האוניברסיטה העברית בירושלים
The Hebrew University of Jerusalem

DIRAC
Detection and Identification of Rare Audio-visual Cues

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

Discriminative Keyword Spotting

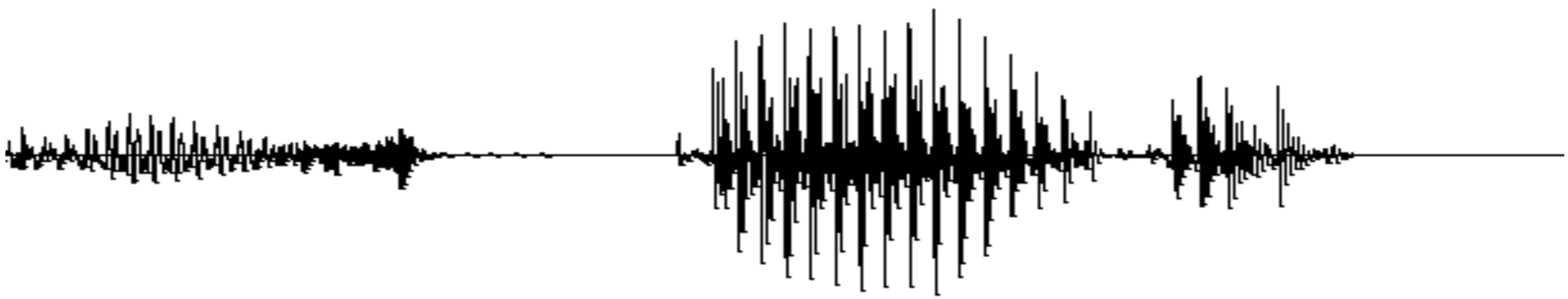
Joseph Keshet, The Hebrew University
David Grangier, IDIAP Research Institute
Samy Bengio, Google Inc.

Outline

- Problem Definition
- Keyword Spotting with HMMs
- Discriminative Keyword Spotting
  - derivation
  - analysis
  - feature functions
- Experimental Results

Problem Definition

<u>Goal</u>: find a keyword in a speech signal

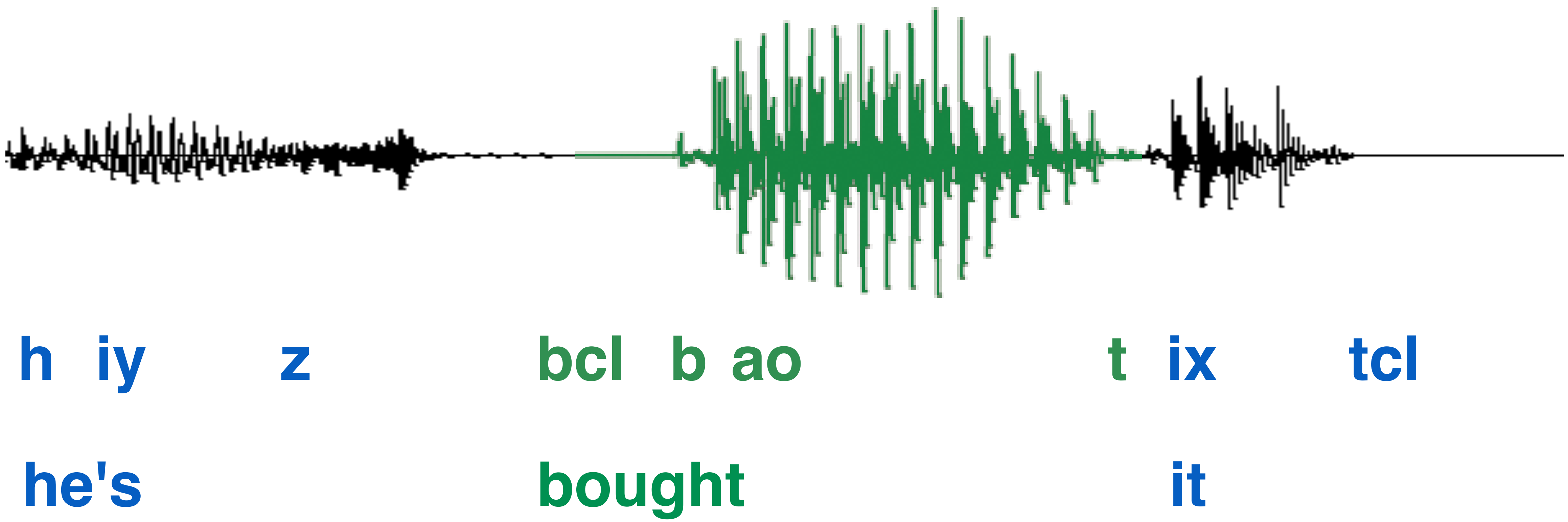


h iy z bcl b ao t ix tcl

he's bought it

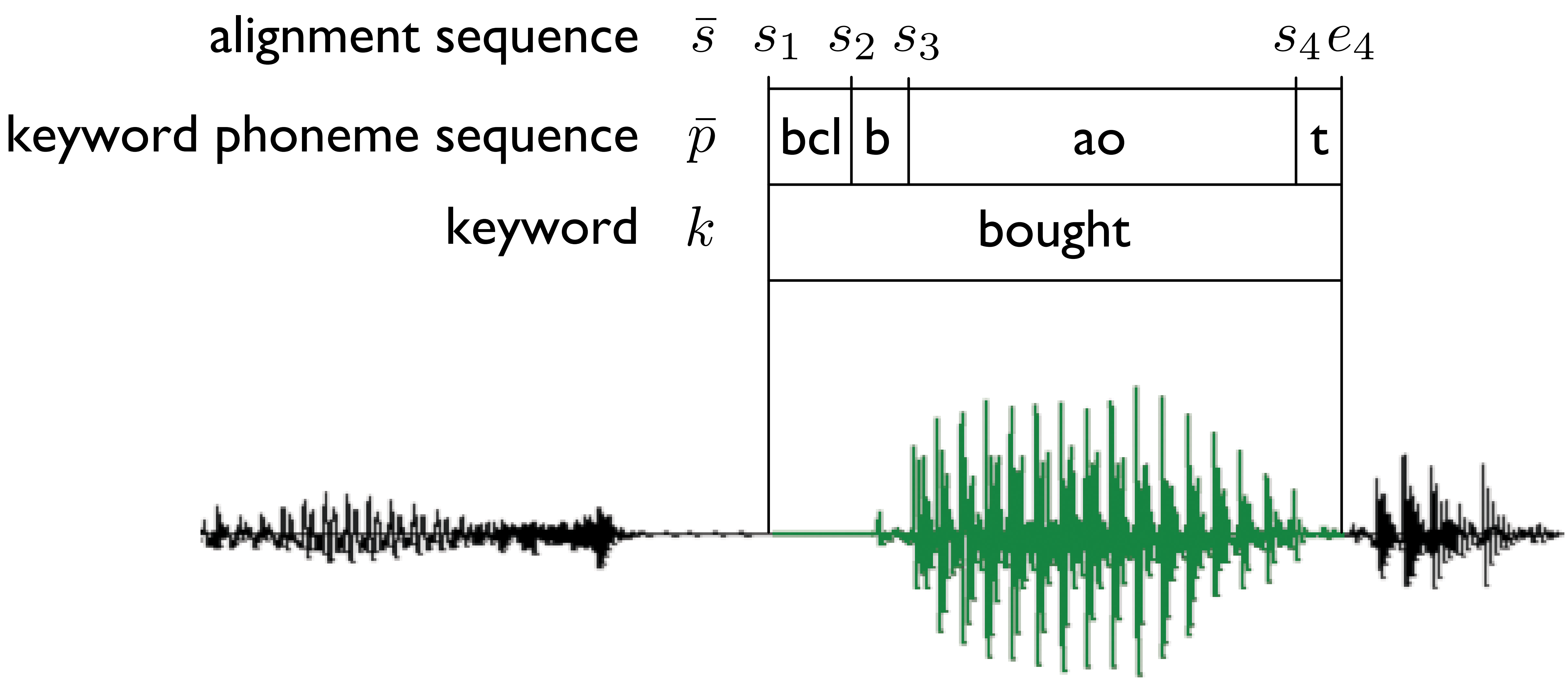
Problem Definition

Goal: find a keyword in a speech signal



Problem Definition

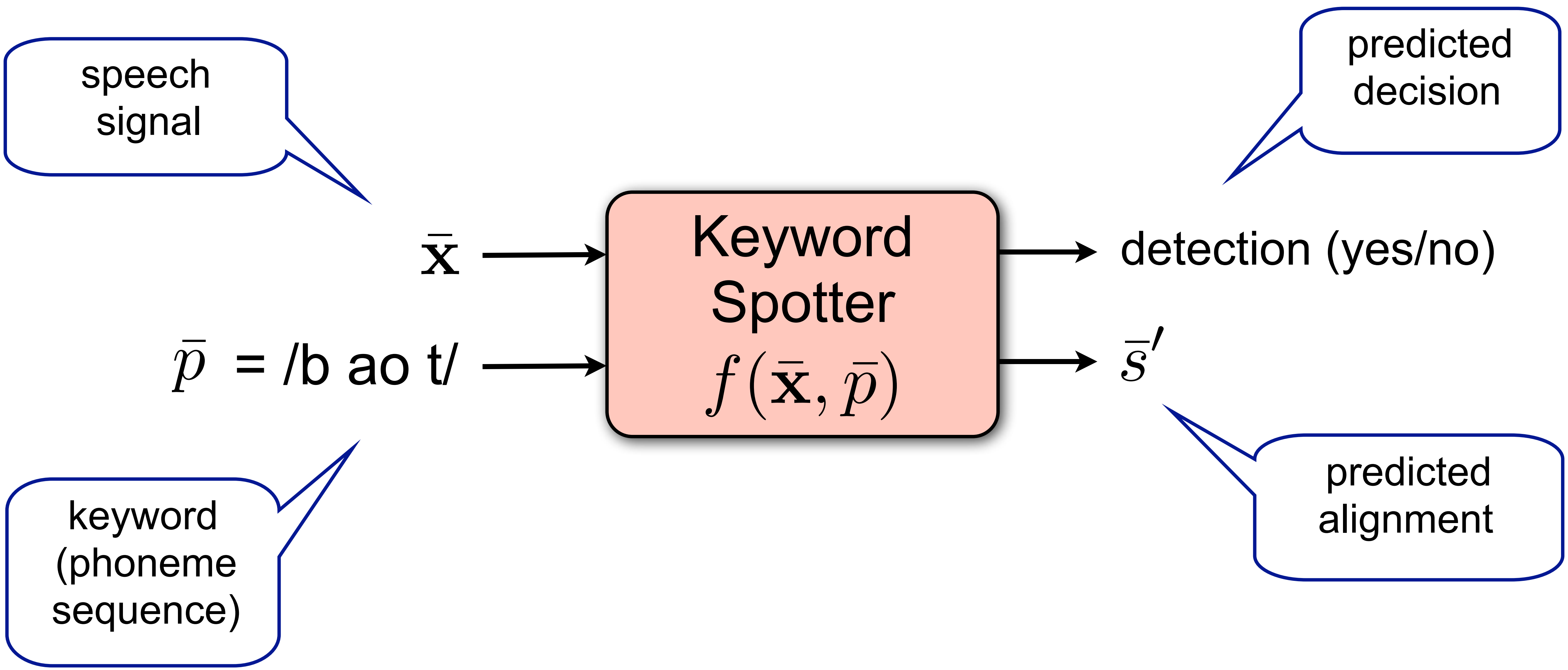
Notation:



$$\bar{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \quad \cdots \quad \mathbf{x}_T)$$

acoustic feature vectors

Problem Definition



Fat is Good

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

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

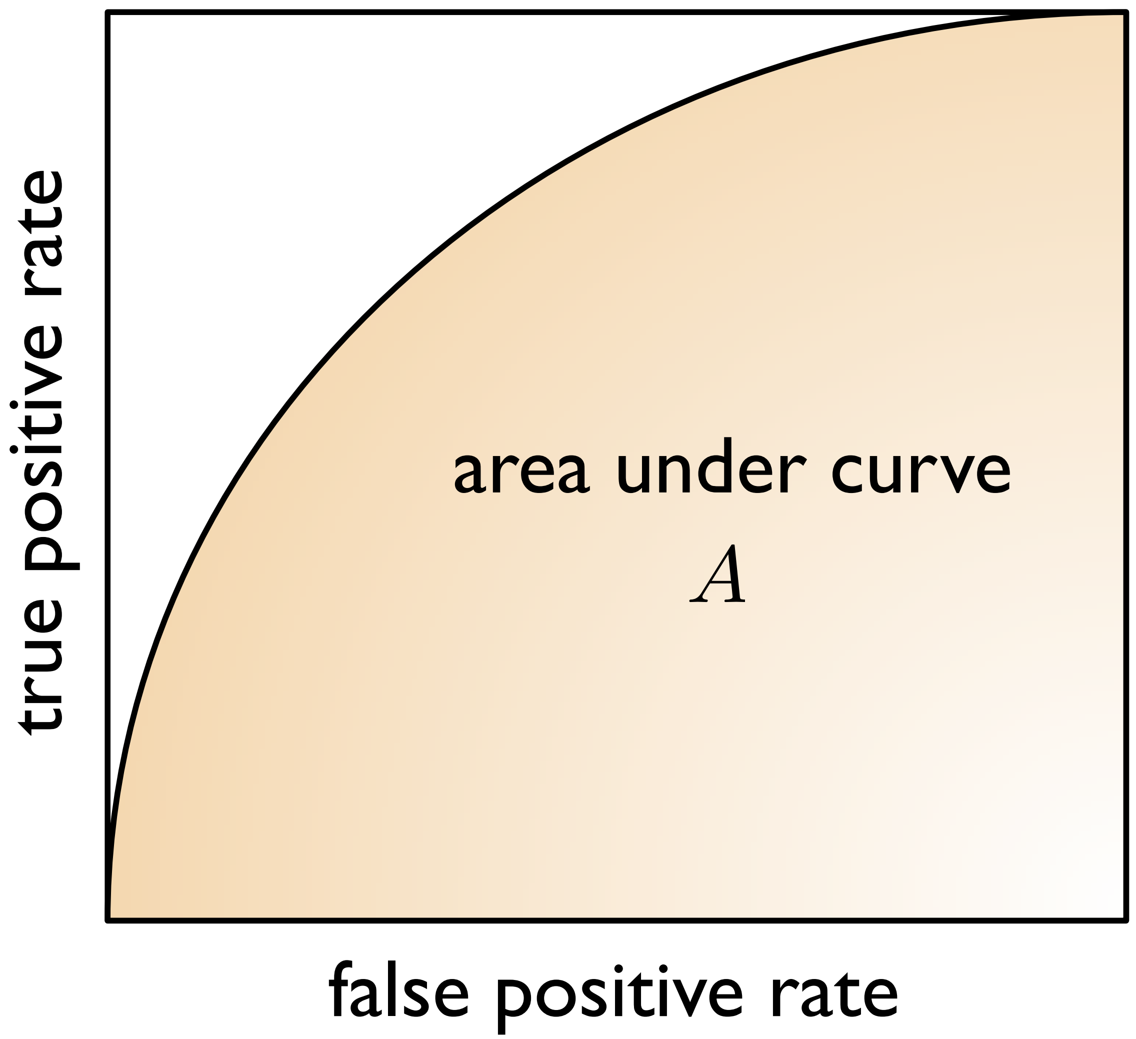
false positive = $\dfrac{\text{detected utterances without keywords}}{\text{total utterances without keywords}}$

