Spoken Term Discovery for Language Documentation using Translations

Antonios Anastasopoulos, Sameer Bansal, Sharon Goldwater, Adam Lopez, and David Chiang

Funding: JS McDonnell Foundation, NSF

Steps in documentation

1. Data collection
2. Manual annotation
3. Automatic annotation/analysis

Steps in documentation

1. Data collection
2. Manual annotation
3. Automatic annotation/analysis

Steps in documentation

1. Data collection
2. Manual annotation
3. Automatic annotation/analysis

Target scenario

Gila aburun ferma hamišaluğ güğüna amuq’dač

Target scenario

Gila aburun ferma hamišaluğ güğüna amuq’dač

Target scenario

Gila aburun ferma hamišaluğ güğüna amuq’dač

Now their farm will not stay behind forever.

Target scenario

Gila aburun ferma hamišaluğ güğüna amuq’dač

Now their farm will not stay behind forever.

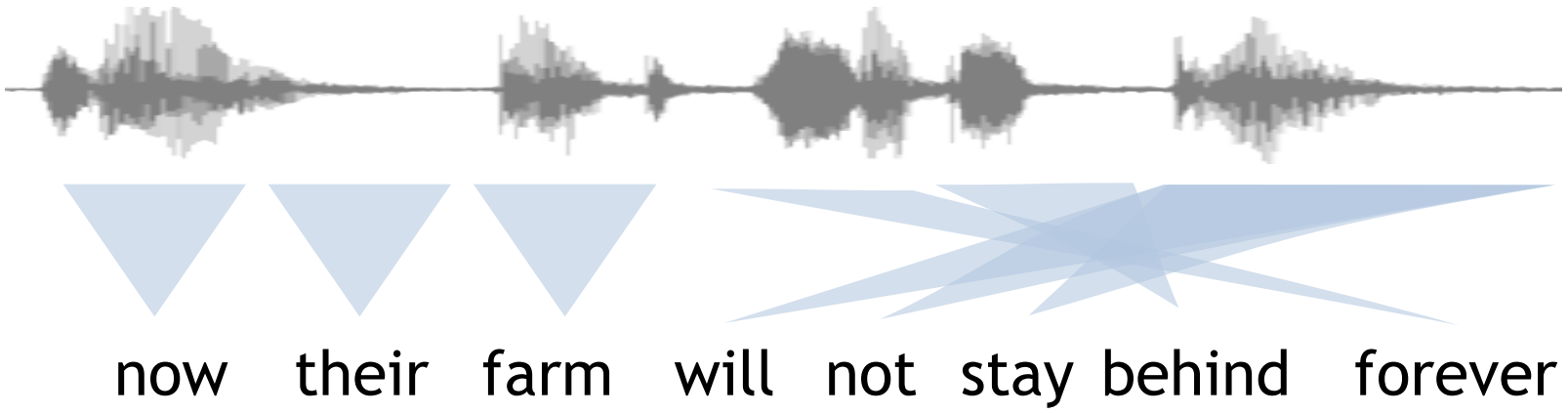
Target scenario

Gila aburun ferma hamišaluğ güğüna amuq’dač

Now their farm will not stay behind forever.

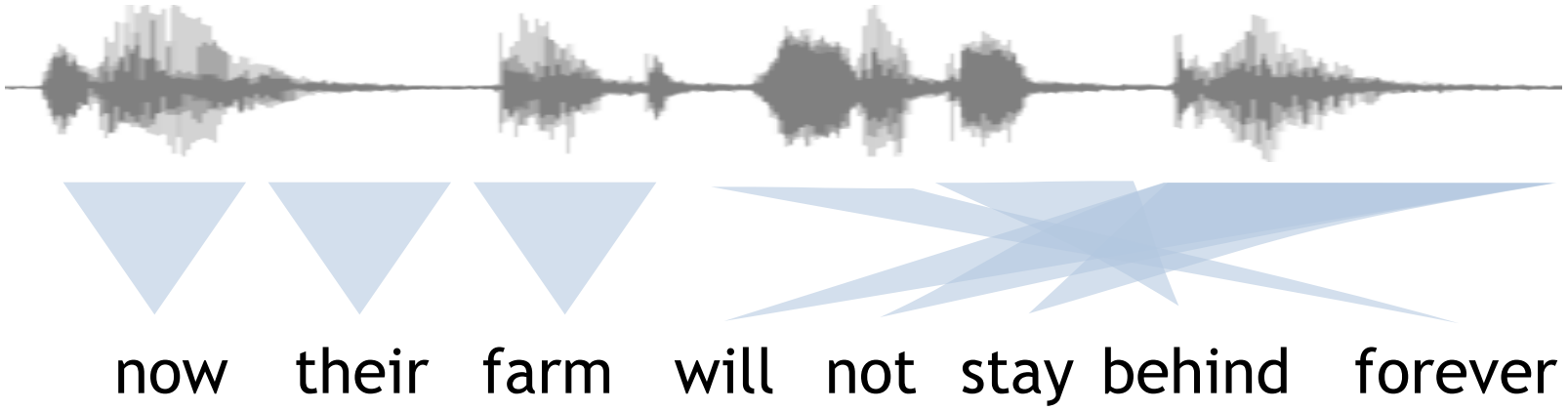
Beyond just alignments

Anastasopoulos et al. (2016), Duong et al. (20

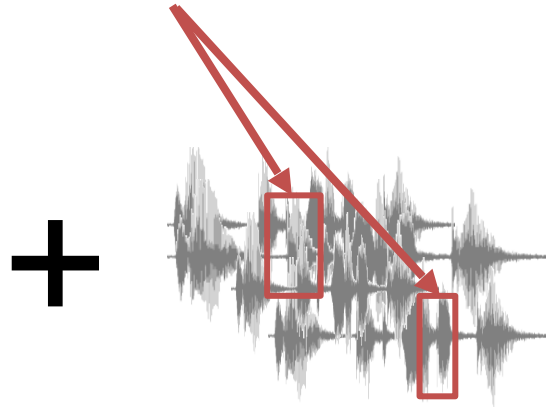


Beyond just alignments

- Previous approaches: Anastasopoulos et al. (2016), Duong et al. (20



- Here:



Realistic scenario

- Predictions on natural speech data
  - Not phones or lattices (Adams et al., 2016; Godard et al., 2016)
  - Not synthetic speech (Bérard et al., 2016)
- Using real endangered language recordings
  - Ainu (~2.5hrs) and Arapaho (~1hr)
  - Plus Spanish (~20hrs)
- Can predict a few terms with ok precision, better than (the only) previous approach.

Baseline system (utd-align)

Bansal et al. (2017

Input data:

the man saw a dog near the tall tree

the dog ran to the man riding a horse

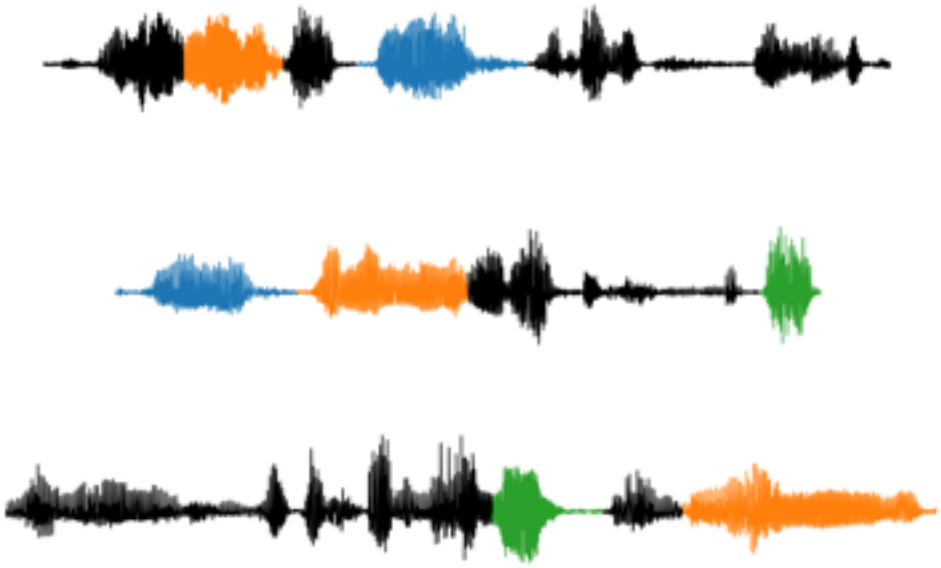
With translations

Speech only

Baseline system (utd-align)

