An Unsupervised Probability Model for Speech-to-Translation Alignment of Low-Resource Languages

Antonios Anastasopoulos[1], David Chiang[1], Long Duong[2]

[1] University of Notre Dame, USA
[2] University of Melbourne, Australia

Motivation

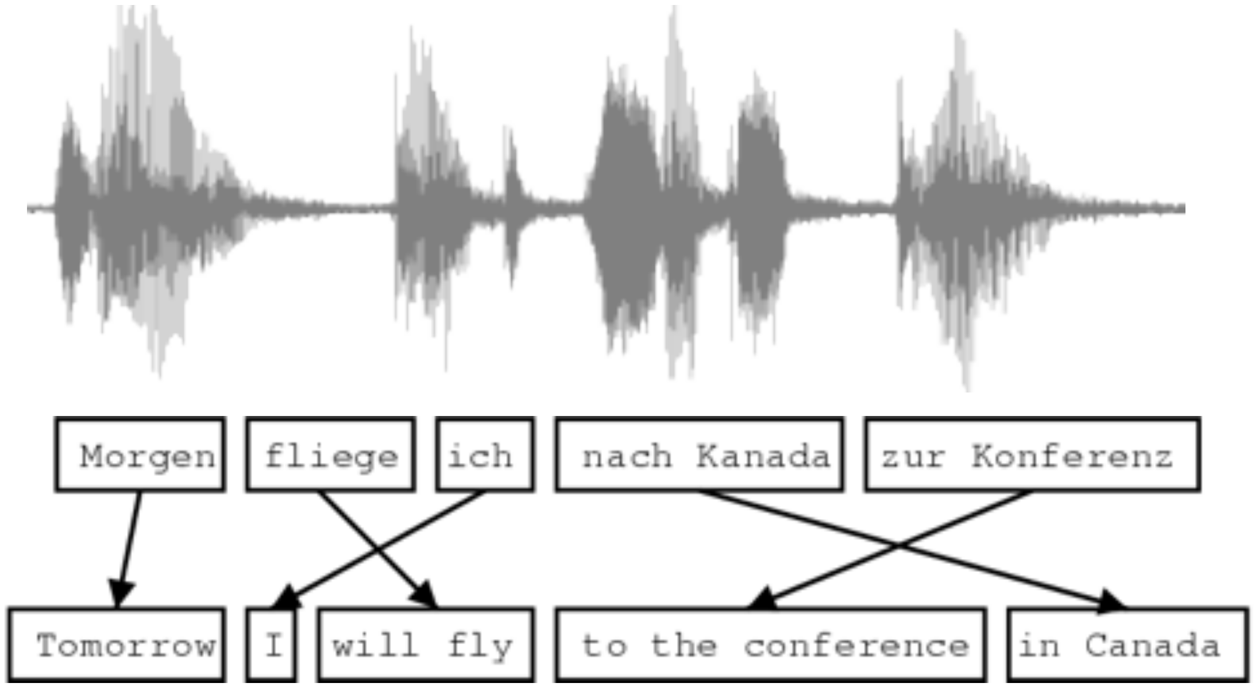
Why Speech-based MT?

90% of languages do not have a writing system

Motivation

Why Speech-based MT?

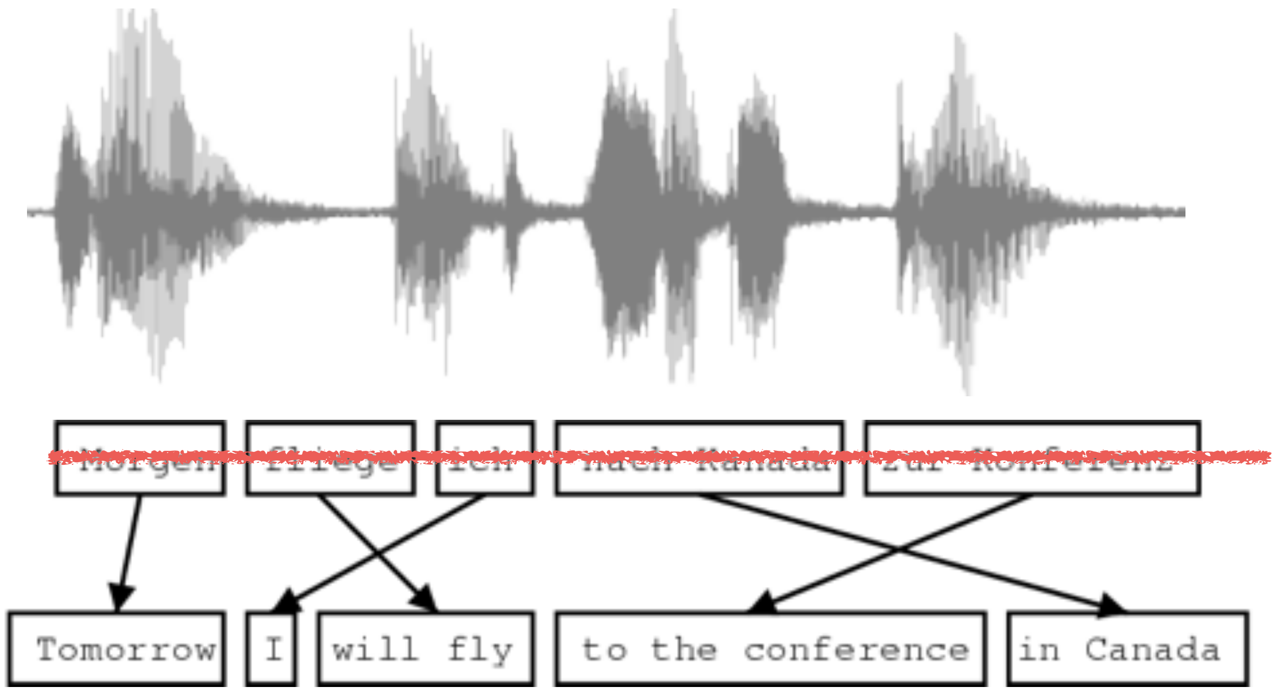
90% of languages do not have a writing system



Motivation

Why Speech-based MT?

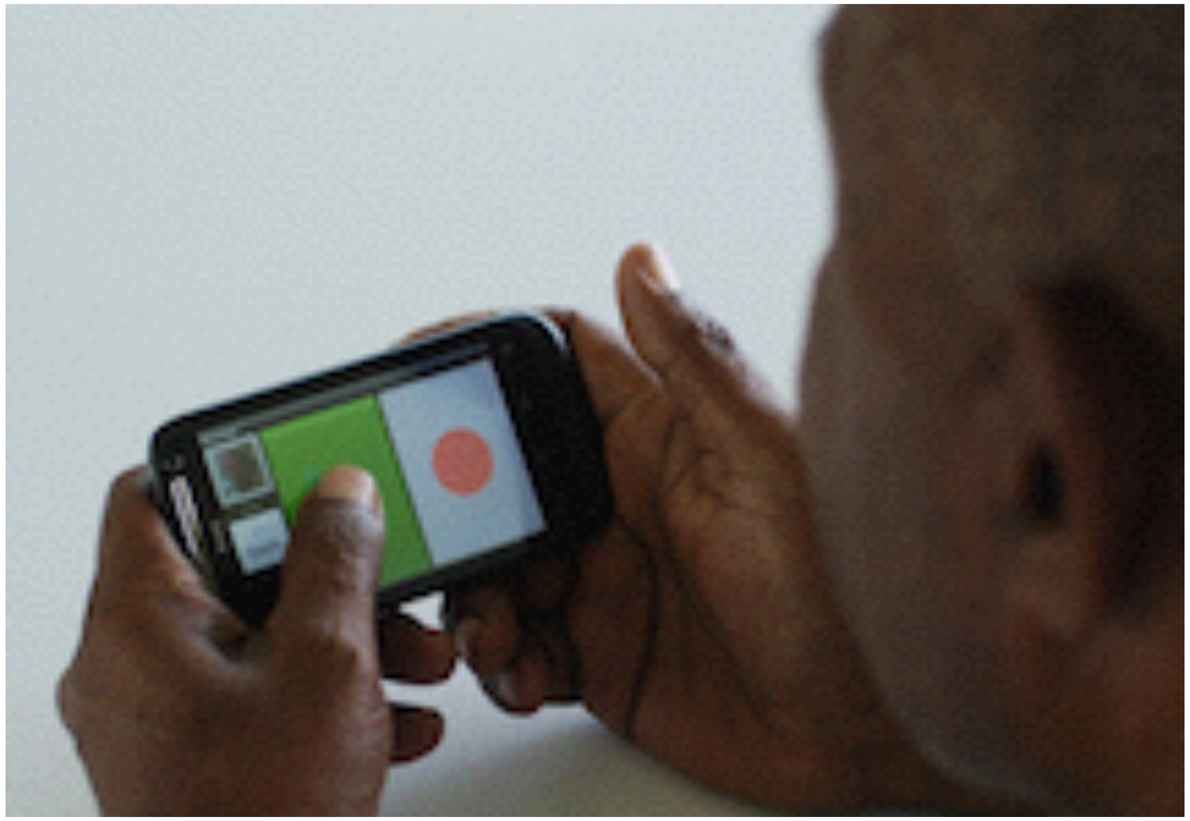
90% of languages do not have a writing system



Motivation

Endangered languages documentation

Use speech with translations



Using the Aikuma (Bird 2010)
app to collect parallel speech

Motivation

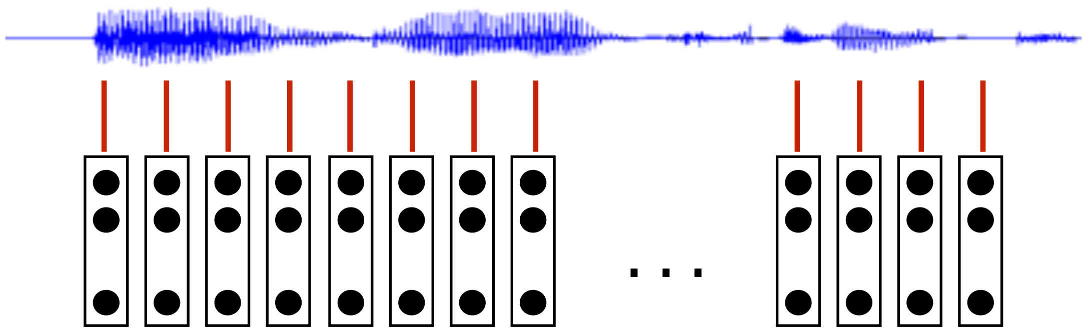
Low-resource languages

Utilize translations rather than transcriptions

Task Description

Task Description

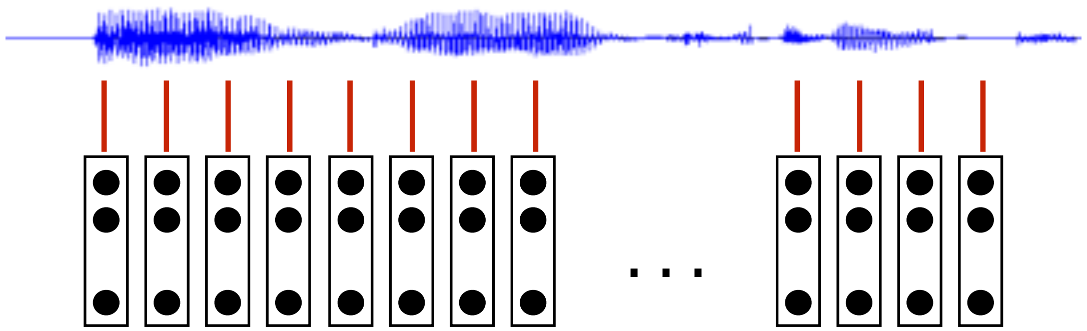
Source side: Frames of the speech signal



Task Description

Source side: Frames of the speech signal

Target side: Translation text



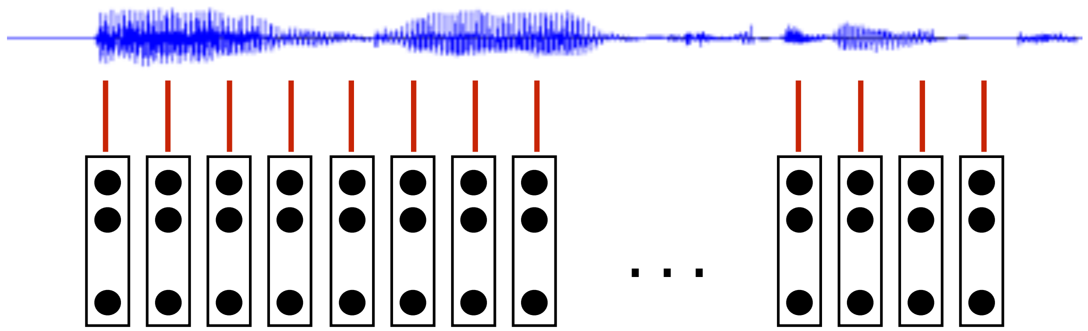
a
little
bit
of
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side



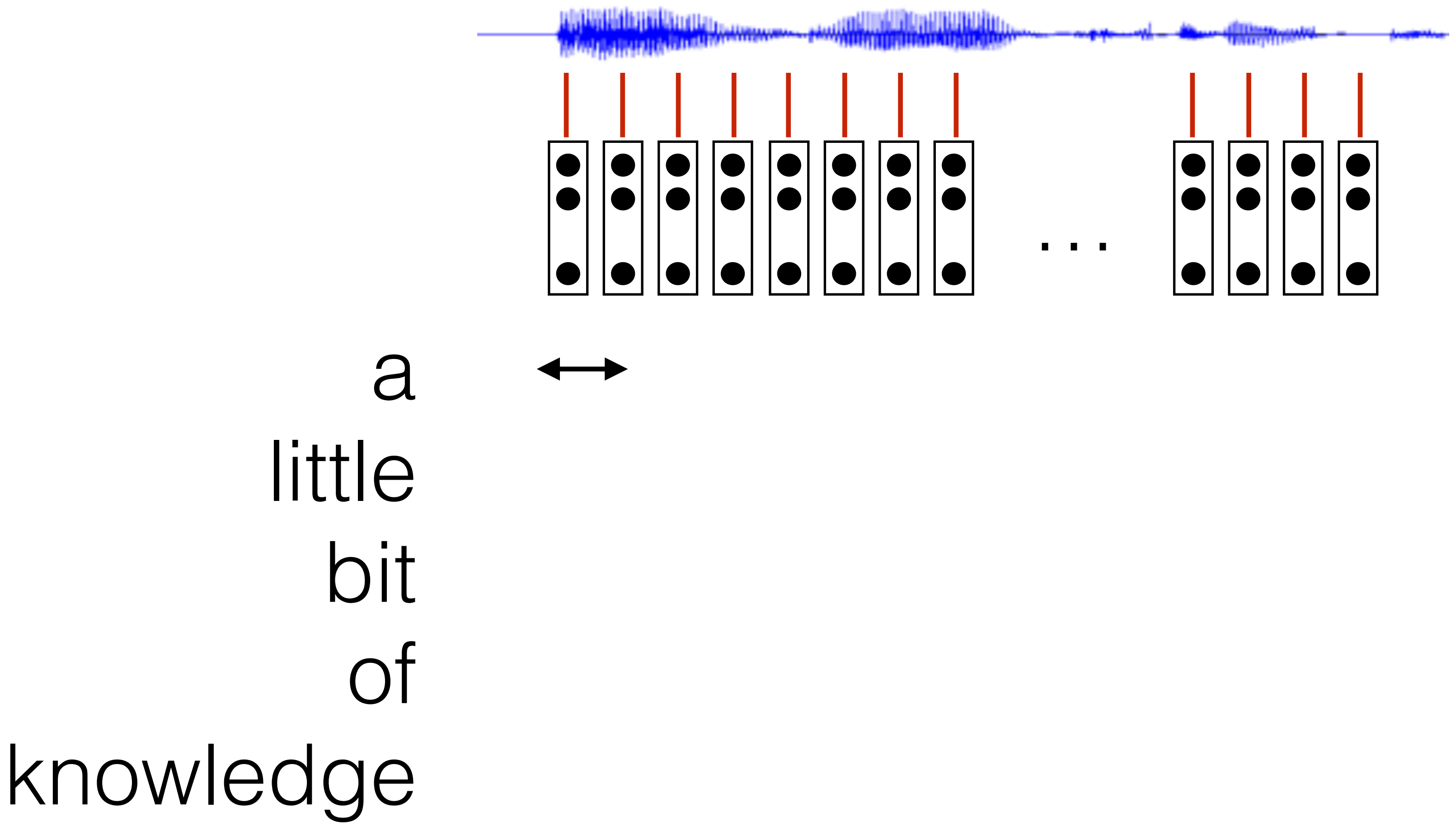
a
little
bit
of
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side

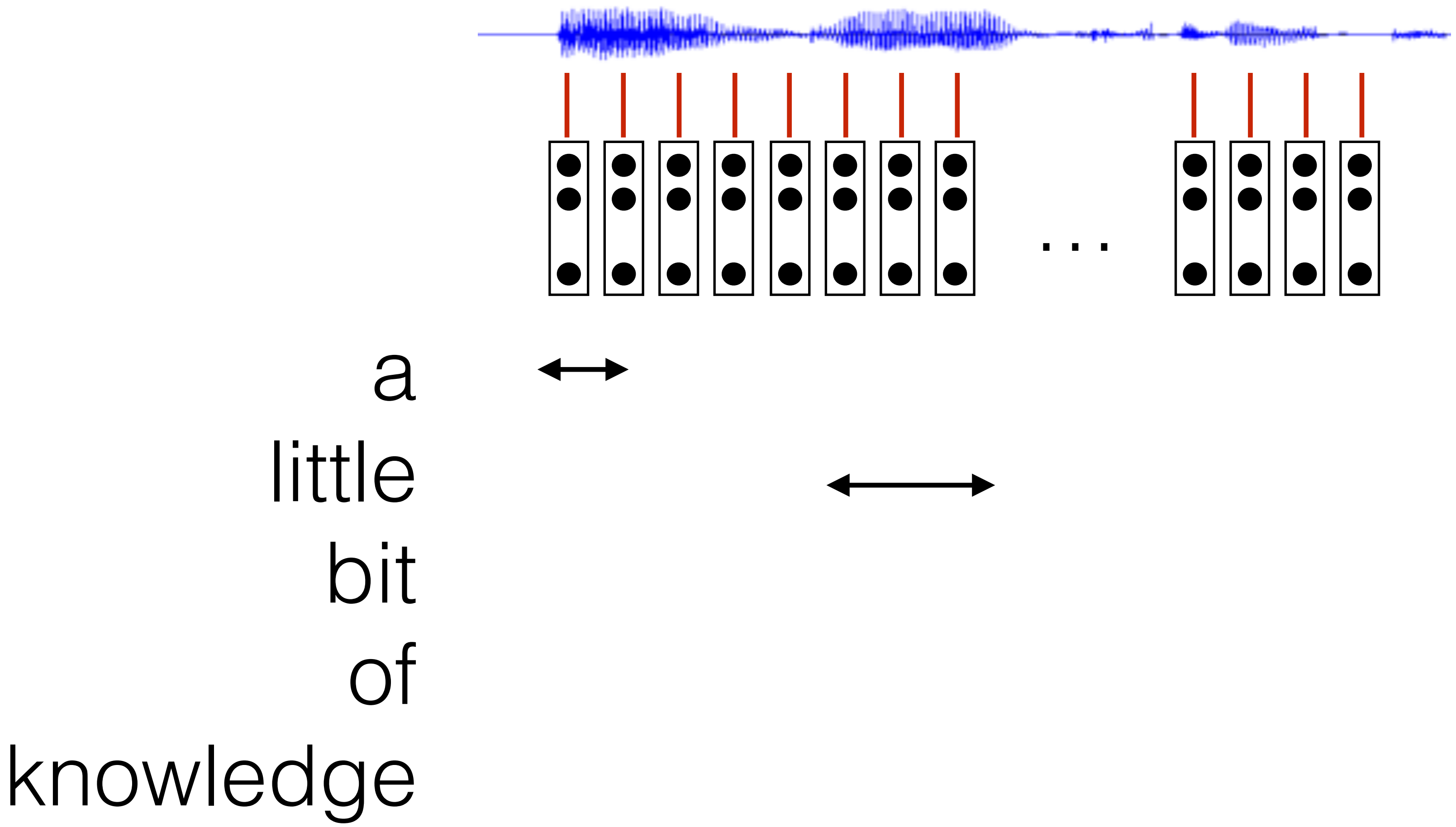


Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side



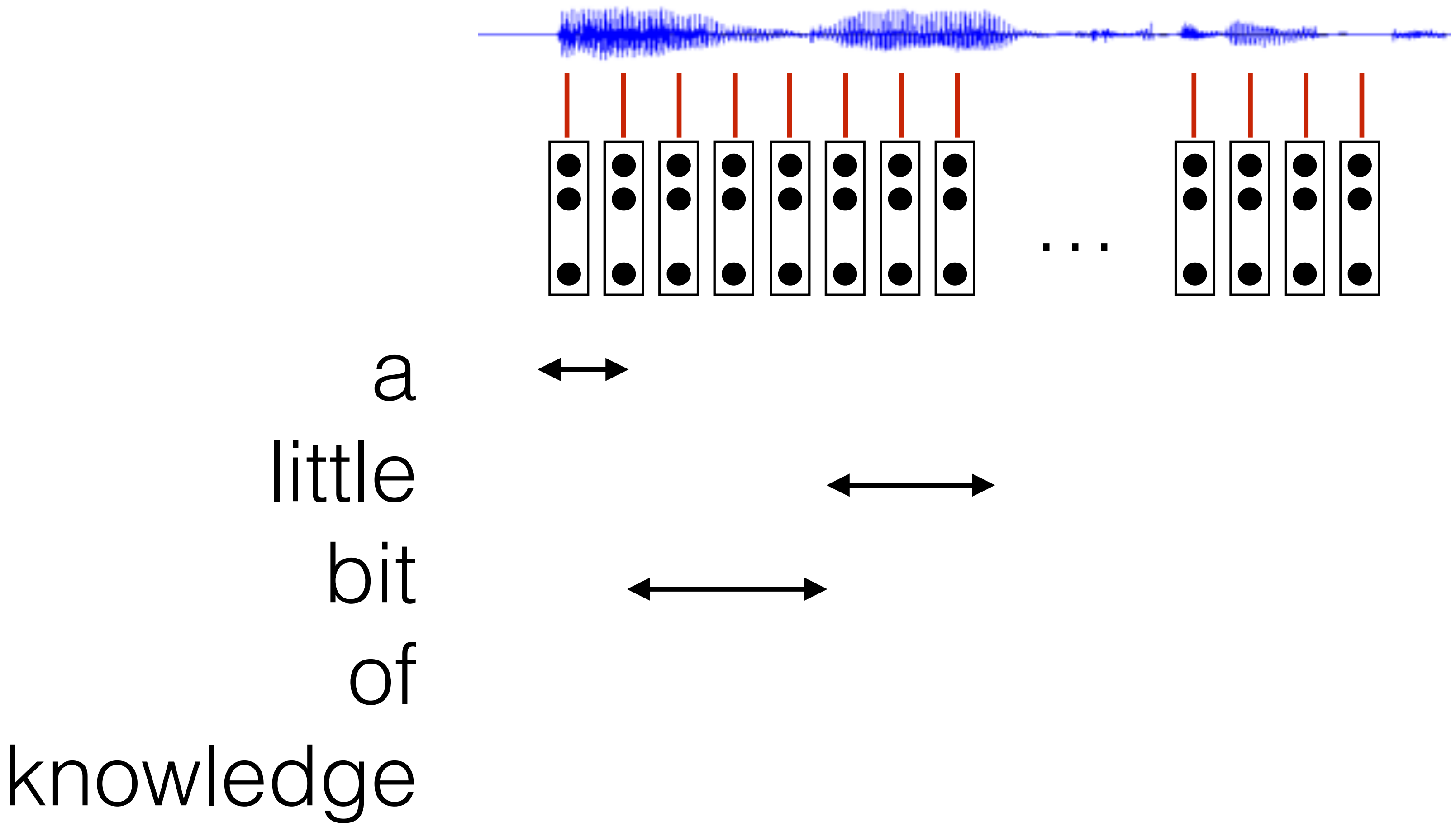
a
little
bit
of
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side



a

little

bit

of

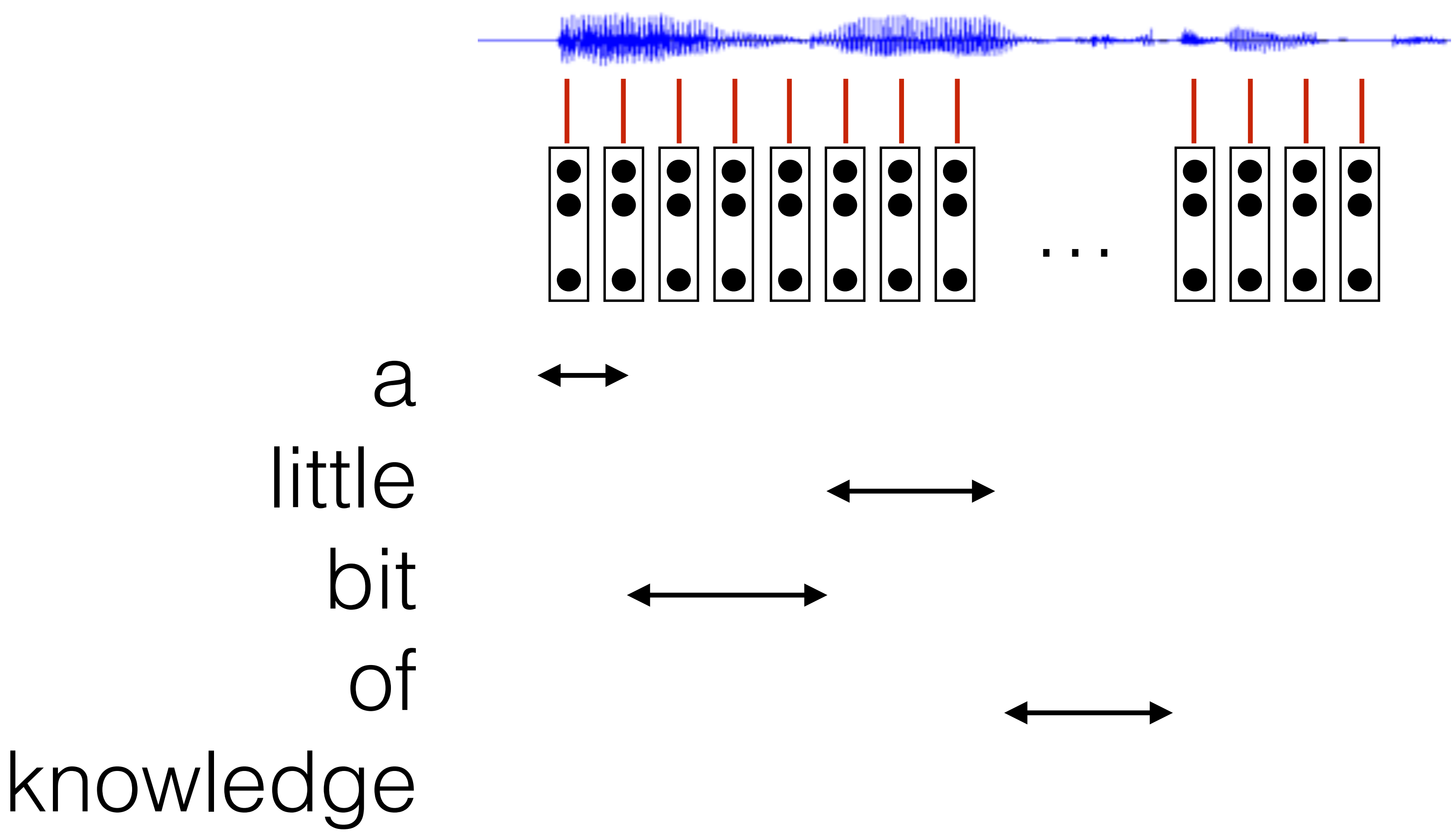
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side



