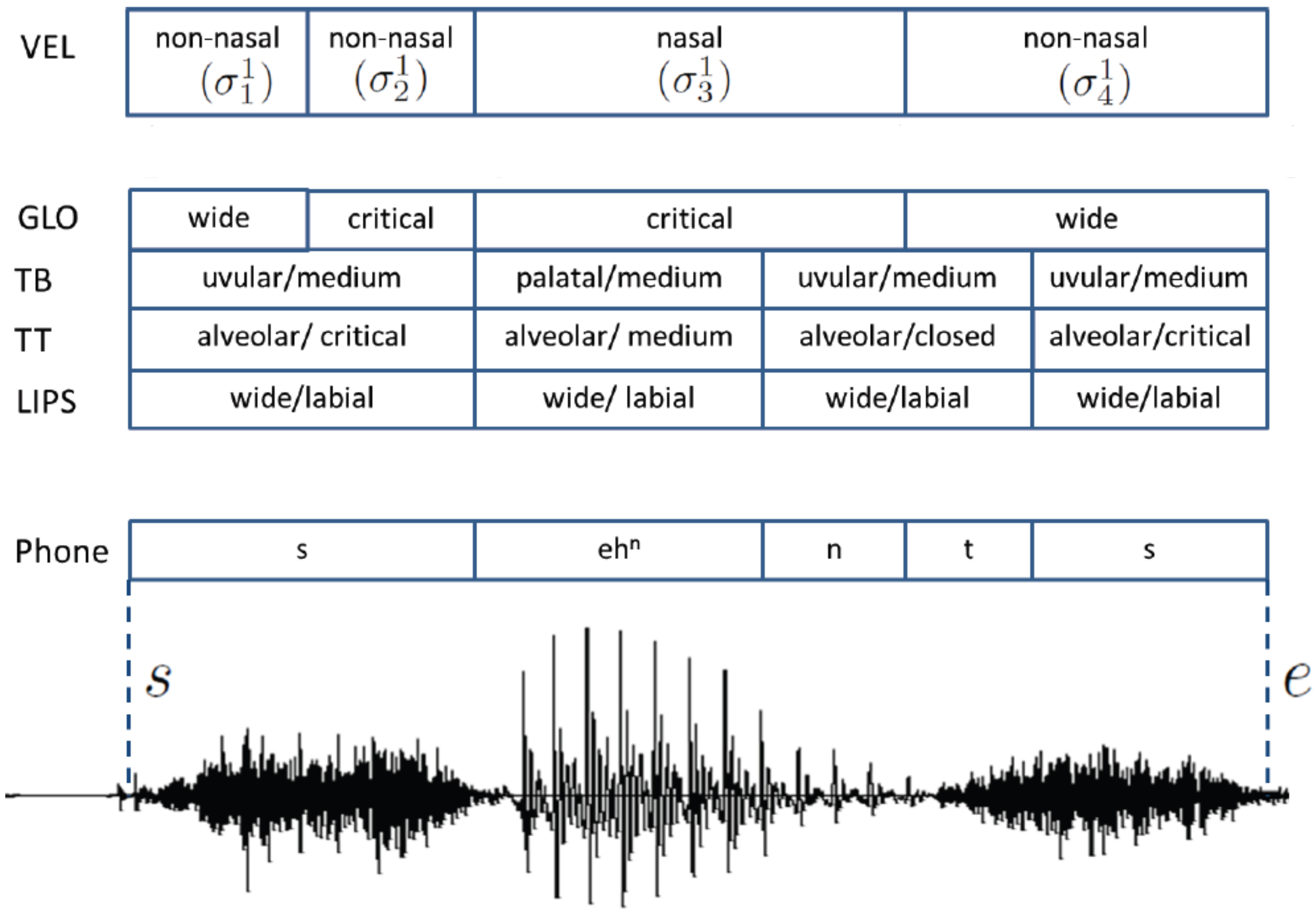
Utterances	500	1000	2500	5000
Training Data	0.8 hrs	1.5 hrs	3.7 hrs	7.4 hours

Experiments

- Creation of "positive" and "negative" examples from training data
  - Each word with at least 5 phonemes in pronunciation chosen as "positive example"
  - Randomly selected utterance not containing word from training data as corresponding "negative example"