Fat is Good

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

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

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

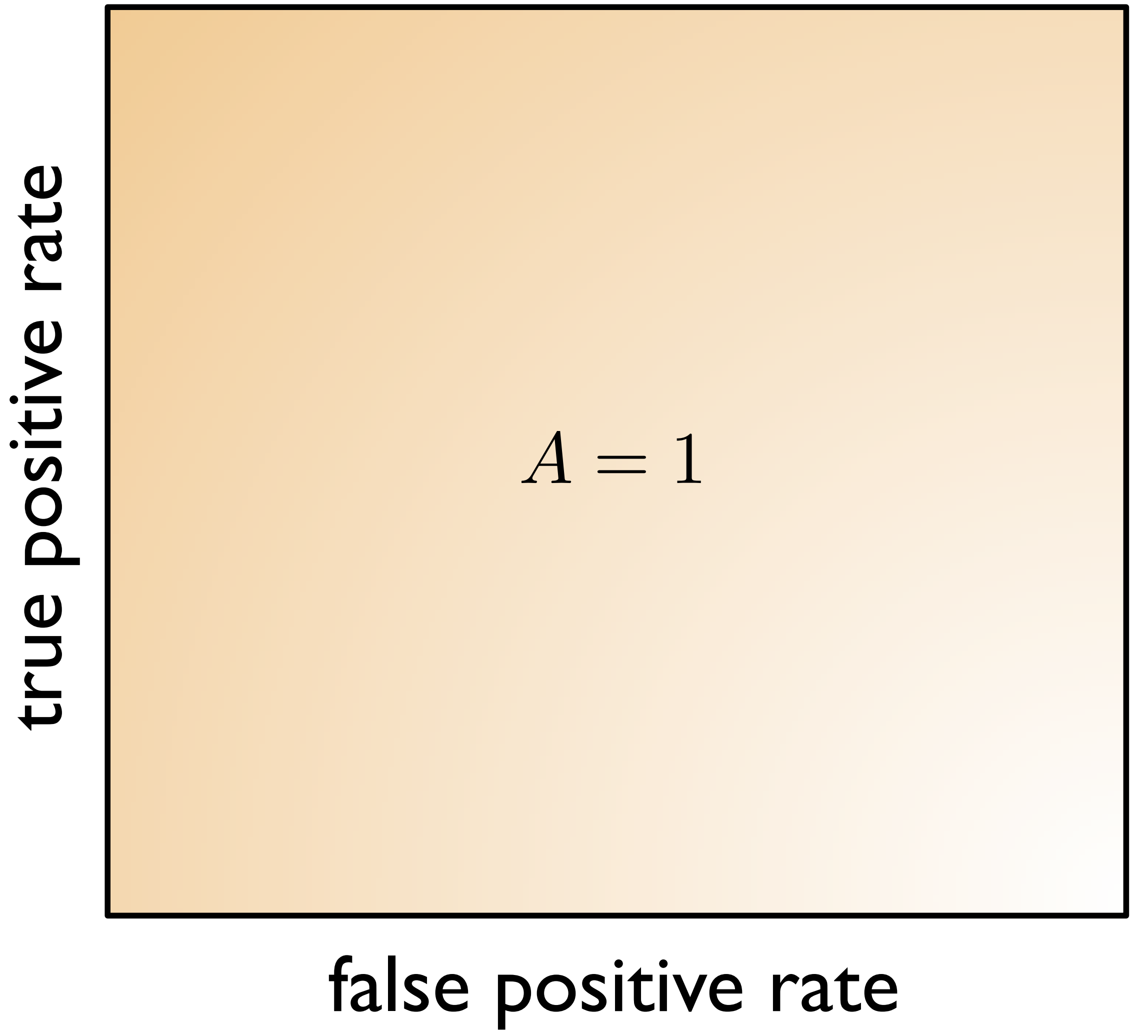


Fat is Good

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

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

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

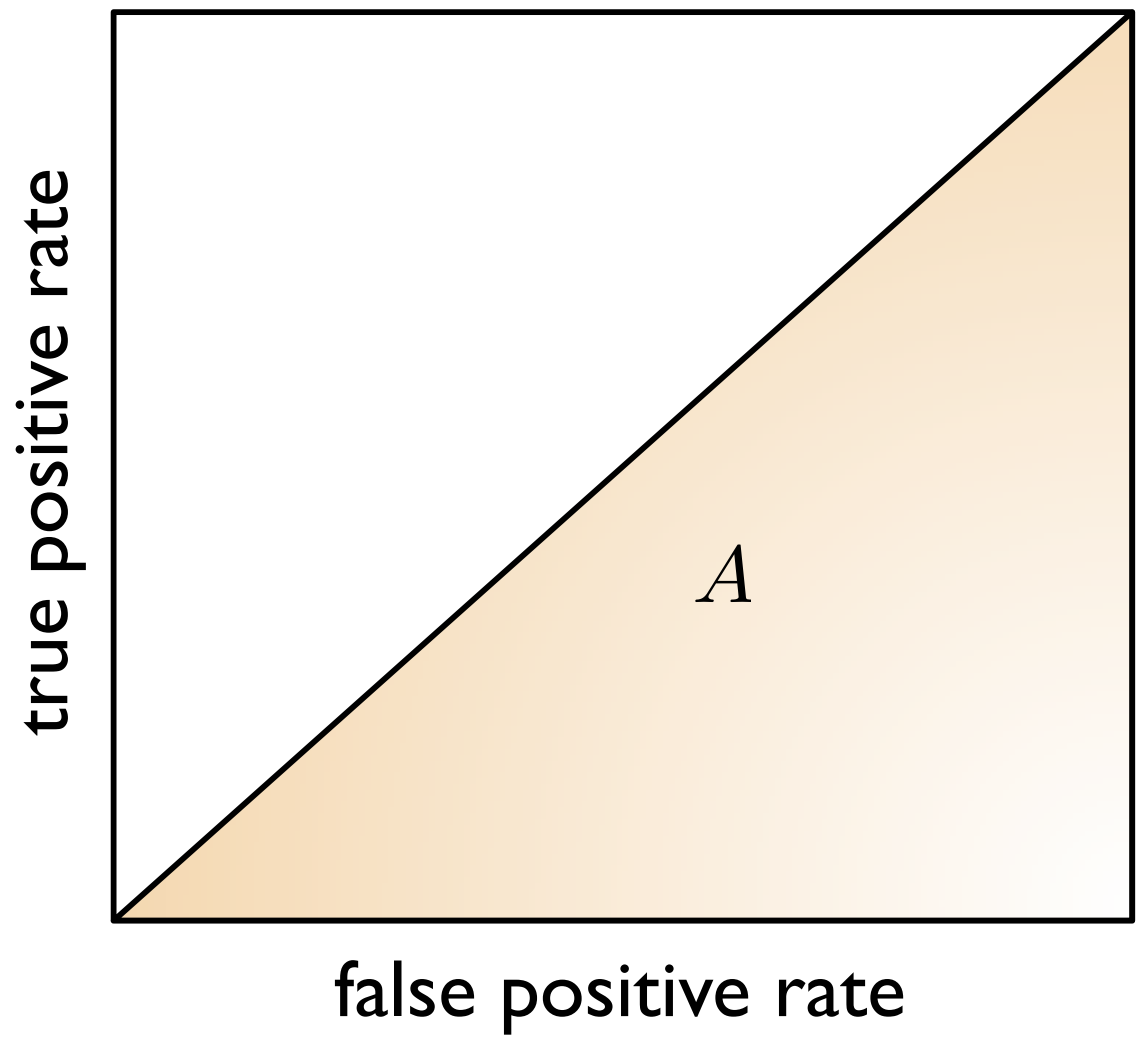


Fat is Good

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

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

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

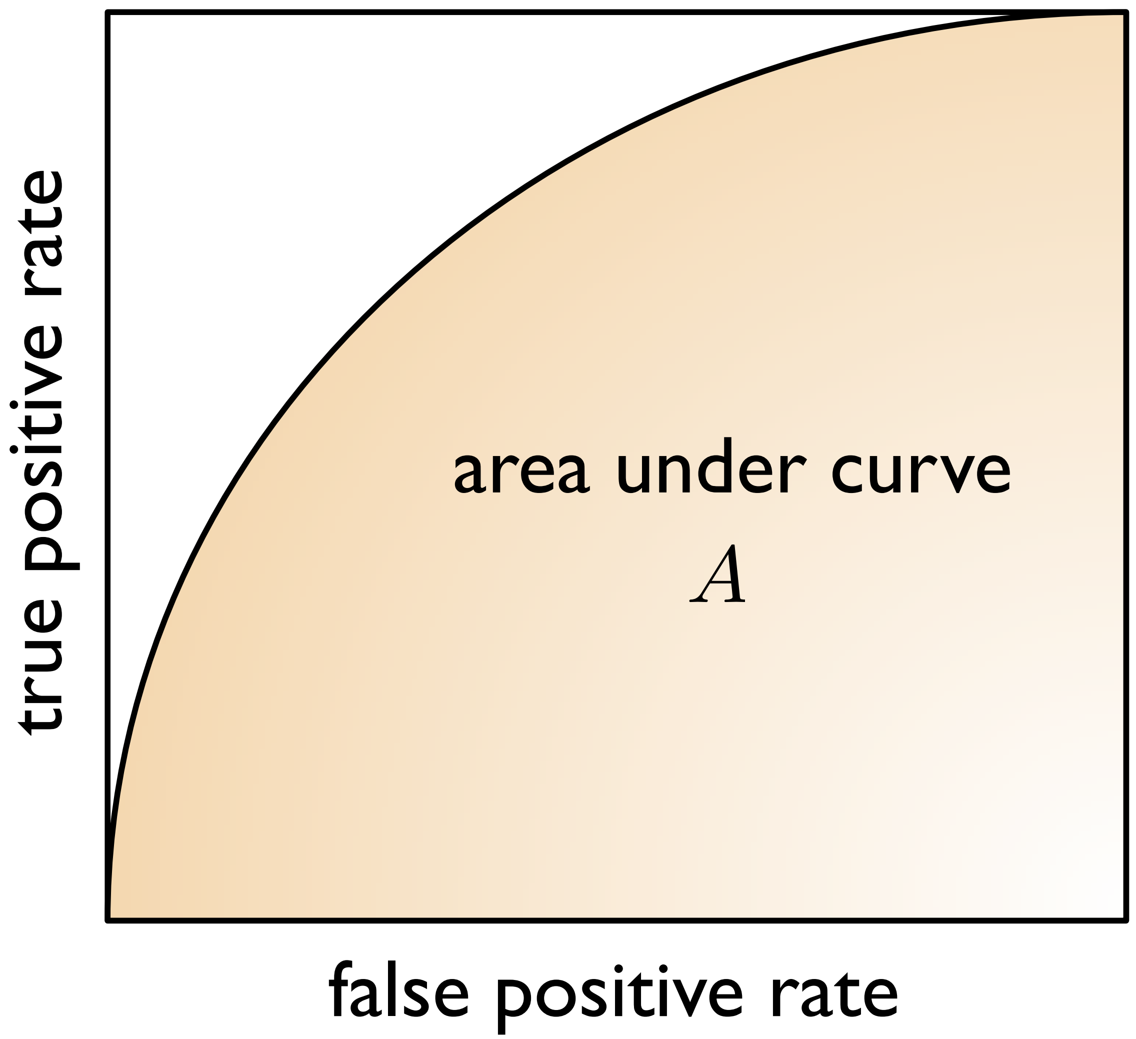


Fat is Good

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

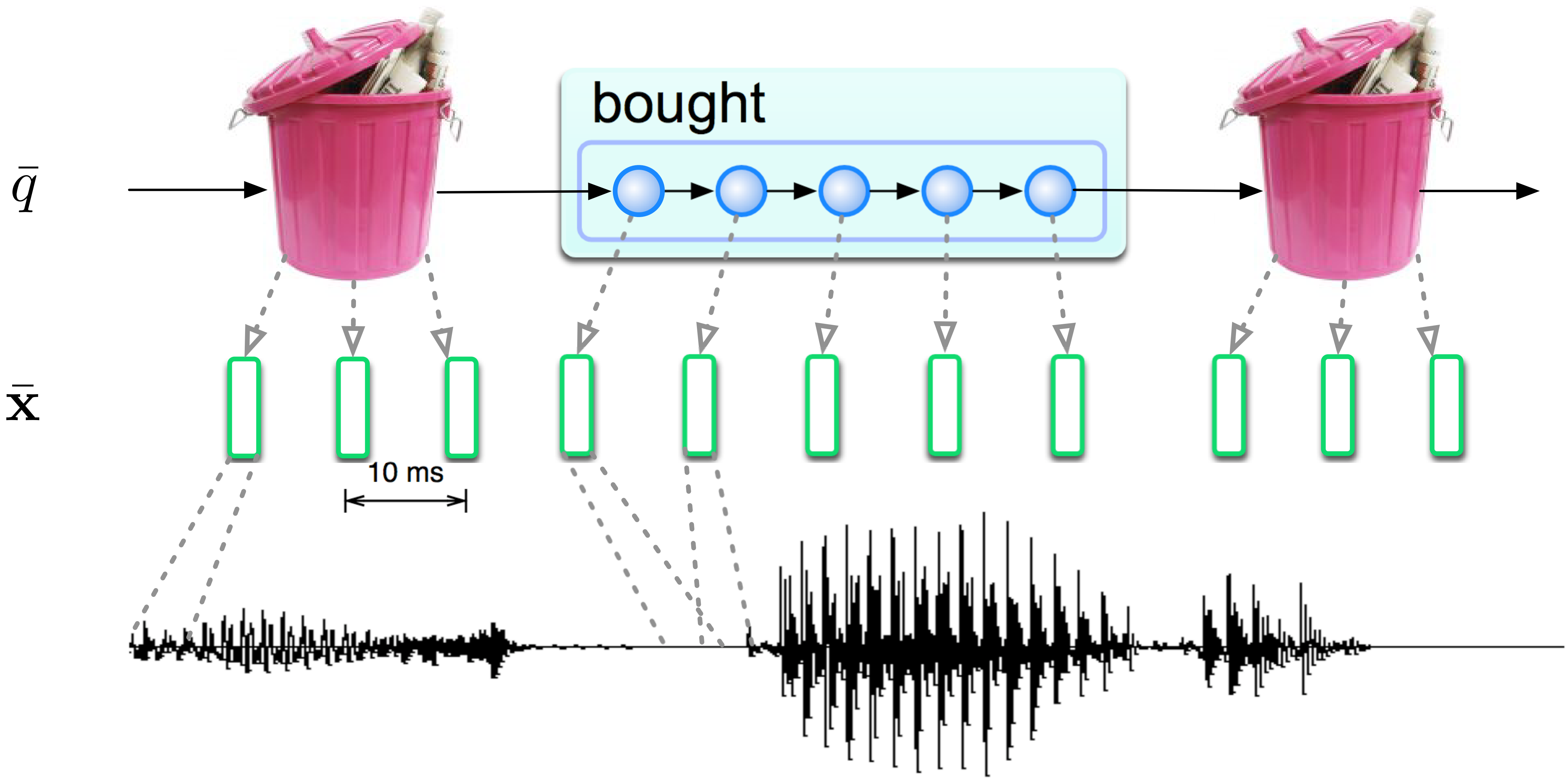
true positive = $\dfrac{\text{detected utterances with keywords}}{\text{total utterances with keywords}}$