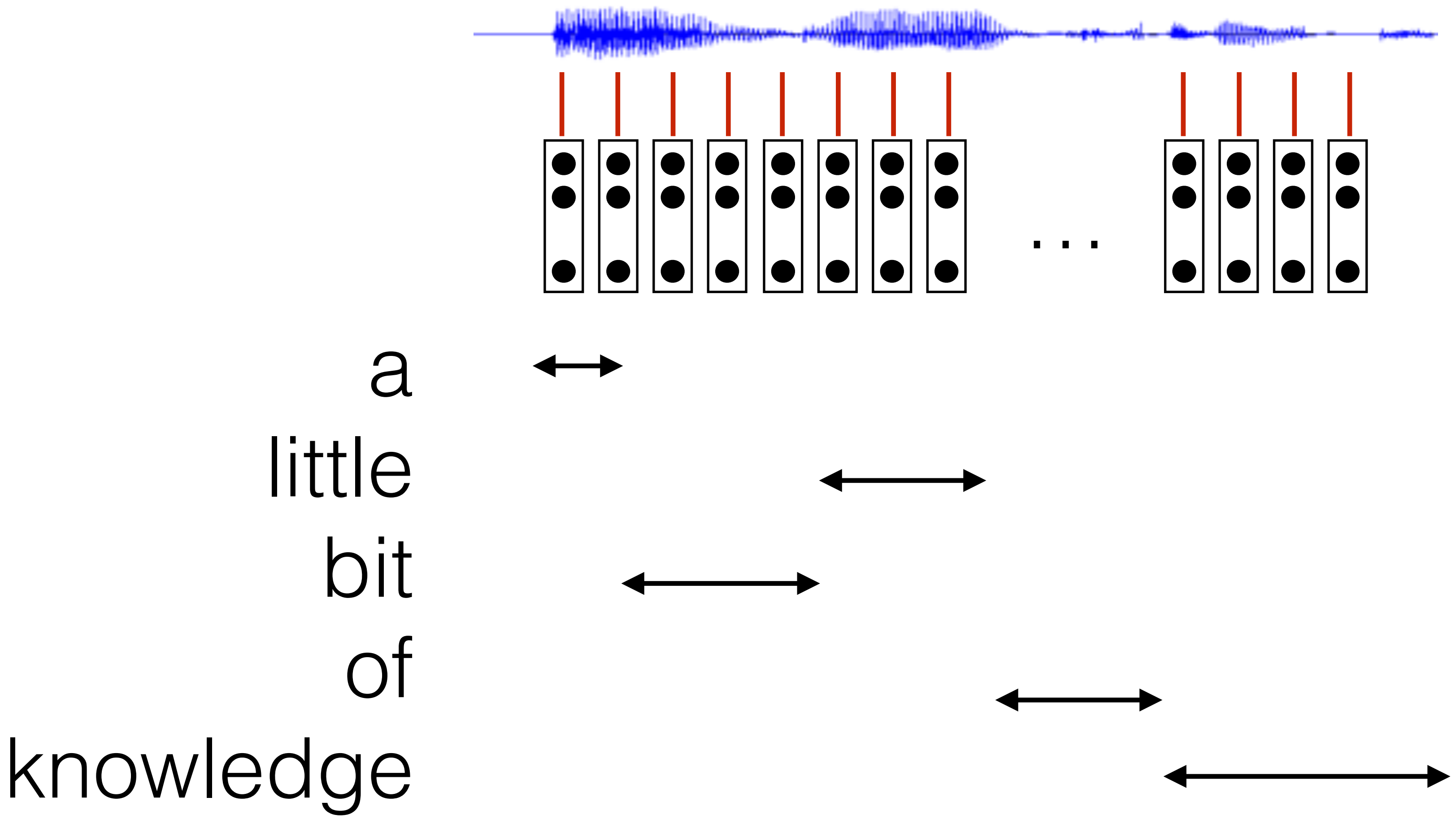
a
little
bit
of
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

Task: find best alignment between source and target side



a

little

bit

of

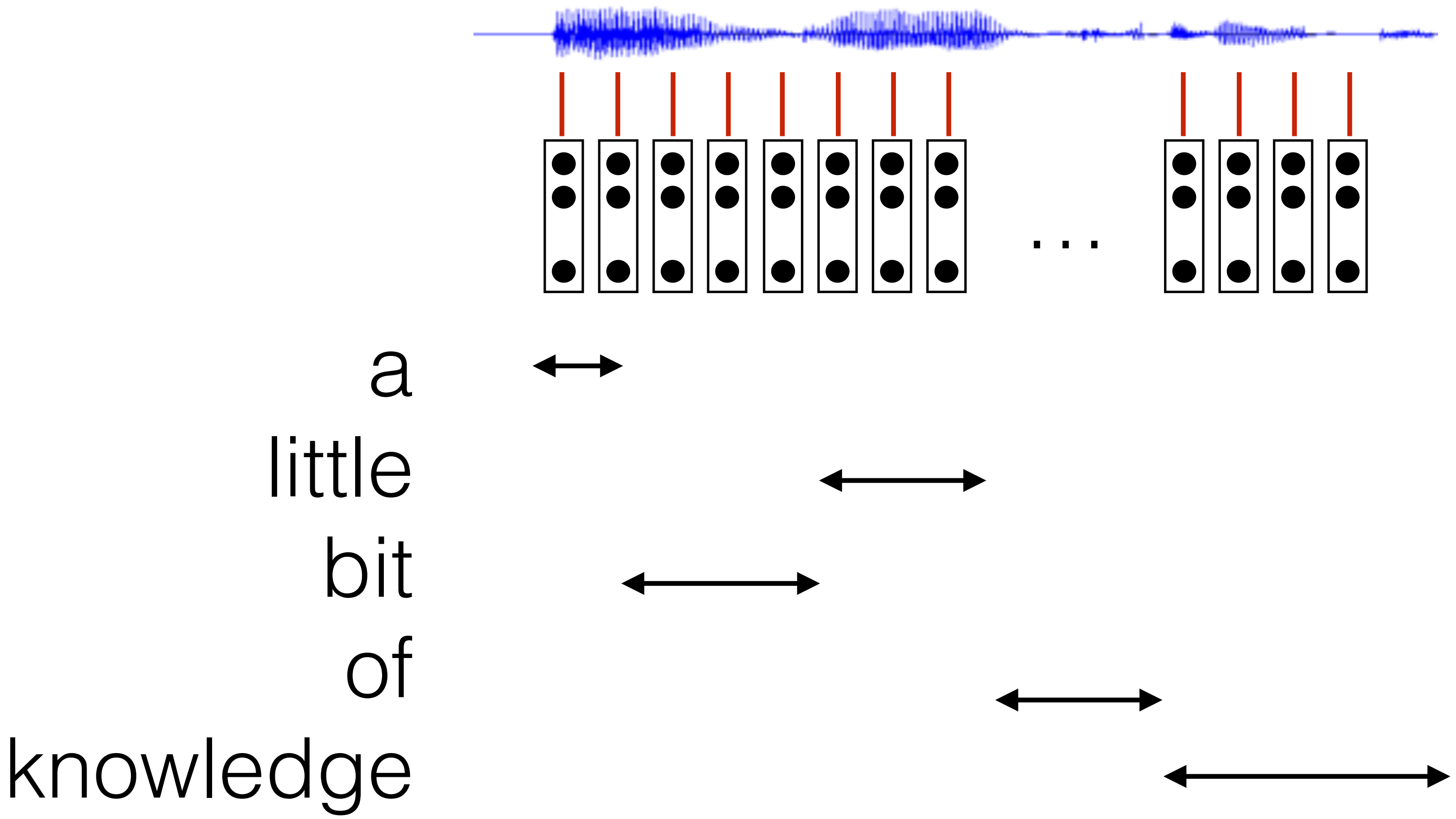
knowledge

Task Description

Source side: Frames of the speech signal

Target side: Translation text

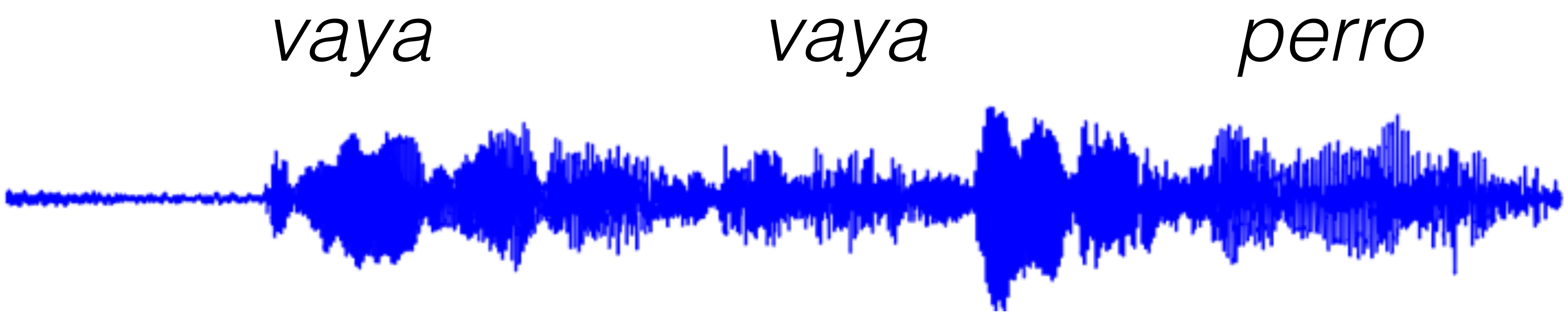
Task: find best alignment between source and target side



Our method outperforms both baselines

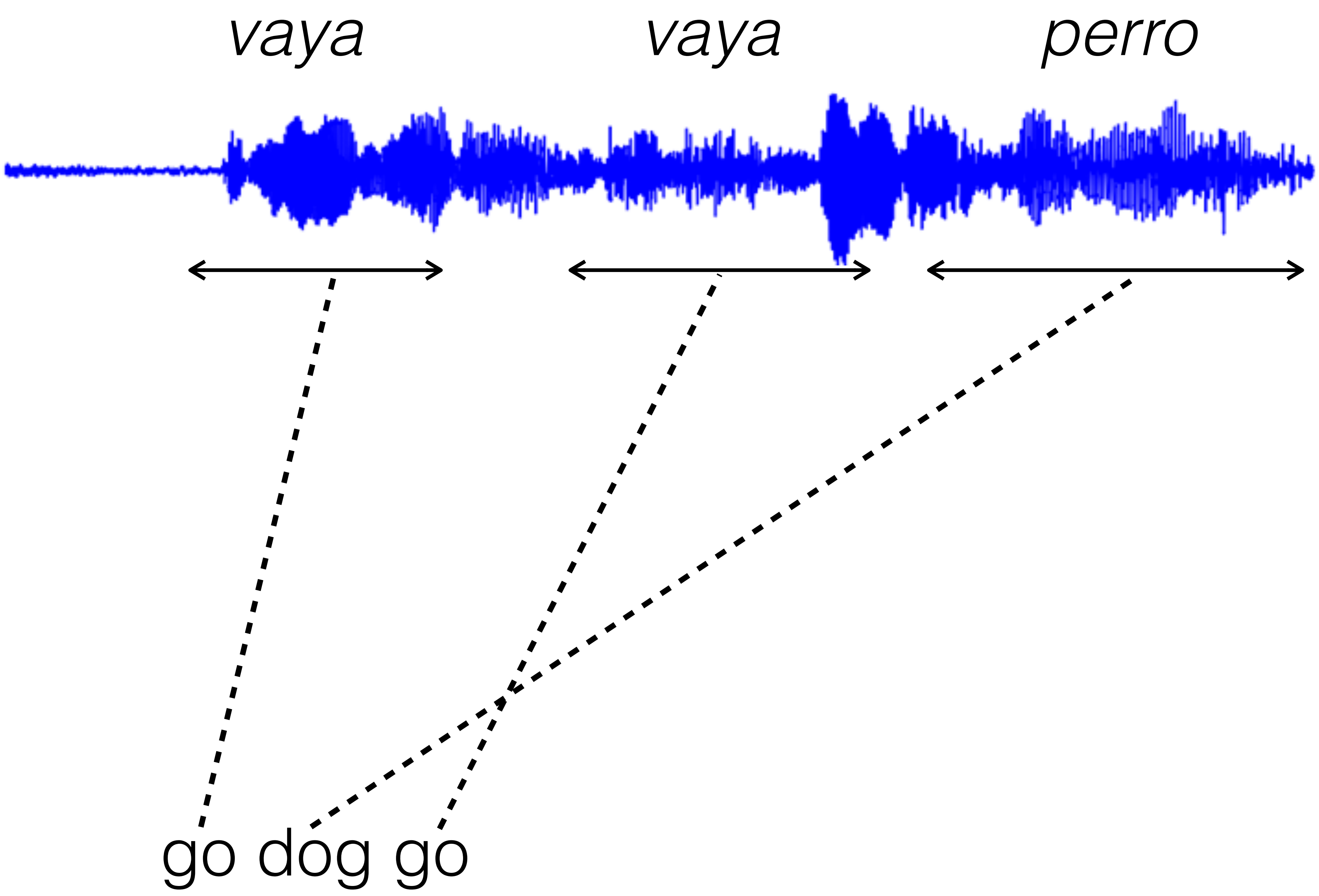
The big picture

The big picture

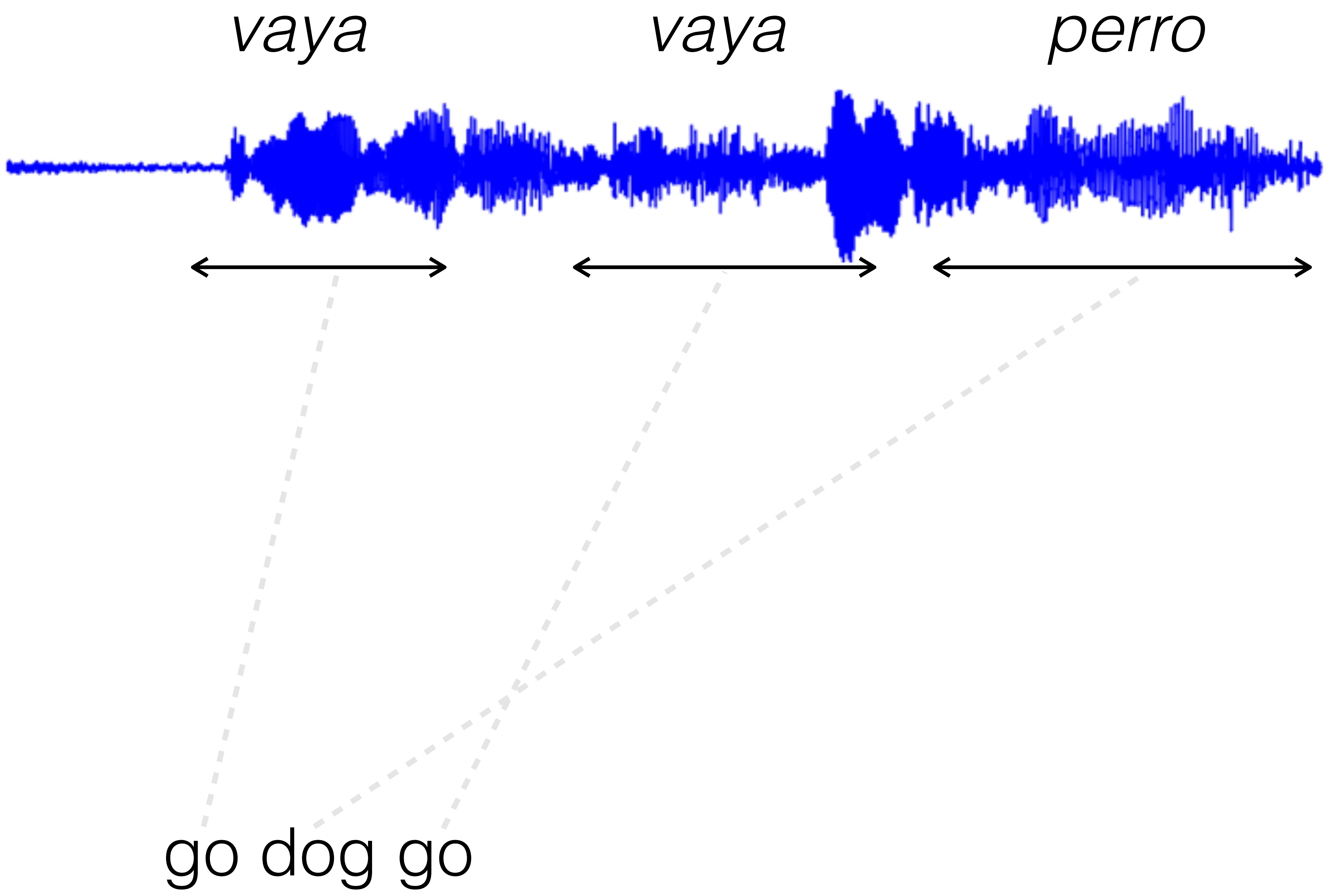


go dog go

The big picture



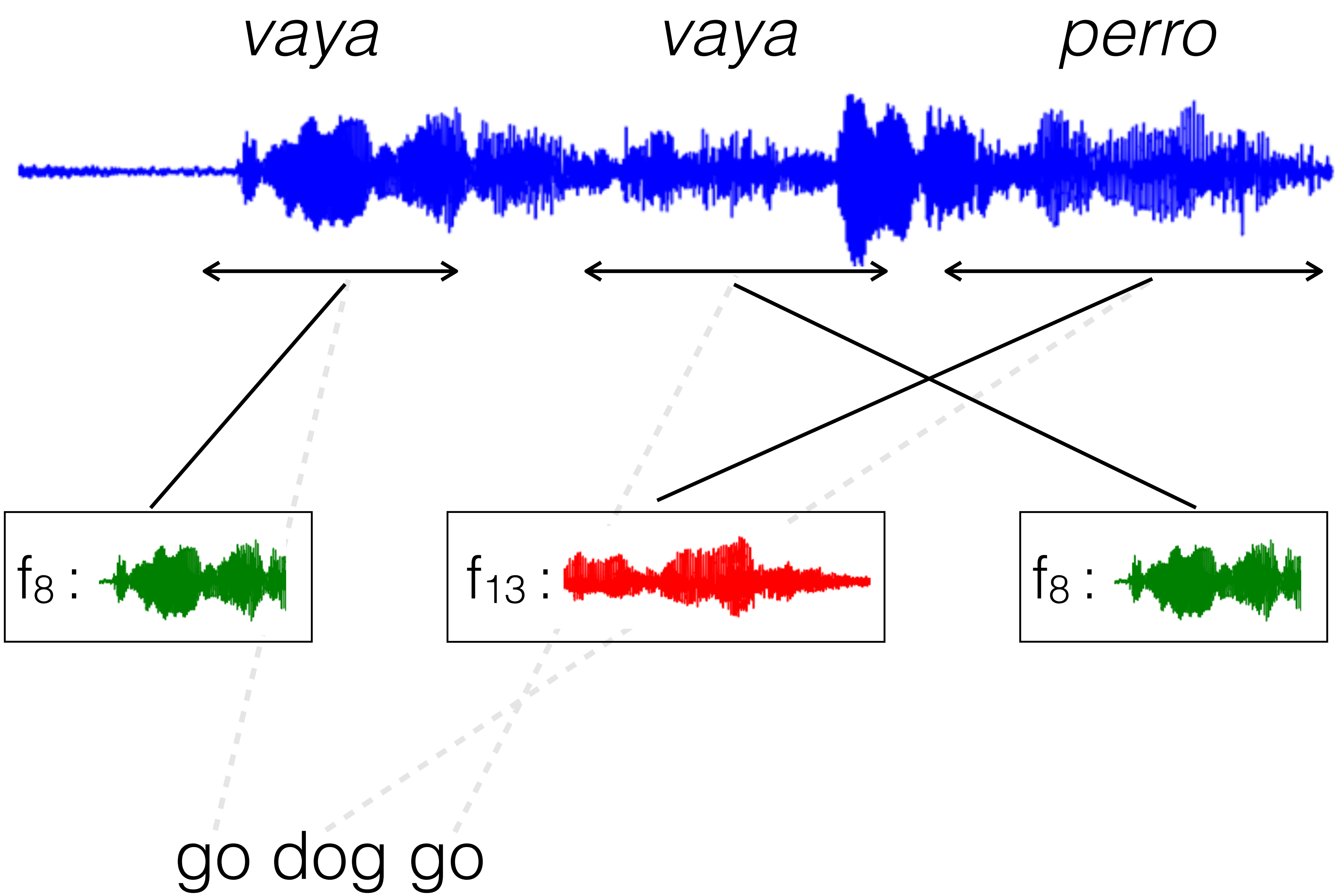
The big picture



vaya *vaya* *perro*

go dog go

The big picture



vaya *vaya* *perro*

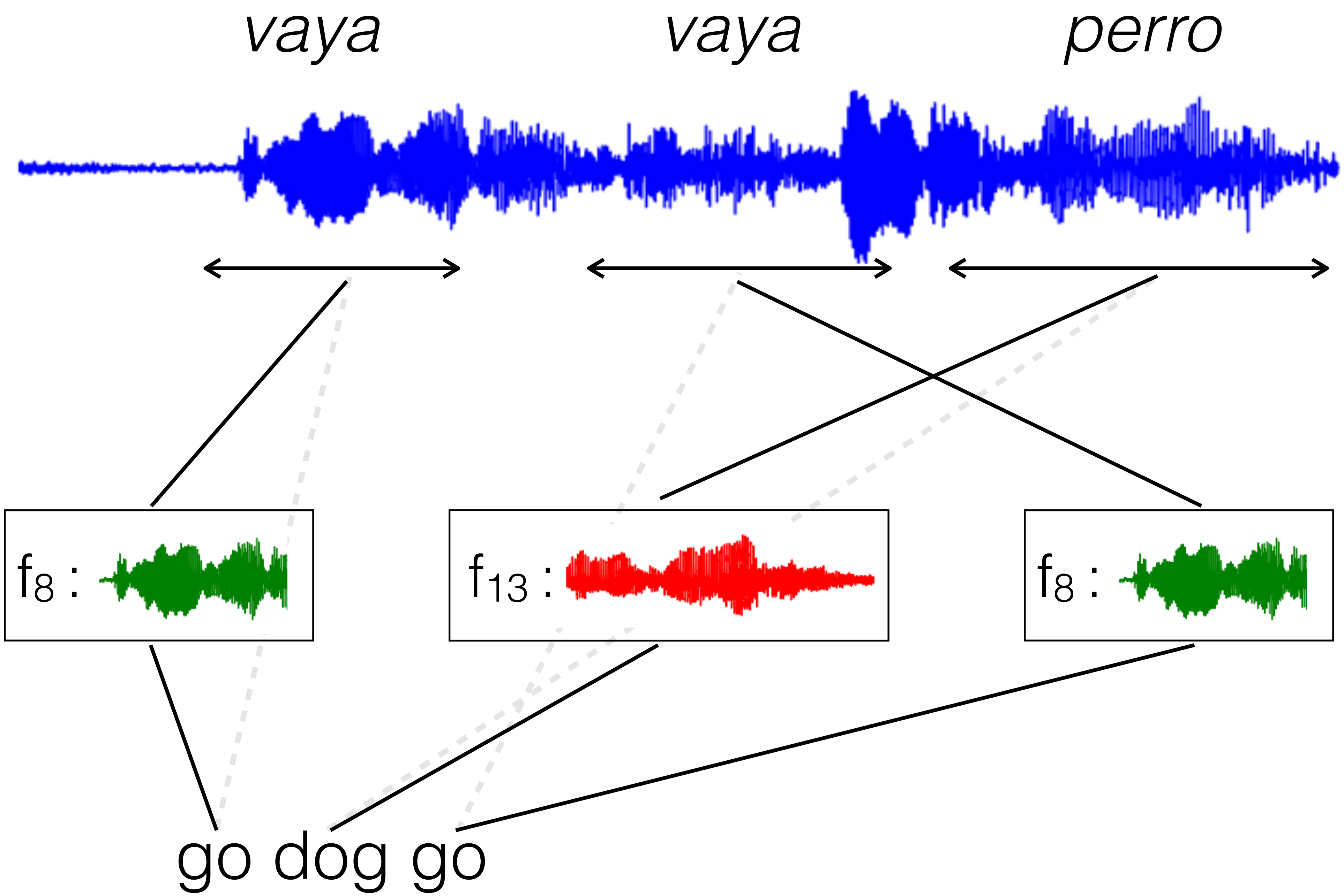
f_8 :

f_{13} :