Bansal et al. (2017

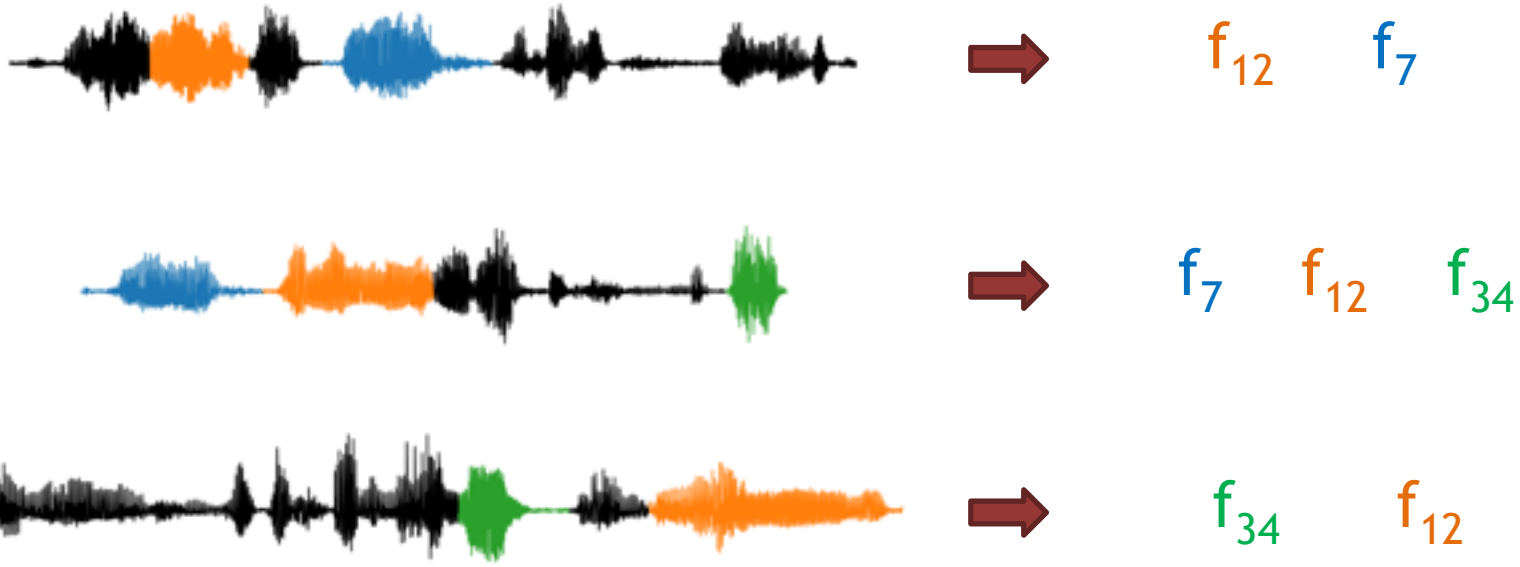
1-2. Segment and cluster (UTD)



Baseline system (utd-align)

Bansal et al. (2017

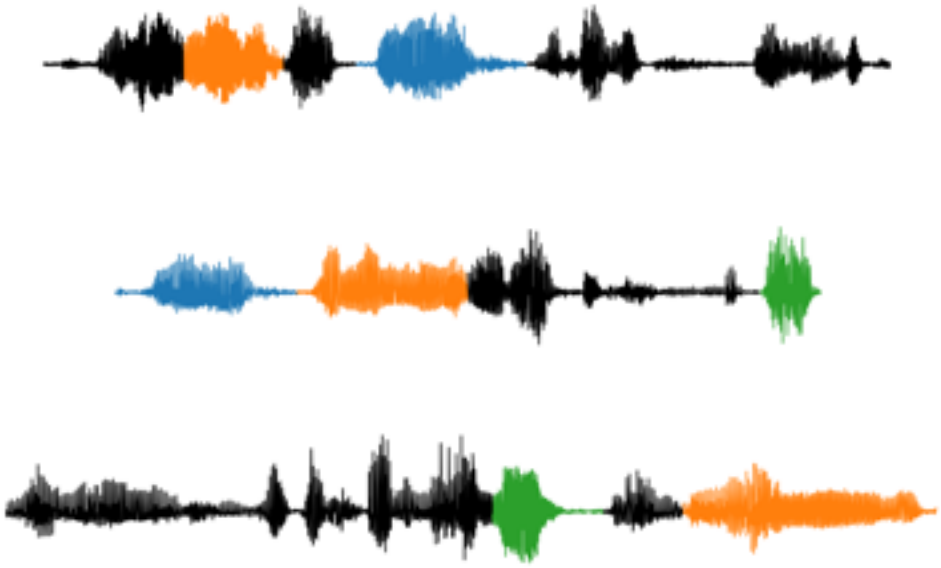
1-2. Segment and cluster (UTD)



Baseline system (utd-align)

Bansal et al. (2017

1-2. Segment and cluster (UTD)



f_{12} f_7

man saw dog tall tree

f_7 f_{12} f_{34}

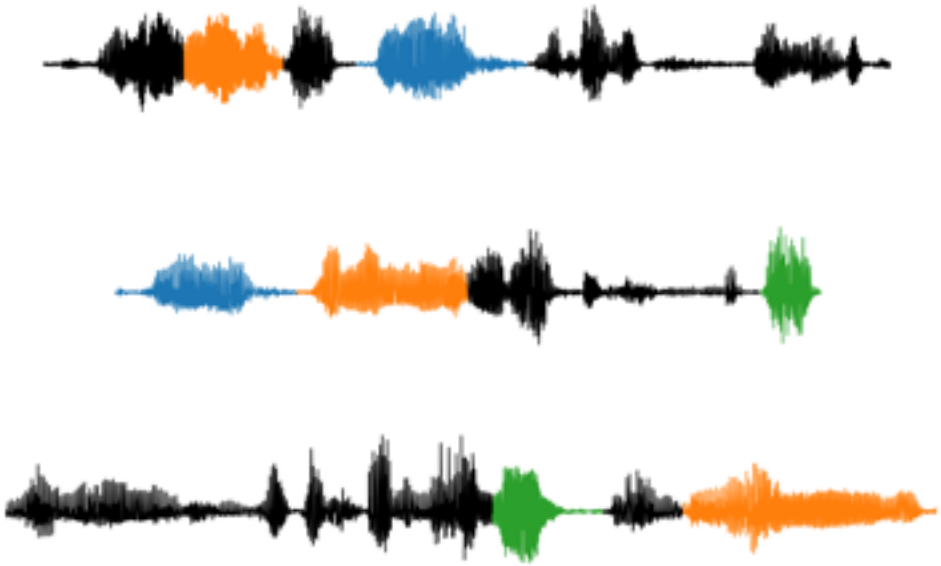
dog ran man riding horse

f_{34} f_{12}

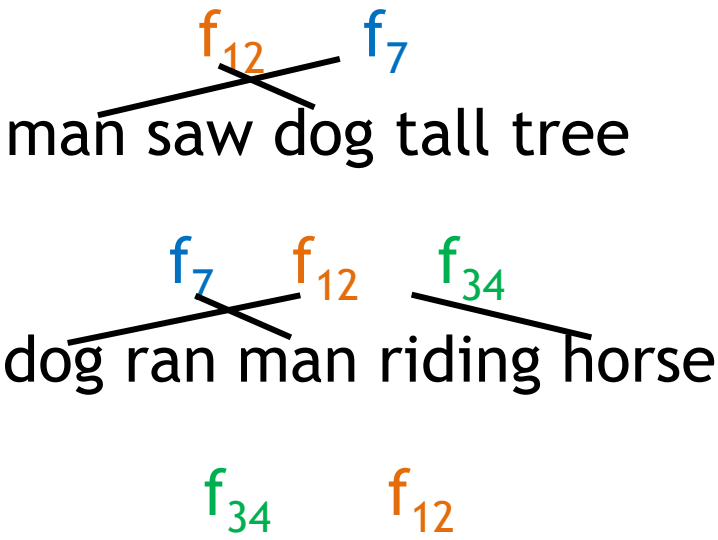
Baseline system (utd-align)

Bansal et al. (2017

1-2. Segment and cluster (UTD)