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



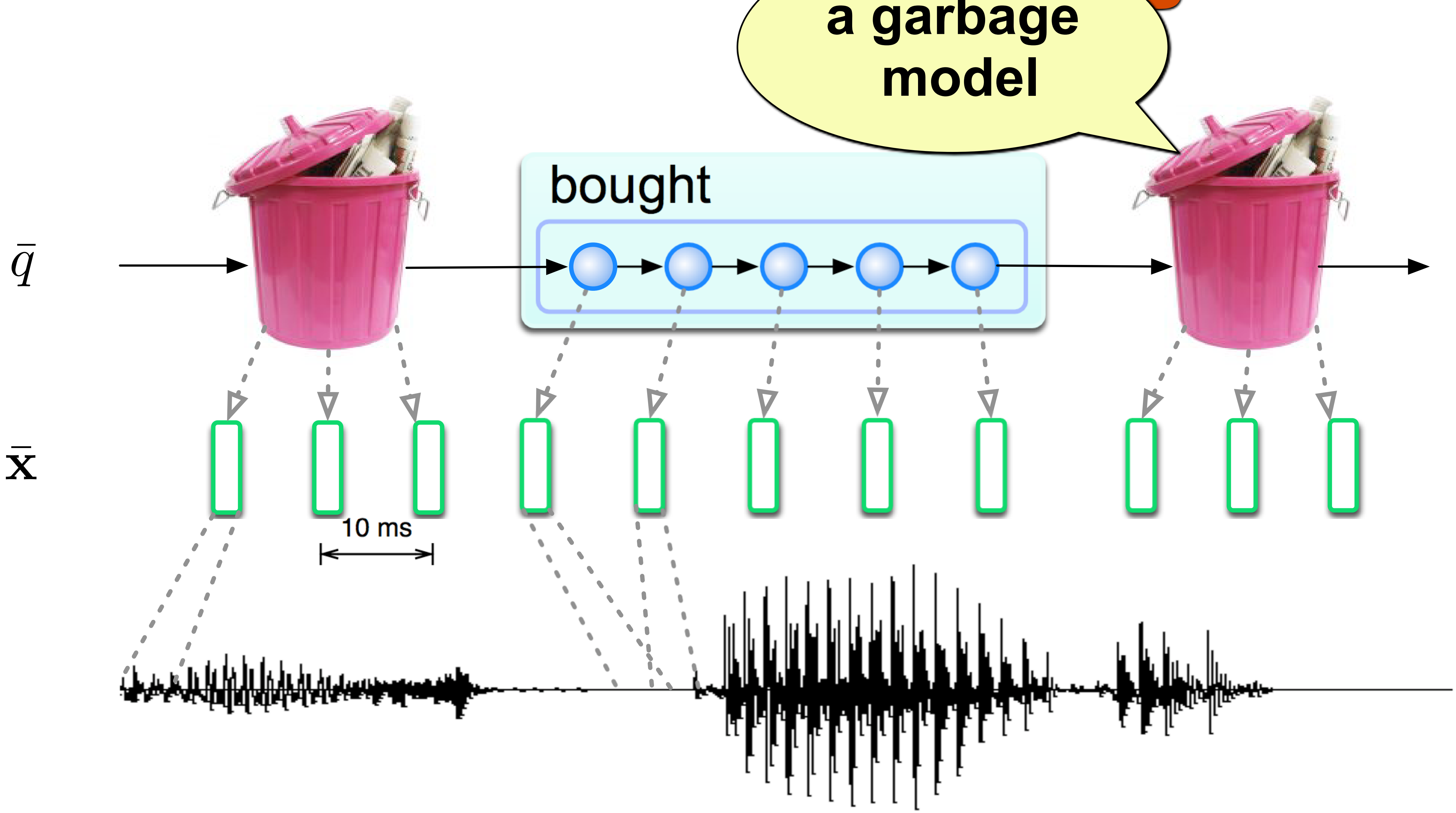
HMM-based Keyword Spotting

HMM-based Keyword Spotting
Whole Word Modeling



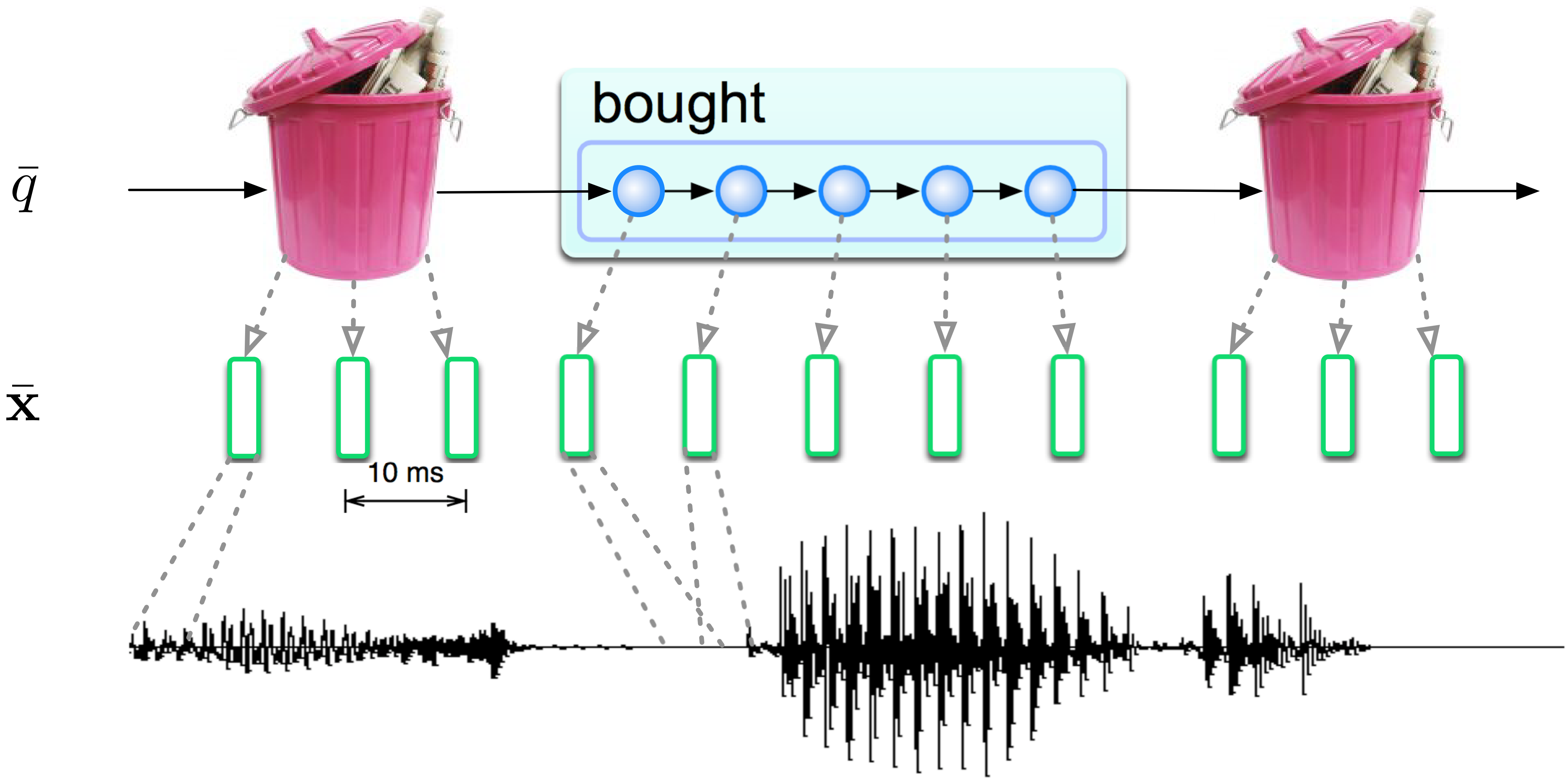
[Rahim et al, 1997; Rohlicek et al, 1989]

HMM-based Keyword Spotting
Whole Word Modeling



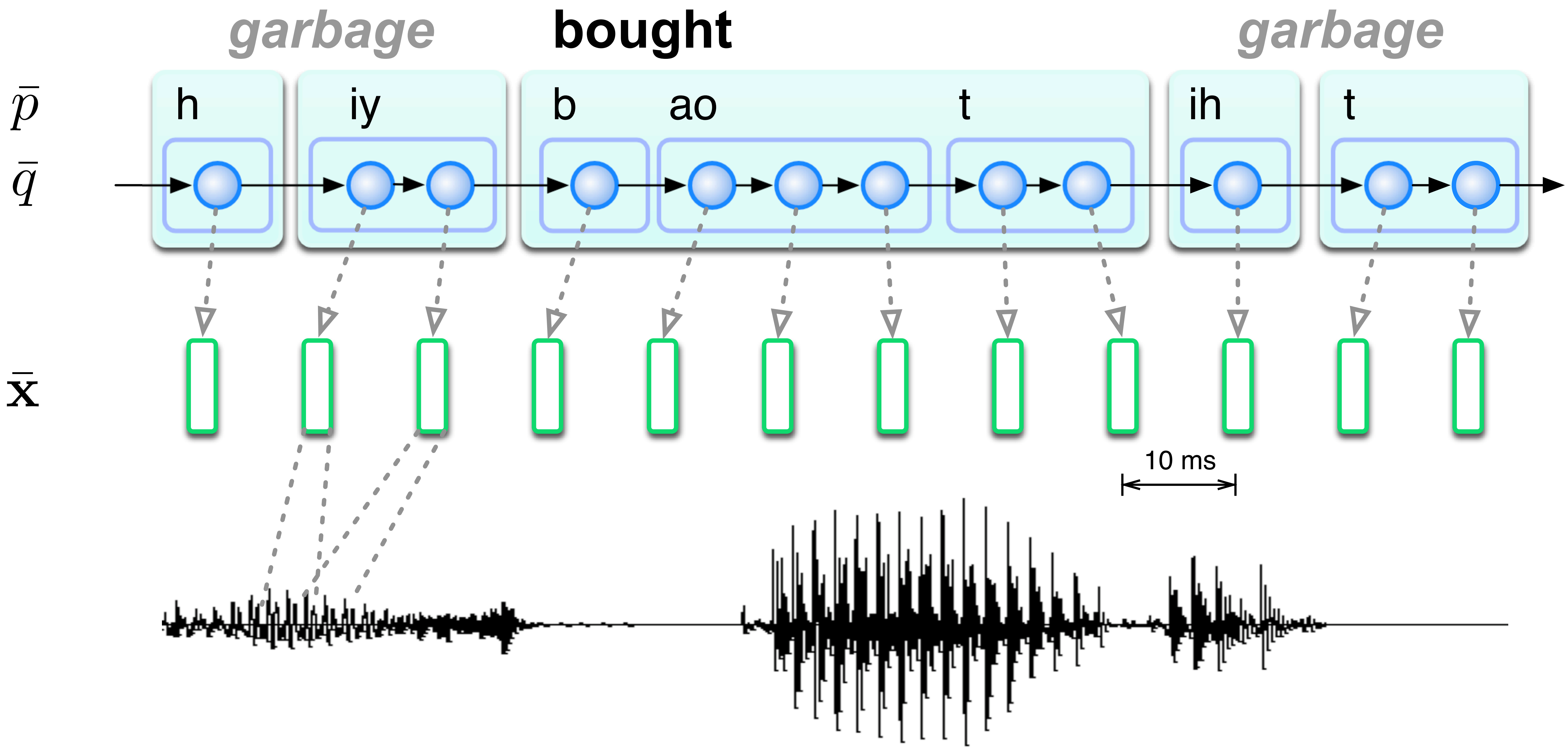
[Rahim et al, 1997; Rohlicek et al, 1989]

HMM-based Keyword Spotting
Whole Word Modeling



[Rahim et al, 1997; Rohlicek et al, 1989]

HMM-based Keyword Spotting
Phoneme-Based

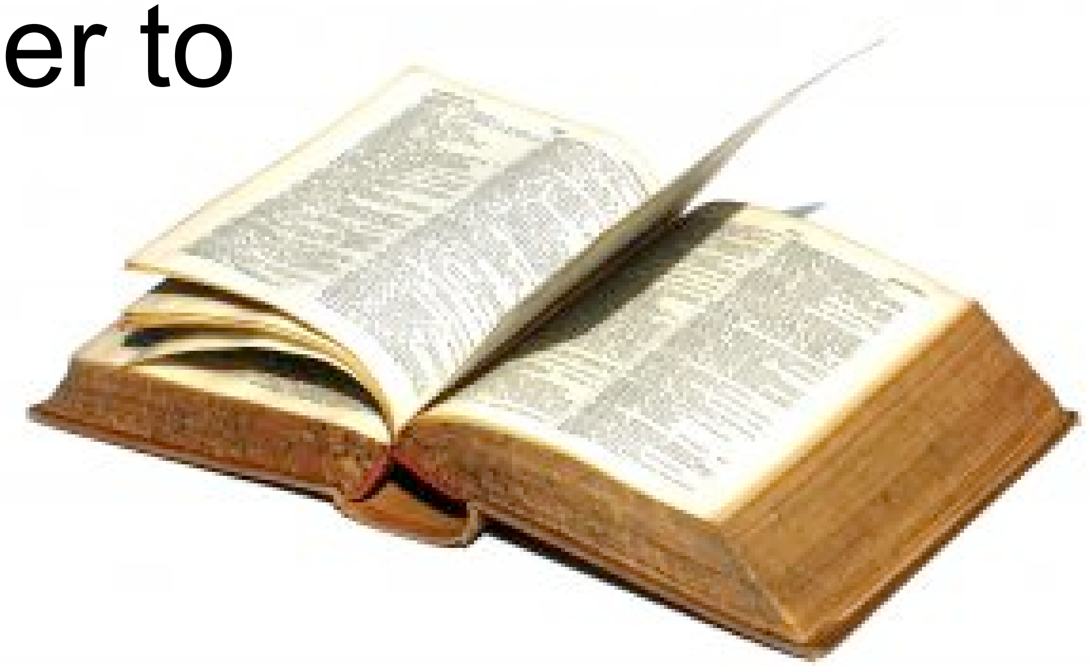


[Bourlard et al, 1994; Manos & Zue, 1997; Rohlicek et al, 1993]

HMM-based Keyword Spotting
Large Vocabulary Based

- Linguistic constraints on the garbage model

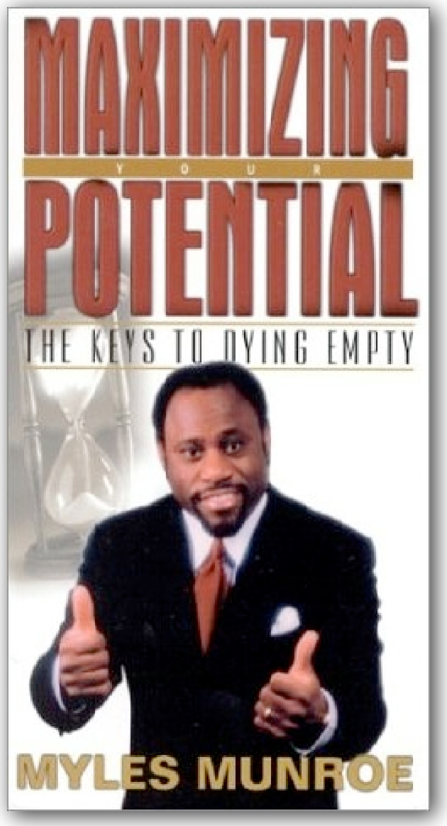
- Does a human listener need to have a large vocabulary in order to recognize one word?



(Cardillo et al, 2002; Rose & Paul, 1990; Szoke et al, 2005; Weintraub, 1995)

HMM Approaches to Keyword Spotting

- Do <u>not</u> address specifically the goal of <u>maximizing the area under the ROC curve</u> for the task of keyword spotting



Discriminative Approach

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

keyword (phoneme sequence)

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

utterance in which the keyword is uttered

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

utterance in which the keyword is not uttered

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

alignment of the keyword and the utterance with keyword

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

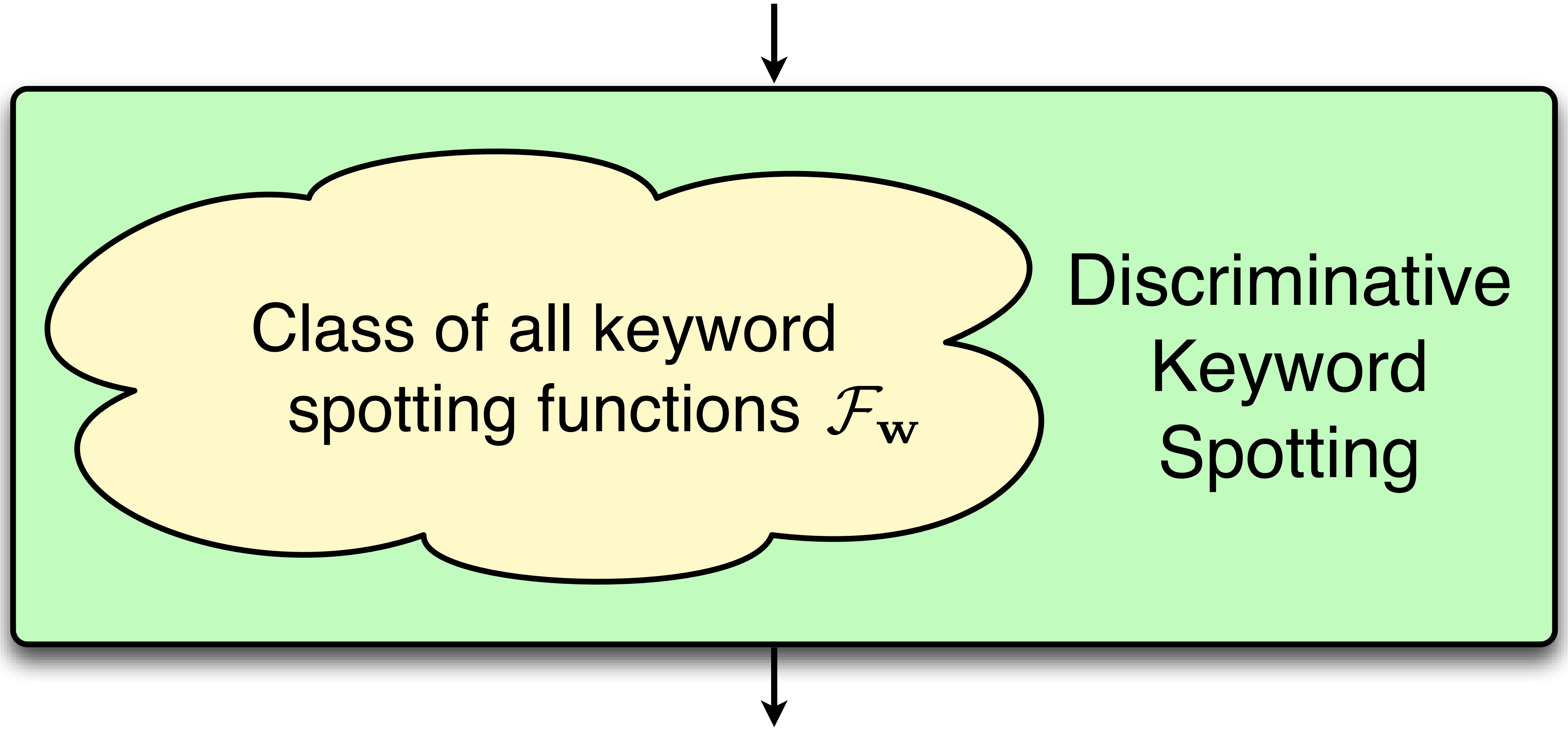
Discriminative Keyword Spotting

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$



Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$f(\bar{\mathbf{x}}, \bar{p}) = \max_{\bar{s}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, \bar{p}, \bar{s})$$