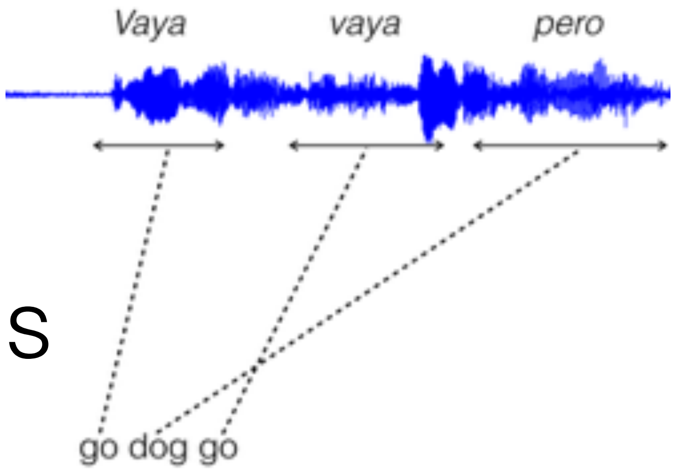
f_8 :

go dog go

The big picture



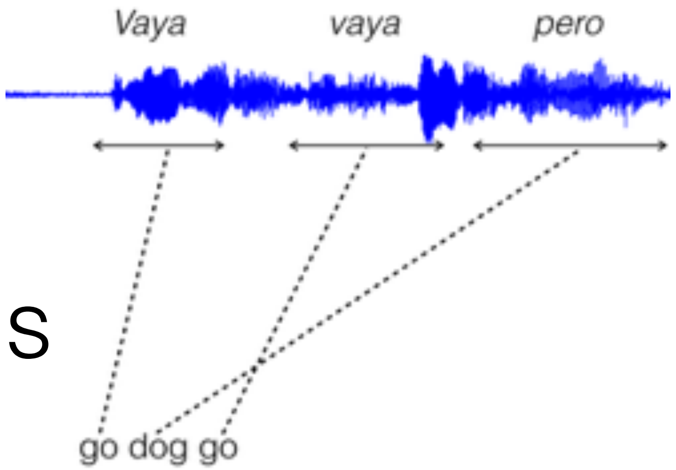
Distortion model



Controls the reordering of the target words

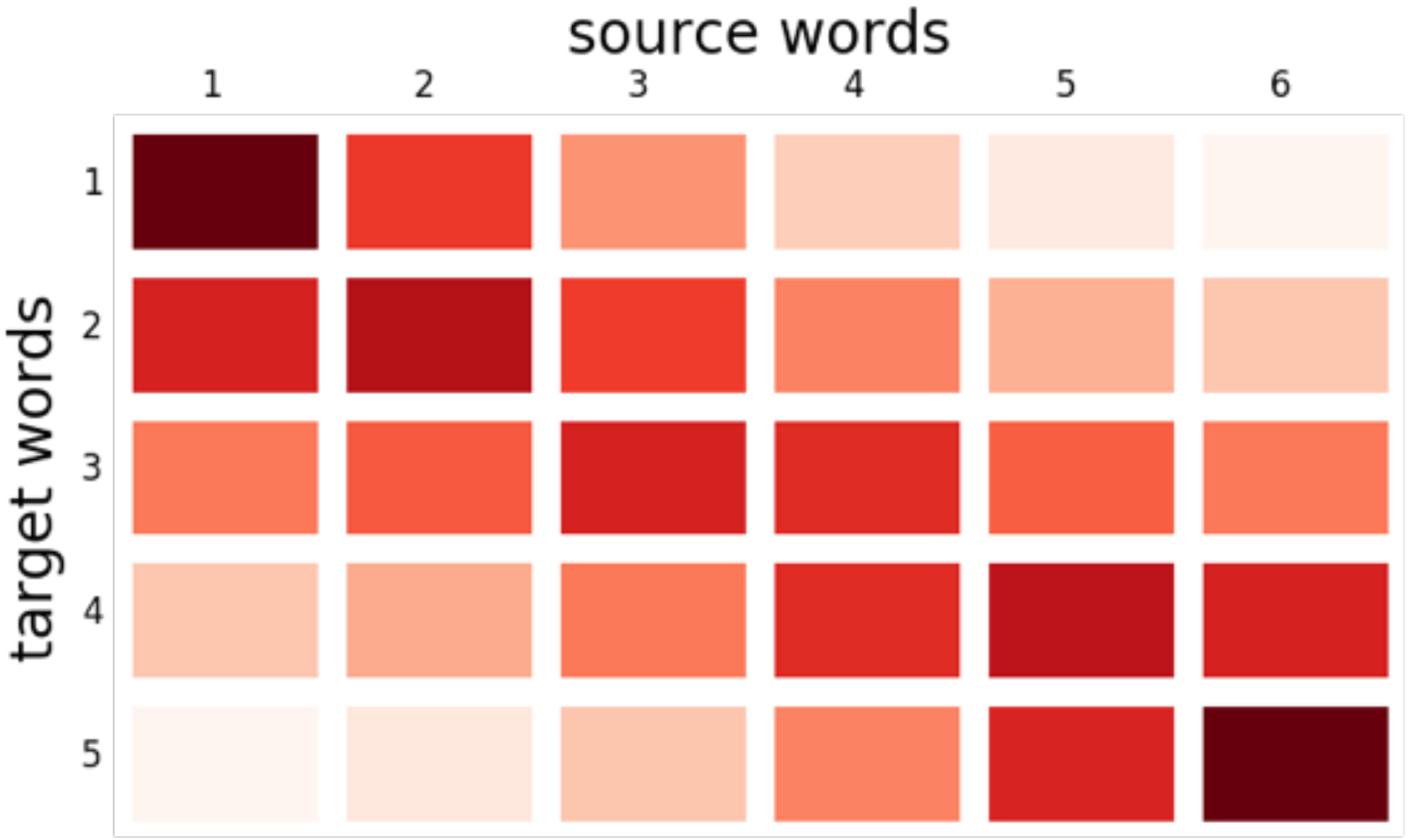
- based on fast-align [Dyer et al.]

Distortion model

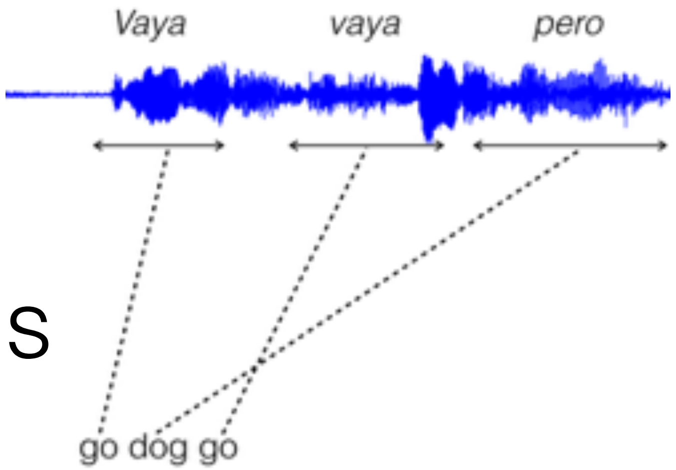


Controls the reordering of the target words
- based on fast-align [Dyer et al.]

Original

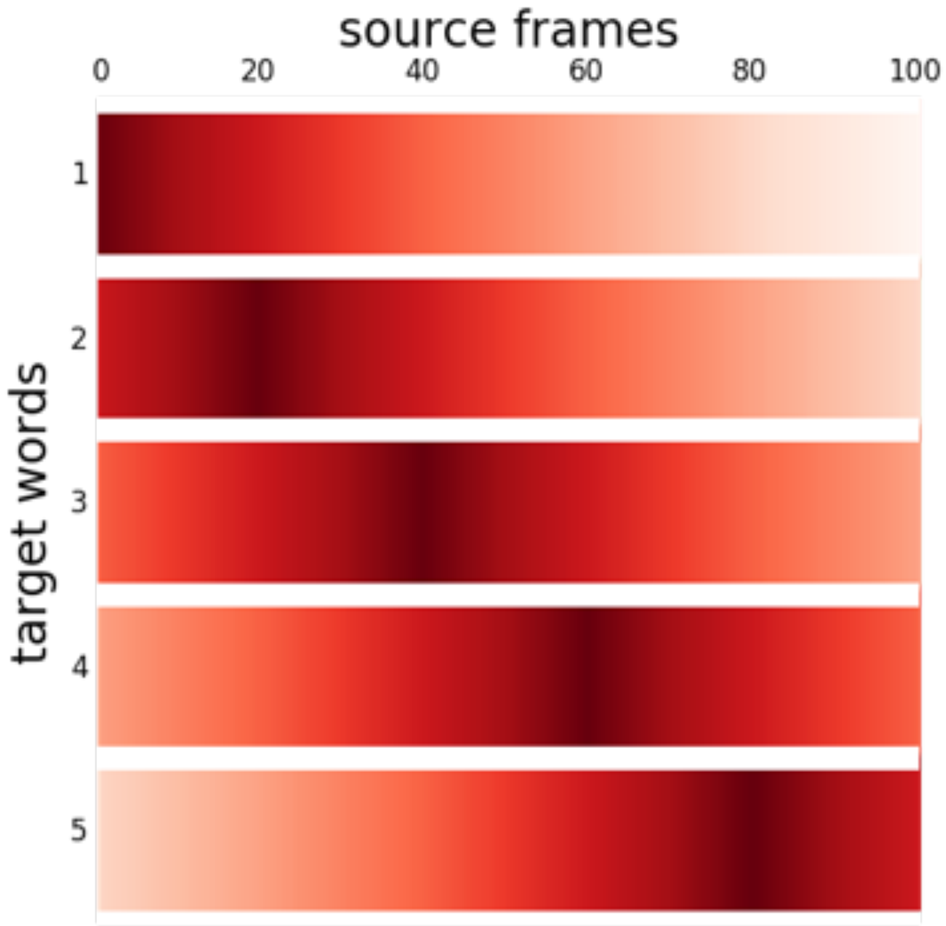


Distortion model

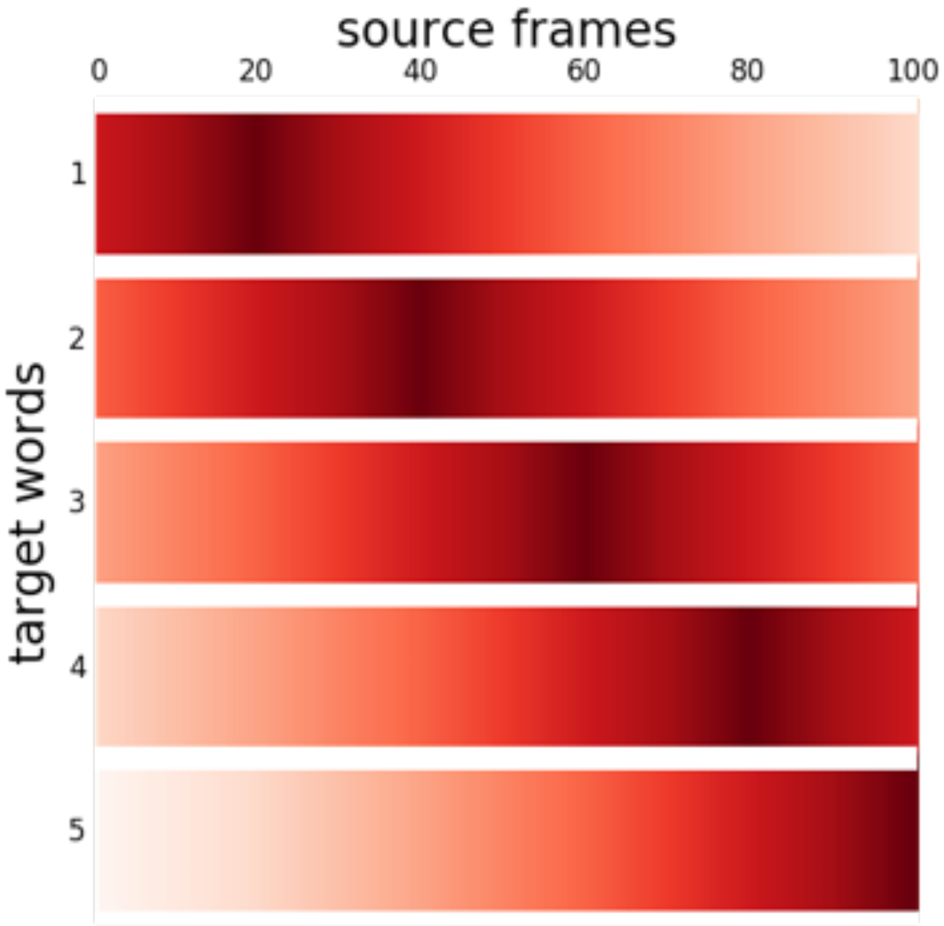


Controls the reordering of the target words
- based on fast-align [Dyer et al.]

Modification

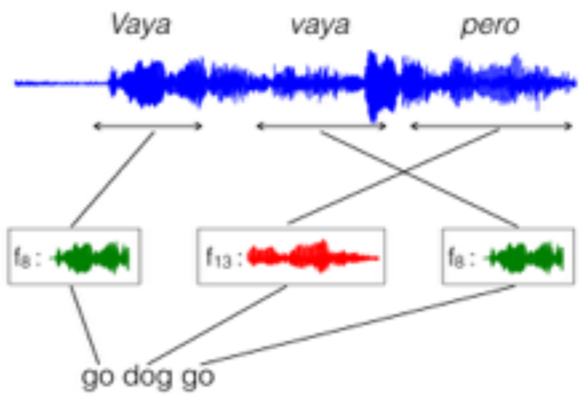


Span Start



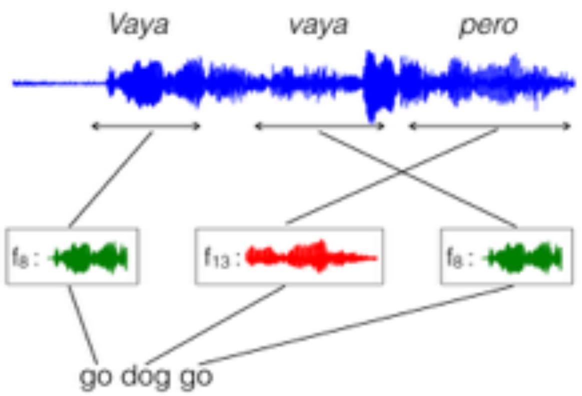
Span End

Clustering model



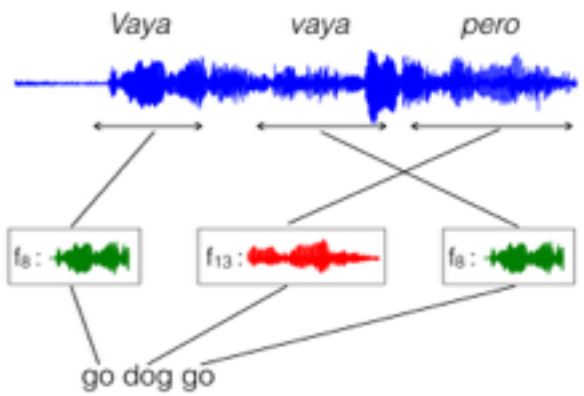
Assuming a *“prototype”* for each cluster

Clustering model



Assuming a *“prototype”* for each cluster

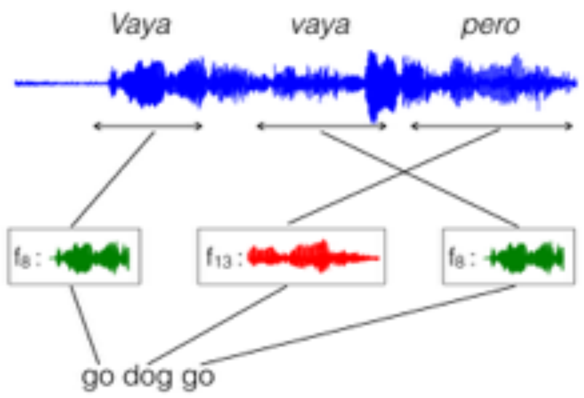
Clustering model



Assuming a *“prototype”* for each cluster

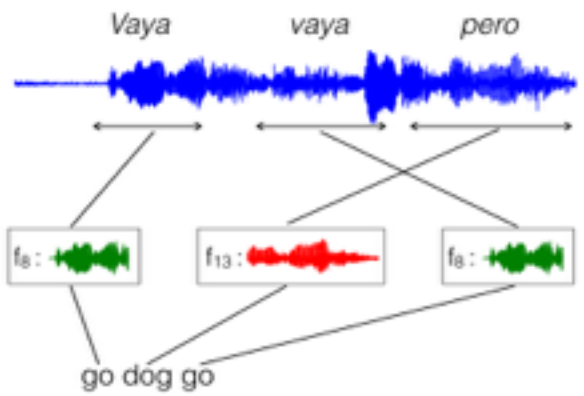
$\boxed{f_8}$

Clustering model

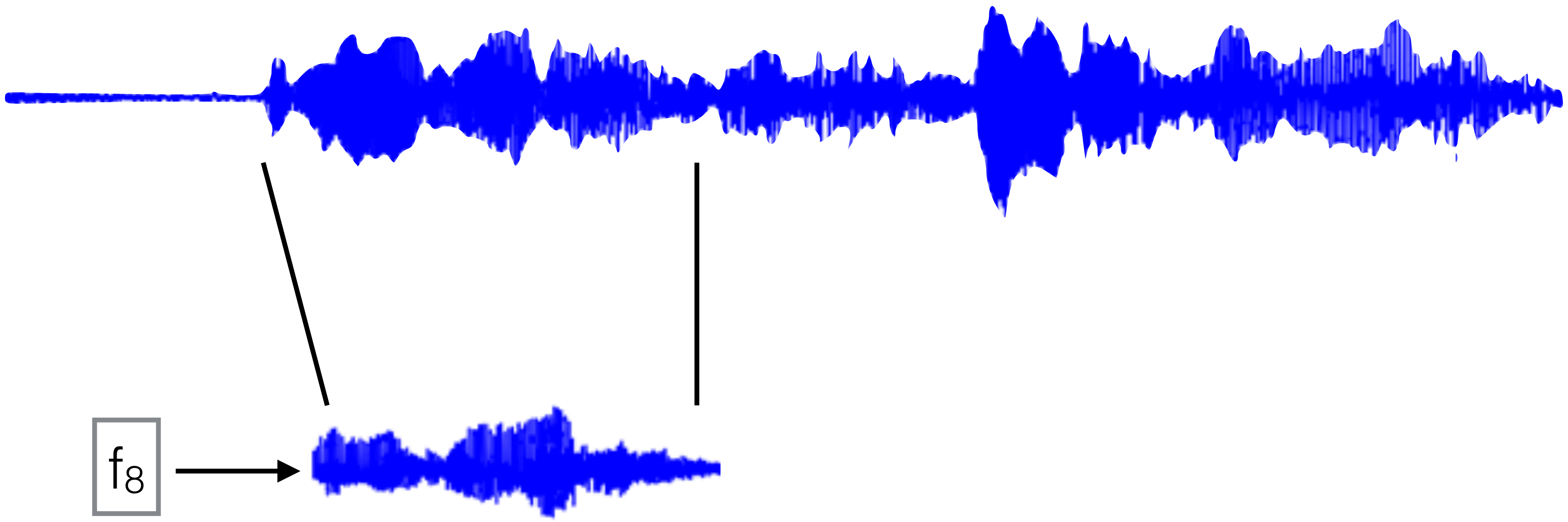


Assuming a *“prototype”* for each cluster

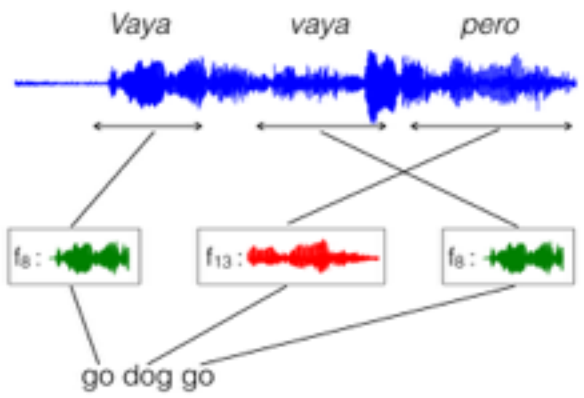
Clustering model



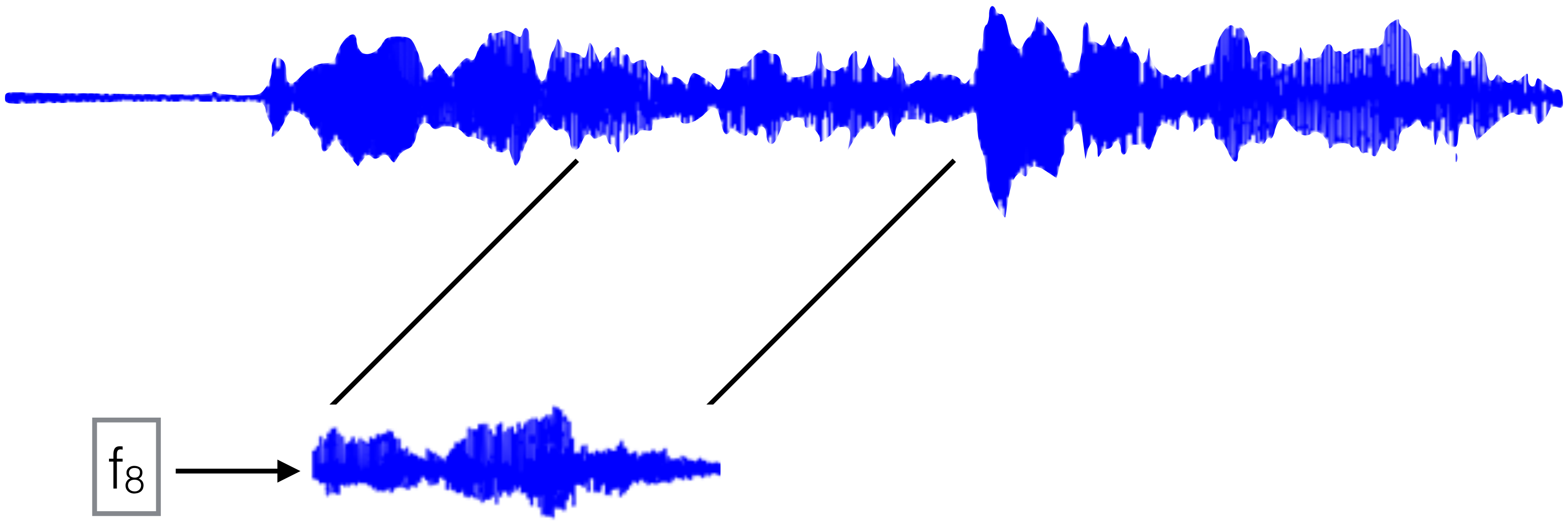
Assuming a *“prototype”* for each cluster