Utterances	500	1000	2500	5000
Positive Examples	1538	2876	7245	14570

Experiments



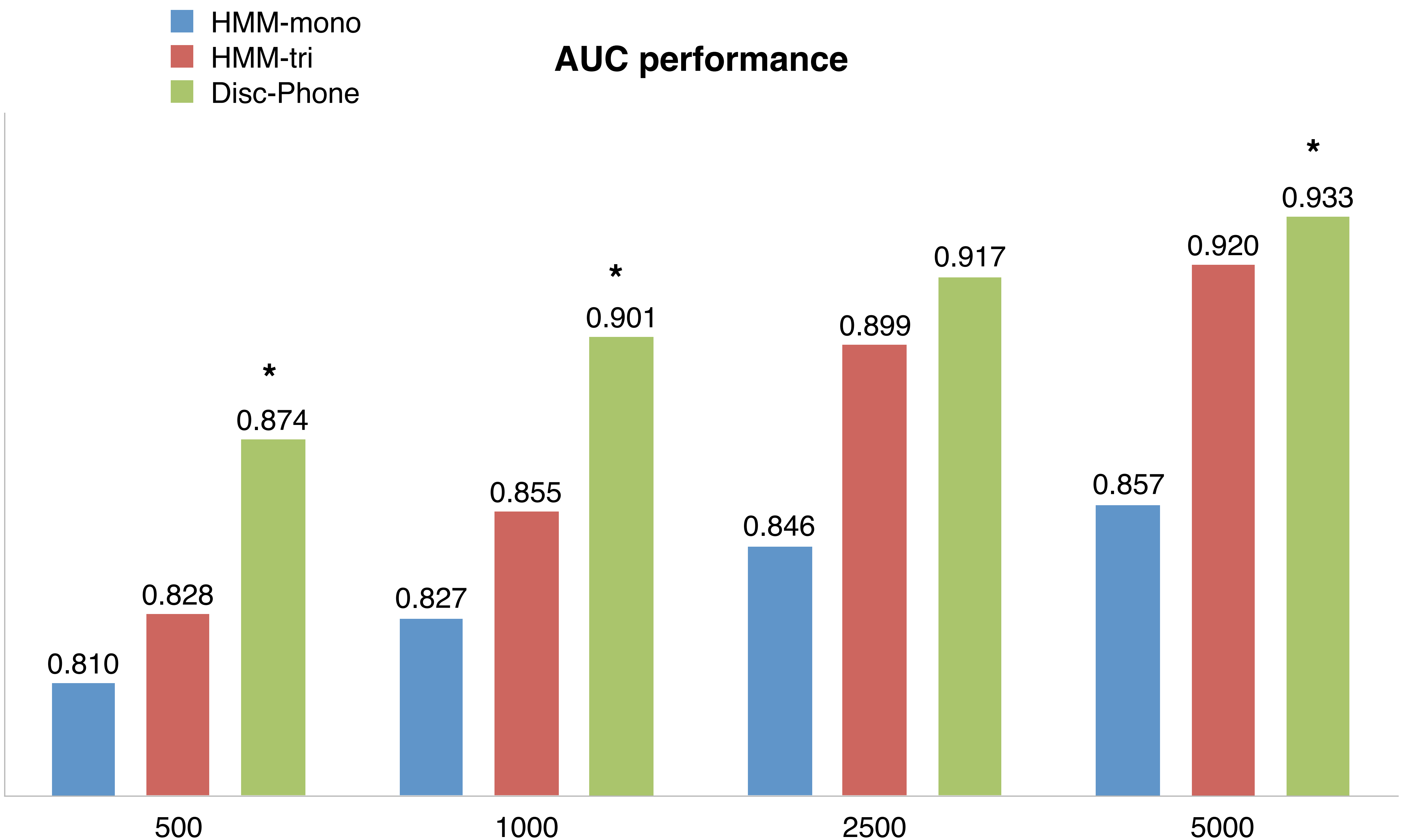
Articulatory Stream	State Space Size
Lips (L)	8
Tongue (T)	25
Glottis/Velum (G)	5

- Enforce synchrony for Lip features (L); Tongue features (T); combination of Glottis and velum (G)
- Allow at most one state of asynchrony between streams