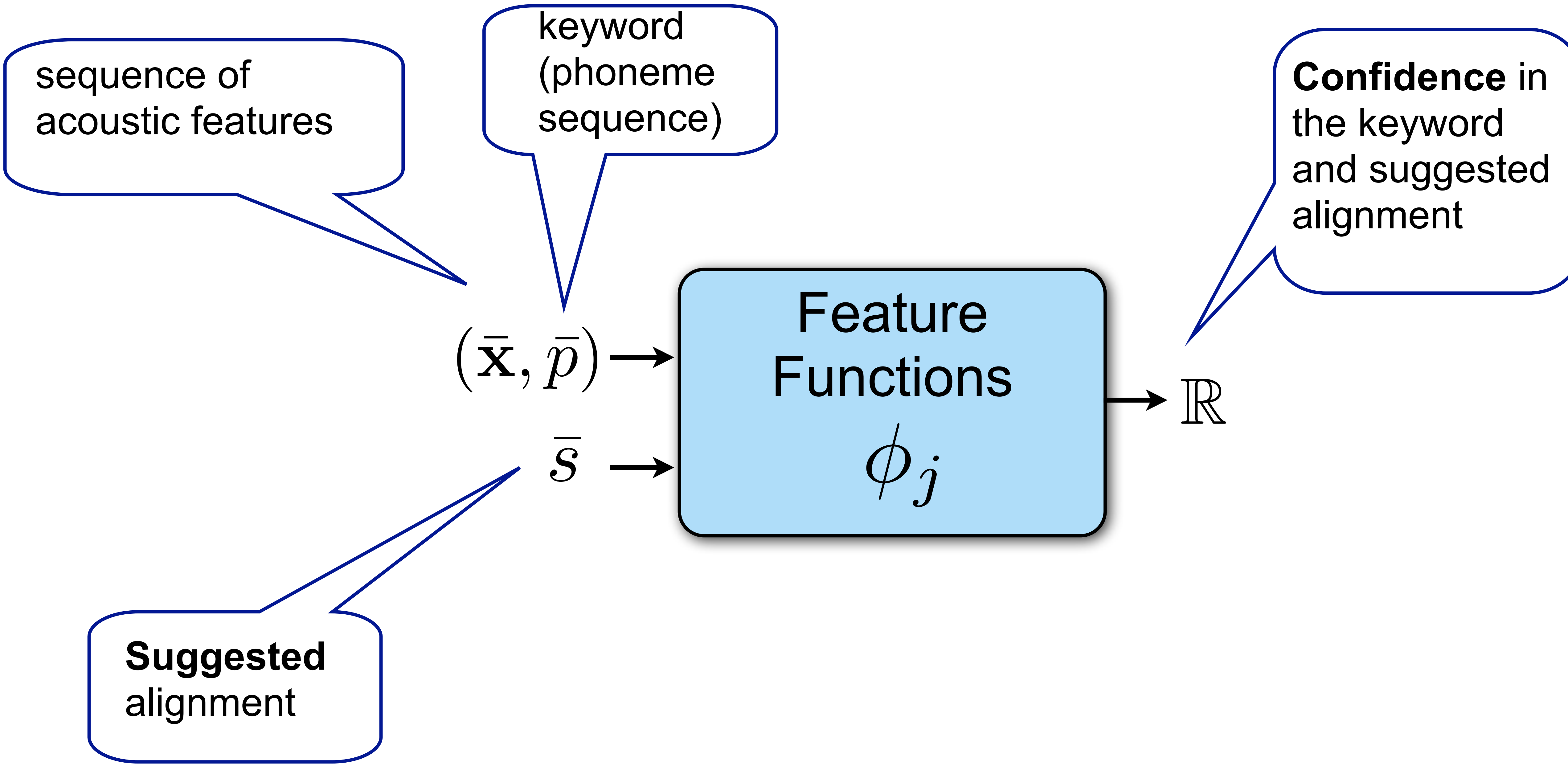
$$\mathbf{w} \in \mathbb{R}^n$$

Discriminative Keyword Spotting

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Feature Functions

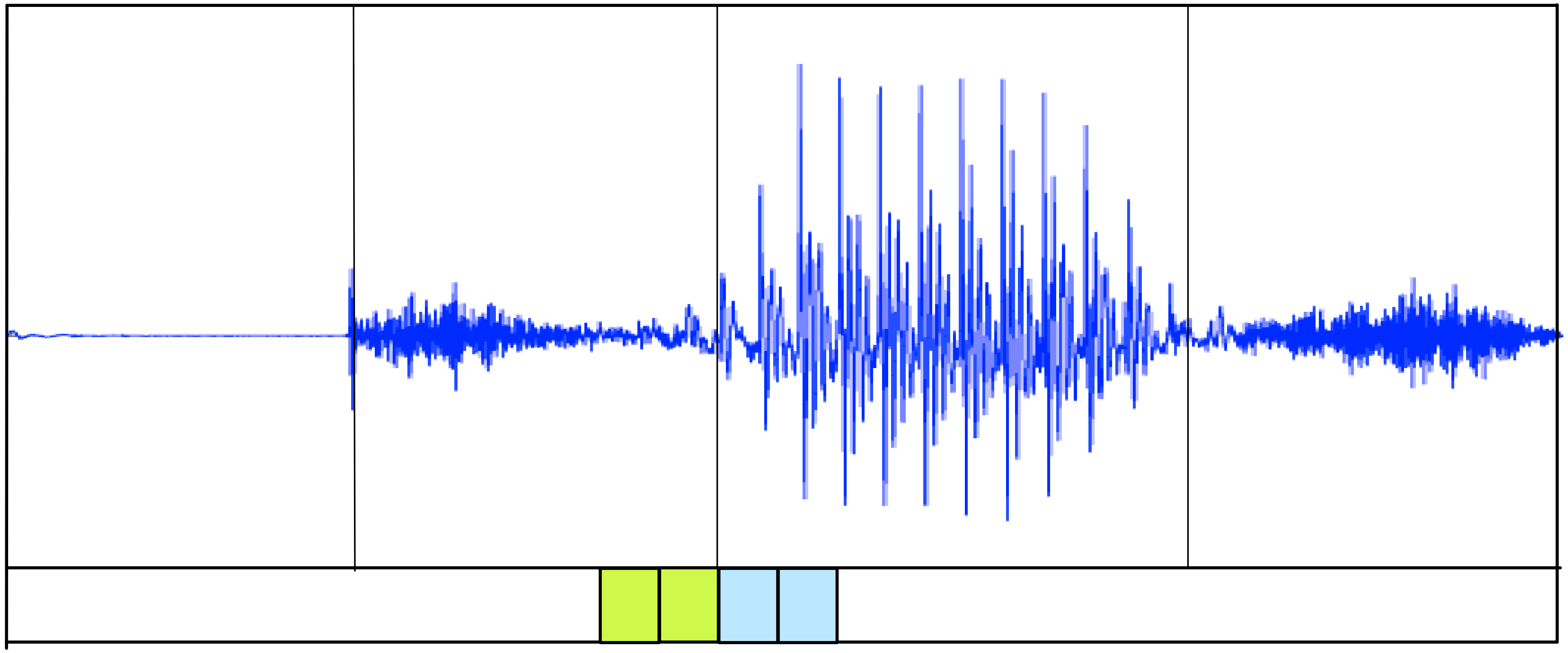
We define 7 feature functions of the form:



Feature Functions I

Cumulative spectral change around the boundaries

$$\phi_j(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=2}^{|\bar{p}|-1} d(\mathbf{x}_{-j+s_i}, \mathbf{x}_{j+s_i}), \quad j \in \{1,2,3,4\}$$

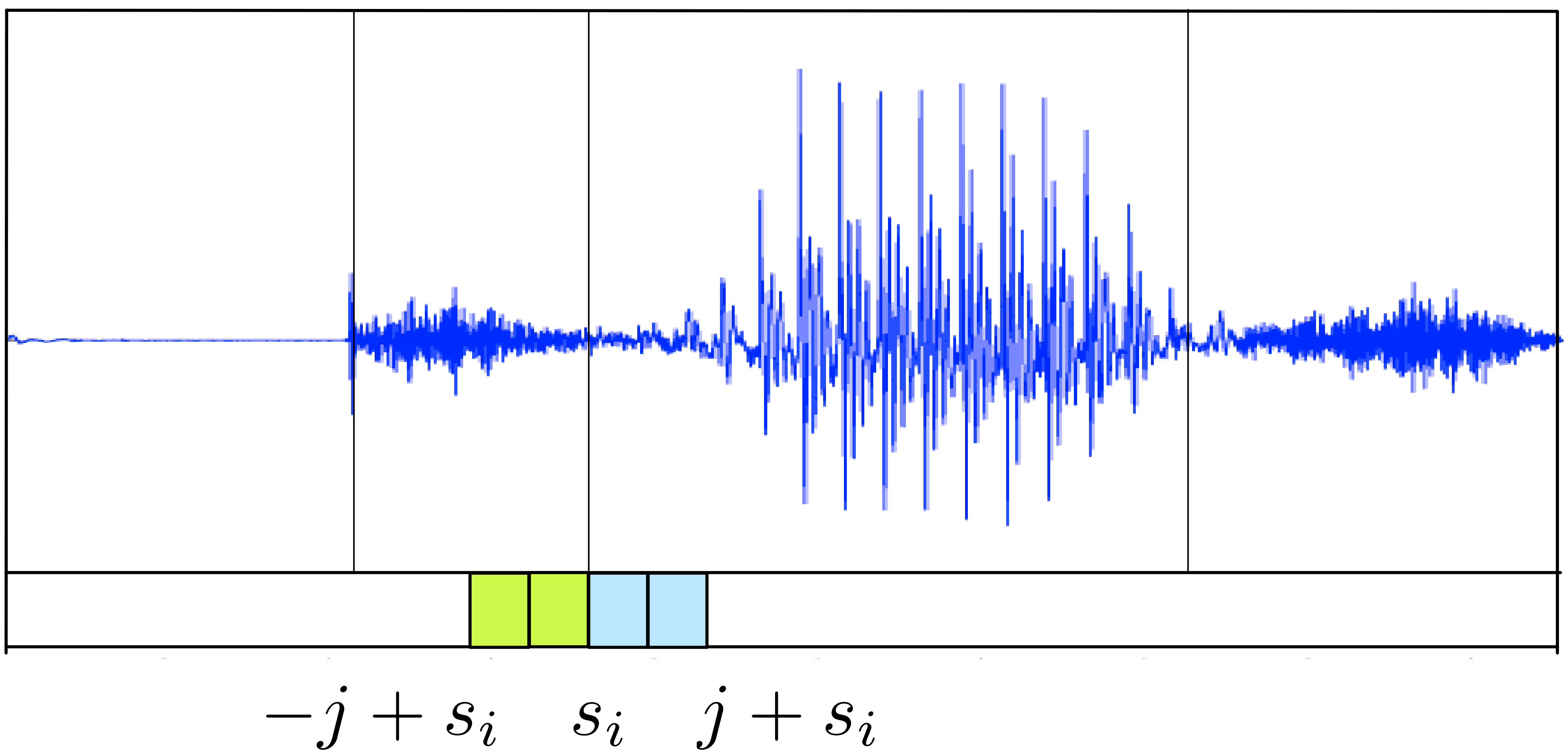


$-j + s_i \quad s_i \quad j + s_i$

Feature Functions I

Cumulative spectral change around the boundaries

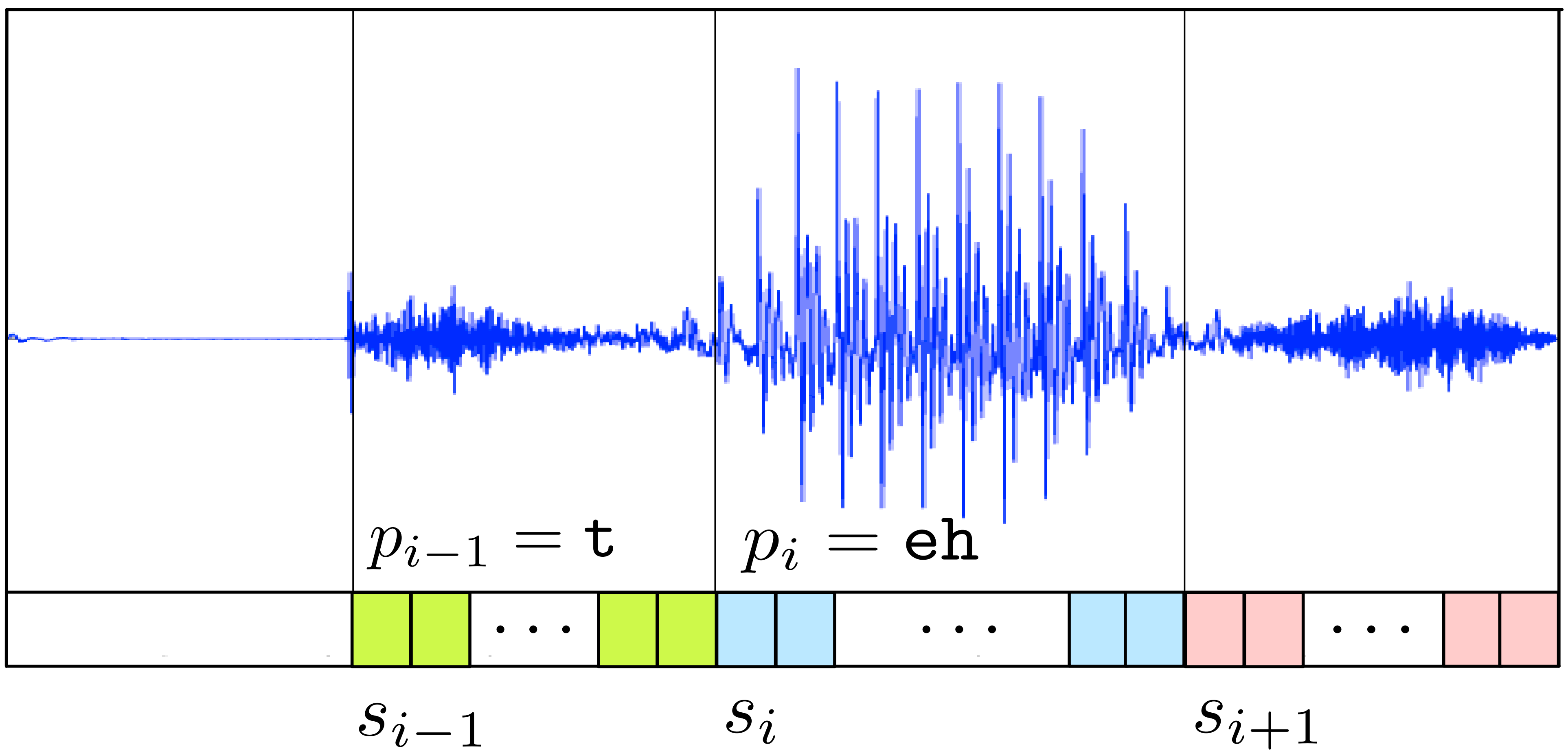
$$\phi_j(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=2}^{|\bar{p}|-1} d(\mathbf{x}_{-j+s_i}, \mathbf{x}_{j+s_i}), \quad j \in \{1,2,3,4\}$$



Feature Functions II

Cumulative confidence in the phoneme sequence

$$\phi_5(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \sum_{t=s_i}^{s_{i+1}-1} g(\mathbf{x}_t, p_i)$$



Feature Functions II

Cumulative confidence in the phoneme sequence

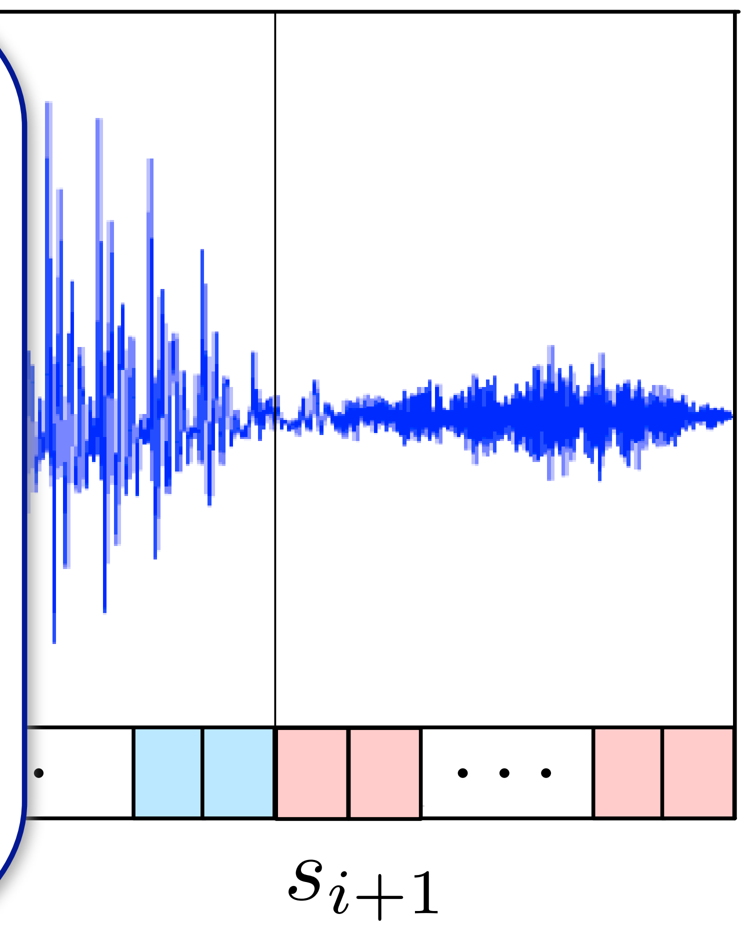
$$\phi_5(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \sum_{t=s_i}^{s_{i+1}-1} g(\mathbf{x}_t, p_i)$$

We build a static frame-based phoneme classifier

$$g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

$g(\mathbf{x}_t, p_i)$ is the confidence that phoneme p_i was uttered at frame $\mathbf{x}_t$

[Dekel, Keshet, Singer, ‘04]

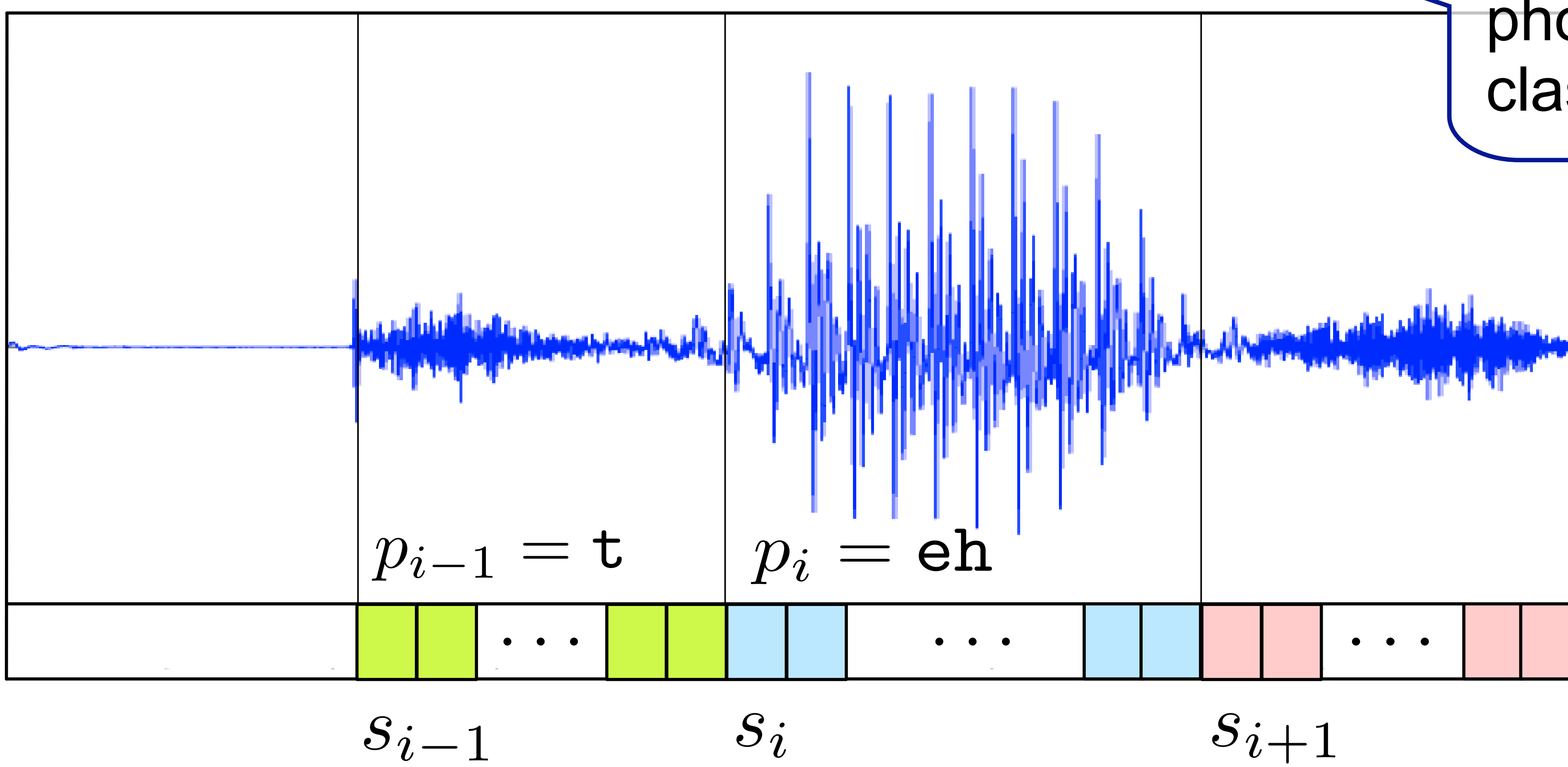


Feature Functions II

Cumulative confidence in the phoneme sequence

$$\phi_5(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \sum_{t=s_i}^{s_{i+1}-1} g(\mathbf{x}_t, p_i)$$