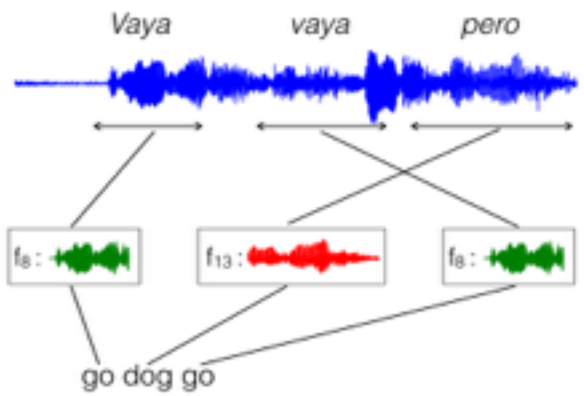
Clustering model



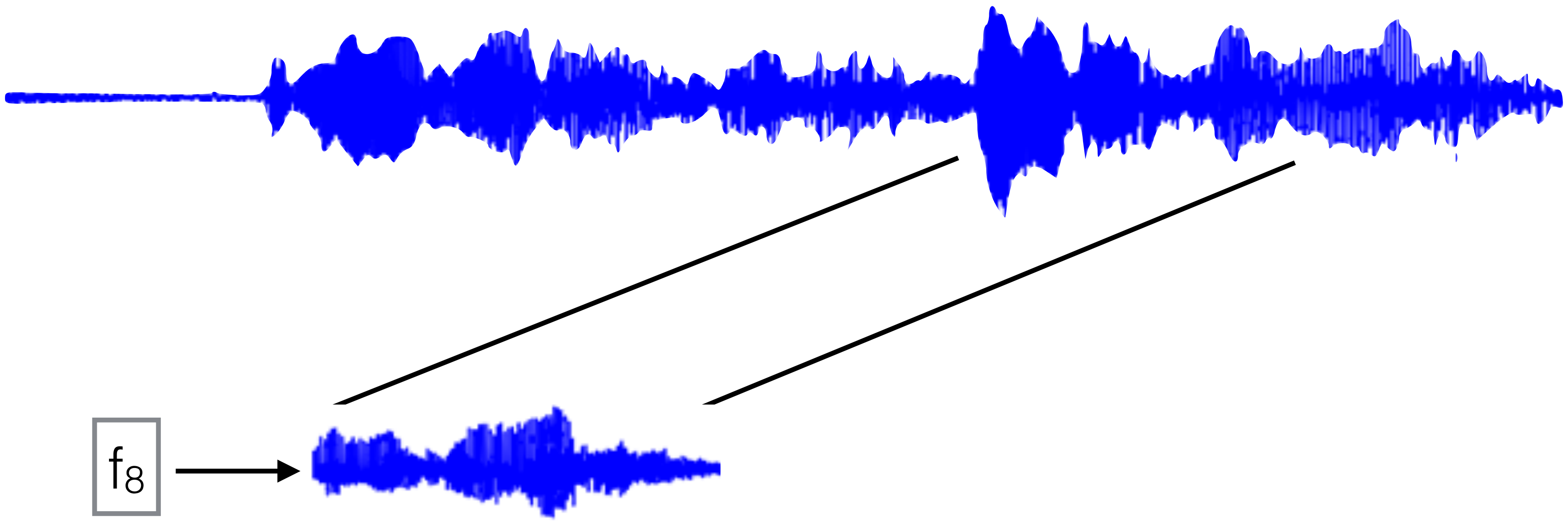
Assuming a *“prototype”* for each cluster



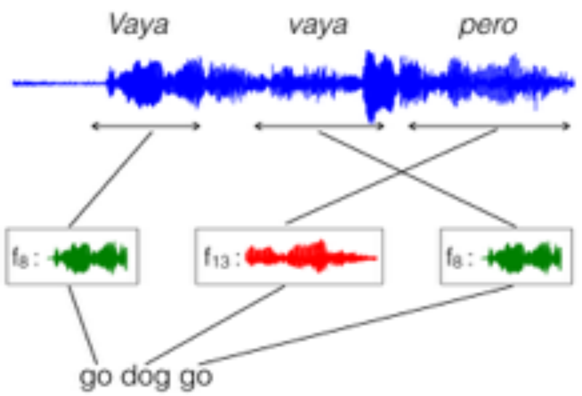
Clustering model



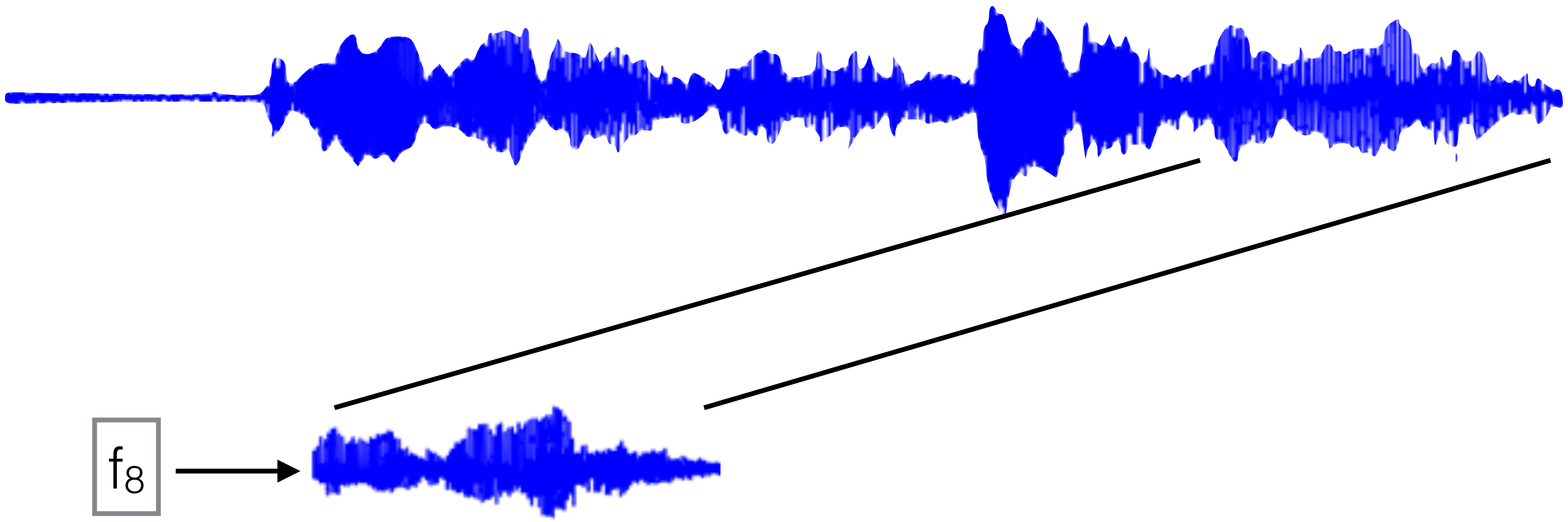
Assuming a *"prototype"* for each cluster



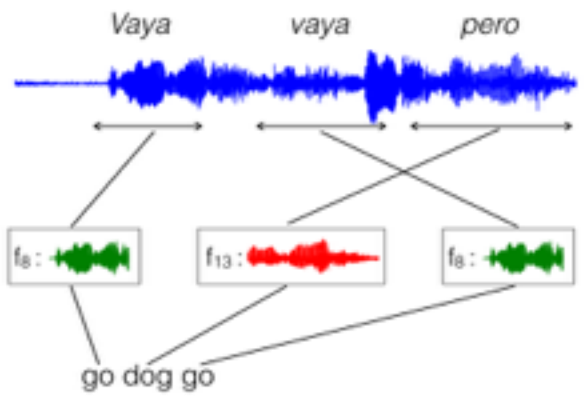
Clustering model



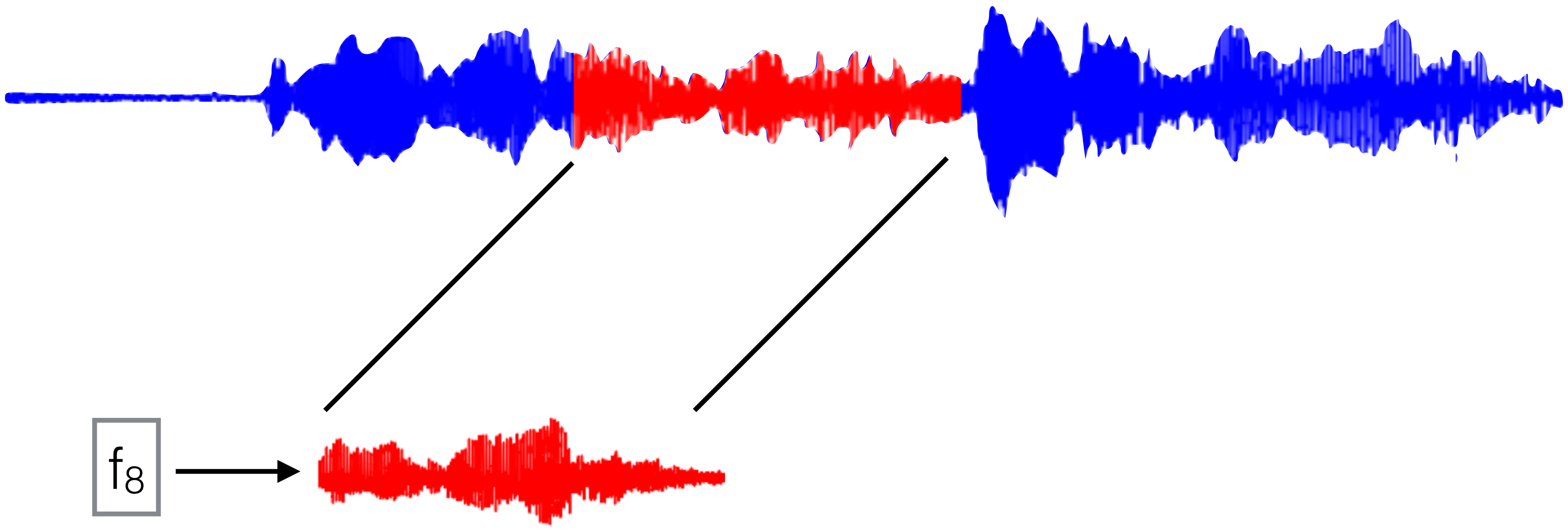
Assuming a *"prototype"* for each cluster