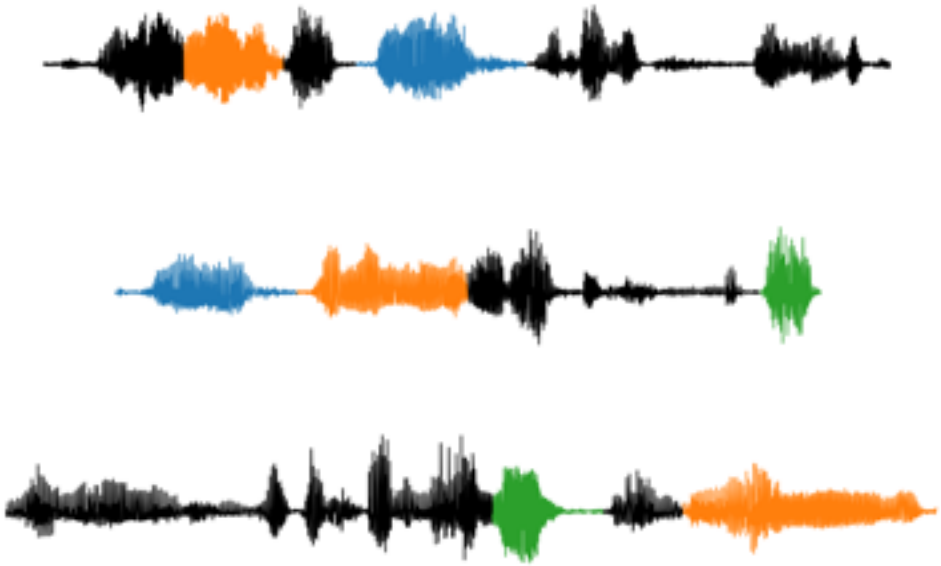
3. Align clusters to words



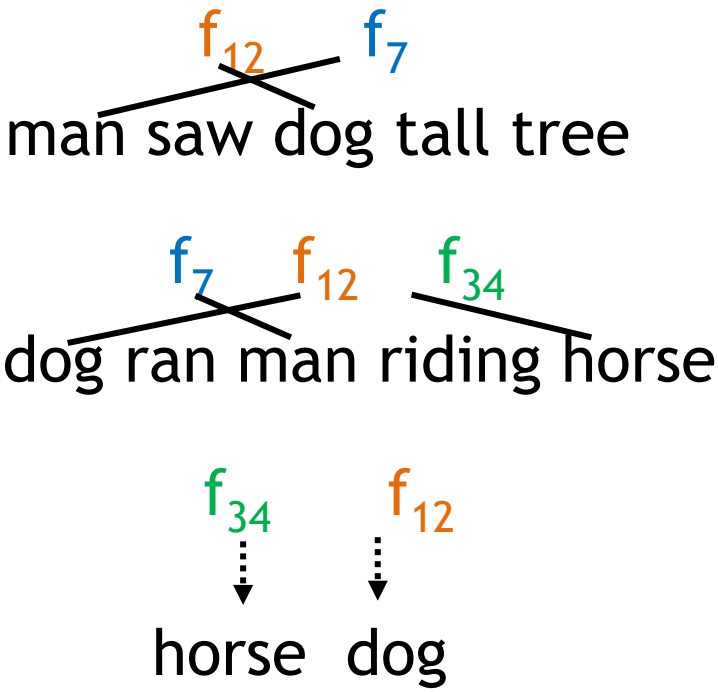
Baseline system (utd-align)

Bansal et al. (2017

1-2. Segment and cluster (UTD)



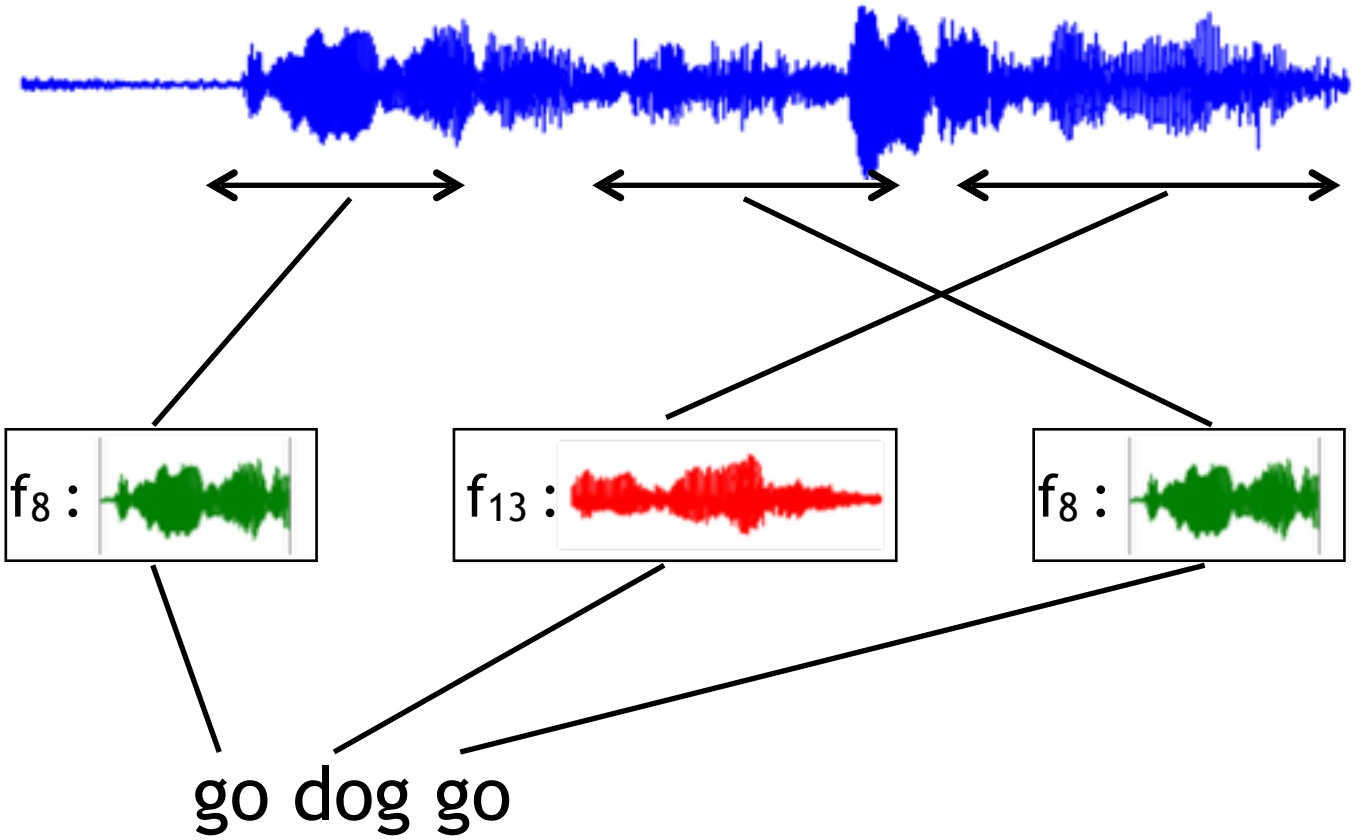
3. Align clusters to words



4. Predict translations for unaligned clusters

Here: joint system plus predictions

- Builds on previous work that jointly learns to segment, cluster, and align (Anastasopoulos et al., 2016)