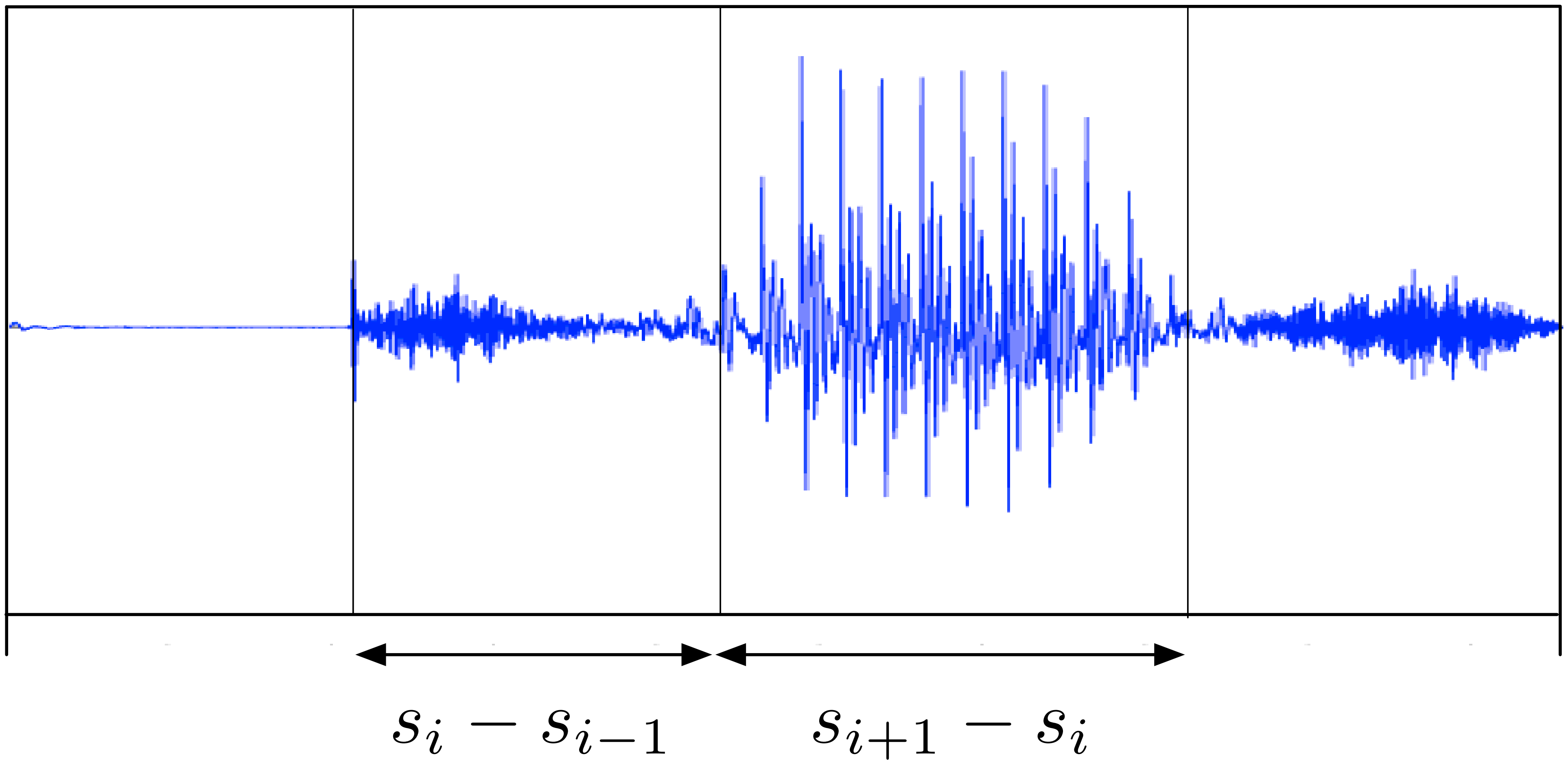
frame based phoneme classifier



Feature Functions III

Phoneme duration model

$$\phi_6(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \log\ \mathcal{N}(s_{i+1} - s_i; \hat{\mu}_{p_i}\ , \hat{\sigma}_{p_i})$$



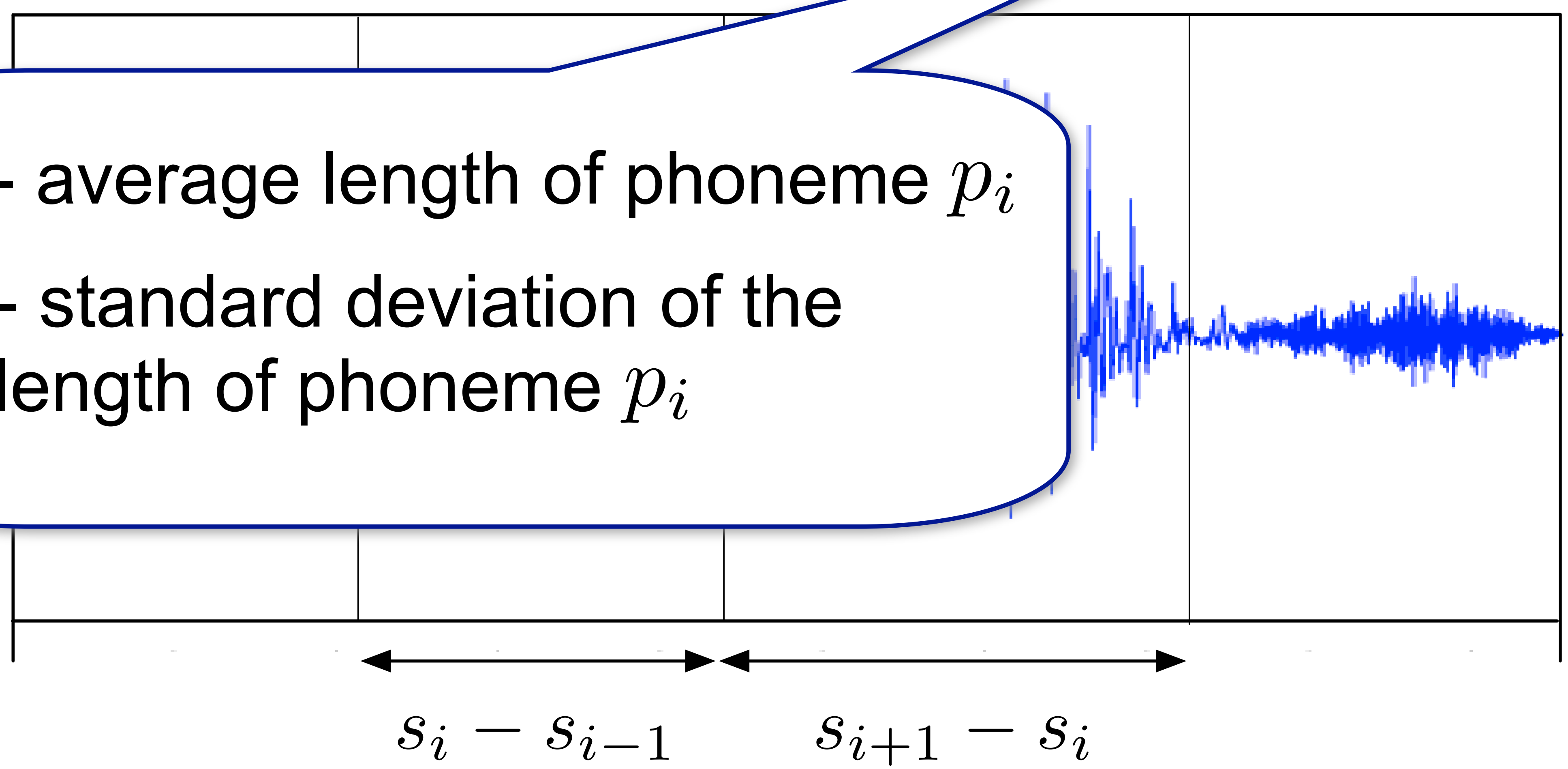
Feature Functions III

Phoneme duration model

$$\phi_6(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \log\ \mathcal{N}(s_{i+1} - s_i; \hat{\mu}_{p_i}, \hat{\sigma}_{p_i})$$

$\hat{\mu}_{p_i}$ - average length of phoneme p_i

$\hat{\sigma}_{p_i}$ - standard deviation of the length of phoneme p_i

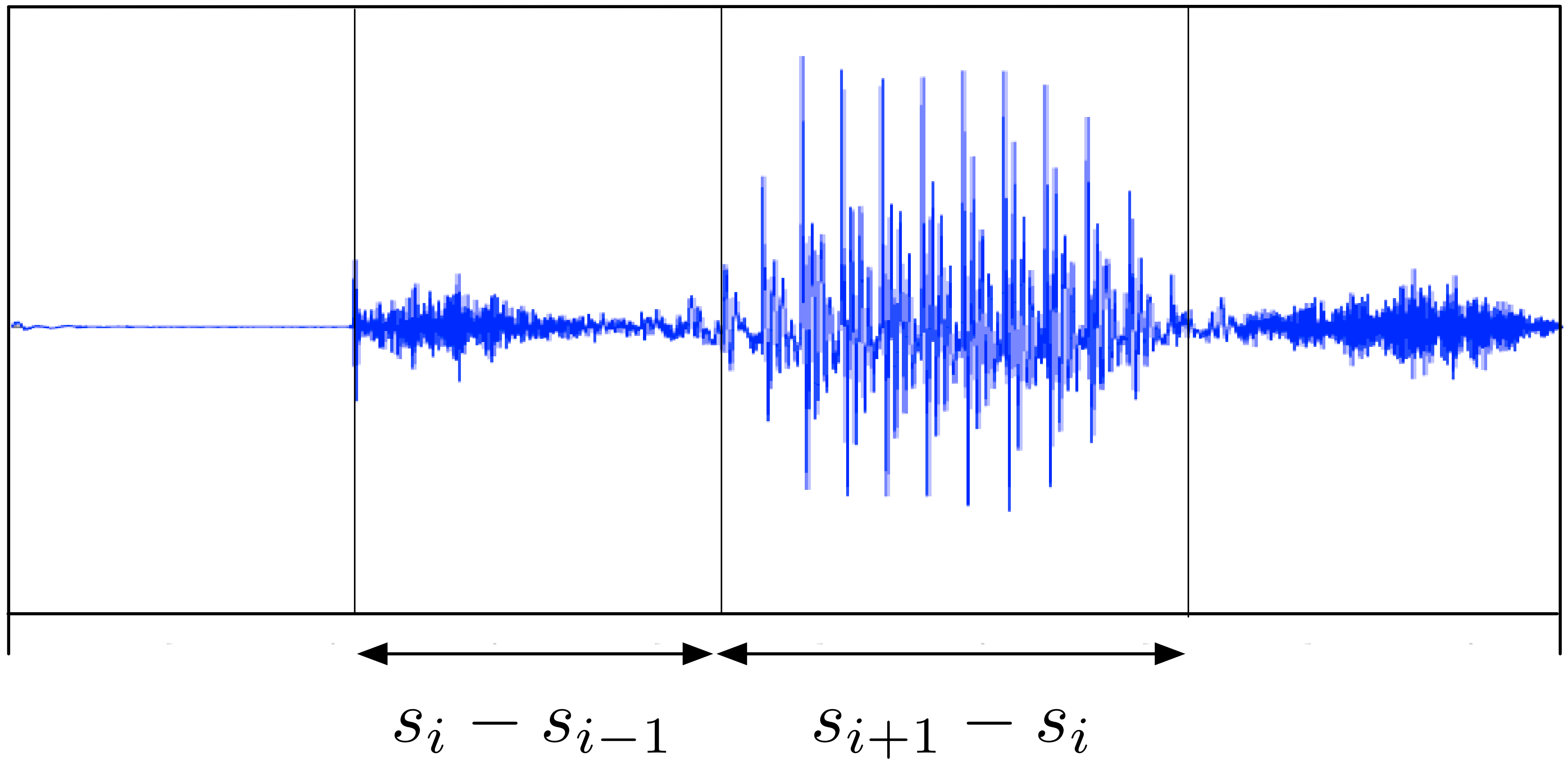


Feature Functions III

Phoneme duration model

Statistics of phoneme p_i

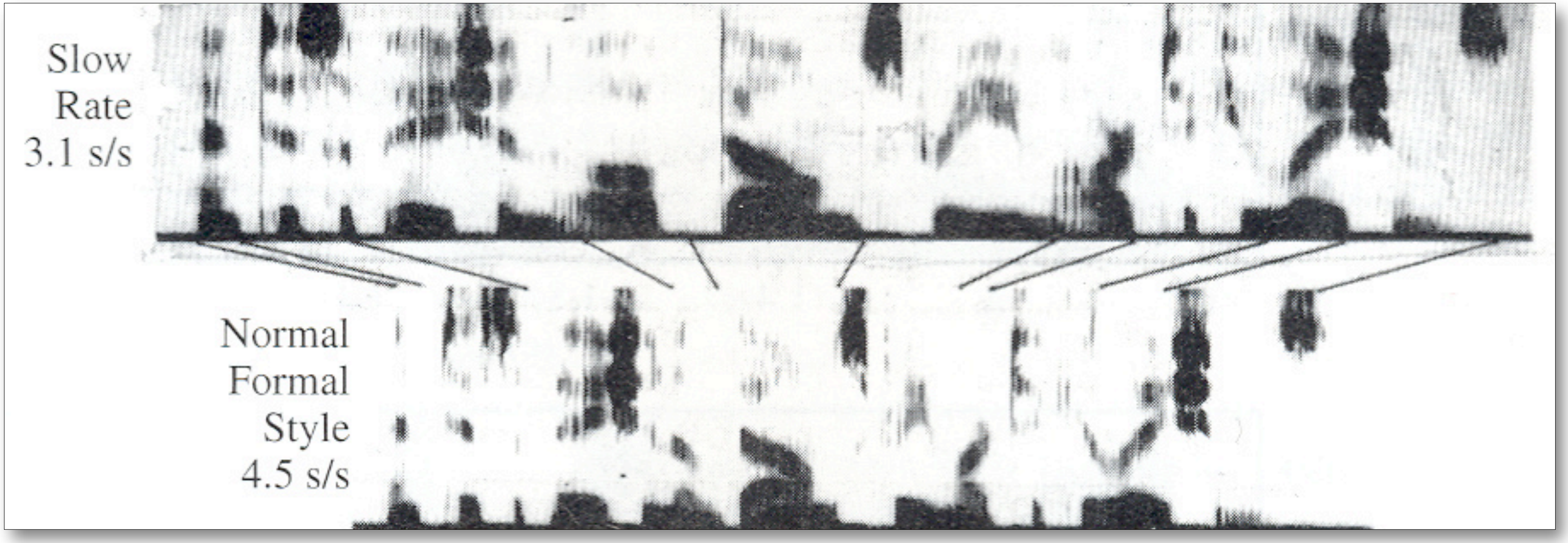
$$\phi_6(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = \sum_{i=1}^{|\bar{p}|} \log \mathcal{N}(s_{i+1} - s_i; \hat{\mu}_{p_i}, \hat{\sigma}_{p_i})$$



Feature Functions IV

Speaking-rate modeling ("dynamics")

$$\phi_7(\bar{\mathbf{x}}, \bar{p}, \bar{s}) = -\sum_{i=2}^{|\bar{p}|-1} \left(\frac{s_{i+1} - s_i}{\hat{\mu}_{p_i}} - \frac{s_i - s_{i-1}}{\hat{\mu}_{p_{i-1}}} \right)^2$$



Spectogram at different rates of articulation (after Pickett, 1980)

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\downarrow$$

$$\downarrow$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\downarrow$$

$$f(\bar{\mathbf{x}}, \bar{p}) = \max_{\bar{s}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, \bar{p}, \bar{s})$$