Clustering model



Assuming a *“prototype”* for each cluster

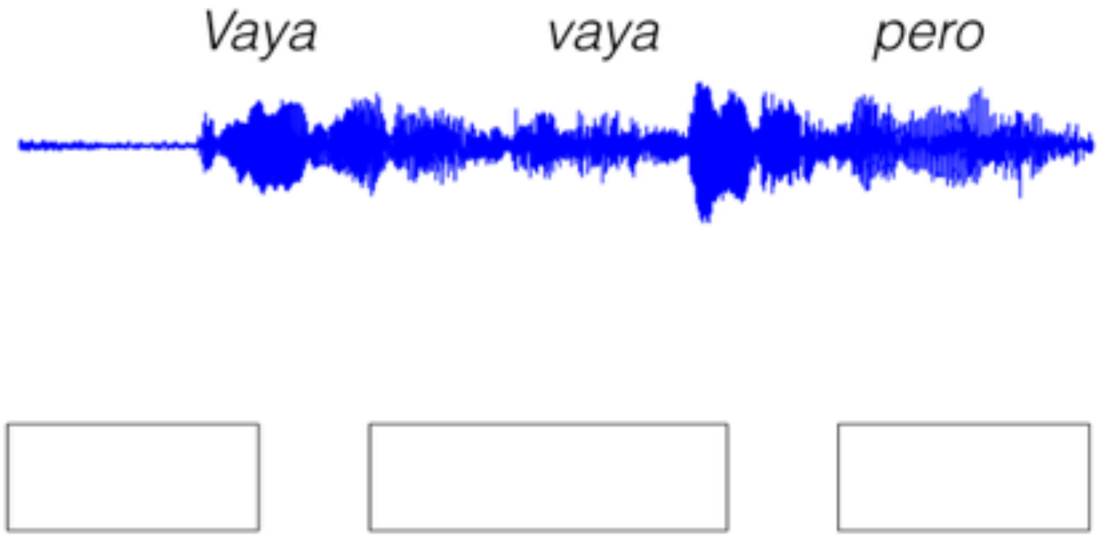


Training

Expectation-Maximization

Training

Expectation-Maximization

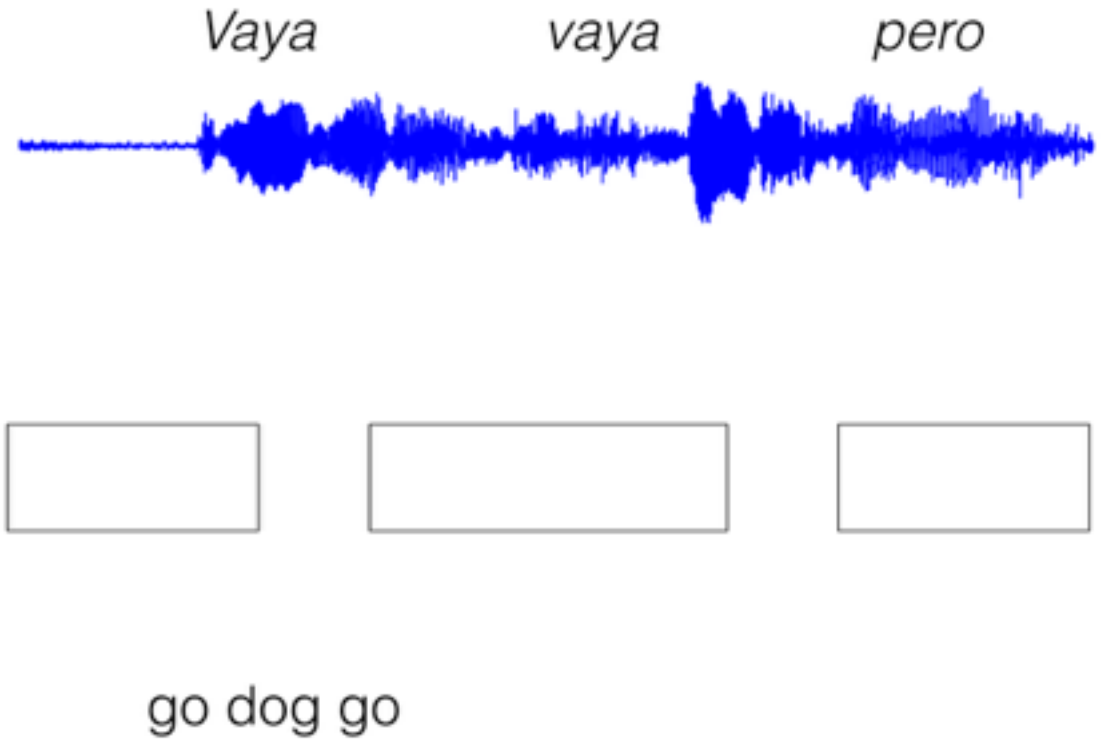


go dog go

Training

Expectation-Maximization

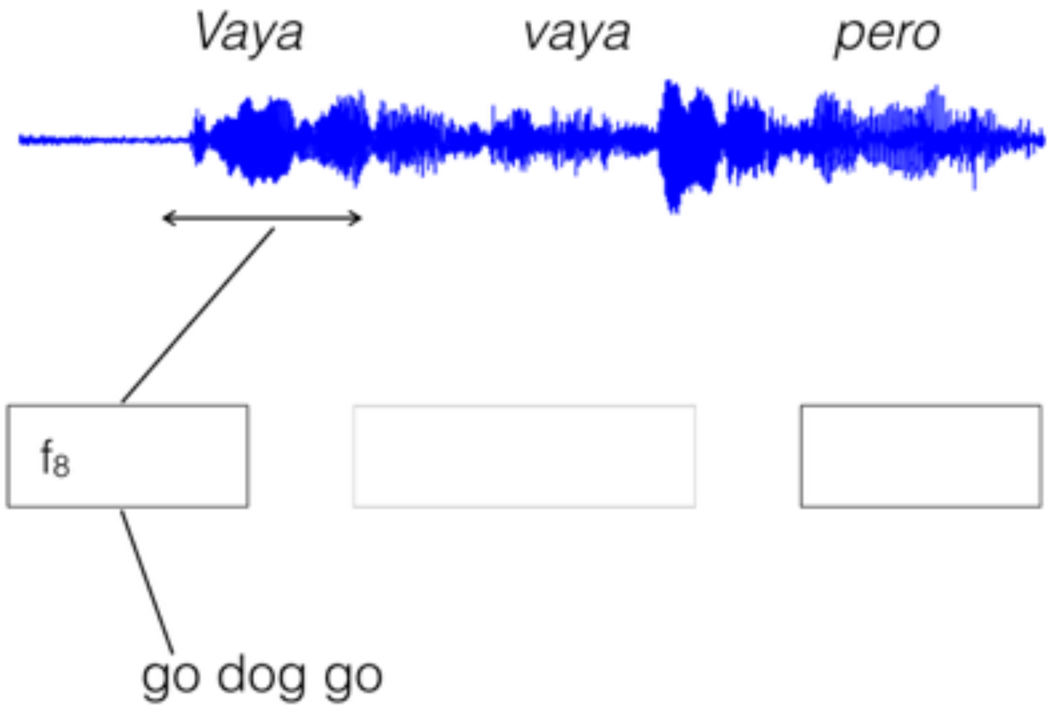
Initialize spans and clusters



Training

Expectation-Maximization

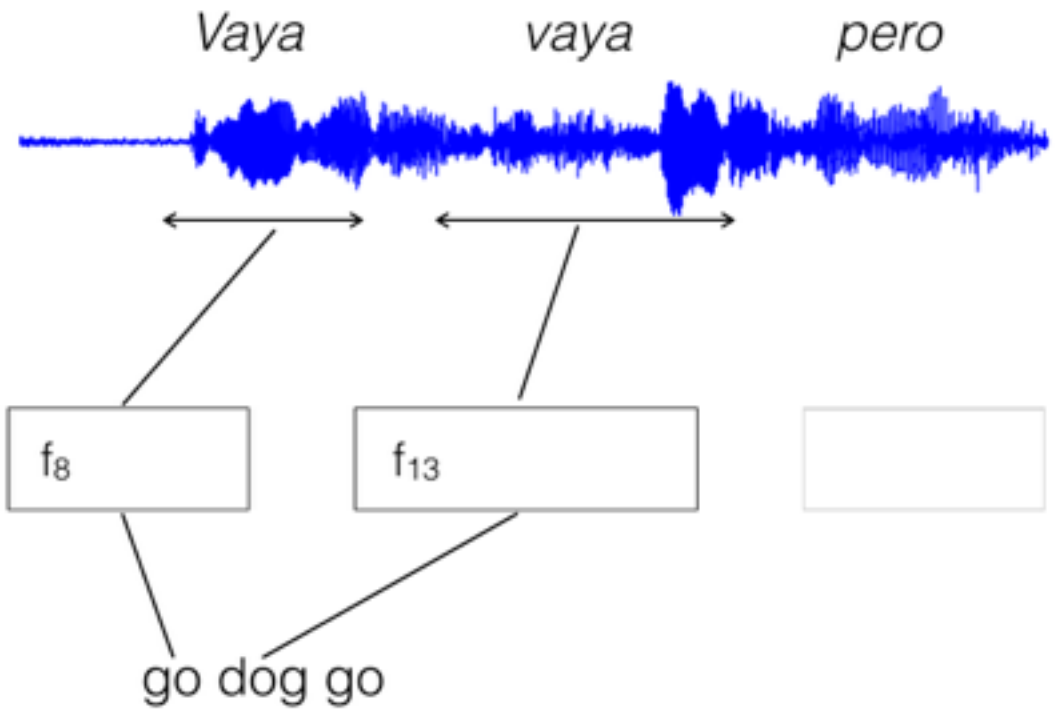
Initialize spans and clusters



Training

Expectation-Maximization

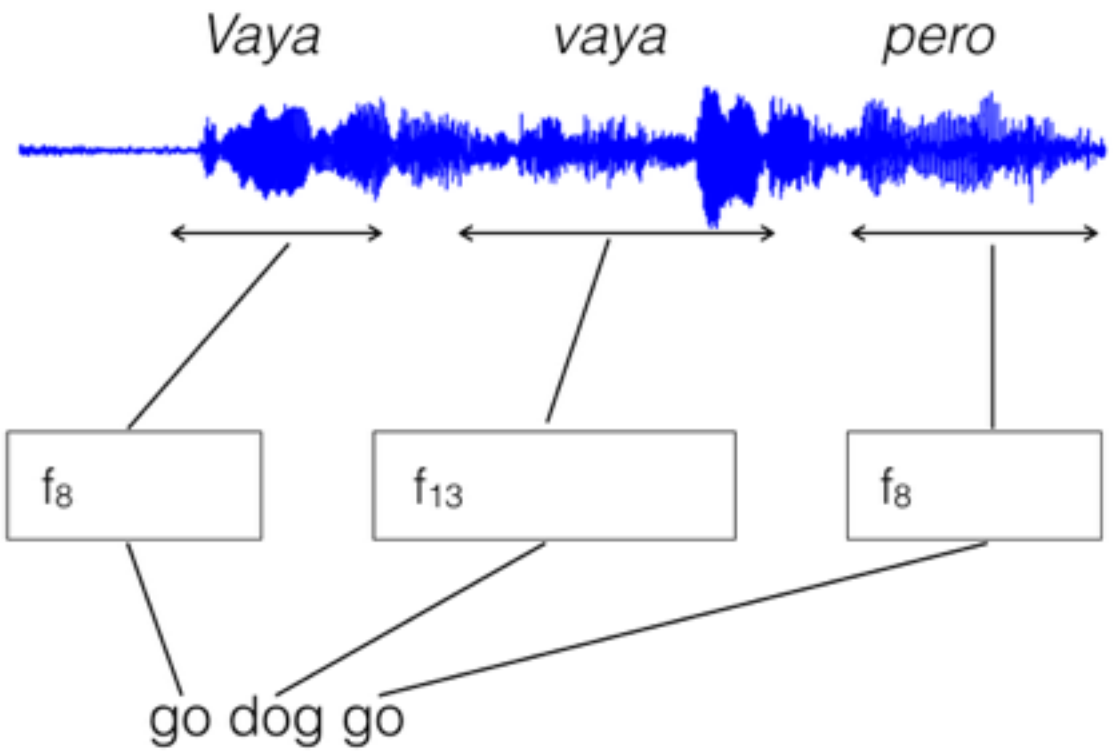
Initialize spans and clusters



Training

Expectation-Maximization

Initialize spans and clusters

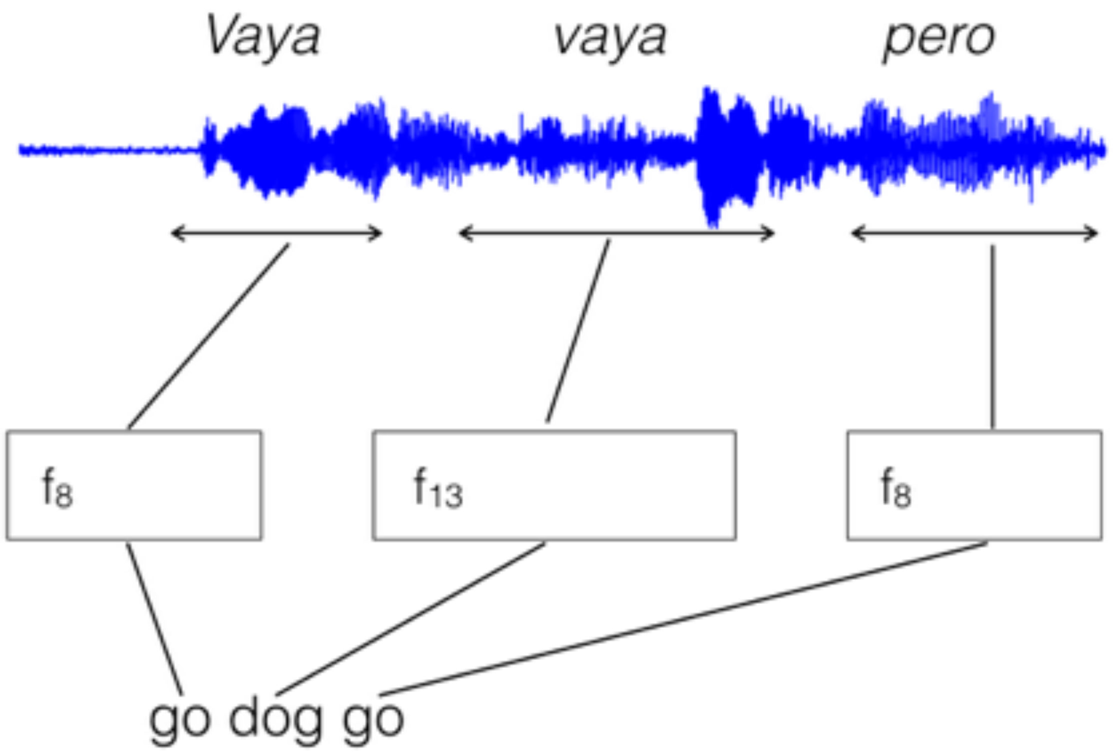


Training

Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes

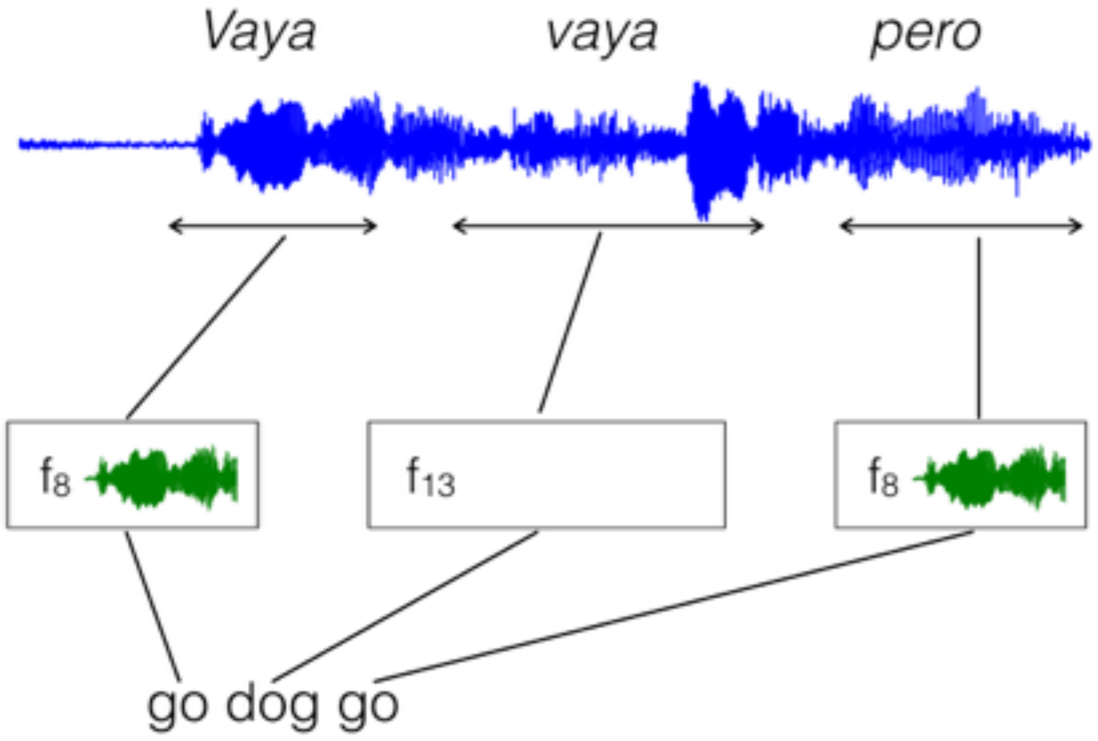


Training

Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes

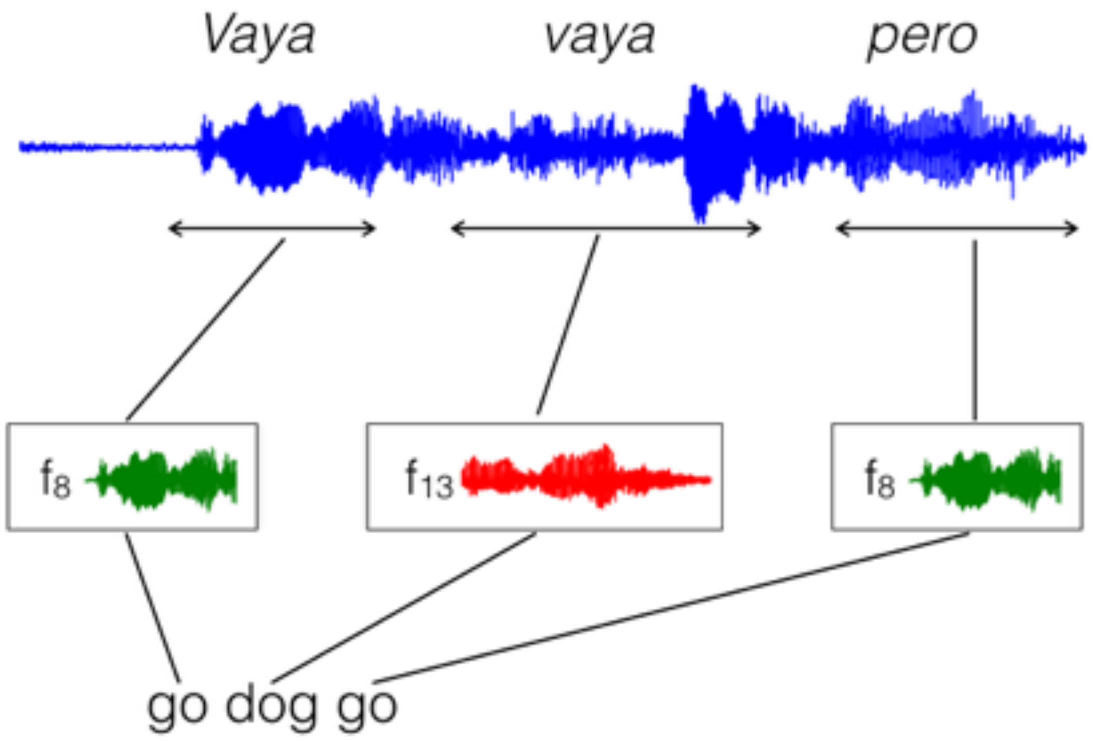


Training

Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes

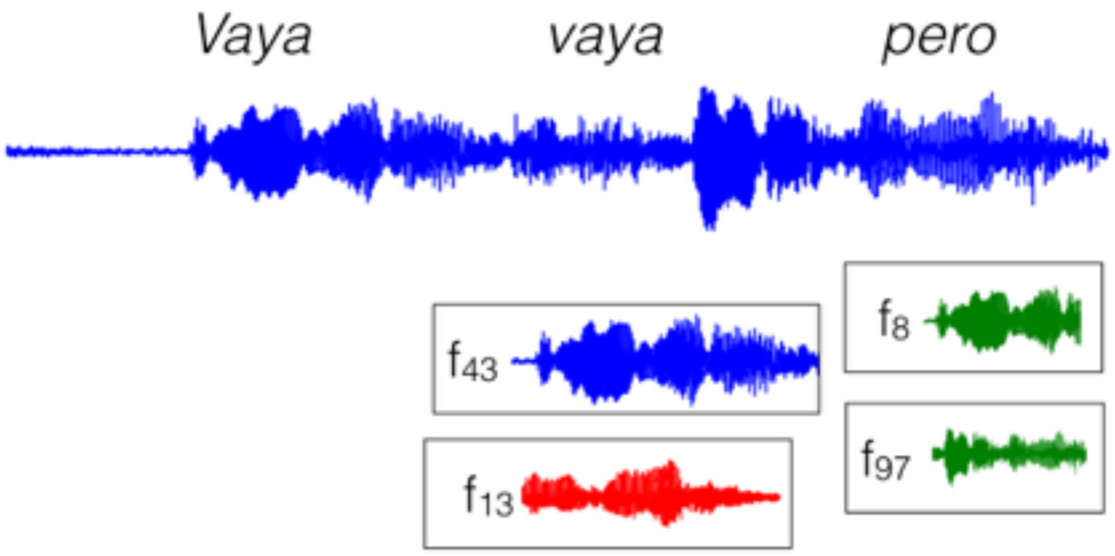


Training

Expectation-Maximization

Initialize spans and clusters

- M step:
    - Re-estimate prototypes
- E step:



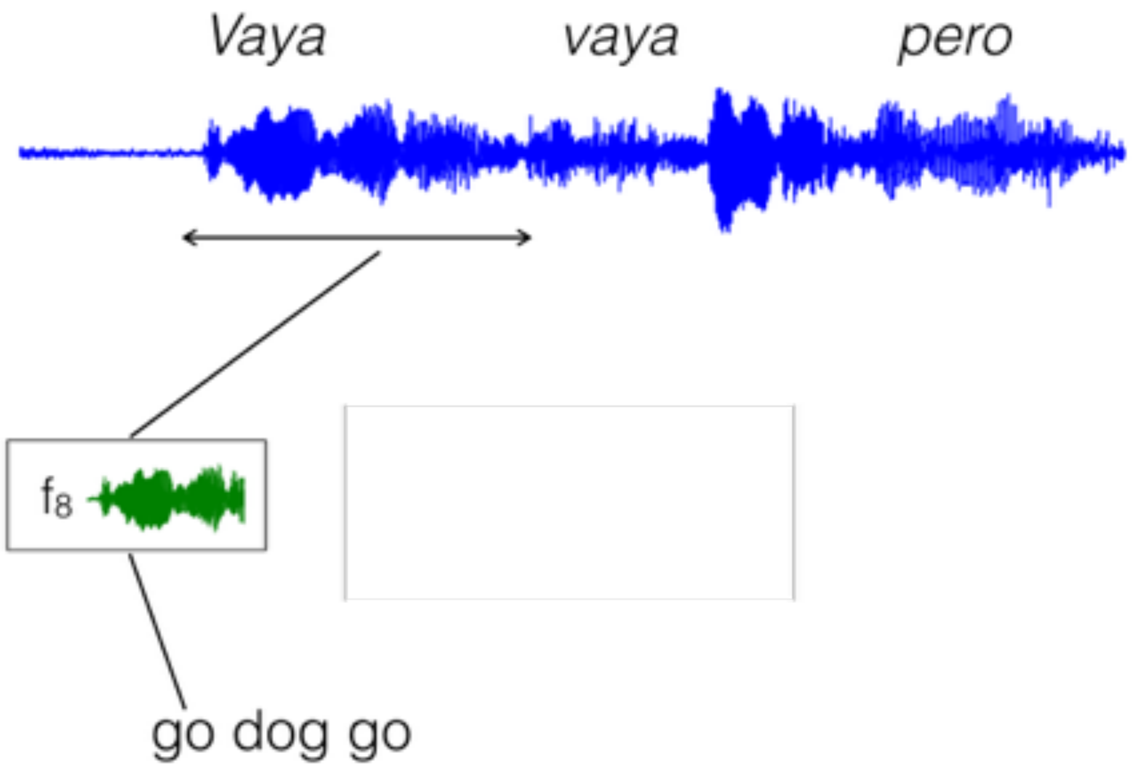
go dog go

Training

Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes
- E step:
  - Assign cluster and align

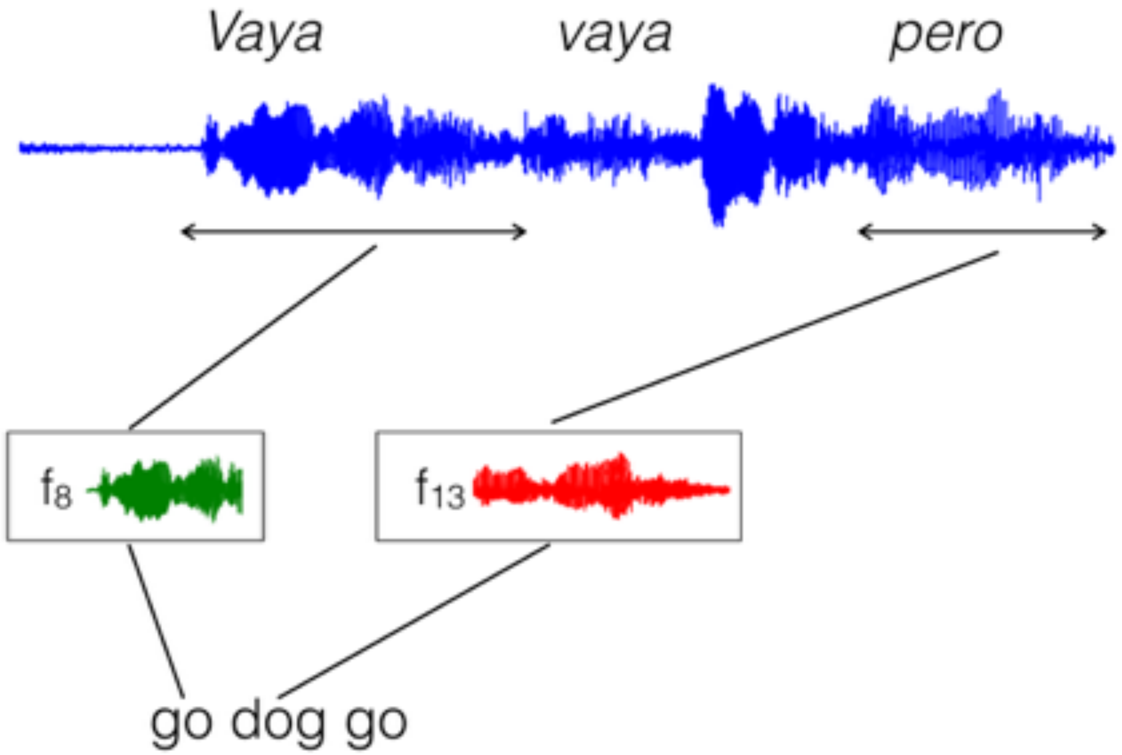


Training

Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes
- E step:
  - Assign cluster and align

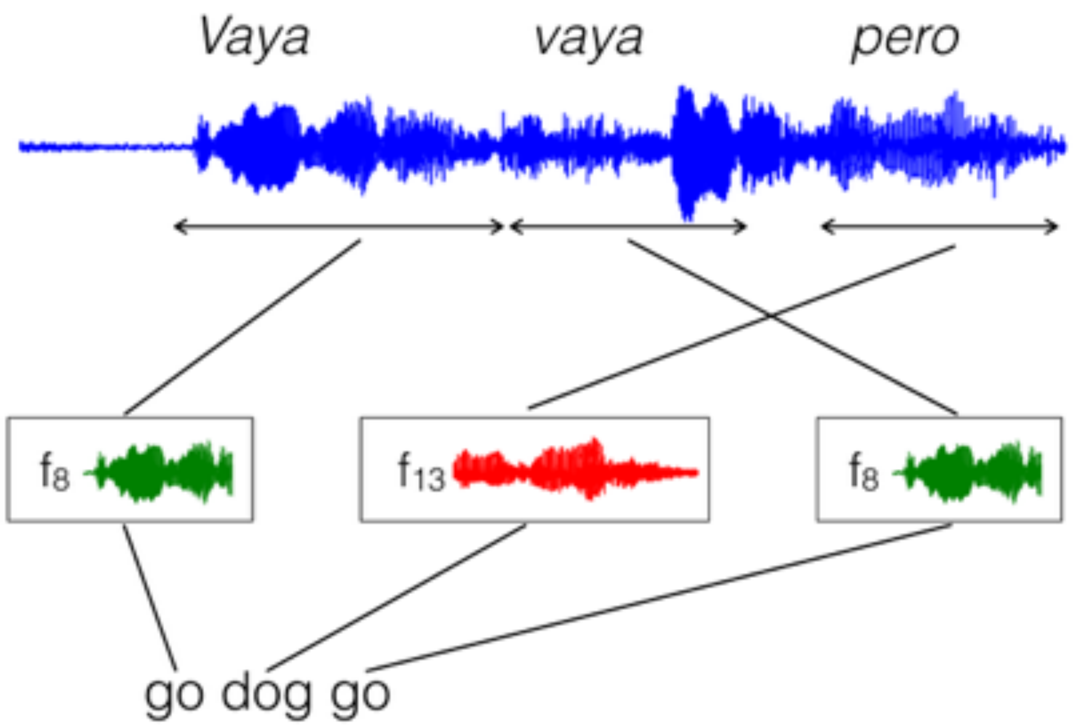


Training

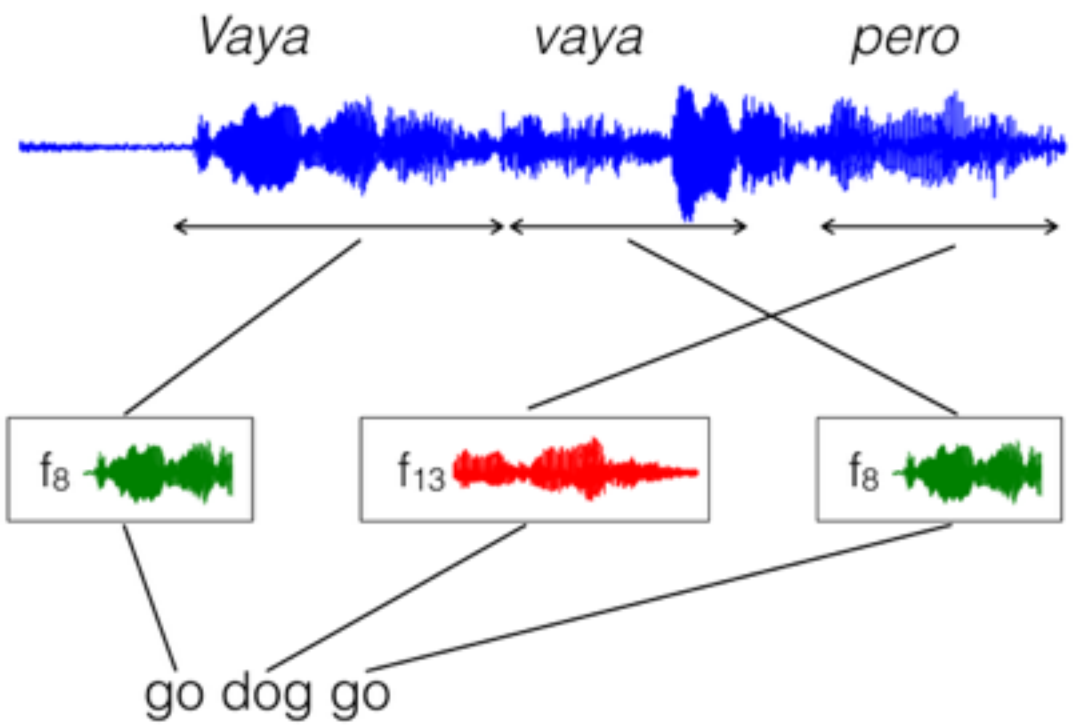
Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes
- E step:
  - Assign cluster and align



Training



Expectation-Maximization

Initialize spans and clusters

- M step:
  - Re-estimate prototypes
- E step:
  - Assign cluster and align
  - We restrict the search space:
    - voice activity detection
    - phone boundary detection [Khanaga et al.]

Experiments

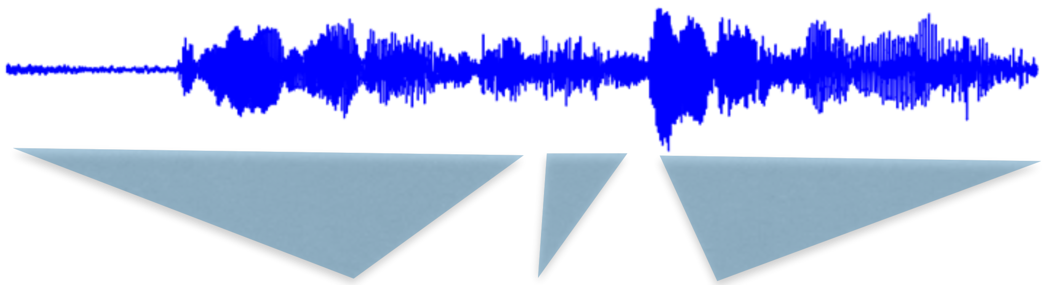
Language Pair	Dataset	Number of utterances
Griko - Italian	[Lekakou et al]	330
Spanish - English	CALLHOME (sample)	2k
	CALLHOME (all)	17k
	Fisher	143k

Baselines

Baselines

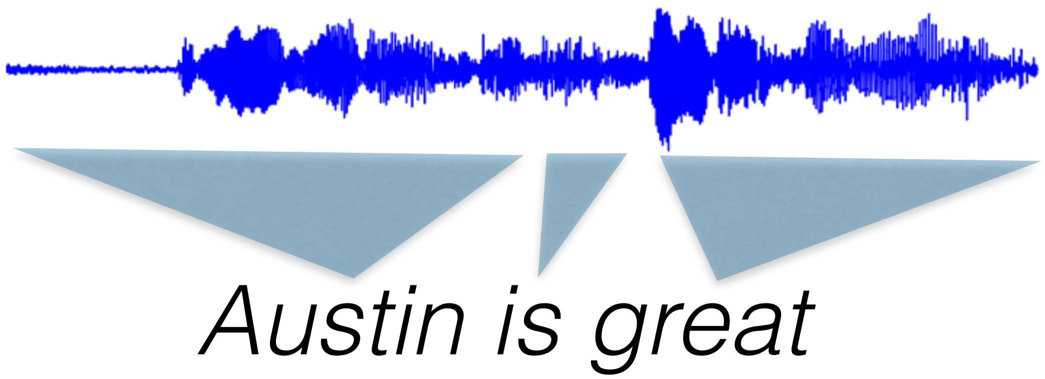
- Naive:
  - frames/word ~ #characters
  - along the diagonal

Baselines



- Naive:
  - frames/word ~ #characters
  - along the diagonal

Baselines



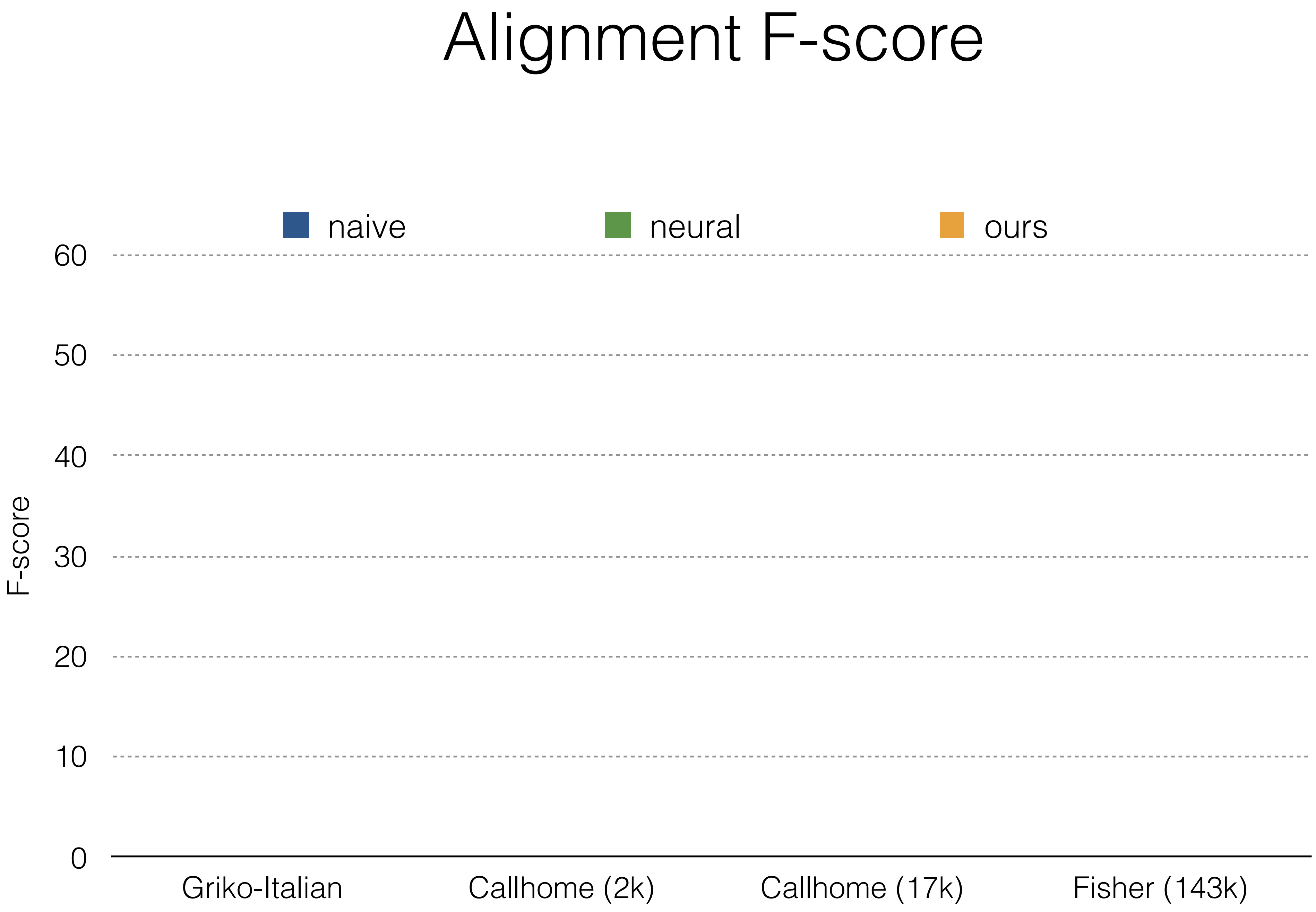
- Naive:
  - frames/word ~ #characters
  - along the diagonal

- Neural [Duong et al]:
  - DNN optimised for direct translation of speech
  - convert attention mechanism weights to alignments