Experiments

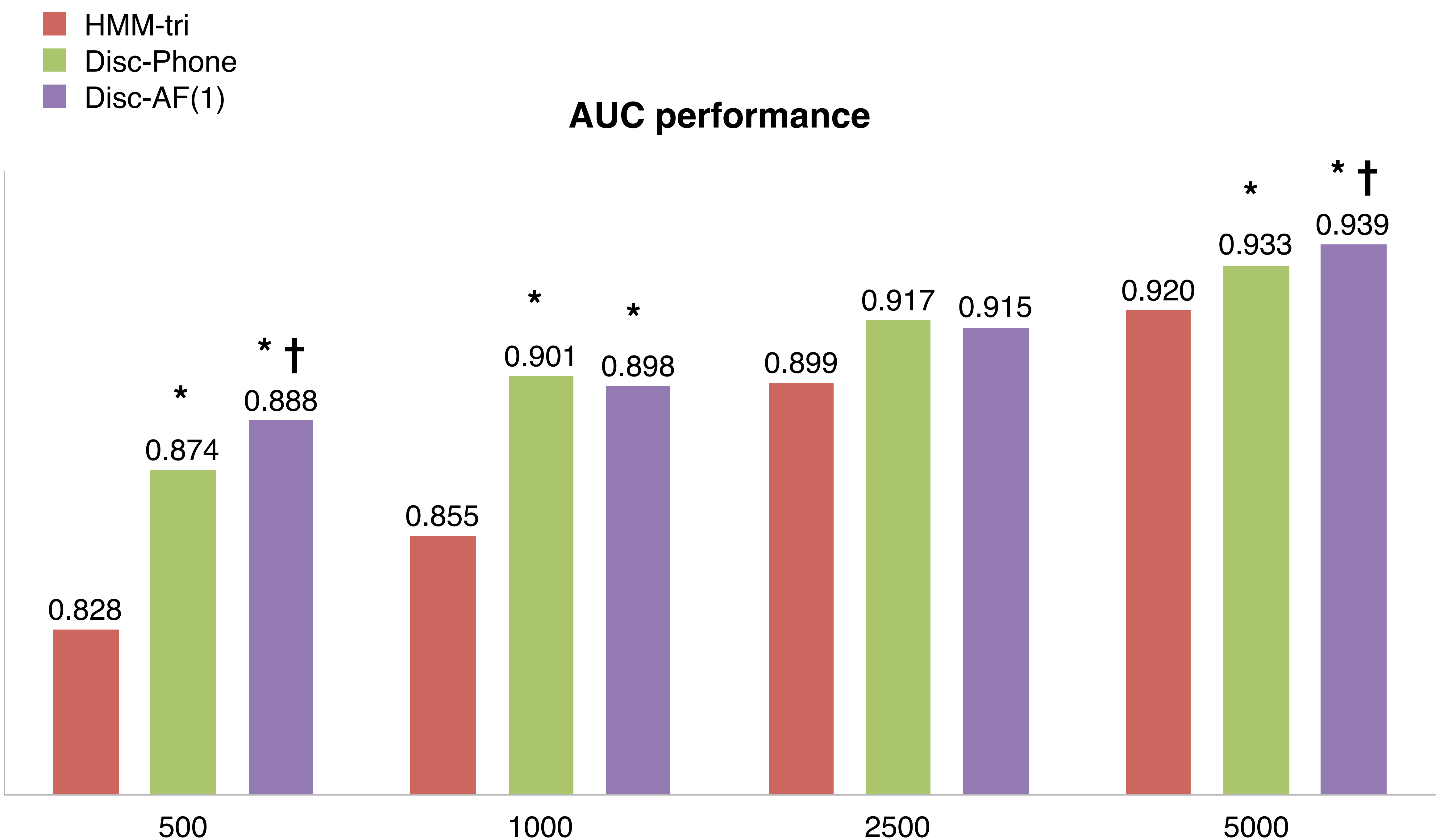
- MLPs trained on Switchboard Transcription Project (STP) (Greenberg et al. 96) data to predict phones and L, T, G labels
- "Tandem" feature extraction: projected computed phone and L, T, G log posteriors on to top 39 principal components using PCA
  - "Tandem" features used as acoustic features in baseline monophone/triphone GMM-HMM keyword-filler and discriminative systems