Training using EM

Training using EM

go dog go

Training using EM

Initialize spans and clusters

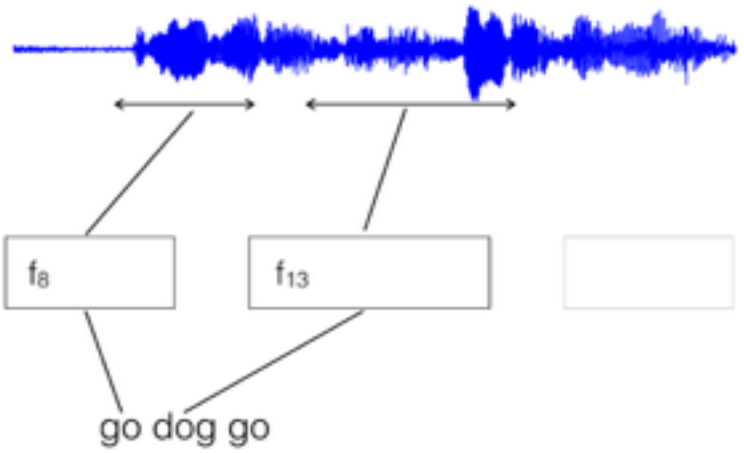
go dog go

Training using EM

Initialize spans and clusters

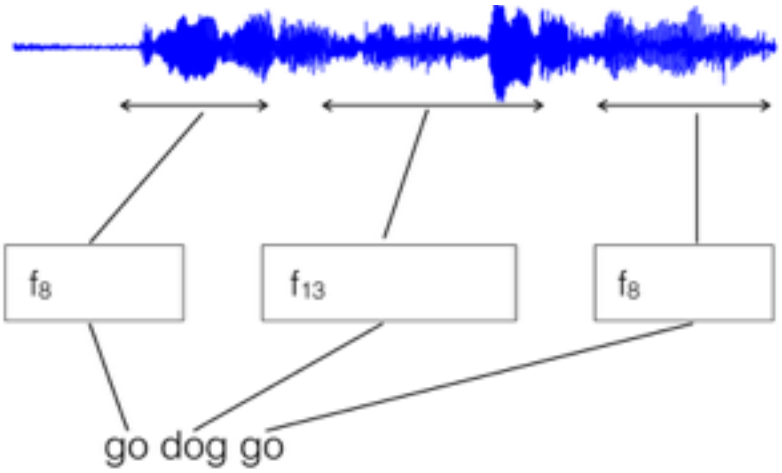
Training using EM

Initialize spans and clusters



Training using EM

Initialize spans and clusters

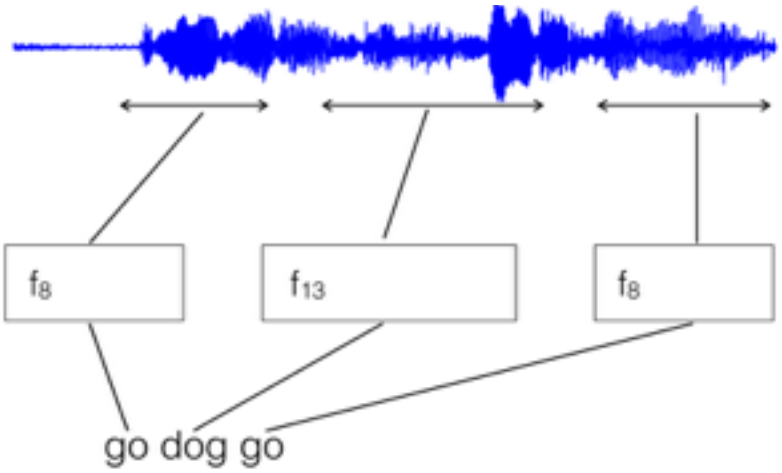


Training using EM

Initialize spans and clusters

M step:

- Re-estimate prototypes

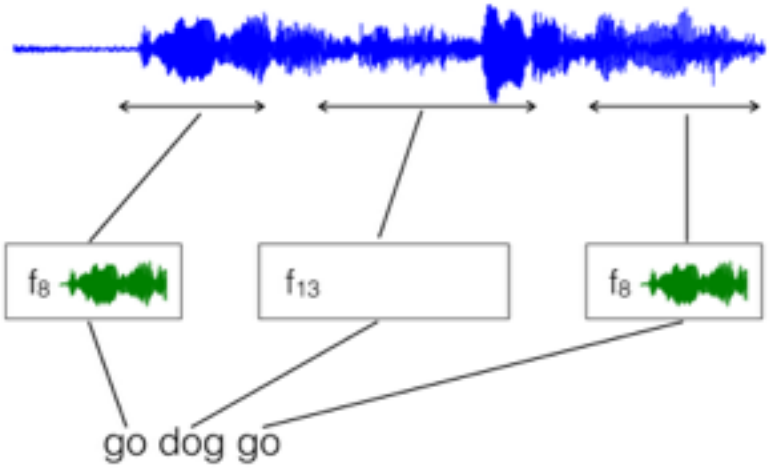


Training using EM

Initialize spans and clusters

M step:

- Re-estimate prototypes

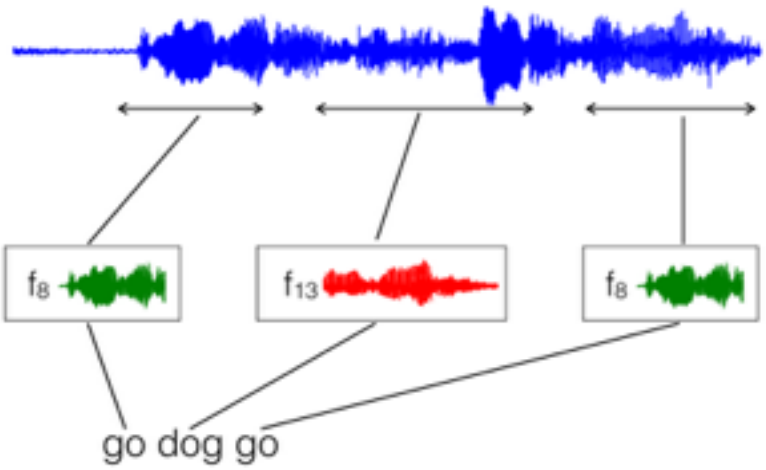


Training using EM

Initialize spans and clusters

M step:

- Re-estimate prototypes



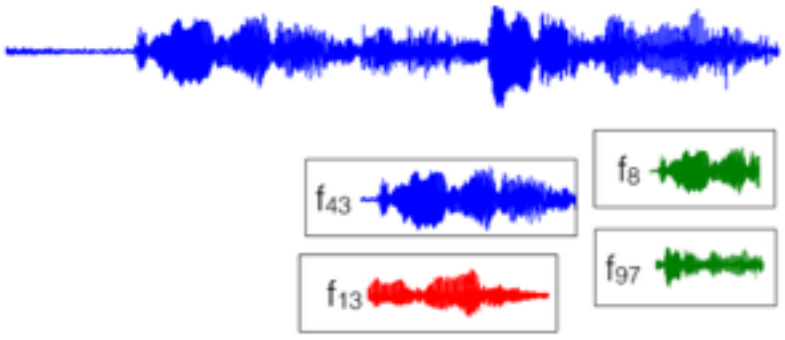
Training using EM

Initialize spans and clusters

M step:

- Re-estimate prototypes

E step:



Training using EM

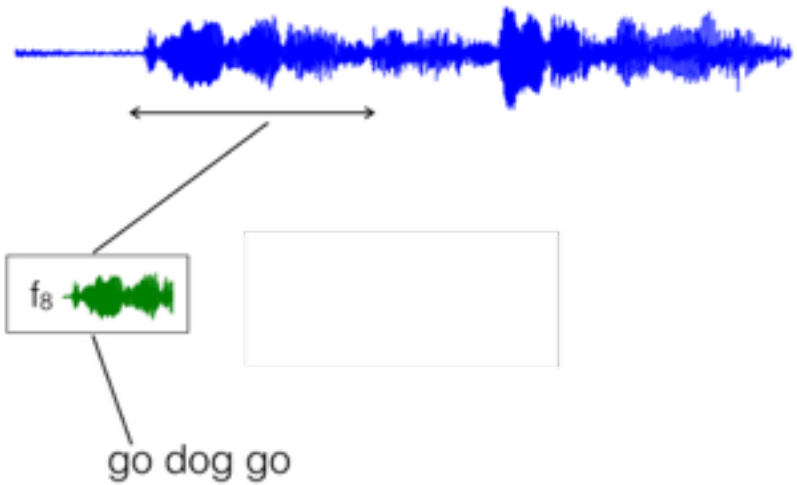
Initialize spans and clusters

M step:

- Re-estimate prototypes

E step:

- Assign cluster and align



Training using EM

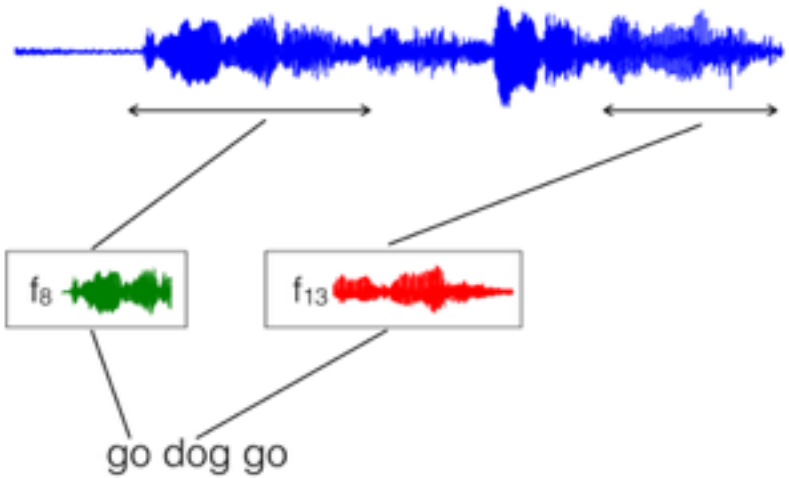
Initialize spans and clusters

M step:

- Re-estimate prototypes

E step:

- Assign cluster and align



Training using EM

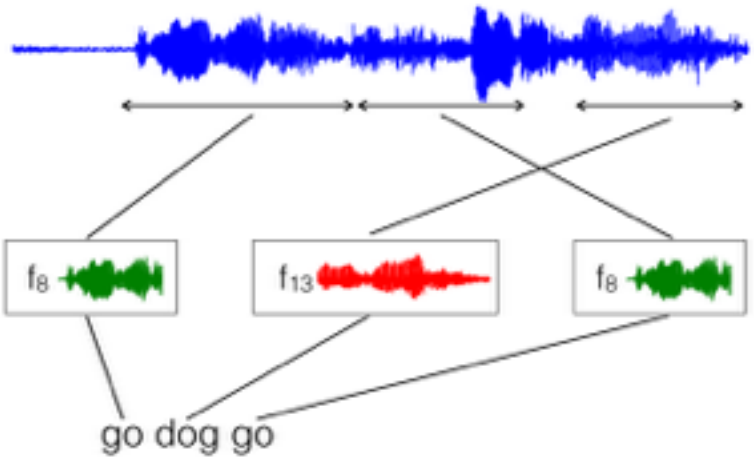
Initialize spans and clusters

M step:

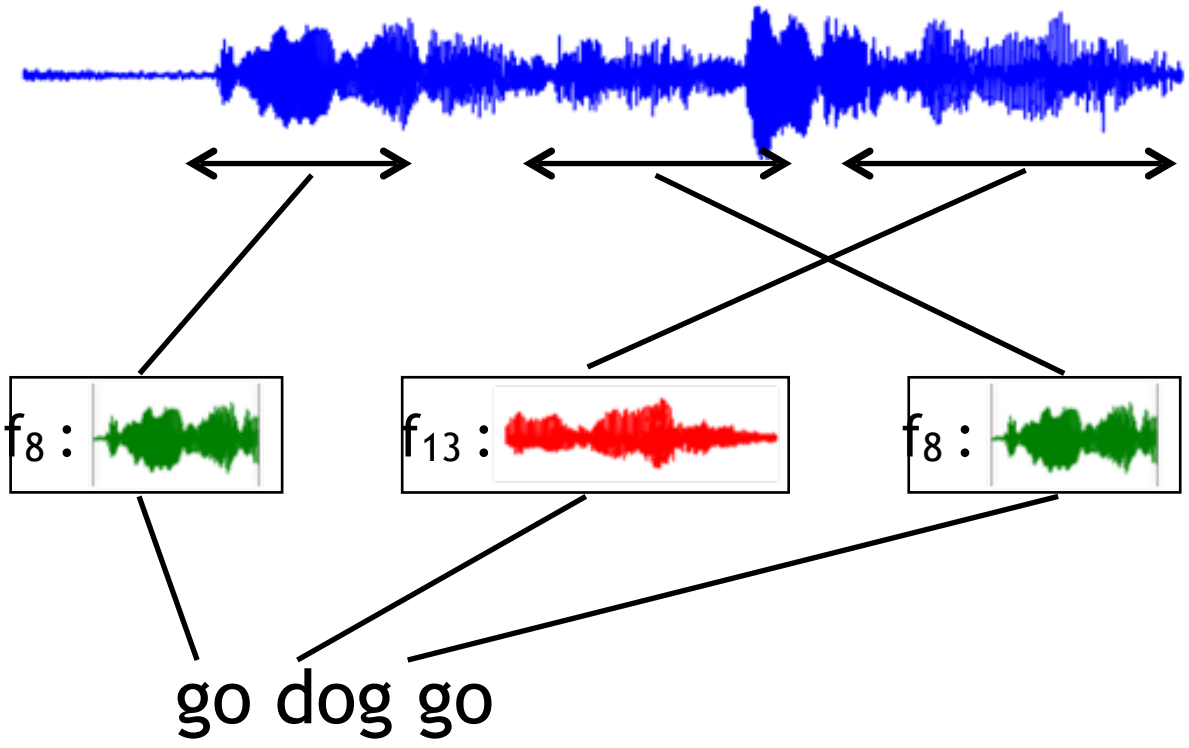
- Re-estimate prototypes

E step:

- Assign cluster and align

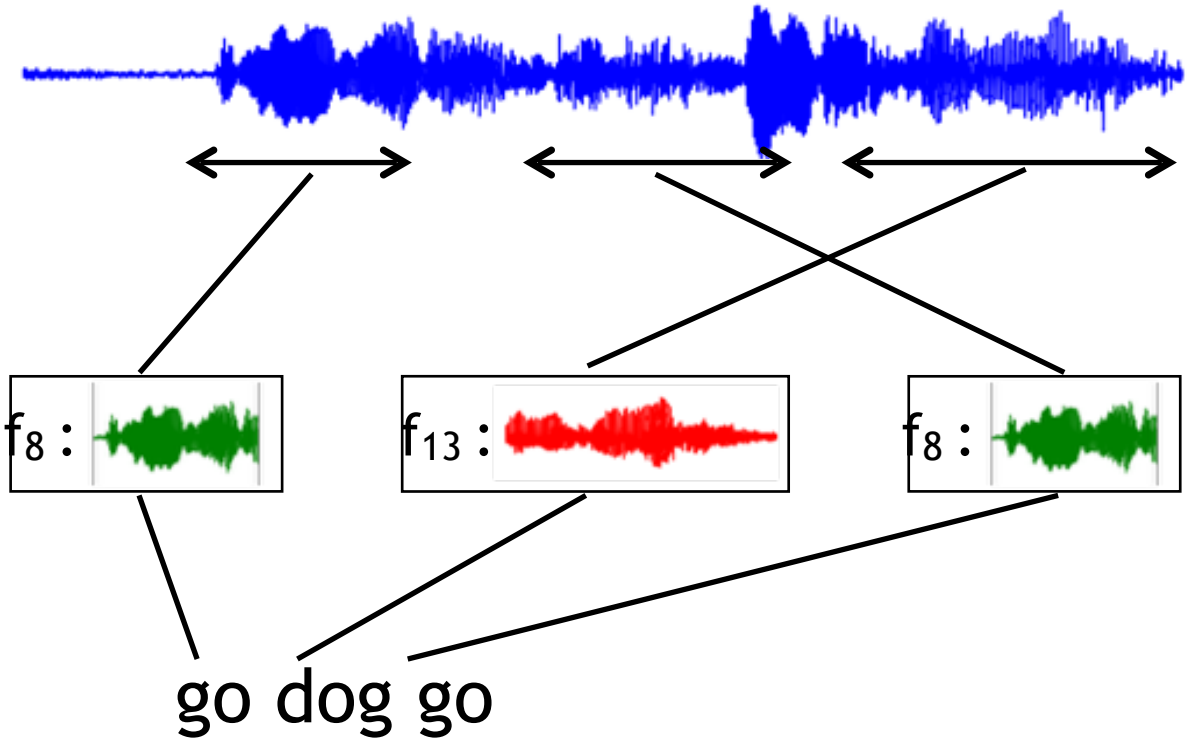


Extension 1: variable # of prototypes

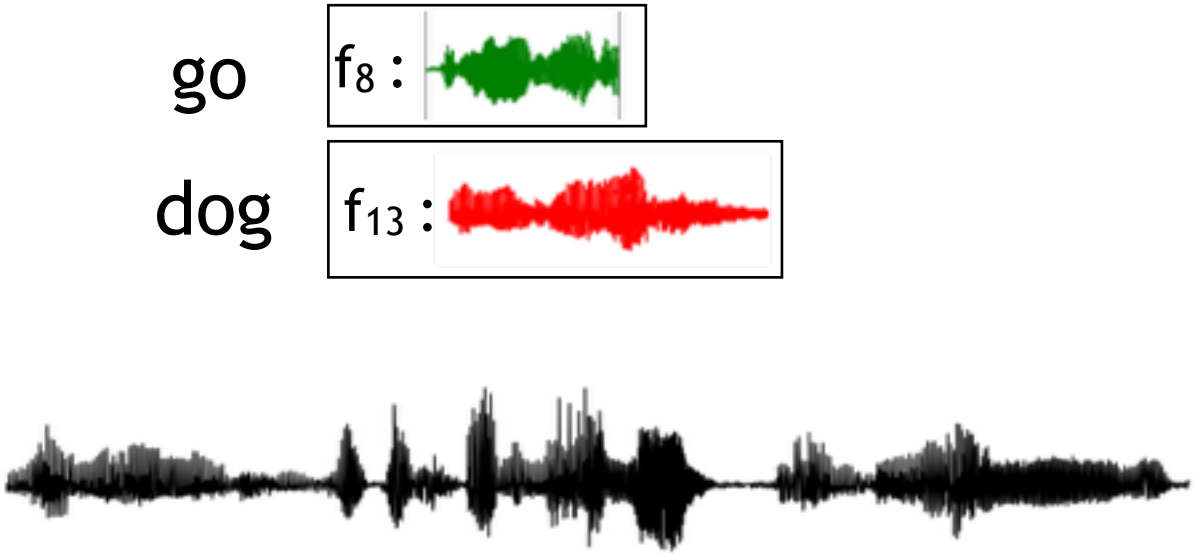


Extension 1: variable # of prototypes

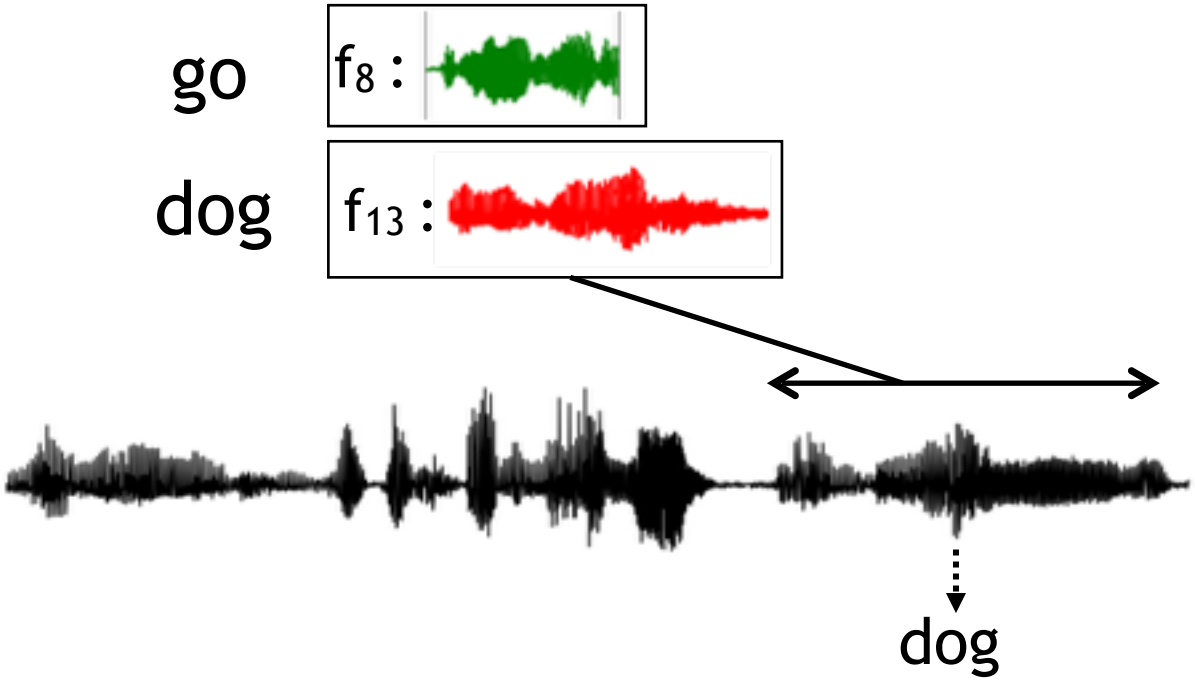
- Before re-estimating prototypes, cluster speech segments aligned to each word using a similarity threshold



Extension 2: use prototypes to predict

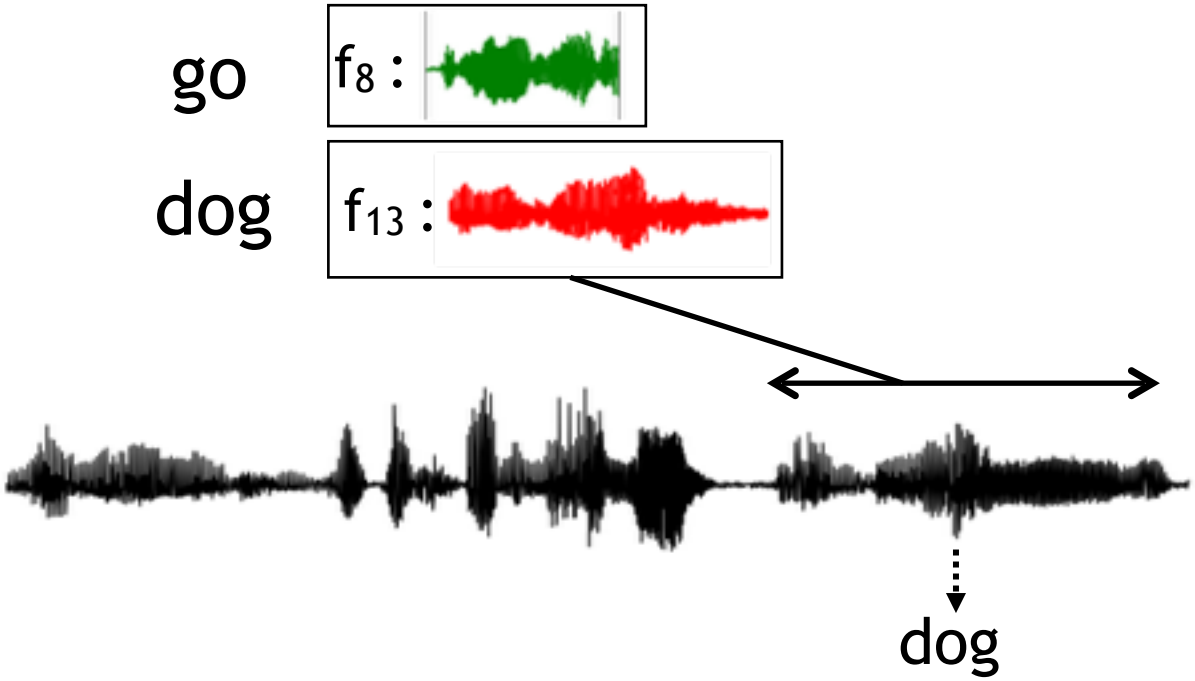


Extension 2: use prototypes to predict



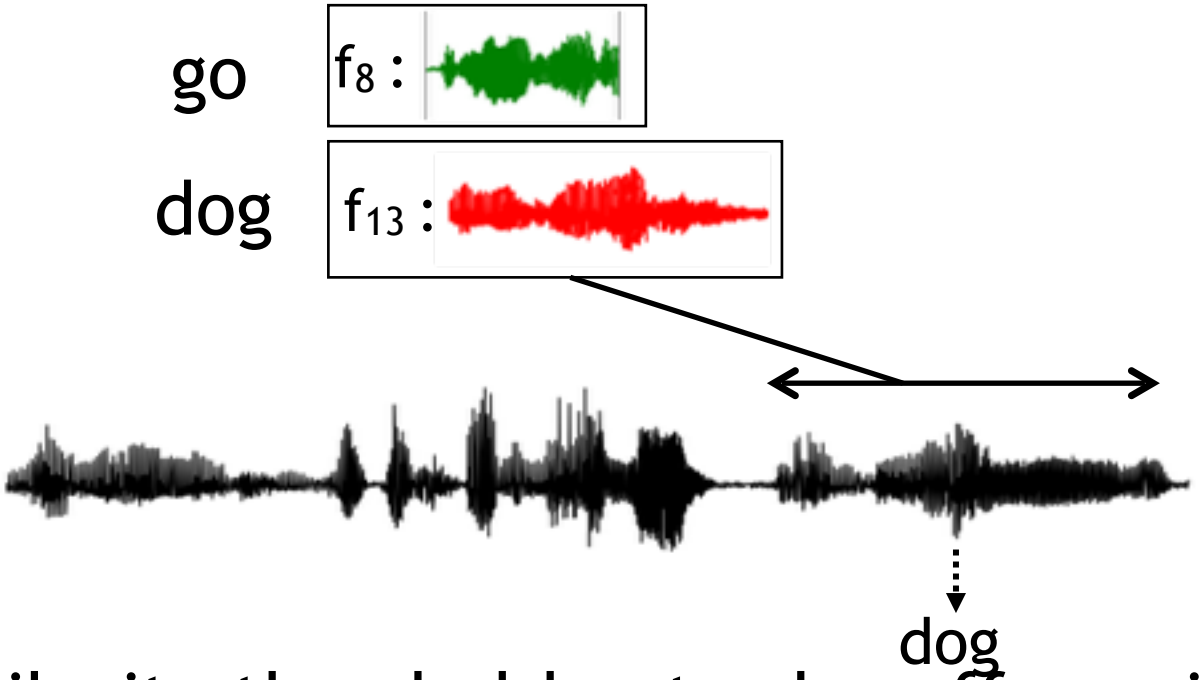
Extension 2: use prototypes to predict

- Modify UTD system (Jansen et al., 2010) to search for prototypes in unlabelled speech.



Extension 2: use prototypes to predict

- Modify UTD system (Jansen et al., 2010) to search for prototypes in unlabelled speech.