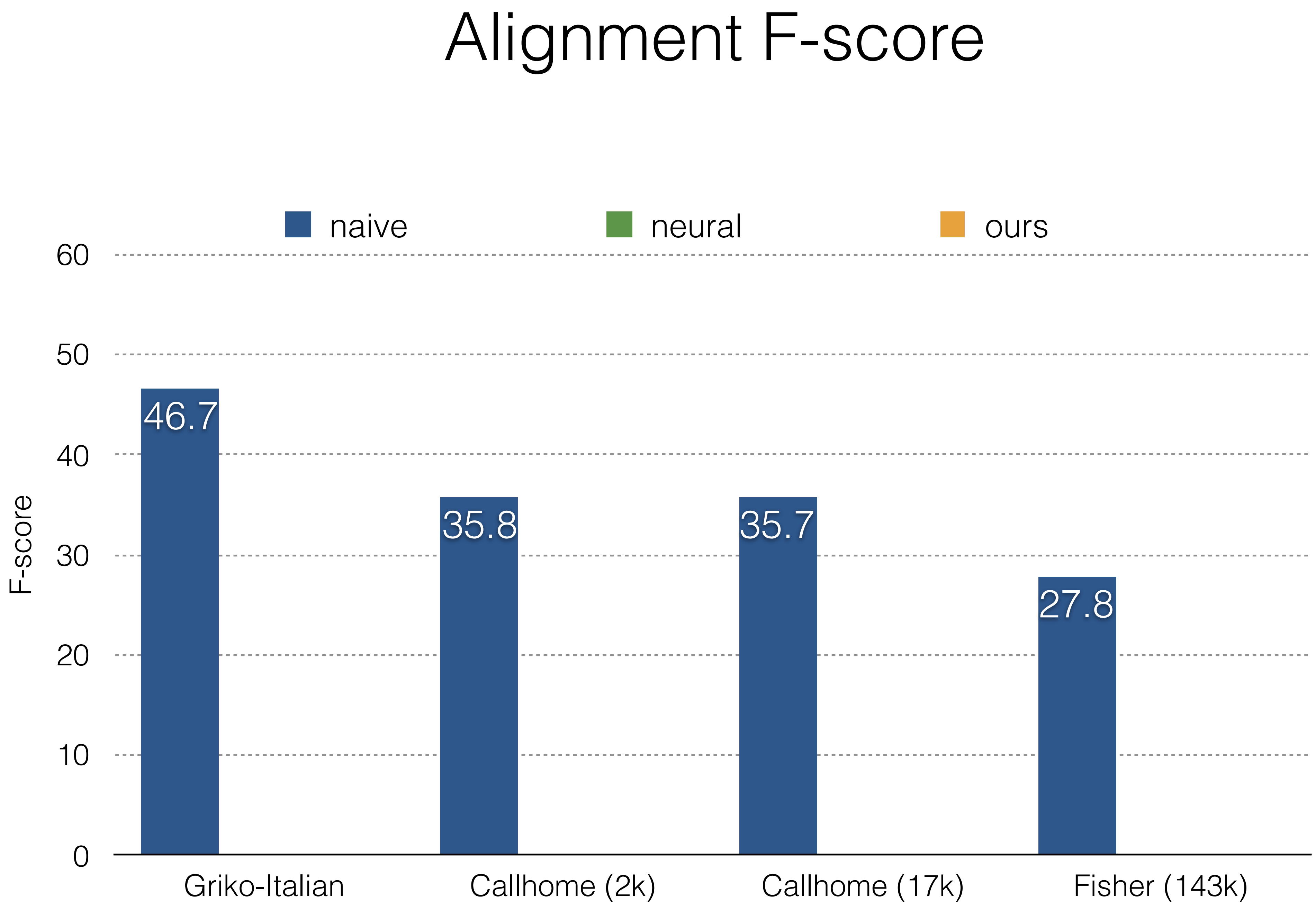
Results

Alignment F-score

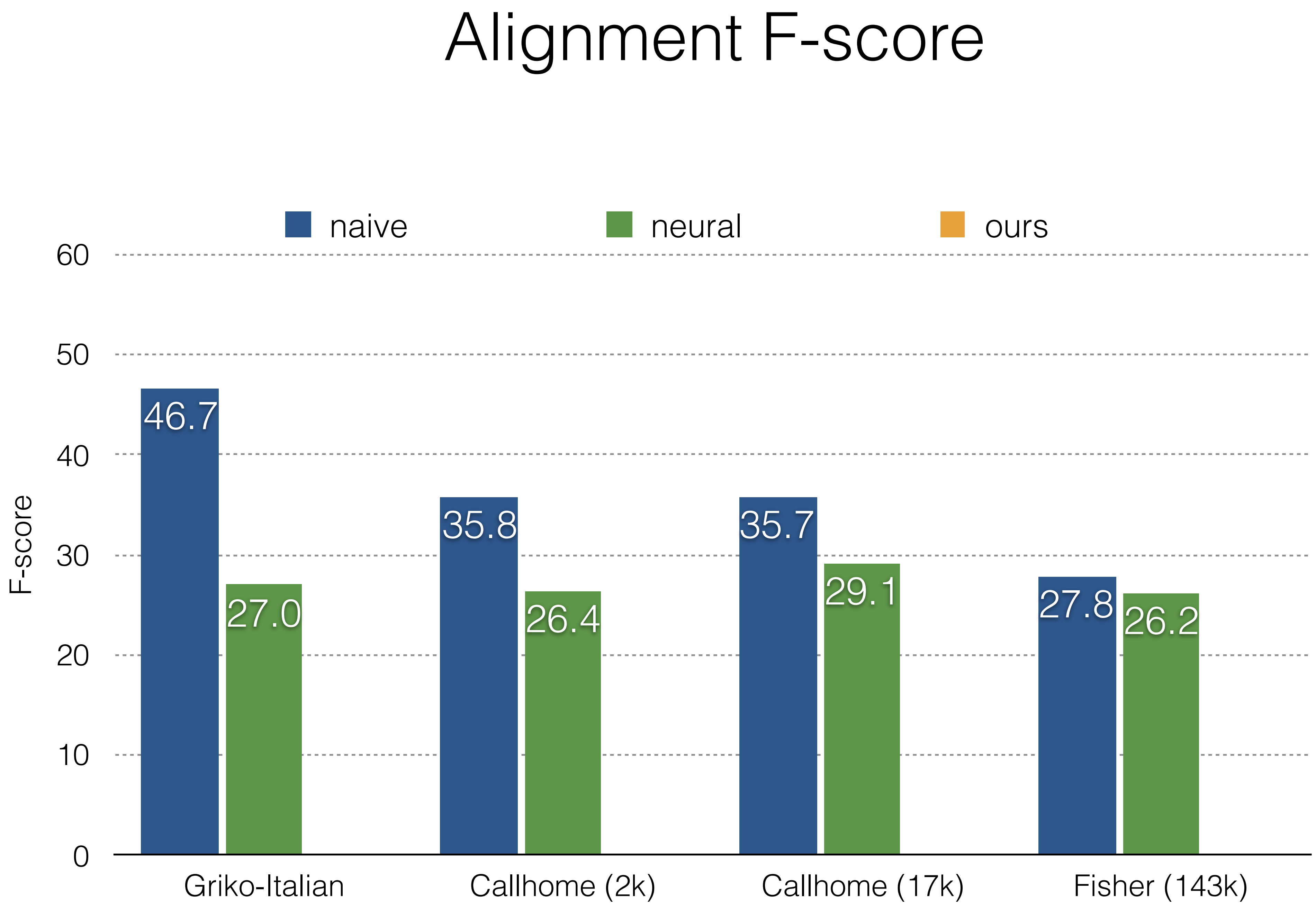
Results


Alignment F-score
🟦 naive
🟩 neural
🟧 ours
F-score
60
50
40
30
20
10
0
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

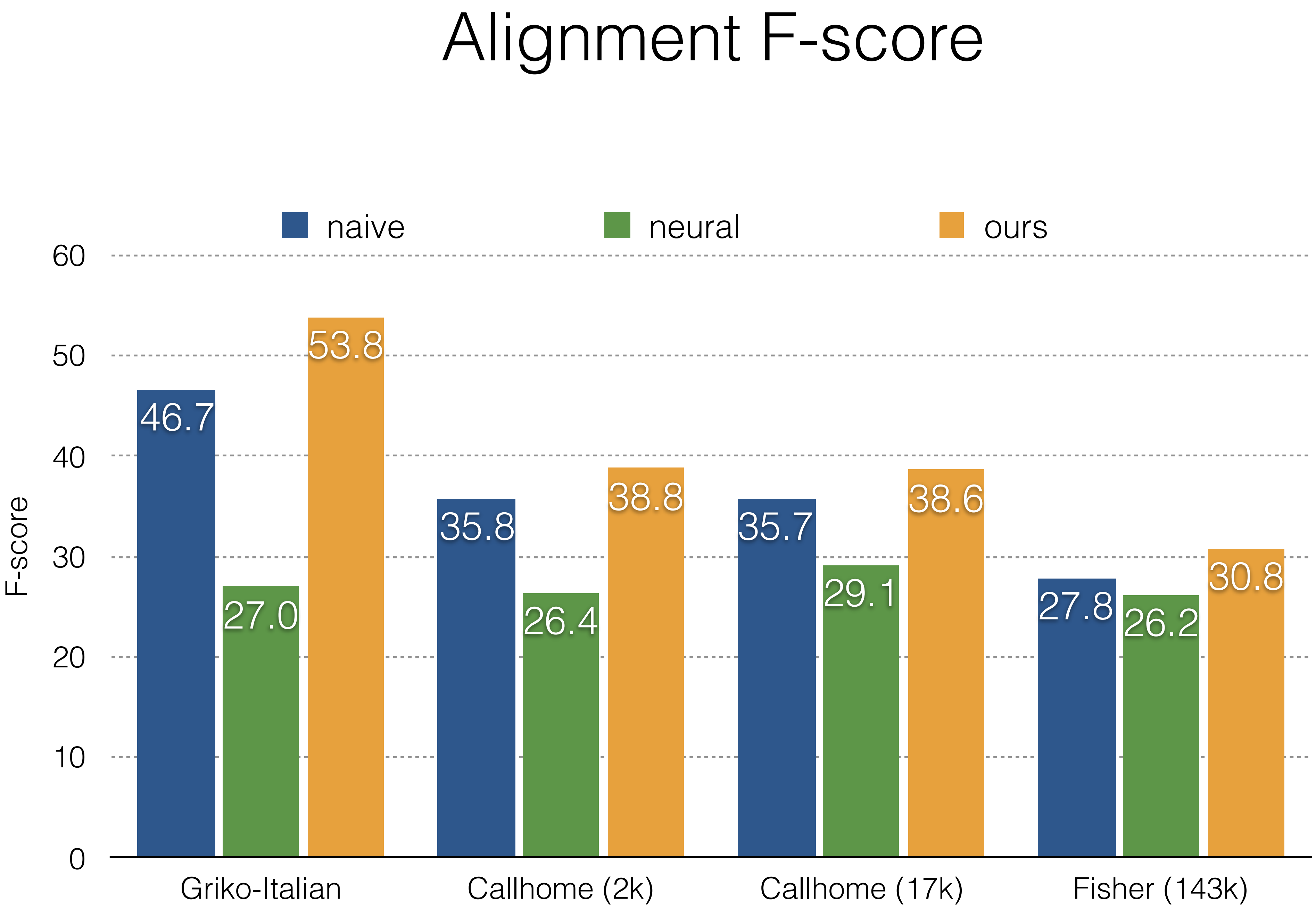
Results


Alignment F-score
naive
neural
ours
F-score
60
50
40
30
20
10
0
46.7
35.8
35.7
27.8
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results


Alignment F-score
naive
neural
ours
F-score
60
50
40
30
20
10
0
46.7
27.0
35.8
26.4
35.7
29.1
27.8
26.2
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results



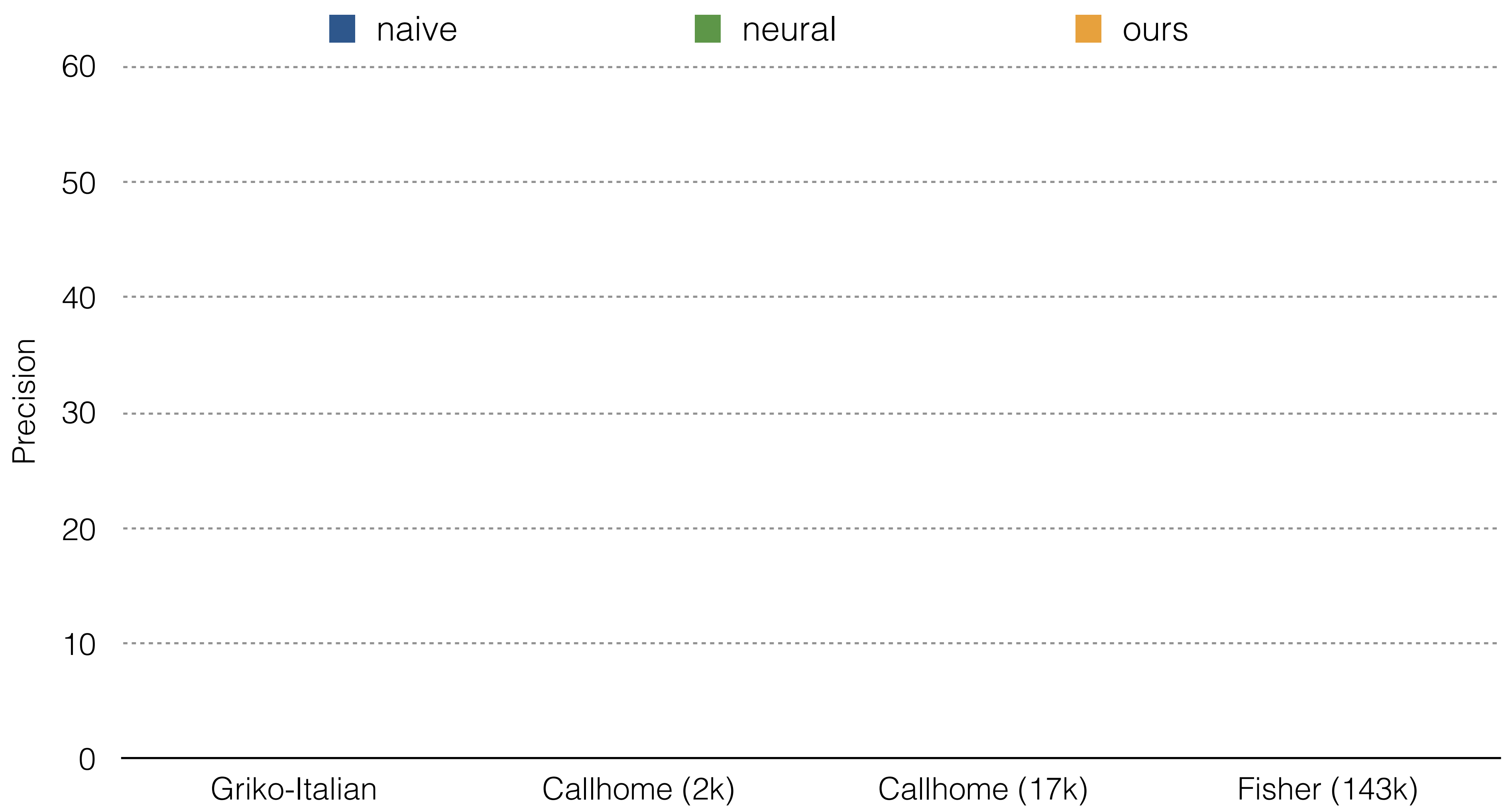
Alignment F-score
naive
neural
ours
F-score
60
50
40
30
20
10
0
46.7
27.0
53.8
35.8
26.4
38.8
35.7
29.1
38.6
27.8
26.2
30.8
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results

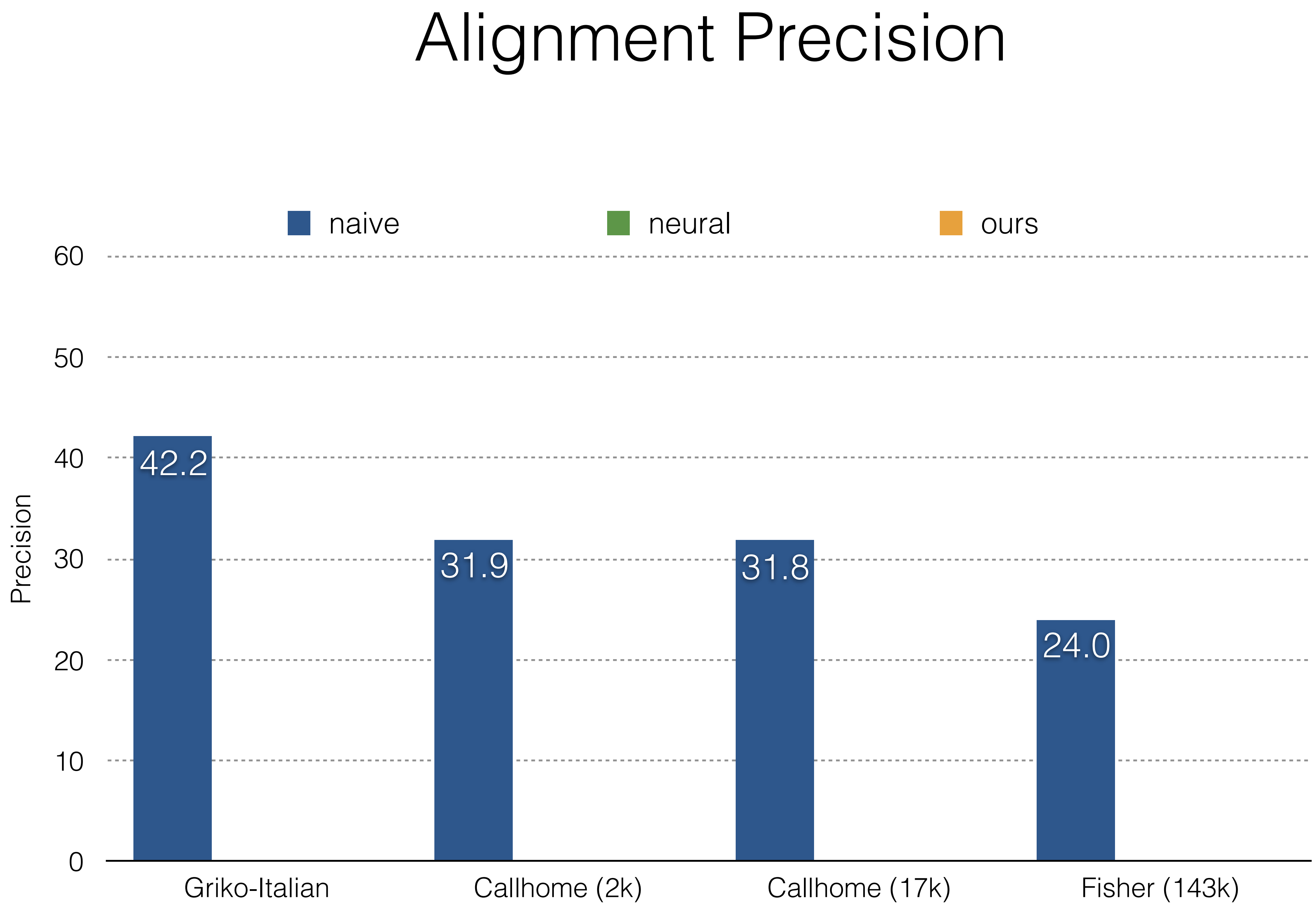
Alignment Precision

Results

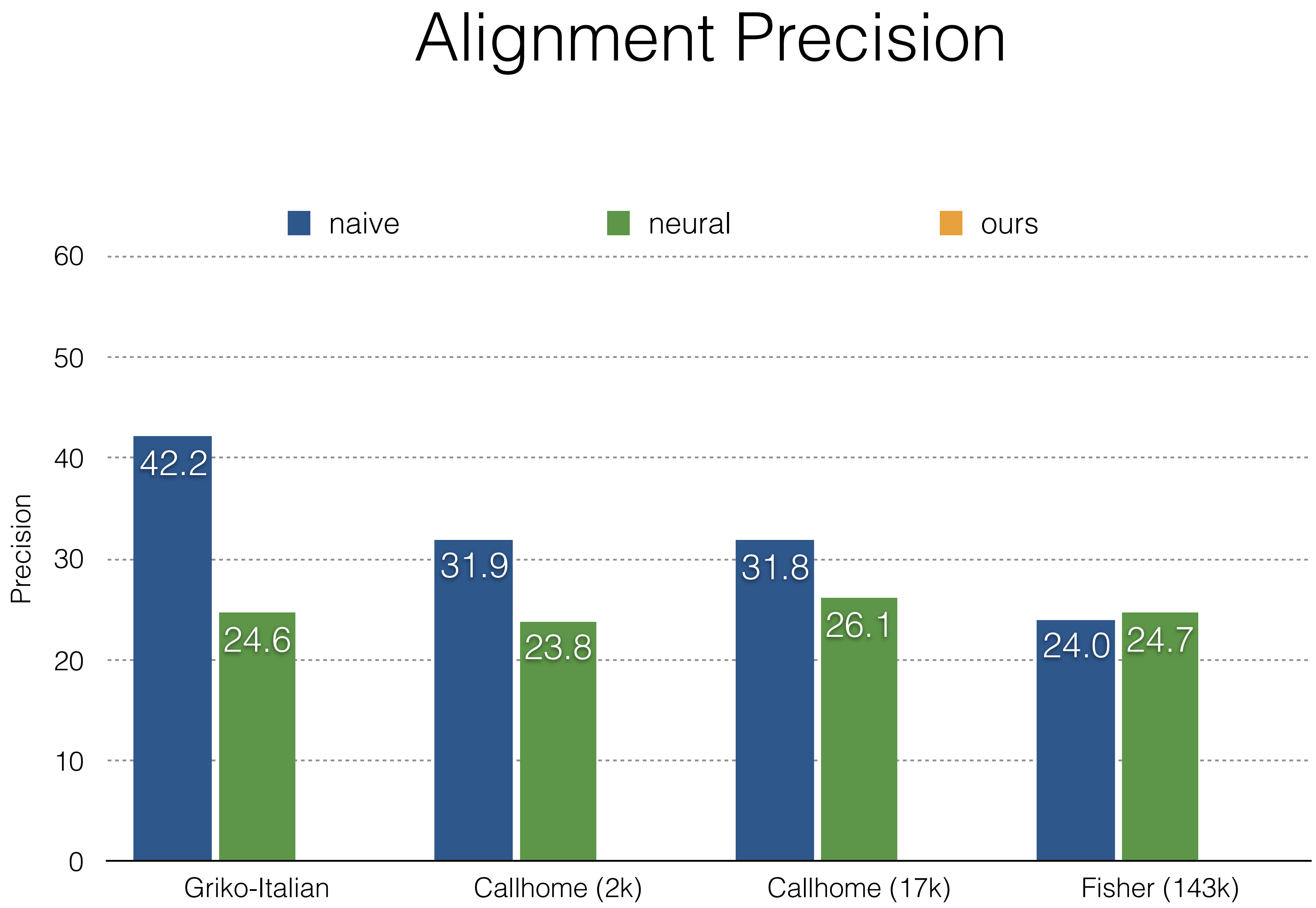
Alignment Precision



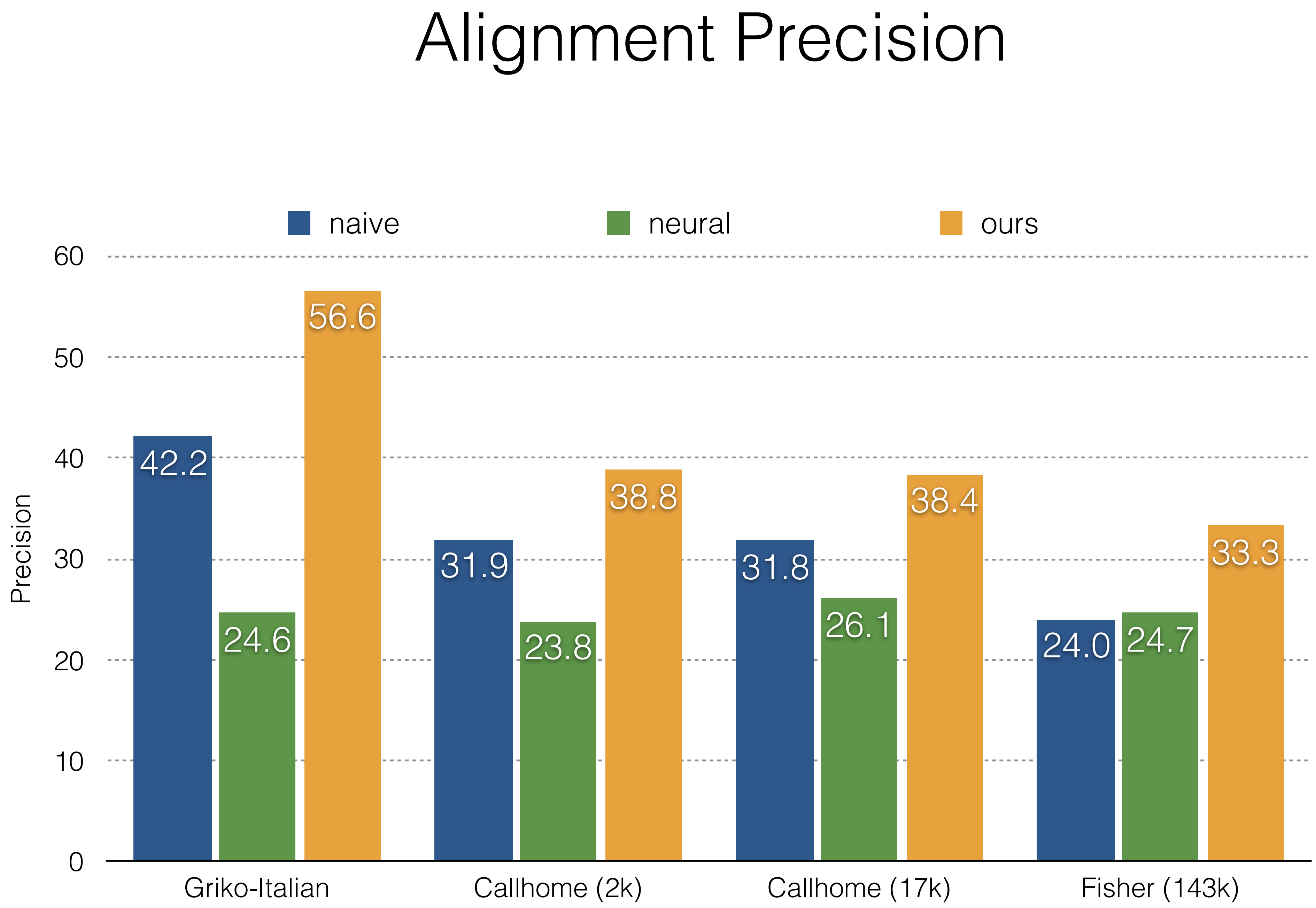
Results


Alignment Precision
naive
neural
ours
Precision
60
50
40
30
20
10
0
42.2
31.9
31.8
24.0
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results