Results: HMM, Disc-Phone



* : significant ($p \leq 0.05$) difference over HMM-tri

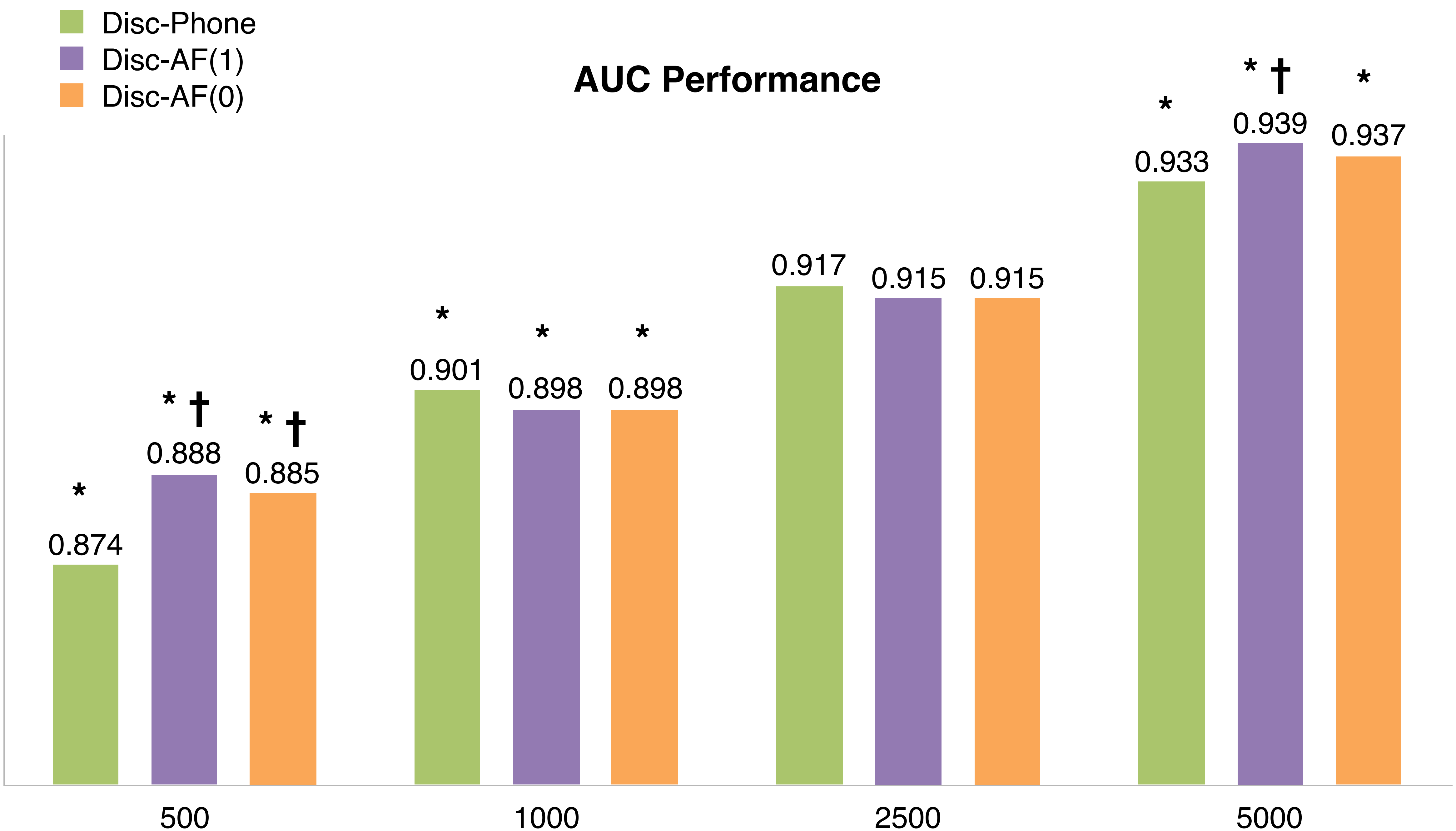
Results: HMM, Disc-Phone, Disc-AF



* : significant ($p \leq 0.05$) difference over HMM-tri

† : significant ($p \leq 0.05$) difference over Disc-Phone

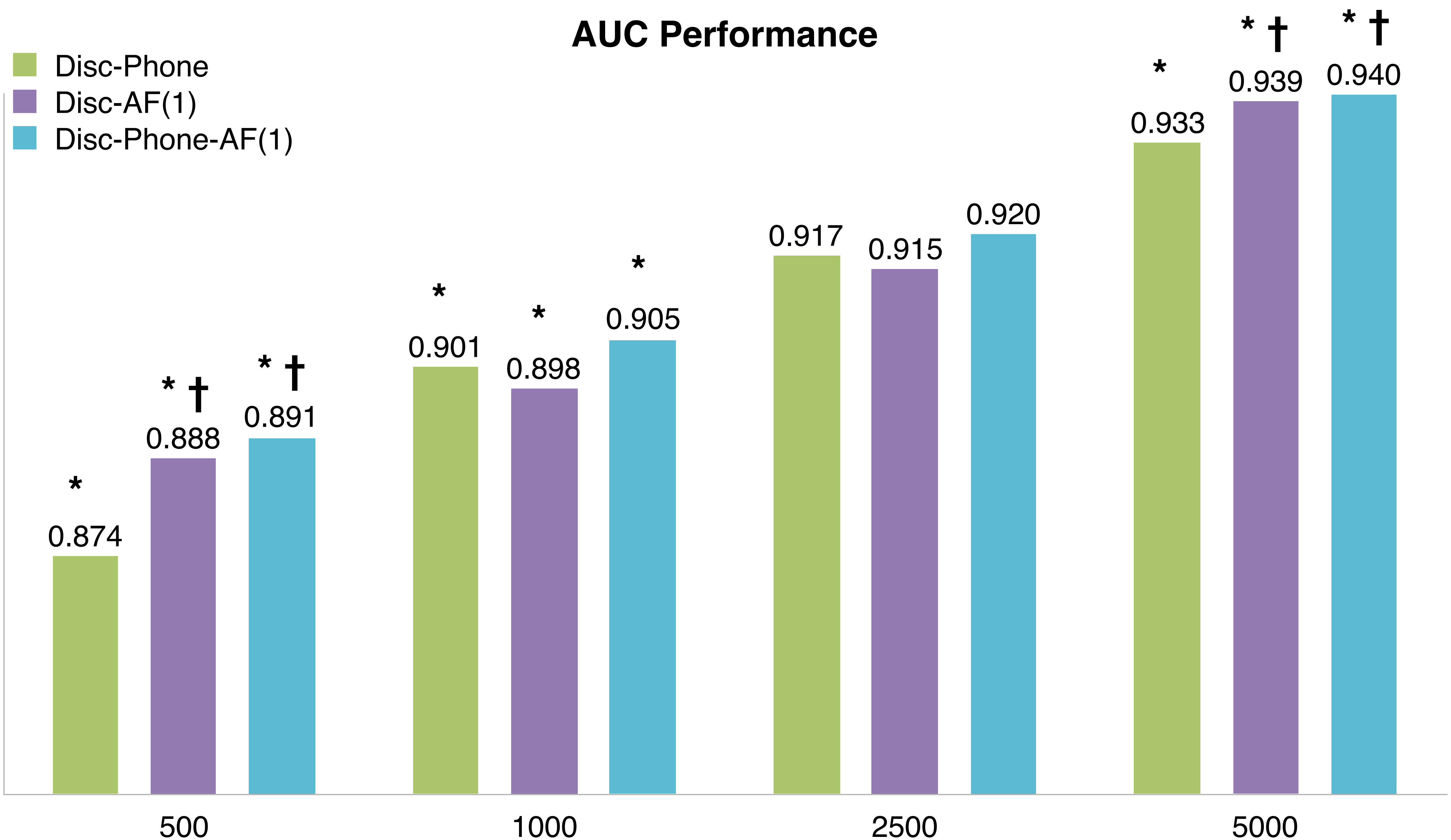
Effect of Asynchrony



* : significant ($p \leq 0.05$) difference over HMM-tri

† : significant ($p \leq 0.05$) difference over Disc-Phone

Combining Phone, AF Models



* : significant ($p \leq 0.05$) difference over HMM-tri
† : significant ($p \leq 0.05$) difference over Disc-Ph

Conclusions

- Discriminative systems outperform the HMM systems by large margins
- AF-based system outperform phone-based systems in very-low-resource conditions
  - System appears to hypothesize greater asynchrony for words with pronunciation variation
- In current work, we are exploring techniques for optimizing ATWV instead

Acknowledgement

articulatory phonology modeling

discriminative keyword spotting

Thanks!