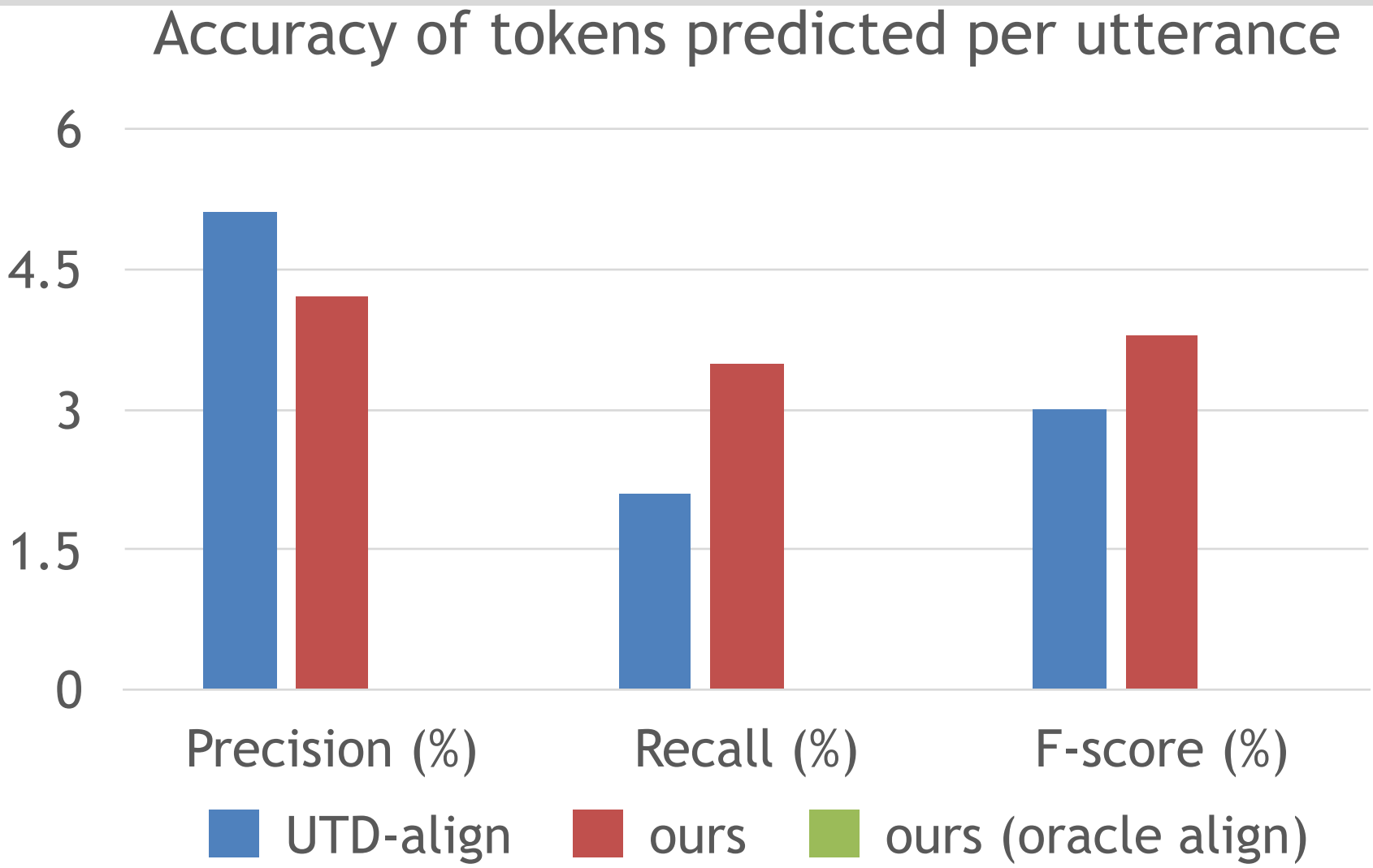
- Similarity threshold *s*: trades off precision/ recall.

Experiments: Spanish

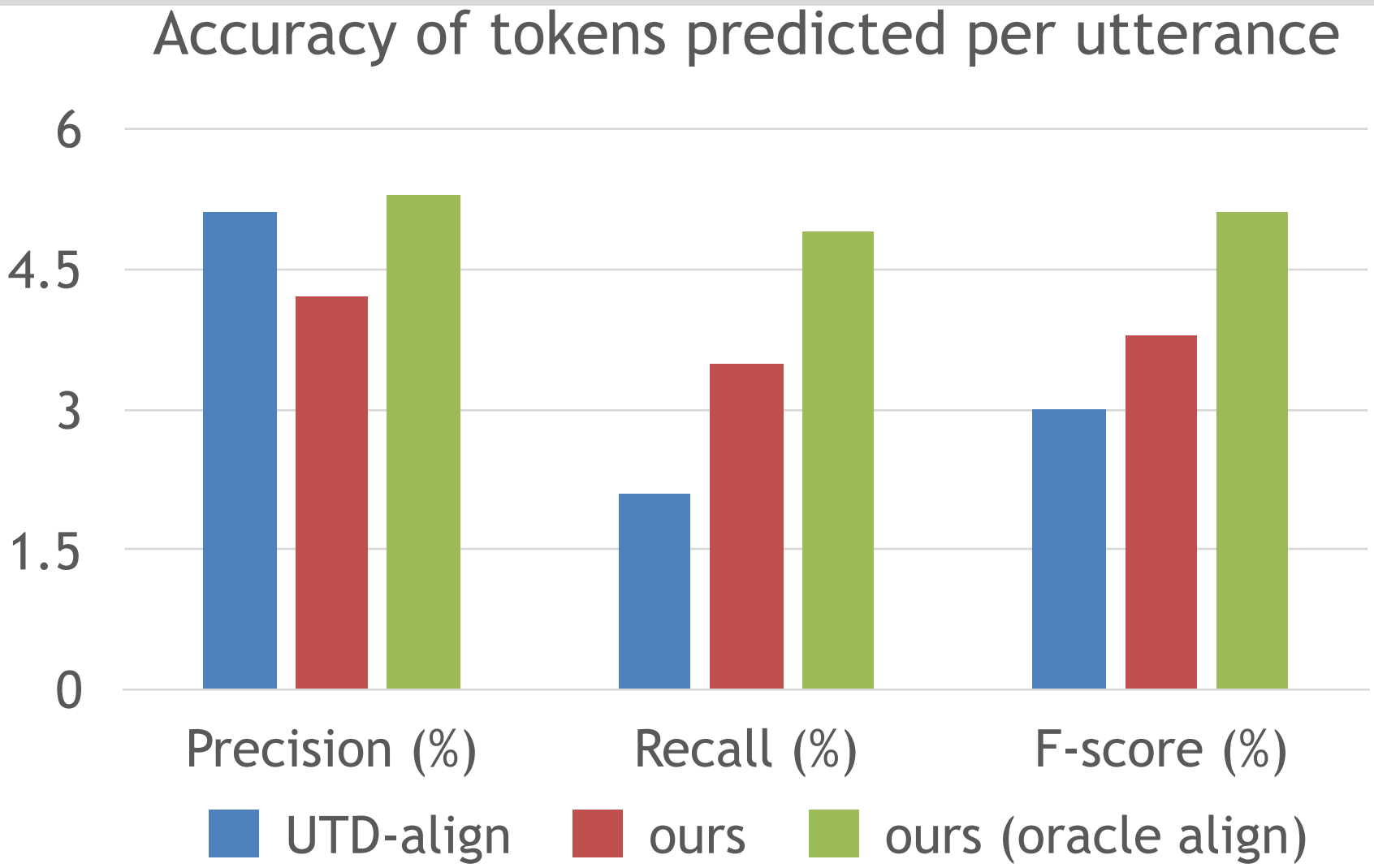
- CALLHOME: 20hr Spanish speech with English translations (Post et al., 2013)
- Random 70% utts training, 10% dev, 20% test
- Tune hyperparameters on dev:
  - Min len of segments used to compute prototypes
  - Sim threshold for creating prototype subclusters
  - % length of prototype to match for predictions

Results: Spanish

Results: Spanish



Results: Spanish



Experiments: endangered languages

- Hokkaido Ainu (10 speakers in 2007)
  - 10 narratives, single speaker (150min)
  - 2 for training: 24min, 490 English word types
- Arapaho (~1000 speakers)
  - 8 narratives, several speakers (40min)
  - 1 for training: 18min, 233 English word types
- No re-tuning of hyperparameters, except threshold for returning matches.

National Institute for Japanese Language and Linguistics (2016). A Glossed Audio Corpus of Ainu Folklore [Software]. http://ainucorpus.ninjal.ac.jp

Arapaho Language Project. http://www.colorado.edu/csilw/alp/index.html

Experiments: endangered languages

- Hokkaido Ainu (10 speakers in 2007)
  - 10 narratives, single speaker (150min)
  - 2 for training: 24min, 490 English word types
- Arapaho (~1000 speakers)
  - 8 narratives, several speakers (40min)
  - 1 for training: 18min, 233 English word types
- No re-tuning of hyperparameters, except threshold for returning matches.