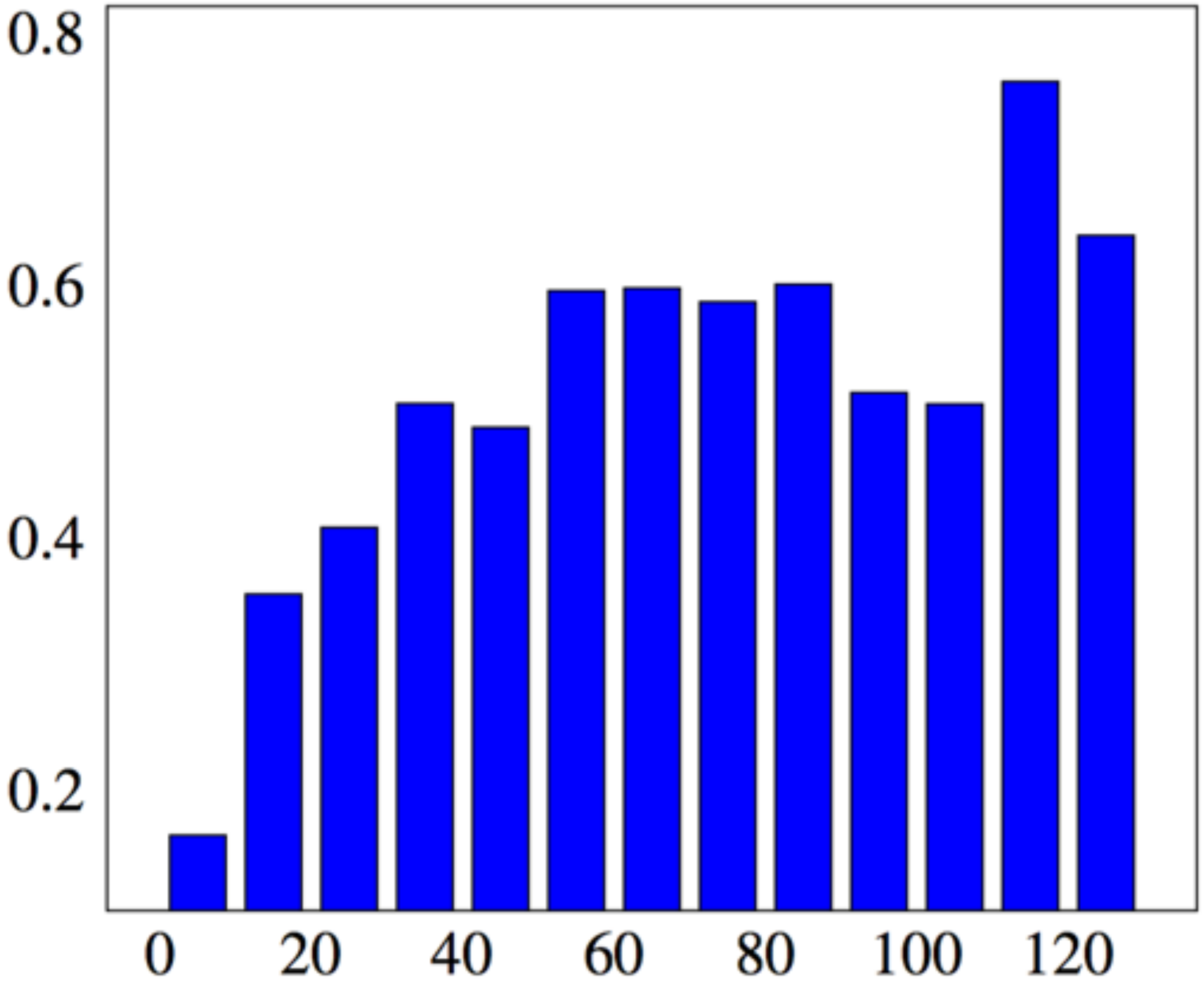

Alignment Precision
naive
neural
ours
Precision
60
50
40
30
20
10
0
42.2
24.6
31.9
23.8
31.8
26.1
24.0
24.7
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results


Alignment Precision
naive
neural
ours
Precision
60
50
40
30
20
10
0
56.6
42.2
24.6
31.9
23.8
38.8
31.8
26.1
38.4
24.0
24.7
33.3
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results

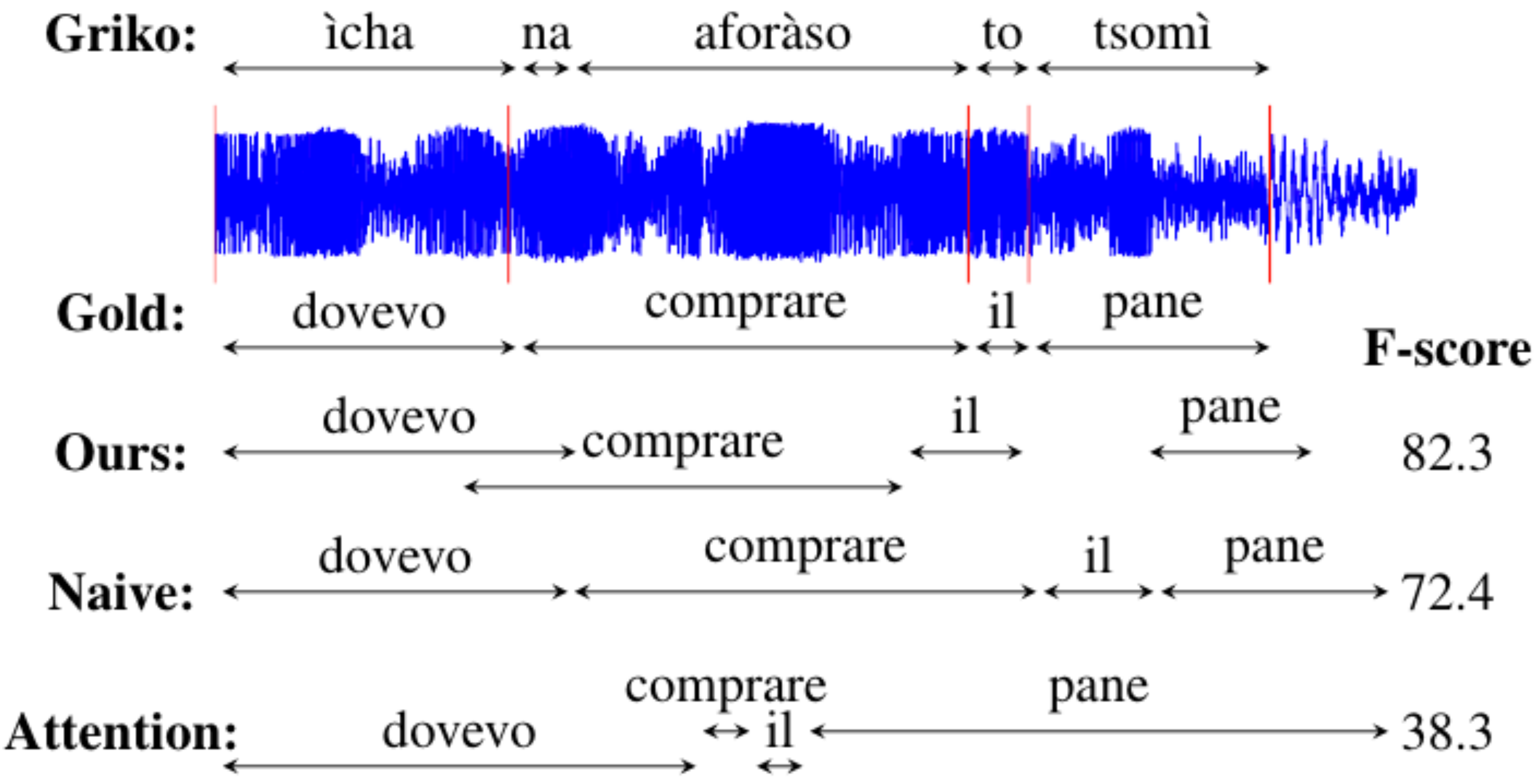
Word-level F-score



average F-score

Word Length (frames)

Example


Griko:
ìcha
na
aforàso
to
tsomì
Gold:
dovevo
comprare
il
pane
F-score
Ours:
dovevo
comprare
il
pane
82.3
Naive:
dovevo
comprare
il
pane
72.4
Attention:
dovevo
comprare
il
pane
38.3

Conclusion

Alignment model

Extension of IBM-2 with fast-align for speech-to-translation

k-means clustering with DTW and DBA

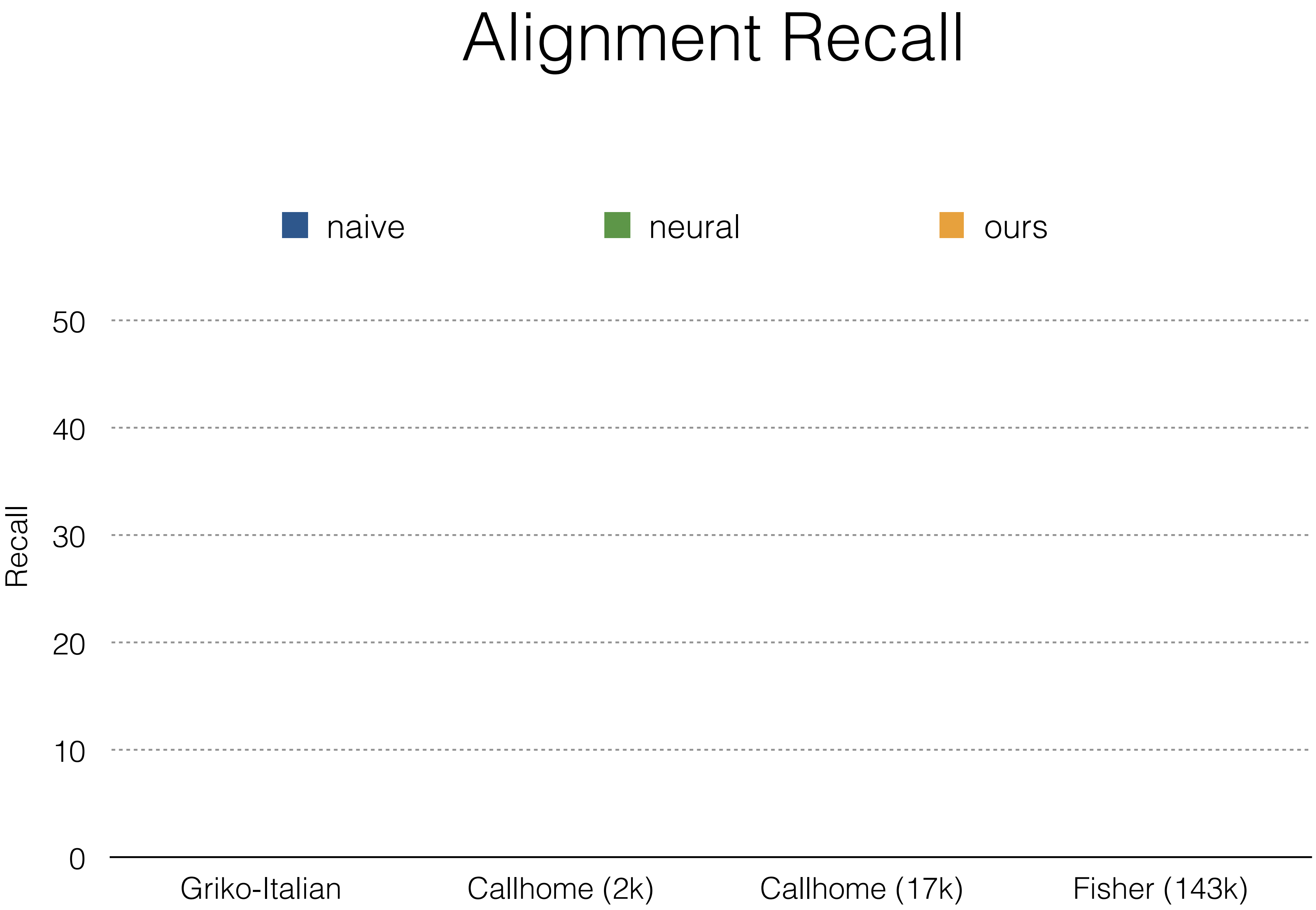
Improvements in F-score and particularly Precision

https://bitbucket.org/ndnlp/speech2translation

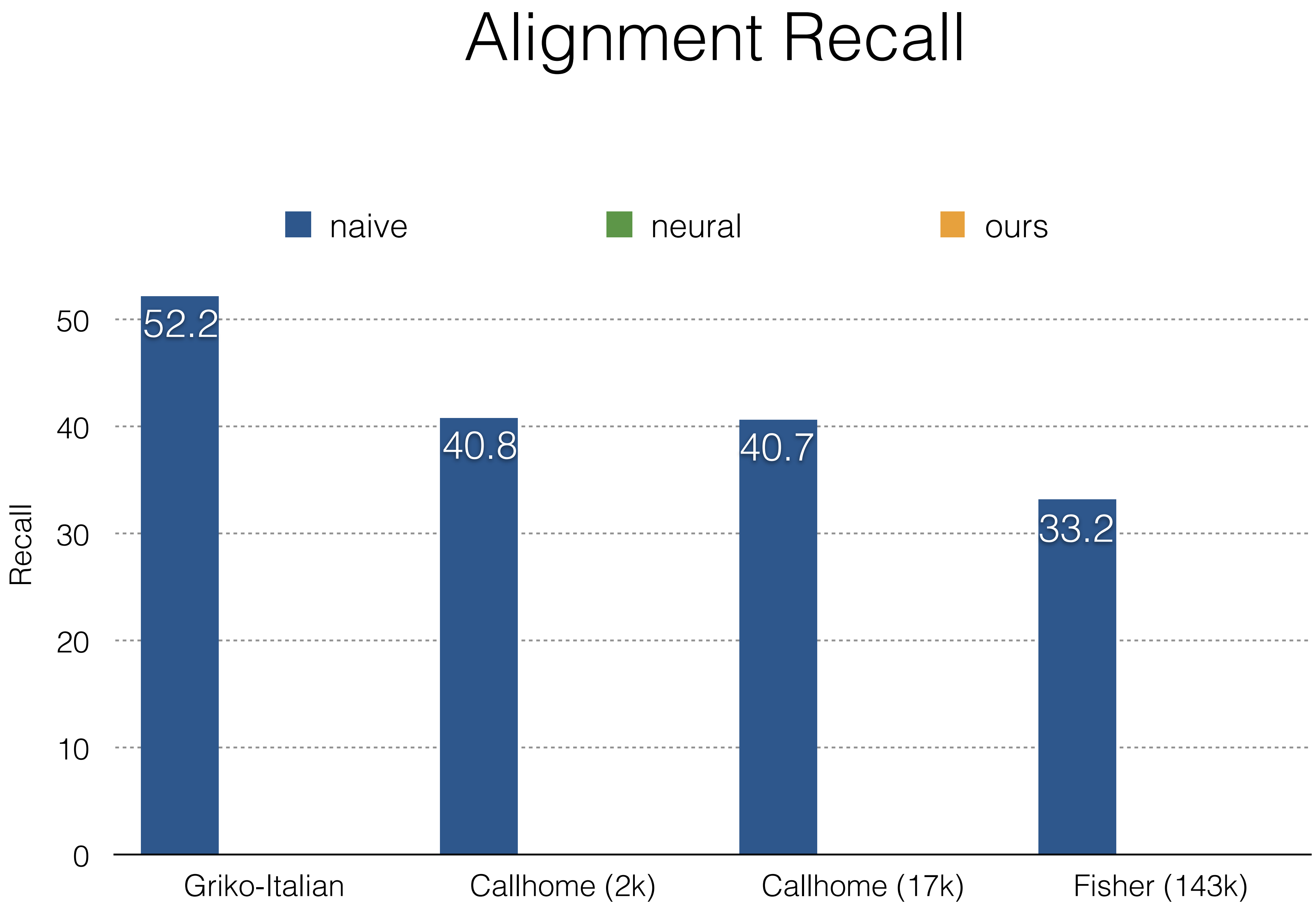
Results

Alignment Recall

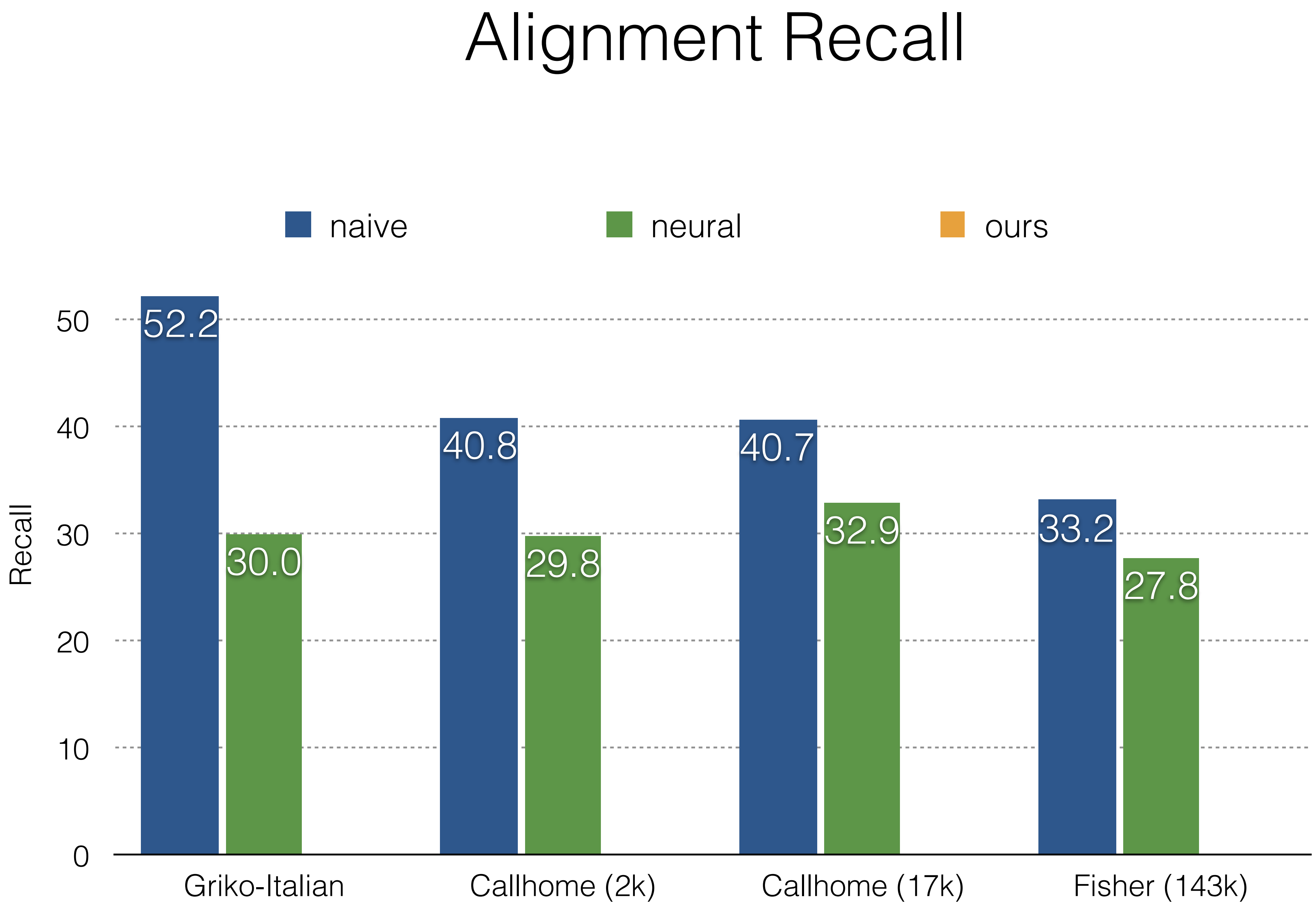
Results


Alignment Recall
🟦 naive
🟩 neural
🟧 ours
Recall
50
40
30
20
10
0
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

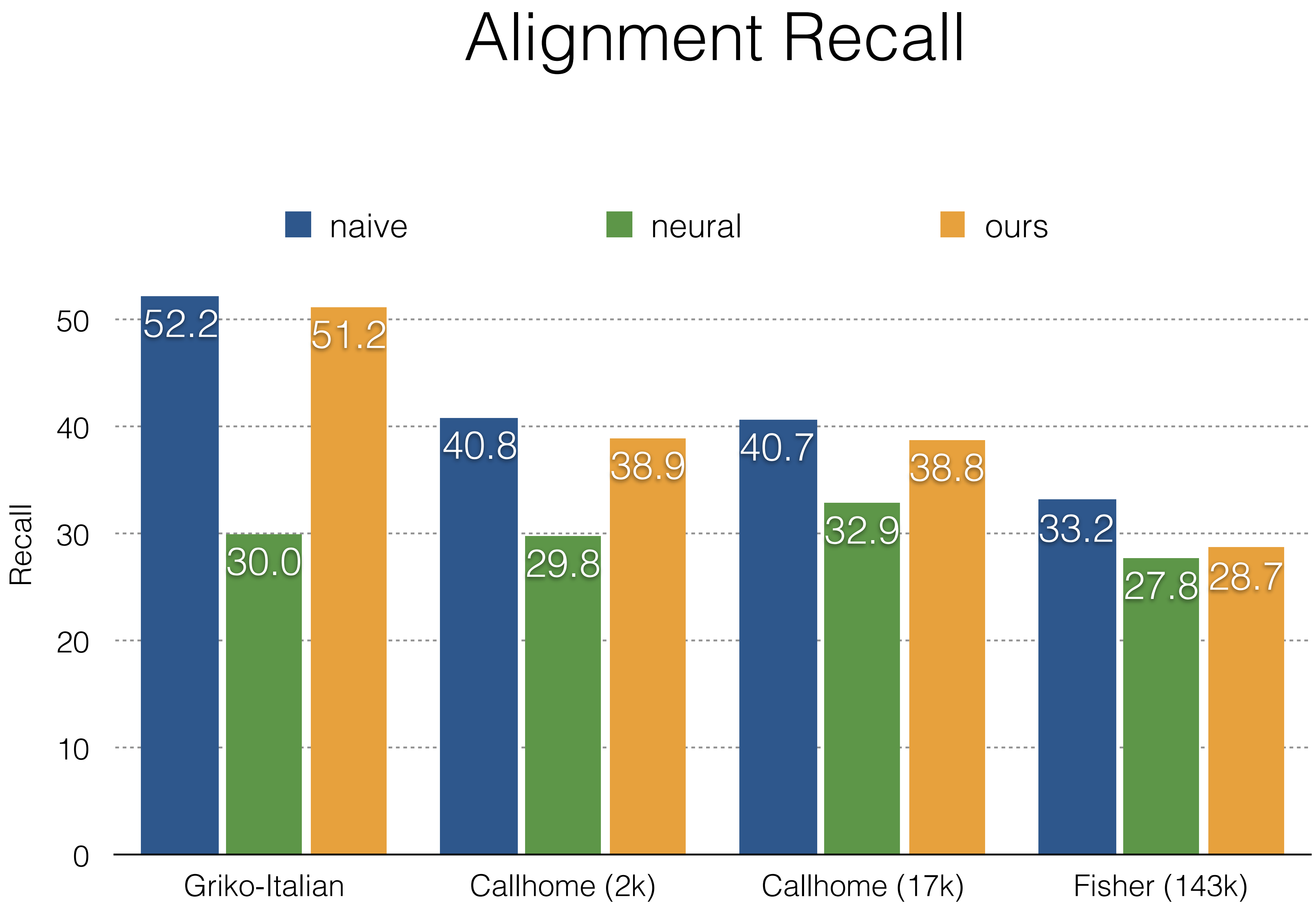
Results


Alignment Recall
naive
neural
ours
Recall
0
10
20
30
40
50
52.2
40.8
40.7
33.2
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results

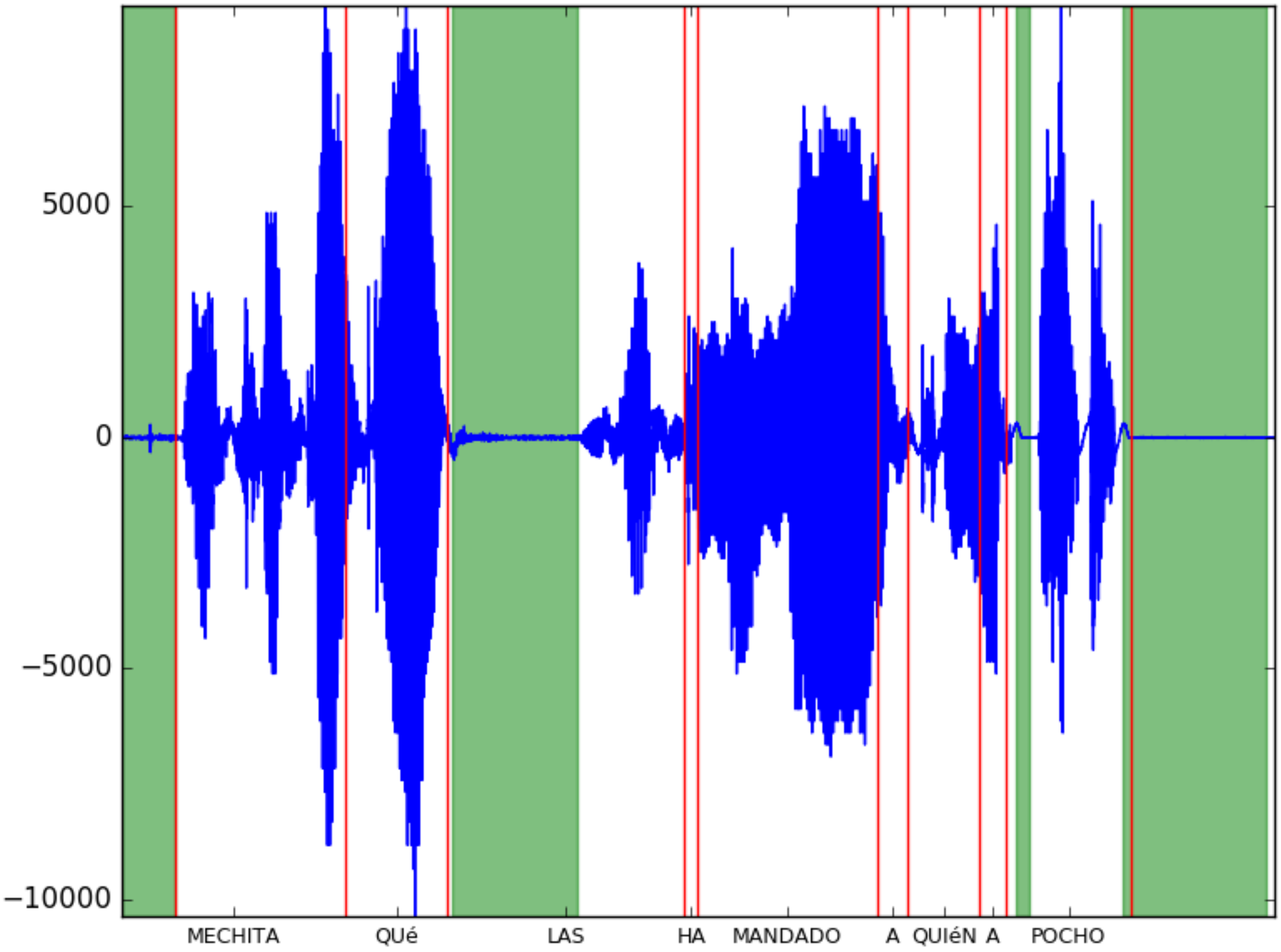

Alignment Recall
naive
neural
ours
Recall
50
40
30
20
10
0
52.2
30.0
40.8
29.8
40.7
32.9
33.2
27.8
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Results