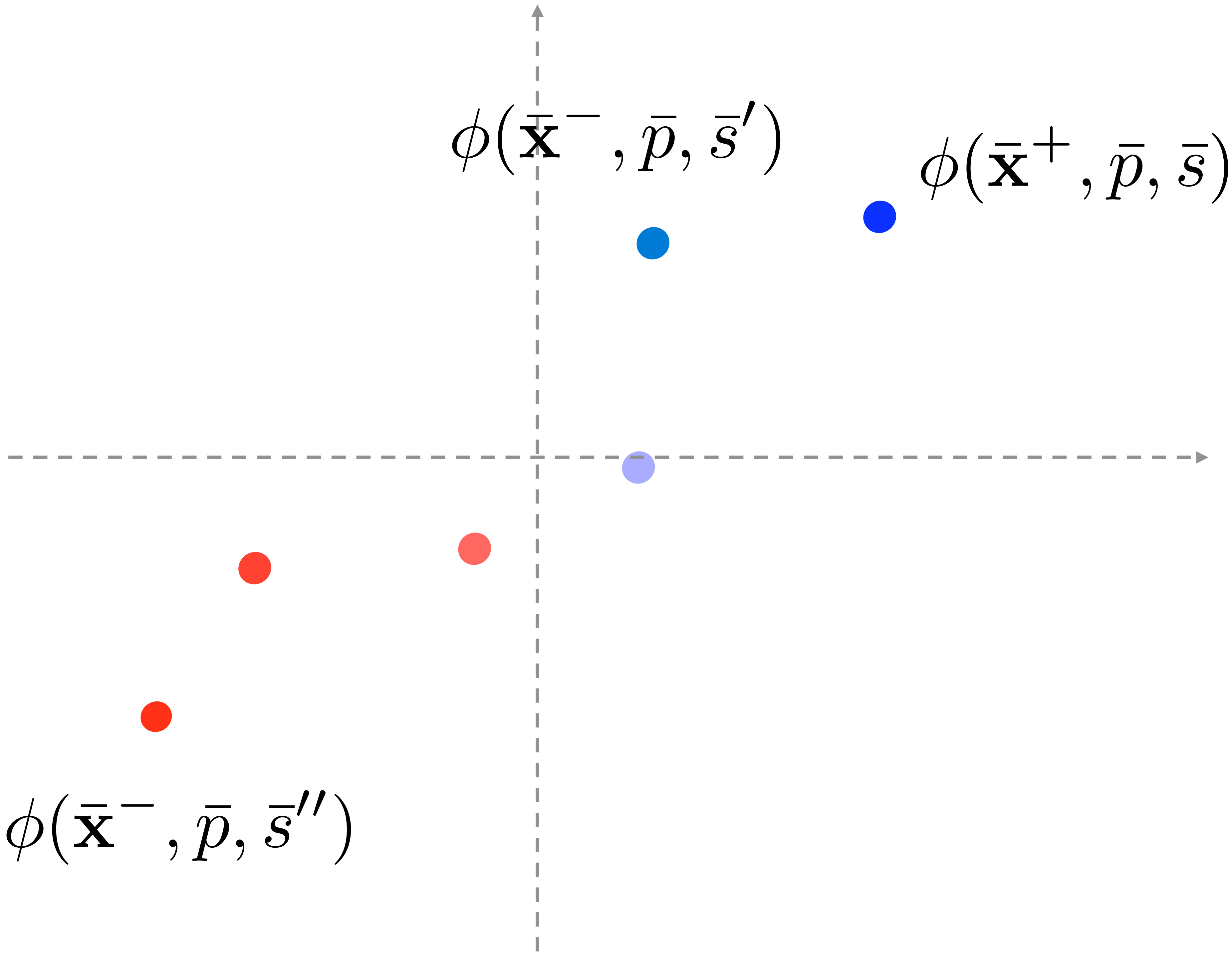
$$\mathbf{w} \in \mathbb{R}^n$$

Discriminative Keyword Spotting

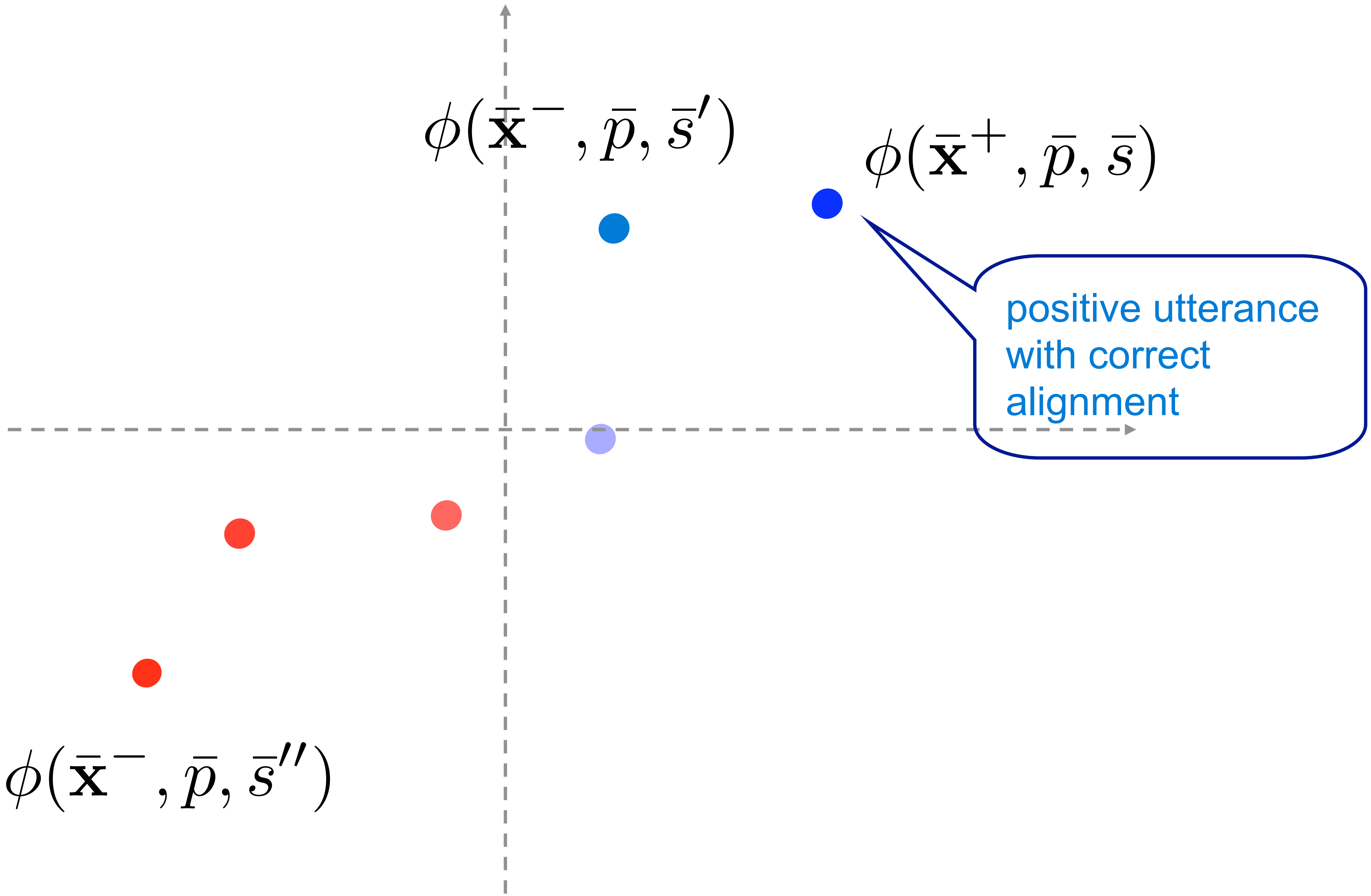
$$\downarrow$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