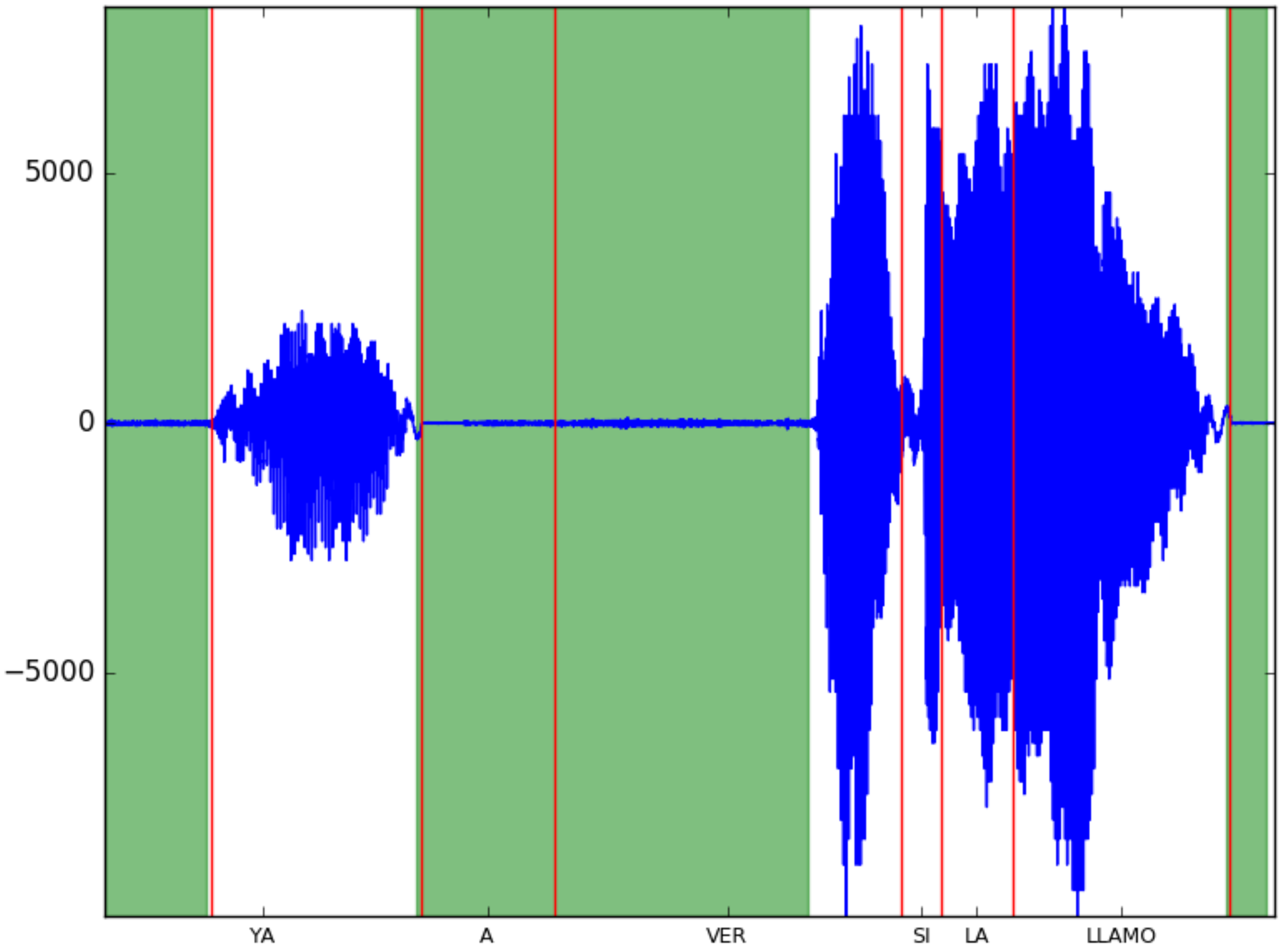


Alignment Recall
naive
neural
ours
Recall
0
10
20
30
40
50
52.2
30.0
51.2
40.8
29.8
38.9
40.7
32.9
38.8
33.2
27.8
28.7
Griko-Italian
Callhome (2k)
Callhome (17k)
Fisher (143k)

Example

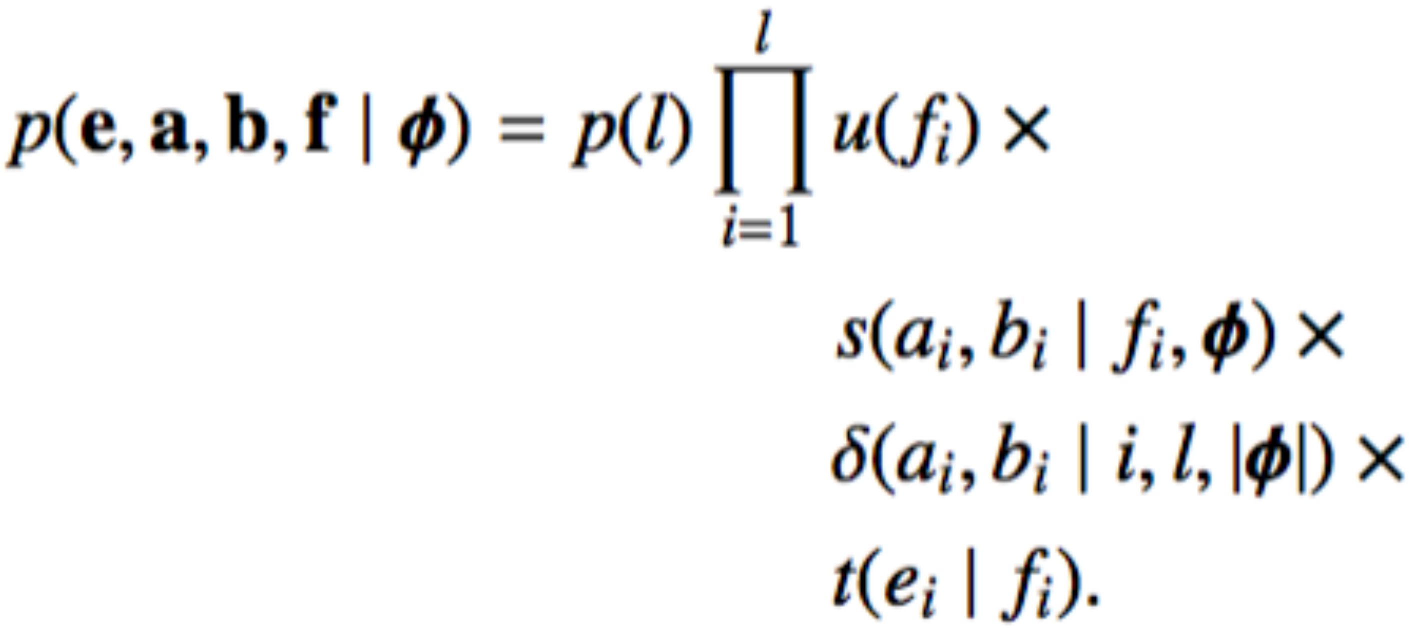


Example



Proper model

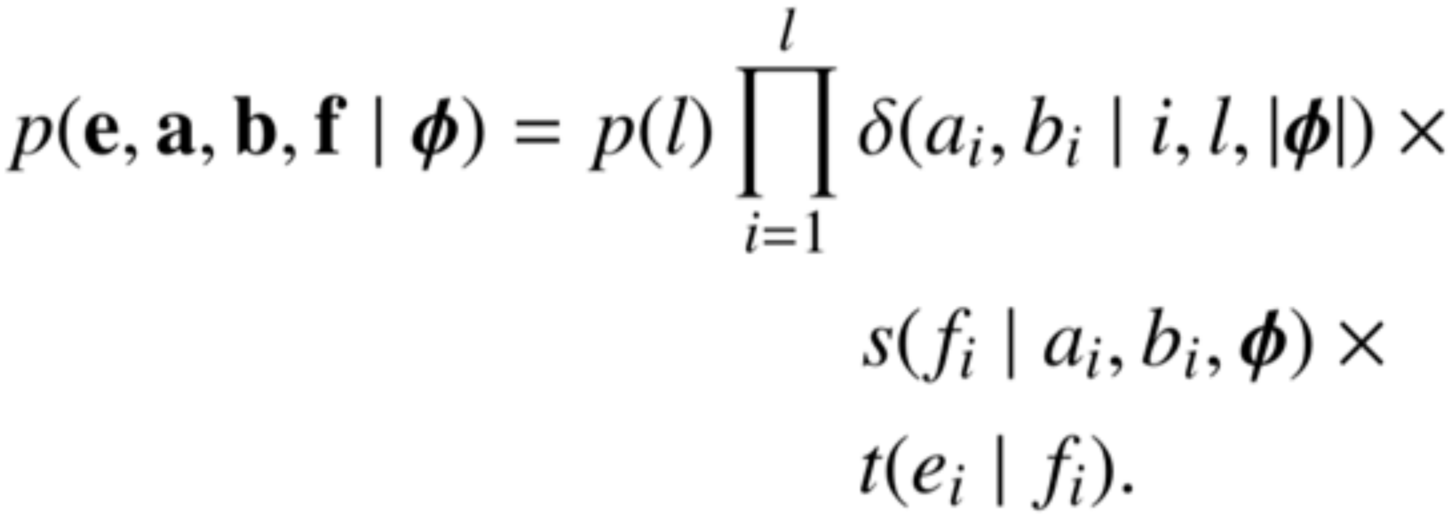
Deficient:



$$
\begin{aligned}
p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi}) = p(l) \prod_{i=1}^{l} & u(f_i) \times \\
& s(a_i, b_i \mid f_i, \boldsymbol{\phi}) \times \\
& \delta(a_i, b_i \mid i, l, |\boldsymbol{\phi}|) \times \\
& t(e_i \mid f_i).
\end{aligned}
$$

Proper model

Proper:



$$p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi}) = p(l) \prod_{i=1}^{l} \delta(a_i, b_i \mid i, l, |\boldsymbol{\phi}|) \times$$

$$s(f_i \mid a_i, b_i, \boldsymbol{\phi}) \times$$

$$t(e_i \mid f_i).$$

Proper model

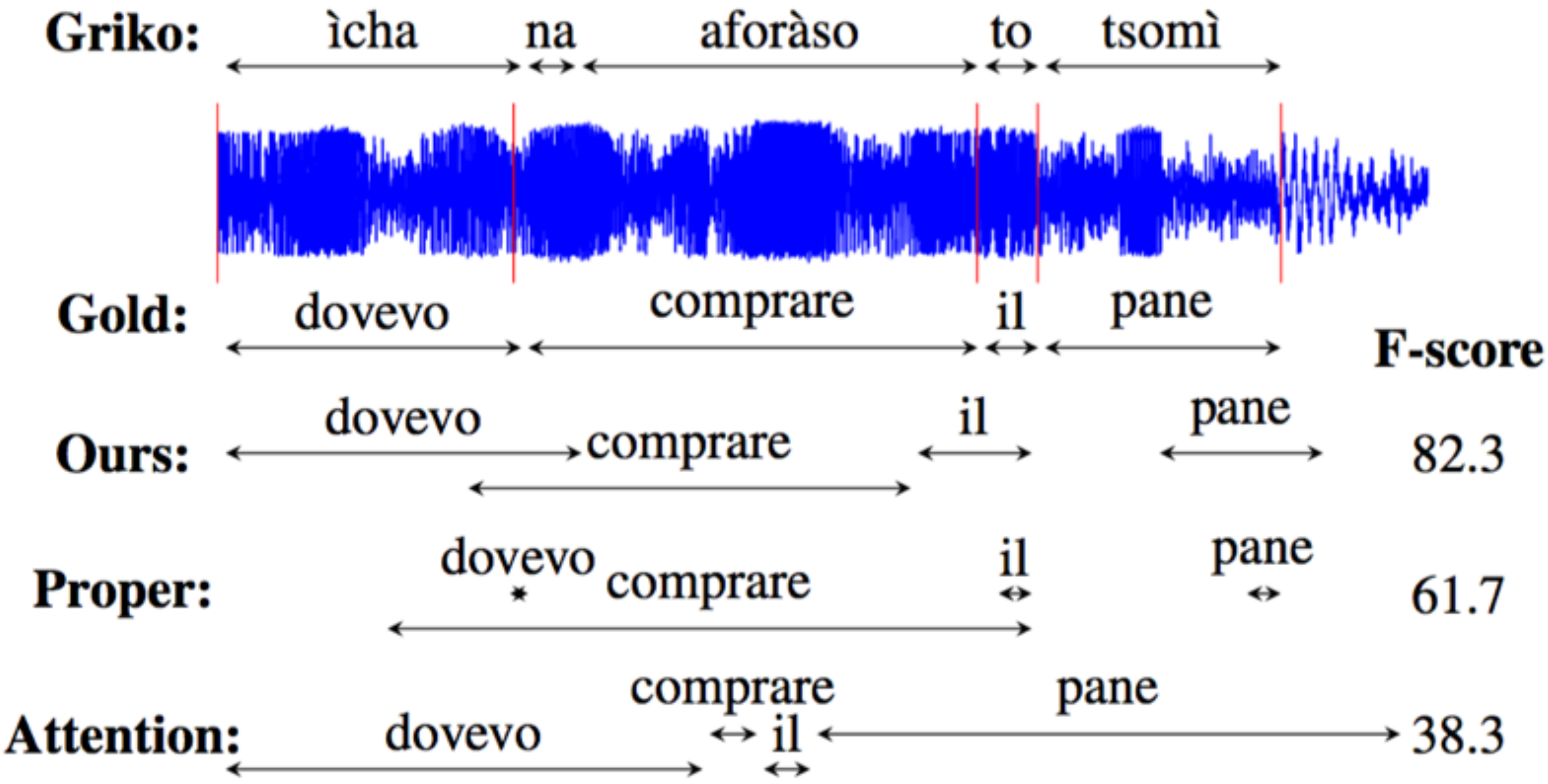
Proper:

$$p(\mathbf{e}, \mathbf{a}, \mathbf{b}, \mathbf{f} \mid \boldsymbol{\phi}) = p(l) \prod_{i=1}^{l} \delta(a_i, b_i \mid i, l, |\boldsymbol{\phi}|) \times s(f_i \mid a_i, b_i, \boldsymbol{\phi}) \times t(e_i \mid f_i).$$

The proper model performs much worse.

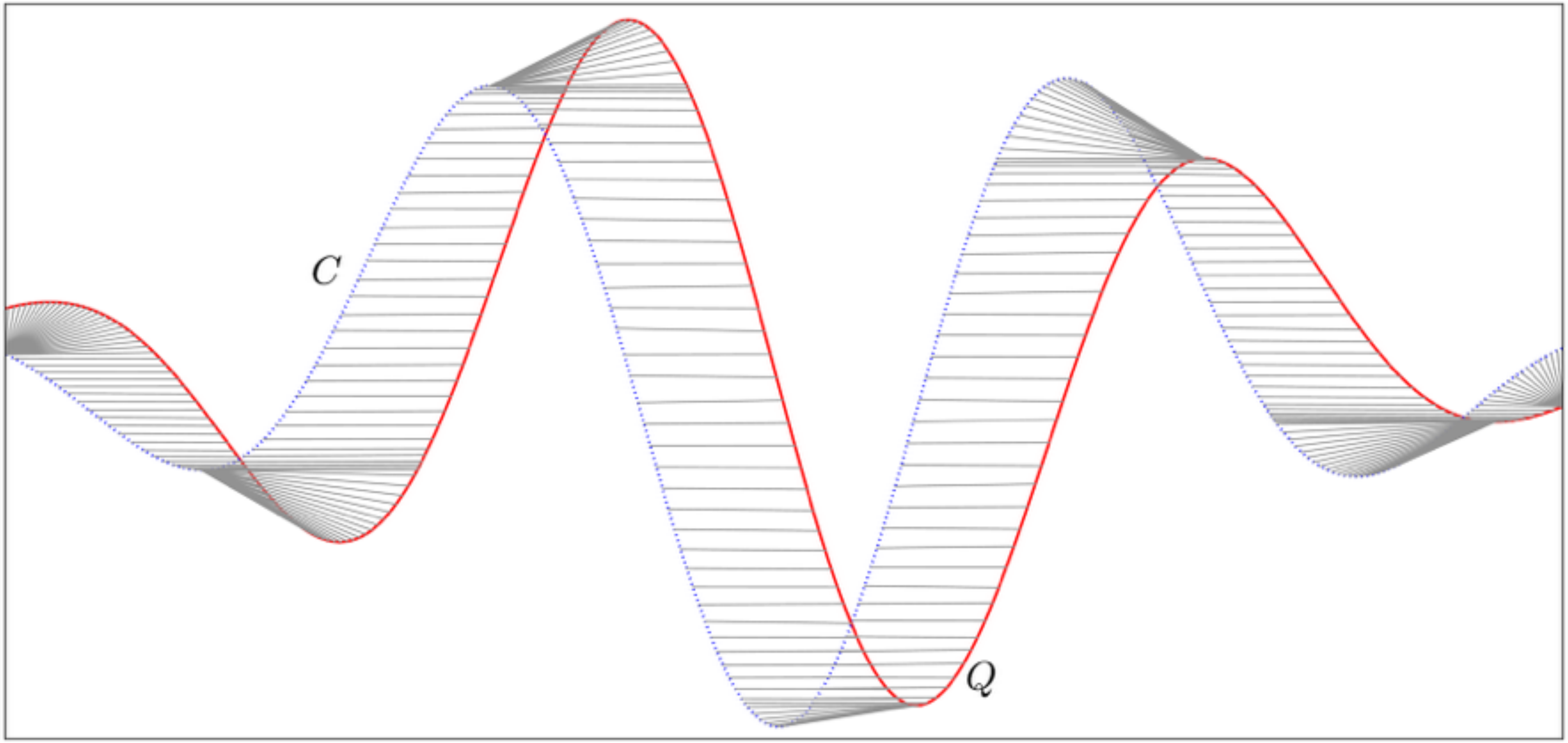
-It favours too long or too short spans

Example


Griko:
ìcha
na
aforàso
to
tsomì
Gold:
dovevo
comprare
il
pane
F-score
Ours:
dovevo
comprare
il
pane
82.3
Proper:
dovevo
comprare
il
pane
61.7
Attention:
dovevo
comprare
il
pane
38.3

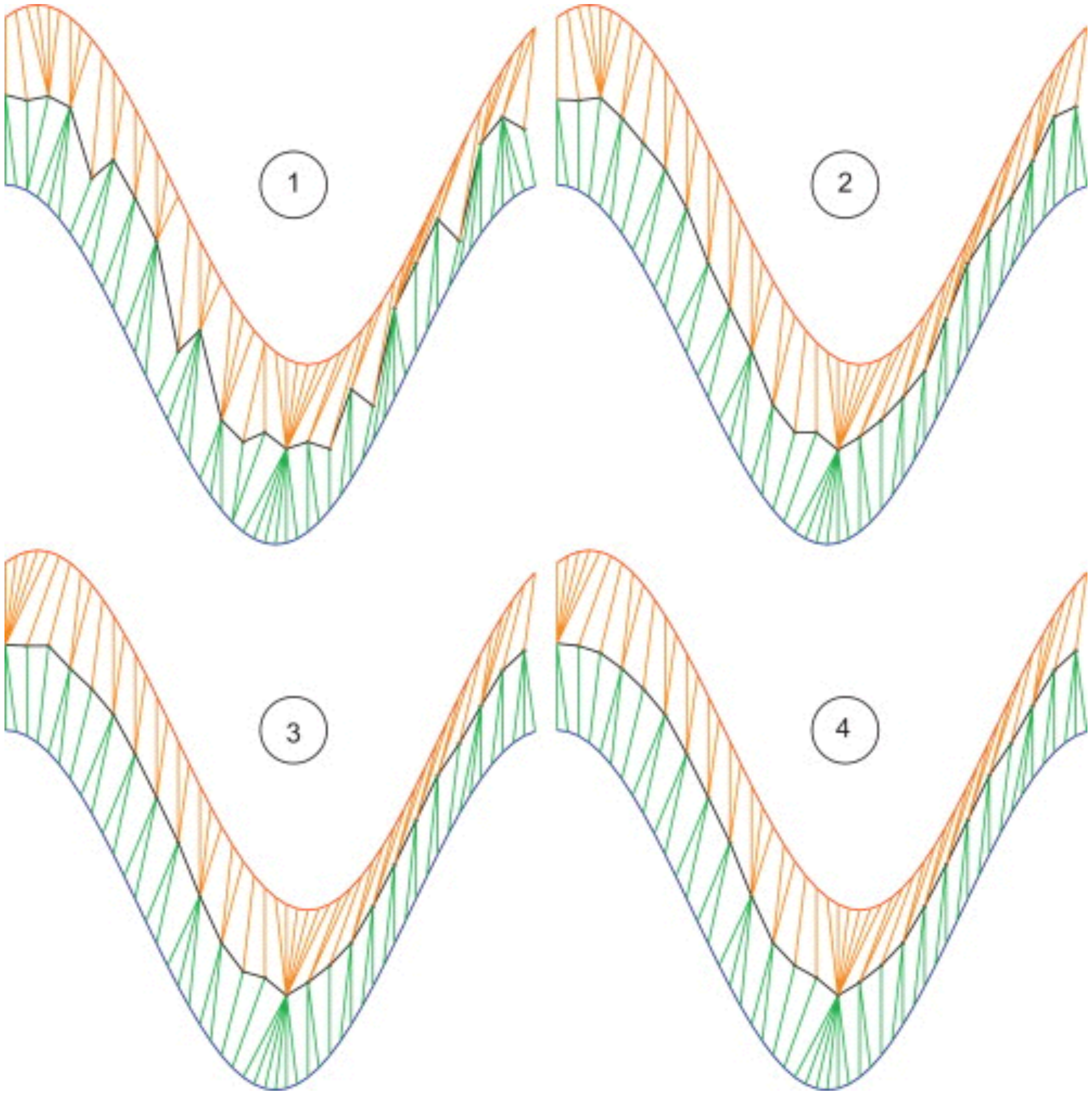
Background: DTW and DBA

Dynamic Time Warping (DTW)



Background: DTW and DBA

DTW Barycenter Averaging (DBA)

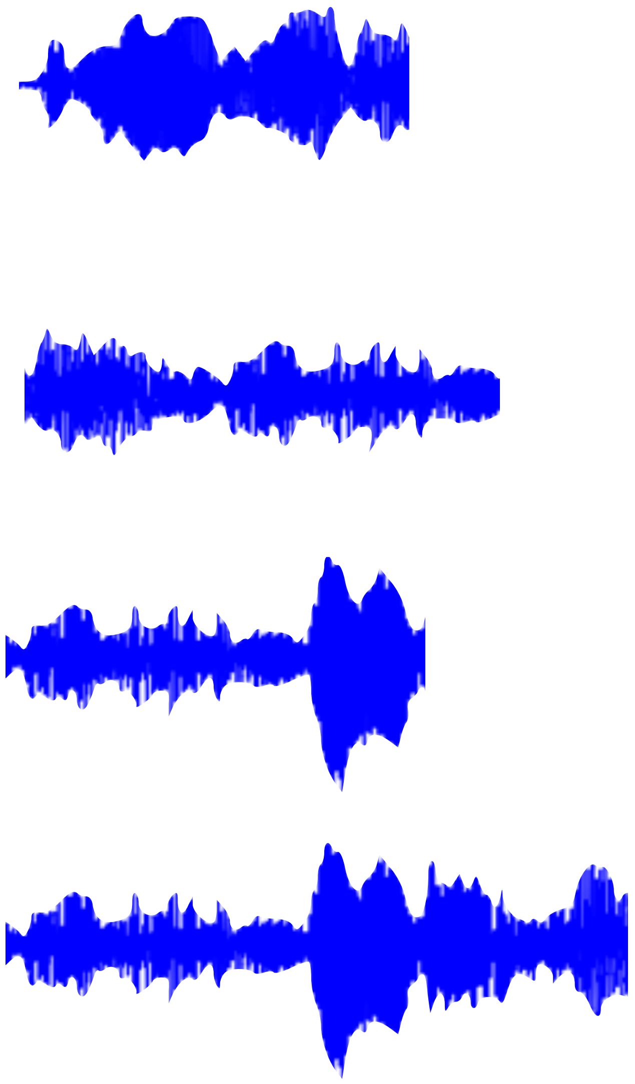


M-step:

Prototype estimation with DTW Barycenter Averaging

M-step:

Prototype estimation with DTW Barycenter Averaging



M-step:

Prototype estimation with DTW Barycenter Averaging