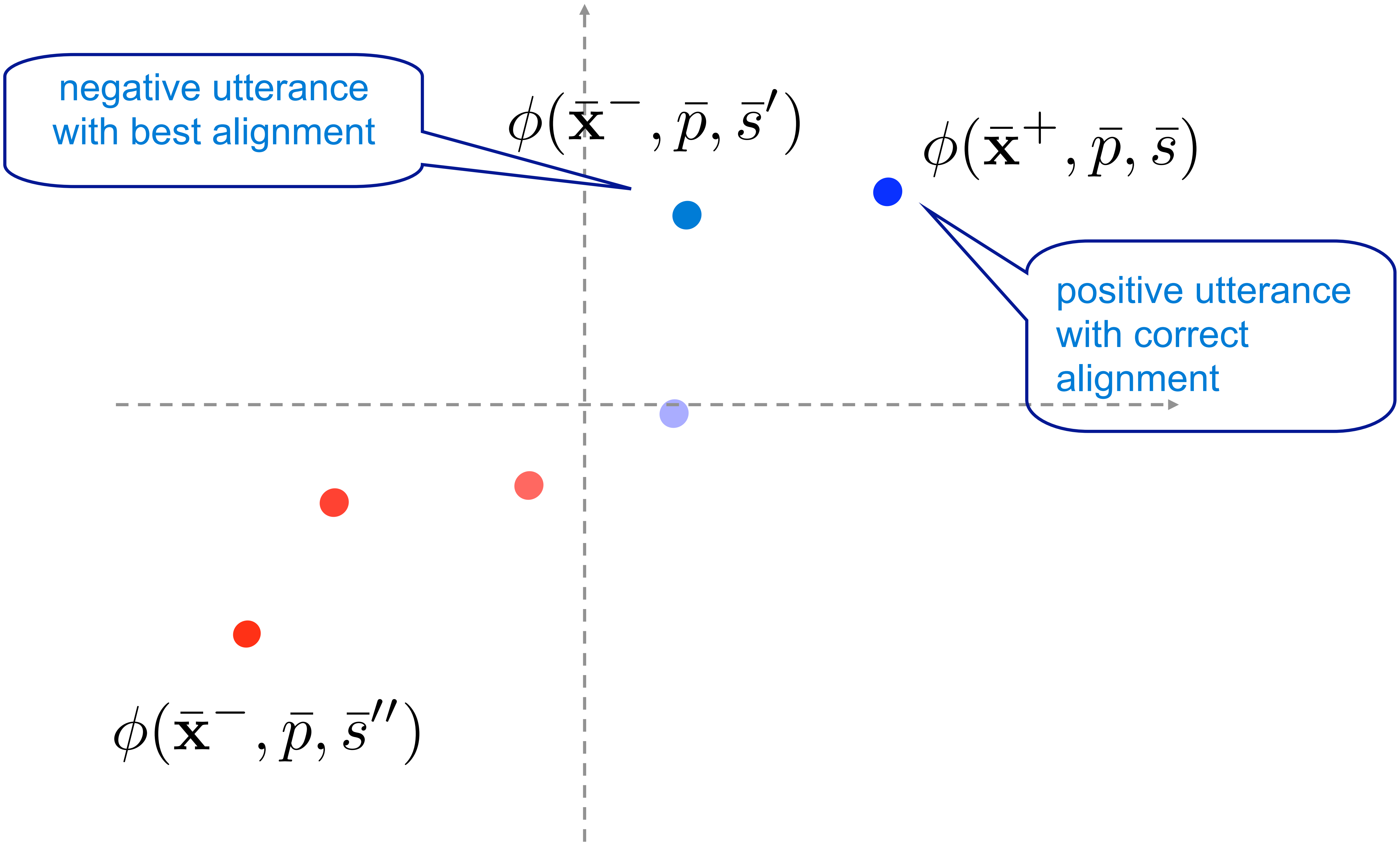
Large-Margin Model



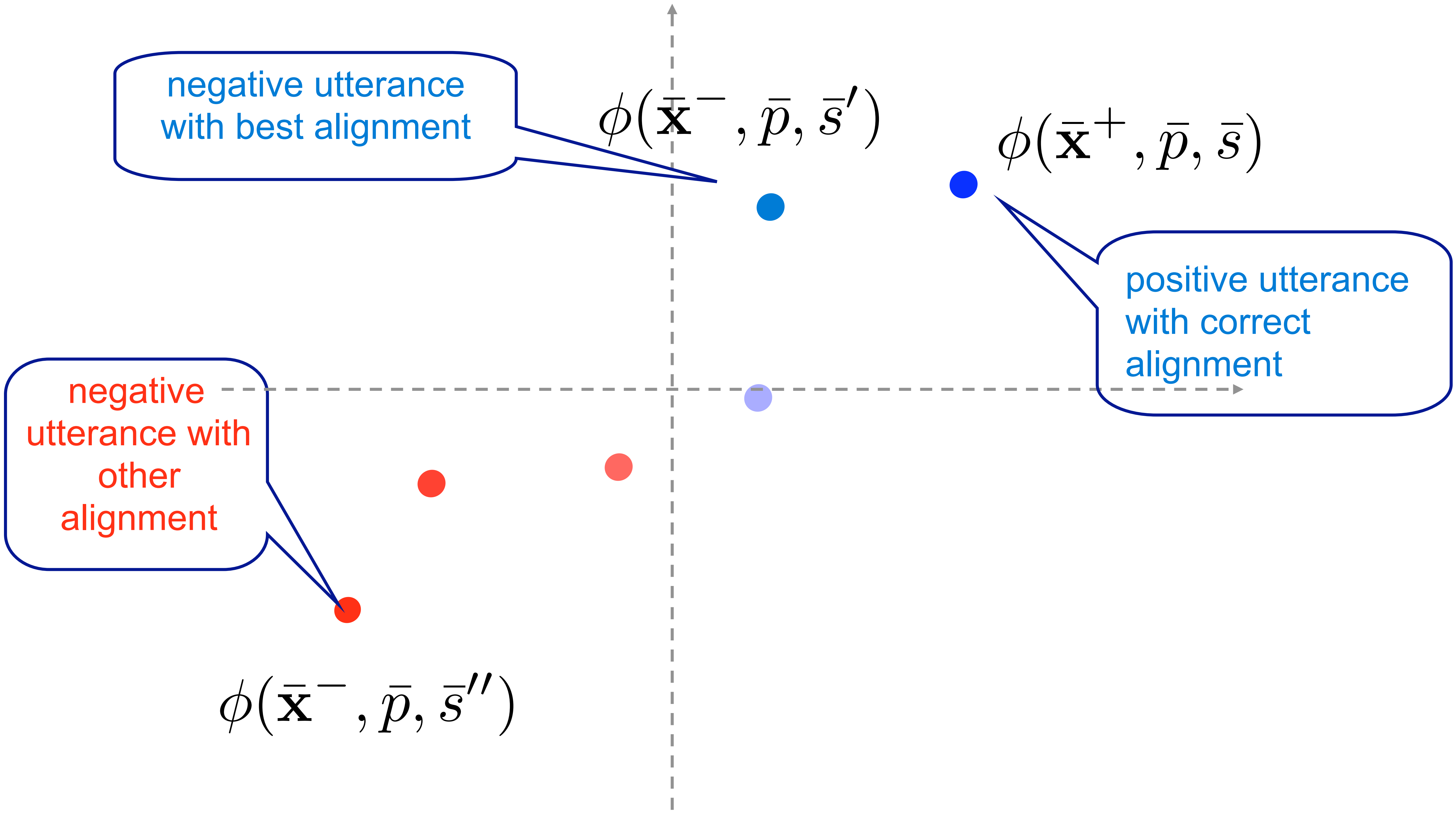
Large-Margin Model



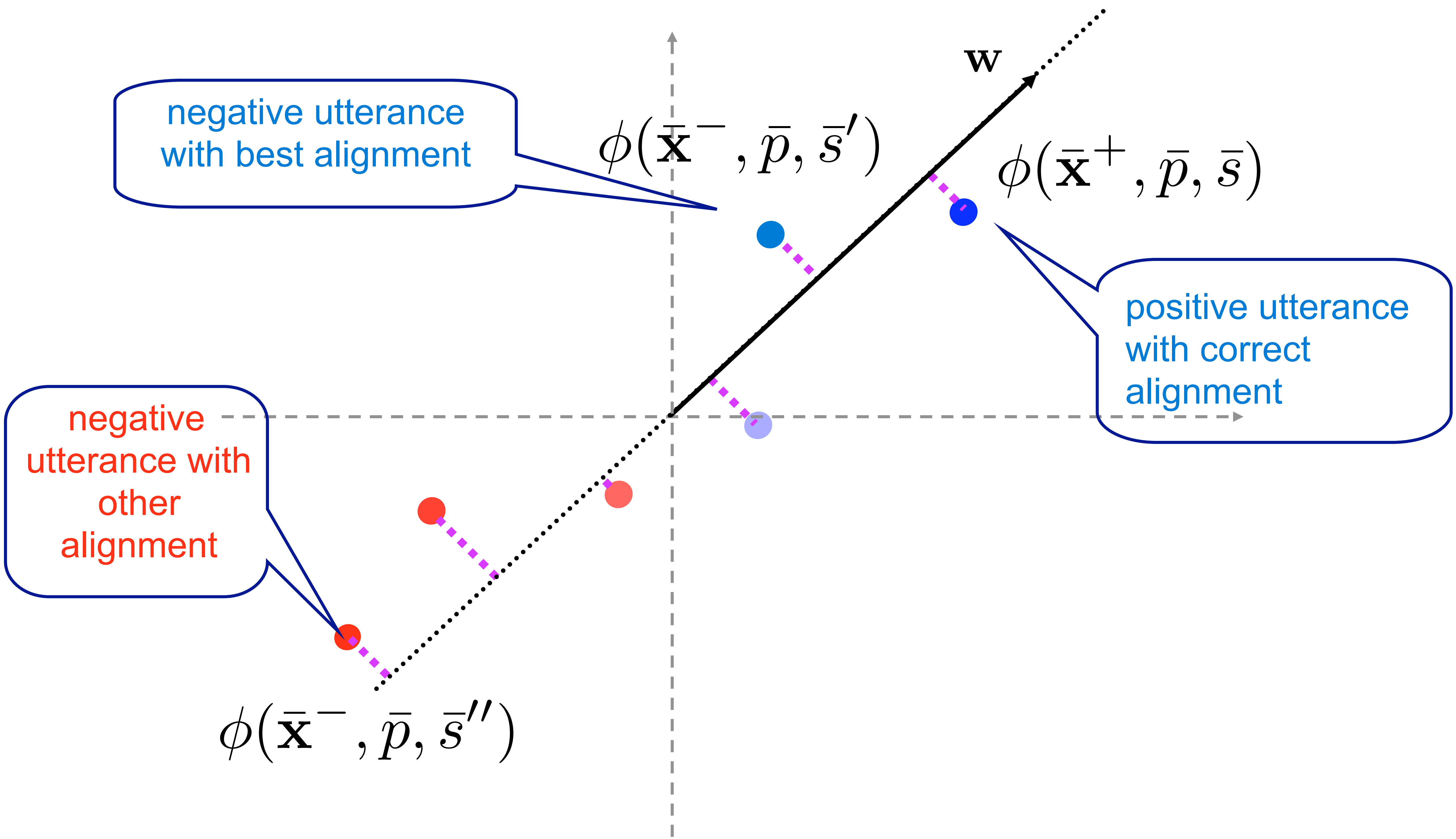
Large-Margin Model



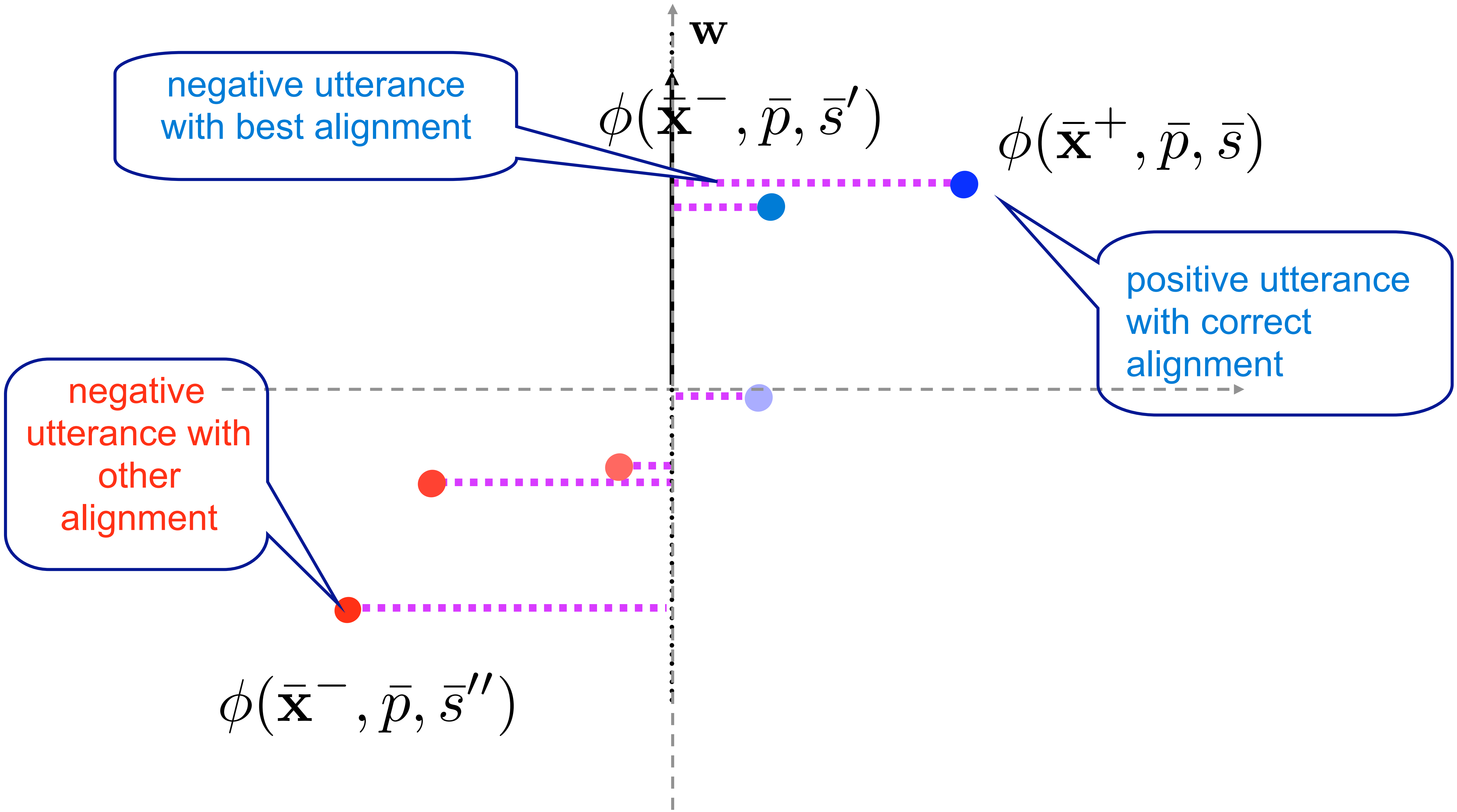
Large-Margin Model



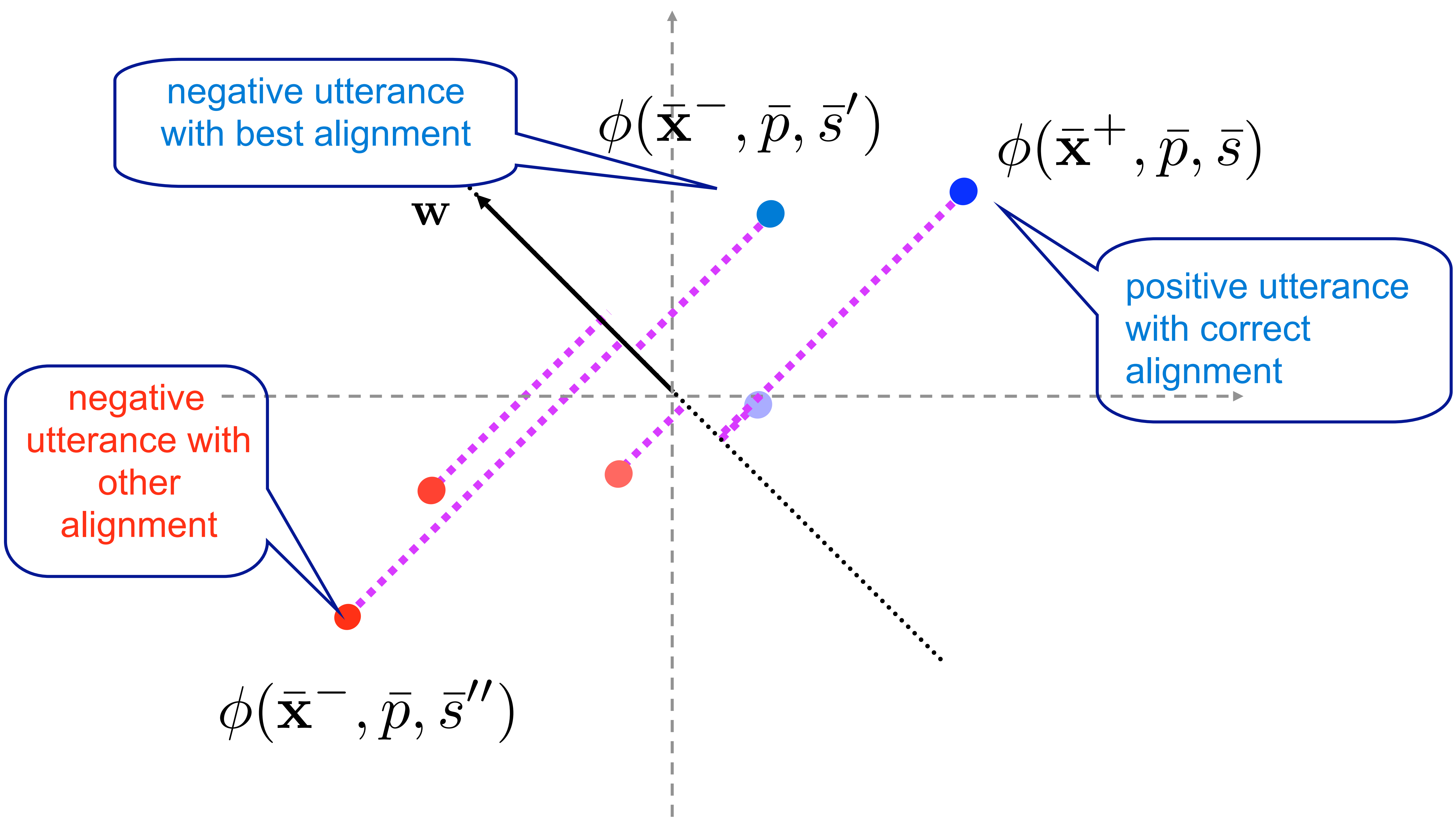
Large-Margin Model



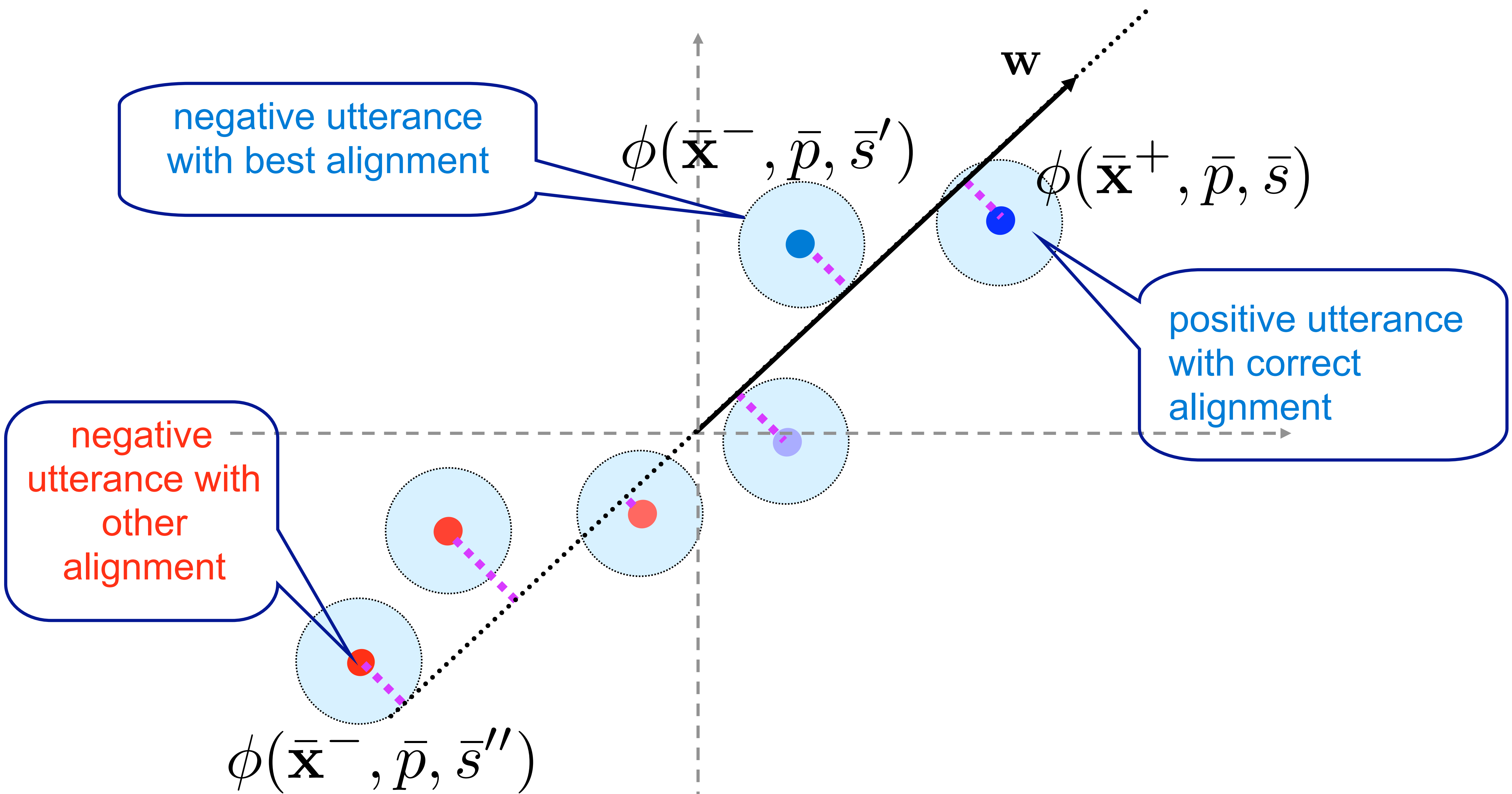
Large-Margin Model



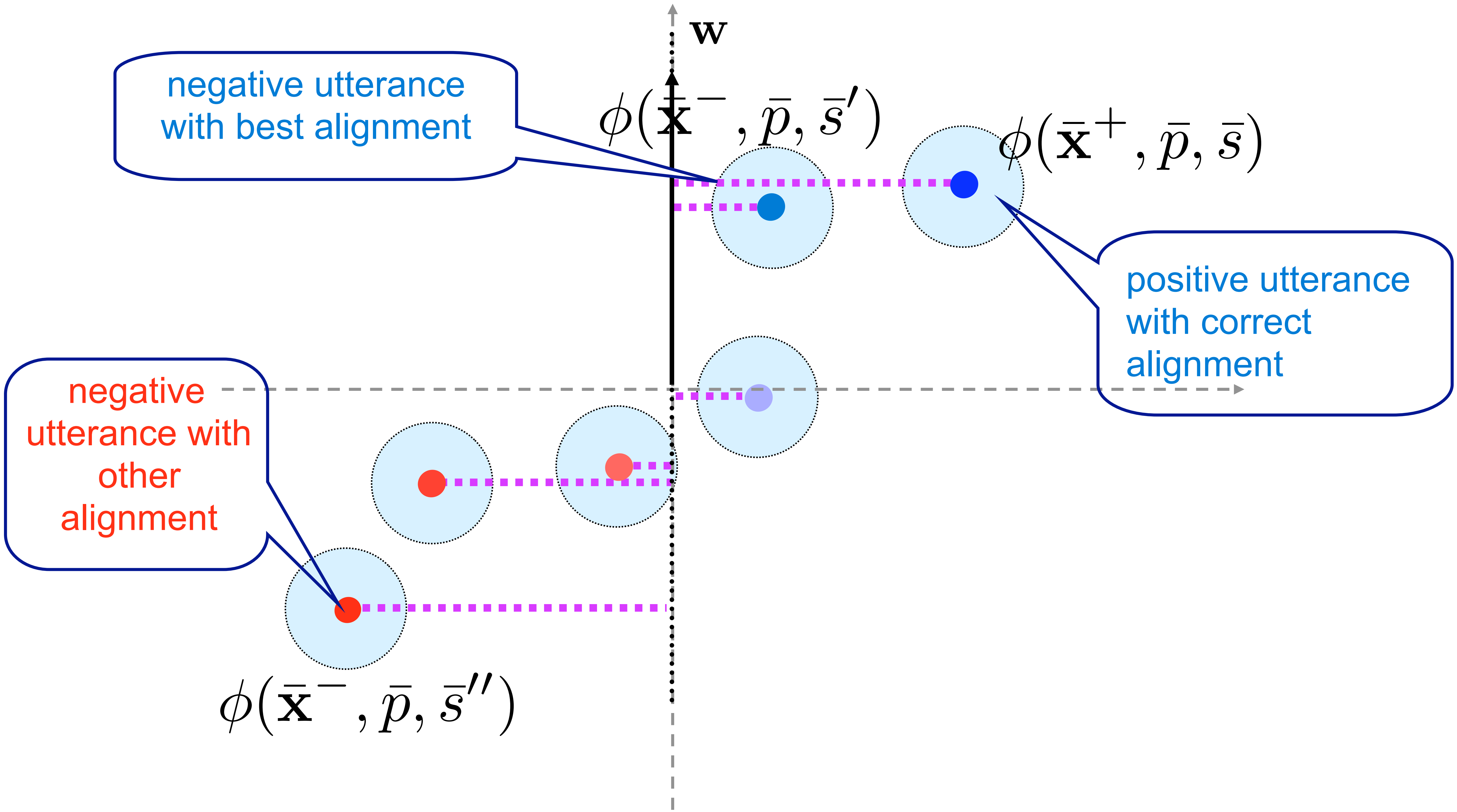
Large-Margin Model



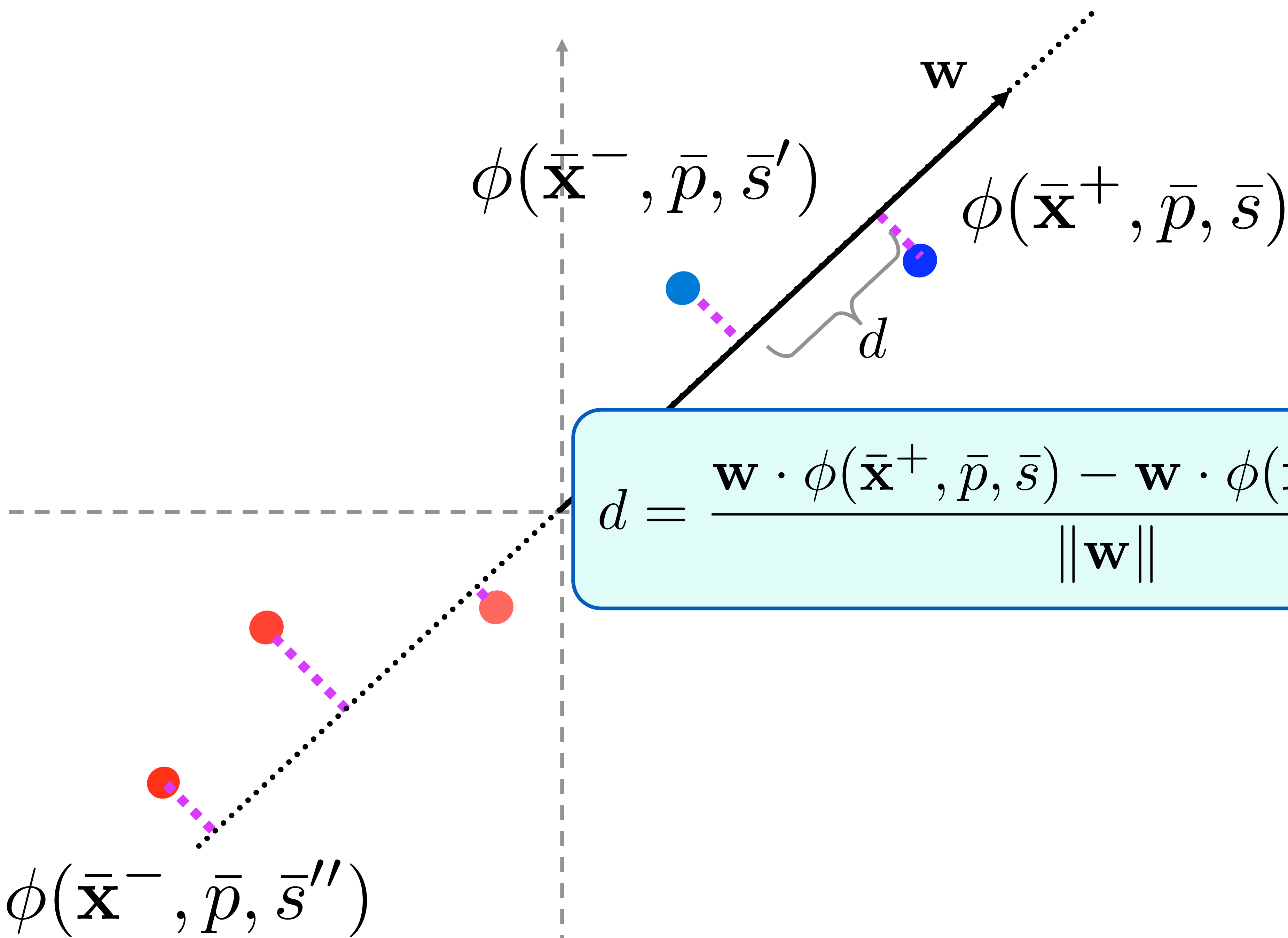
Large-Margin and Noise



Large-Margin and Noise



Large-Margin Derivation



$$d = \frac{\mathbf{w} \cdot \phi(\bar{\mathbf{x}}^+, \bar{p}, \bar{s}) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}^-, \bar{p}, \bar{s}')}{\|\mathbf{w}\|}$$

$$\mathbf{w} \cdot \phi(\bar{\mathbf{x}}^+, \bar{p}, \bar{s}) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}^-, \bar{p}, \bar{s}') \geq 1 \quad \forall \bar{s}'$$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\downarrow$$

$$f(\bar{\mathbf{x}}, \bar{p}) = \max_{\bar{s}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, \bar{p}, \bar{s})$$

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

Discriminative Keyword Spotting

$$\downarrow$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$f(\bar{\mathbf{x}}, \bar{p}) = \max_{\bar{s}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, \bar{p}, \bar{s})$$

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

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$f(\bar{\mathbf{x}}, \bar{p}) = \max_{\bar{s}} \mathbf{w} \cdot \boldsymbol{\phi}(\bar{\mathbf{x}}, \bar{p}, \bar{s})$$

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

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\max_{\mathbf{w}} d$$

$$\text{s.t.} \quad \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall j \ \forall \bar{s}'$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

<u>Discriminative</u> learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall j \ \forall \bar{s}'$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall j \ \forall \bar{s}'$$

Exponential number of constraints

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Learning Paradigm

Discriminative learning from examples

$$S = \{(\bar{p}_1, \bar{\mathbf{x}}_1^+, \bar{\mathbf{x}}_1^-, \bar{s}_1), \ldots, (\bar{p}_m, \bar{\mathbf{x}}_m^+, \bar{\mathbf{x}}_m^-, \bar{s}_m)\}$$

$$\begin{aligned} &\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t.} \quad \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall j \ \forall \bar{s}' \end{aligned}$$

Keyword spotter $f(\bar{\mathbf{x}}, \bar{p})$

Iterative Algorithm

Given a training set: $S = \{(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)\}$

Find $\mathbf{w}$

$$\begin{cases} \mathbf{w}^\star = \arg\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{such that} \\ \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \;\; \forall j \;\; \forall \bar{s}' \end{cases}$$

Iterative Algorithm

Given a training set: $S = \{(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)\}$

Find $\mathbf{w}$

$$\begin{cases} \mathbf{w}^\star = \arg\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{such that} \\ \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \;\; \forall j \;\; \forall \bar{s}' \end{cases}$$

Exponential
number of
constraints

Iterative Algorithm

Given a training set: $S = \{(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)\}$

Find $\mathbf{w}$

$$\begin{cases} \mathbf{w}^\star = \arg\min \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{such that} \\ \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w}\cdot\phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \;\; \forall j \;\; \forall \bar{s}' \end{cases}$$

Iterative Algorithm

Denote current suggestion by $\mathbf{w}_{j-1}$

Process one example $(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)$ at a time

$$\begin{cases} \mathbf{w}_j = \arg\min \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{j-1}\|^2 \quad \text{such that} \\ \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall \bar{s}' \end{cases}$$

Iterative Algorithm

Denote current suggestion by $\mathbf{w}_{j-1}$

Process one example $(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)$ at a time

$$\begin{cases} \mathbf{w}_j = \arg\min \frac{1}{2} \|\mathbf{w} - \mathbf{w}_{j-1}\|^2 \ \ \text{such that} \\ \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall \bar{s}' \end{cases}$$

Exponential number of constraints

Iterative Algorithm

Denote current suggestion by $\mathbf{w}_{j-1}$

Process one example $(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)$ at a time

$$\begin{cases} \mathbf{w}_j = \arg\min \frac{1}{2} \|\mathbf{w} - \mathbf{w}_{j-1}\|^2 \;\; \text{such that} \\ \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \quad \forall \bar{s}' \end{cases}$$

Iterative Algorithm

Approximation: Replace exponentially many constraints with a single (most violated) constraint.

Define: $\bar{s}' = \arg\max_{\bar{s}} \mathbf{w}_{j-1} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s})$

$$\begin{cases} \mathbf{w}_j = \arg\min \dfrac{1}{2} \|\mathbf{w} - \mathbf{w}_{j-1}\|^2 \ \text{ such that} \\ \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \end{cases}$$

Iterative Algorithm

Approximation: Replace exponentially many constraints with a single (most violated) constraint.

Define: $\bar{s}' = \arg\max_{\bar{s}} \mathbf{w}_{j-1} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s})$

$$\begin{cases} \mathbf{w}_j = \arg\min \frac{1}{2} \|\mathbf{w} - \mathbf{w}_{j-1}\|^2 \quad \text{such that} \\ \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}') \geq 1 \end{cases}$$

$$\mathbf{w}_j = \mathbf{w}_{j-1} + \frac{1 - \mathbf{w}_{j-1} \Delta\boldsymbol{\phi}}{\|\Delta\boldsymbol{\phi}\|^2}$$

$$\Delta\boldsymbol{\phi} = \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \mathbf{w} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}')$$

Iterative Algorithm

Input: training set $S = \{(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)\}$

Initialize: $\mathbf{w}_0 = 0$

For each example $(\bar{p}_j, \bar{\mathbf{x}}_j^+, \bar{\mathbf{x}}_j^-, \bar{s}_j)$

Predict: $\bar{s}' = \arg\max_{\bar{s}} \mathbf{w}_{j-1} \cdot \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s})$

Set: $\Delta\phi = \phi(\bar{\mathbf{x}}_j^+, \bar{p}_j, \bar{s}_j) - \phi(\bar{\mathbf{x}}_j^-, \bar{p}_j, \bar{s}')$

If $\mathbf{w} \cdot \Delta\phi \leq 1$

Update: $\mathbf{w}_j = \mathbf{w}_{j-1} + \dfrac{1 - \mathbf{w}_{j-1}\Delta\phi}{\|\Delta\phi\|^2}$

Output Choose $\mathbf{w}_j$ which attains the lowest cost on a validation set.

Formal Properties

- Convex optimization problem - single minimum
- Worse case analysis: Area Under Curve during the training phase is high

$$1 - \tilde{A} \leq \frac{1}{m}\|\mathbf{w}^\star\|^2 + \frac{2C}{m}\sum_{i=1}^{m}\ell(\mathbf{w}^\star)$$

- The expected Area Under Curve on unseen examples is high in probability

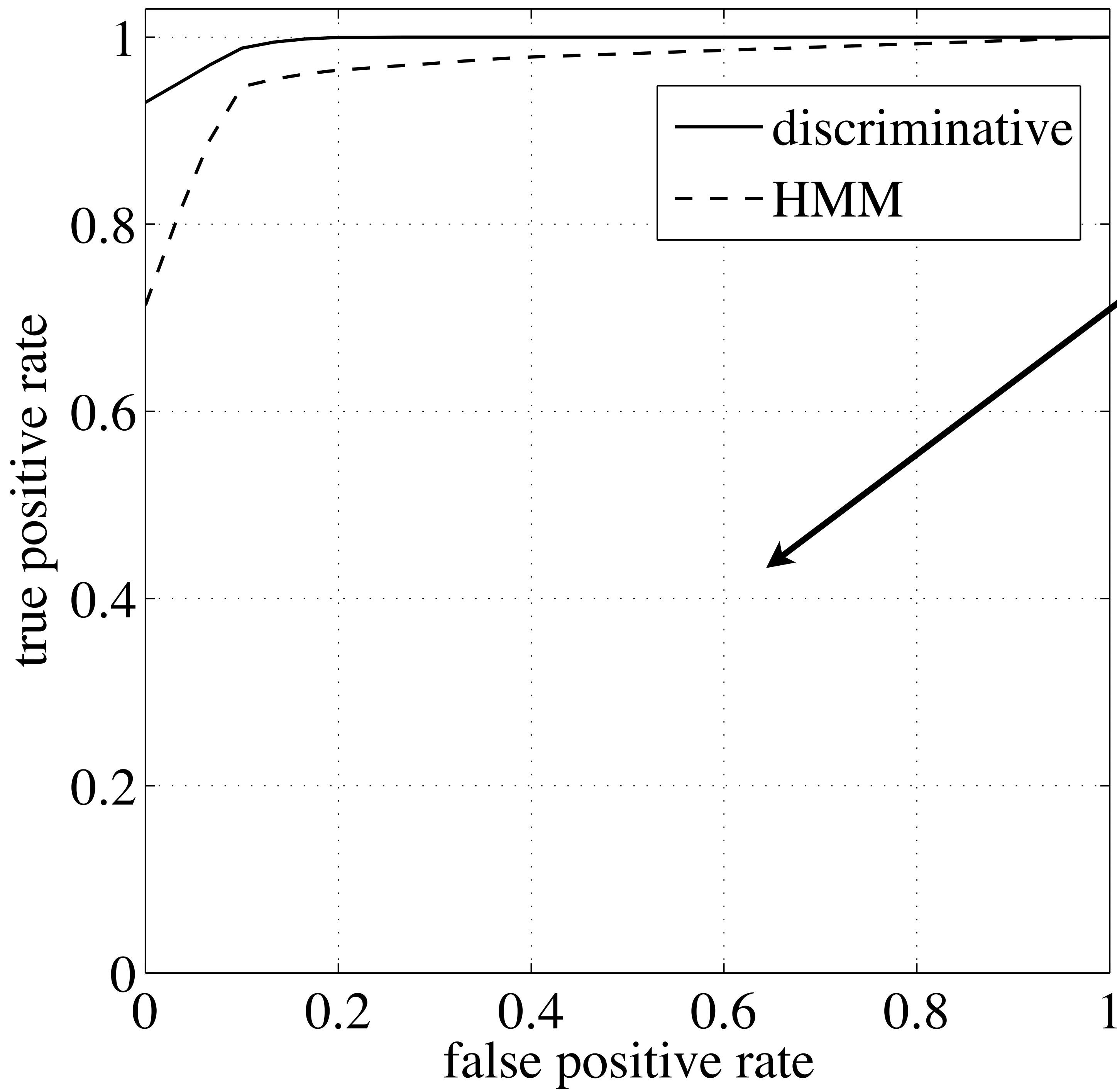
$$1 - A \leq \frac{1}{m}\sum_{i=1}^{m}\ell(\mathbf{w}^\star) + \frac{\|\mathbf{w}^\star\|^2}{m} + \mathcal{O}\left(\ln(m/\delta), \frac{1}{\sqrt{m_{\text{val}}}}\right)$$

Experimental Results

Training Setup

- TIMIT corpus
- Phoneme representation:
  - 39 phonemes (Lee & Hon, 1989)
- Acoustic Representation:
  - MFCC+Δ+$\Delta\Delta$ (ETSI standard)
- TIMIT training set:
  - 500 utterances for training set of the feature functions
  - 3116 utterance used for training set
  - 80 utterances used for validation (40 keywords)

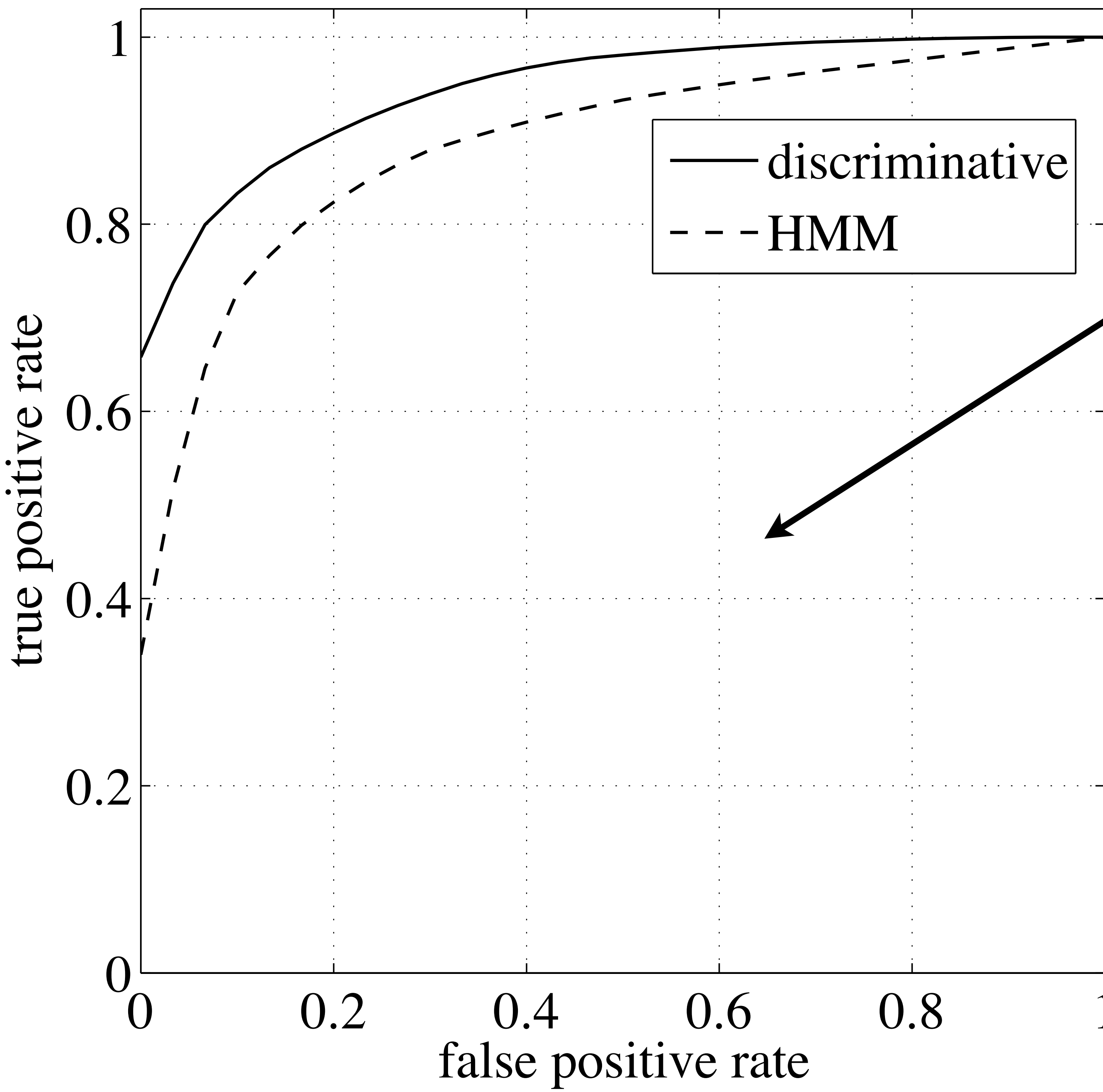
Results on TIMIT



Area under the ROC curve:
0.99 discriminative
0.96 HMM

80 new keywords, and for each, 20 positive and 20 negative utterances

Results on WSJ



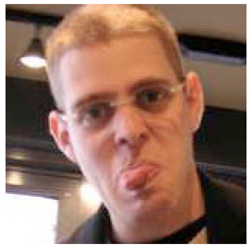
Area under the ROC curve:
0.94 discriminative
0.88 HMM

model trained on TIMIT, same 80 new keywords, and for each, 20 positive and 20 negative utterances from si_tr_s part of WSJ

Practicalities & Algorithms

- The quadratic programming
  - Algorithm for solving the quadratic programming with exponential number of constraints
    [Keshet, Grangier and Bengio, 2006]
- Training the feature function classifiers
  - Hierarchical phoneme classifier
    [Dekel, Keshet and Singer, 2004]
- Non-separable case
  - Common technique in training soft SVM
    [Cristianini & Shawe-Taylor, 2000; Vapnik, 1998]





Thanks!