National Institute for Japanese Language and Linguistics (2016). A Glossed Audio Corpus of Ainu Folklore [Software]. http://ainucorpus.ninjal.ac.jp

Arapaho Language Project. http://www.colorado.edu/csilw/alp/index.html

Experiments: endangered languages

- Hokkaido Ainu (10 speakers in 2007)
  - 10 narratives, single speaker (150min)
  - 2 for training: 24min, 490 English word types
- Arapaho (~1000 speakers)
  - 8 narratives, several speakers (40min)
  - 1 for training: 18min, 233 English word types
- No re-tuning of hyperparameters, except threshold for returning matches.

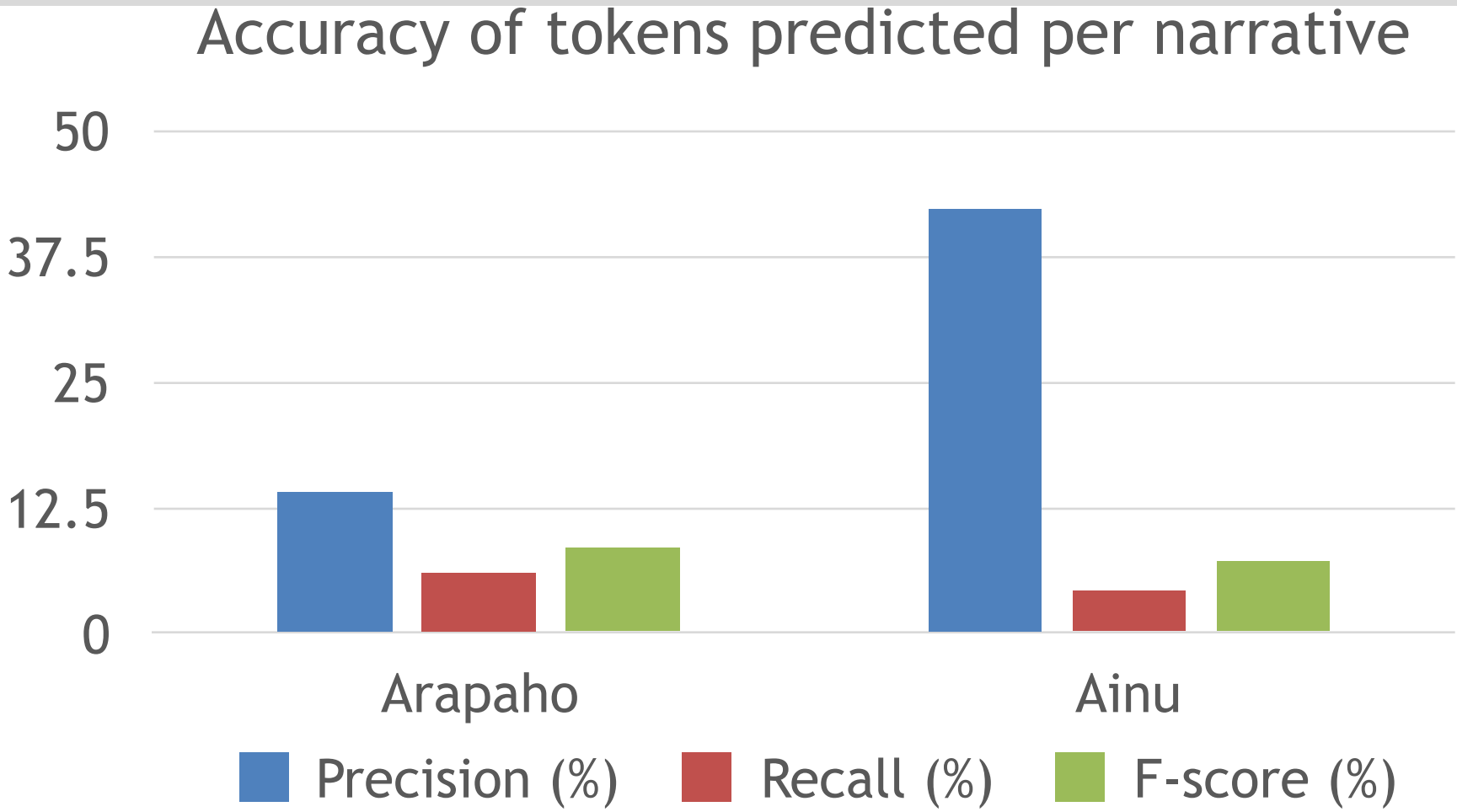
National Institute for Japanese Language and Linguistics (2016). A Glossed Audio Corpus of Ainu Folklore [Software]. http://ainucorpus.ninjal.ac.jp

Arapaho Language Project. http://www.colorado.edu/csilw/alp/index.html

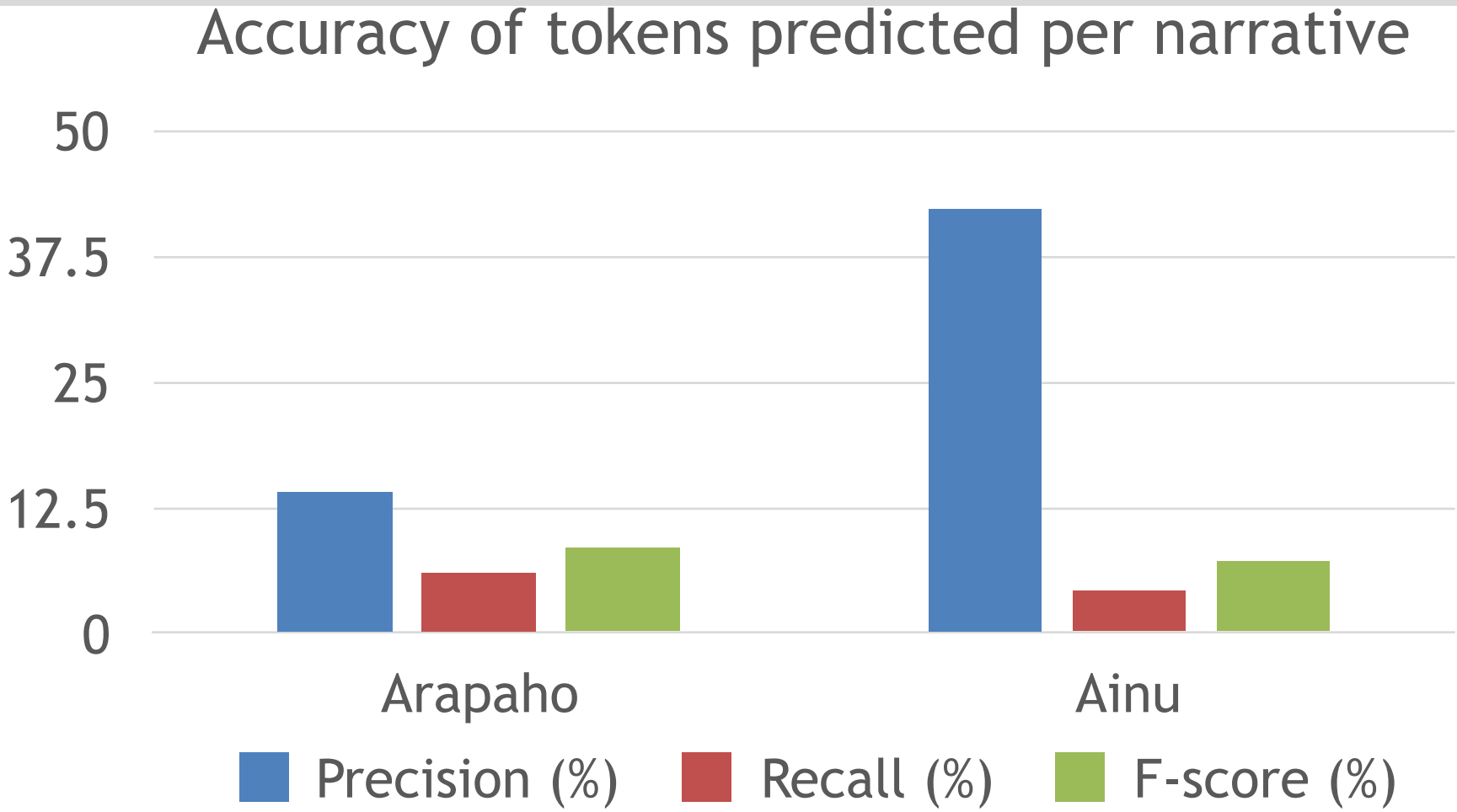
Results: Arapaho / Ainu

- Compute token accuracies over full narratives.
- On average per narrative,
  - UTD-align finds only 2 / 4 tokens (0.4% / 0.1% recall).
  - Our system finds 65 / 122 tokens.

Results: Arapaho / Ainu



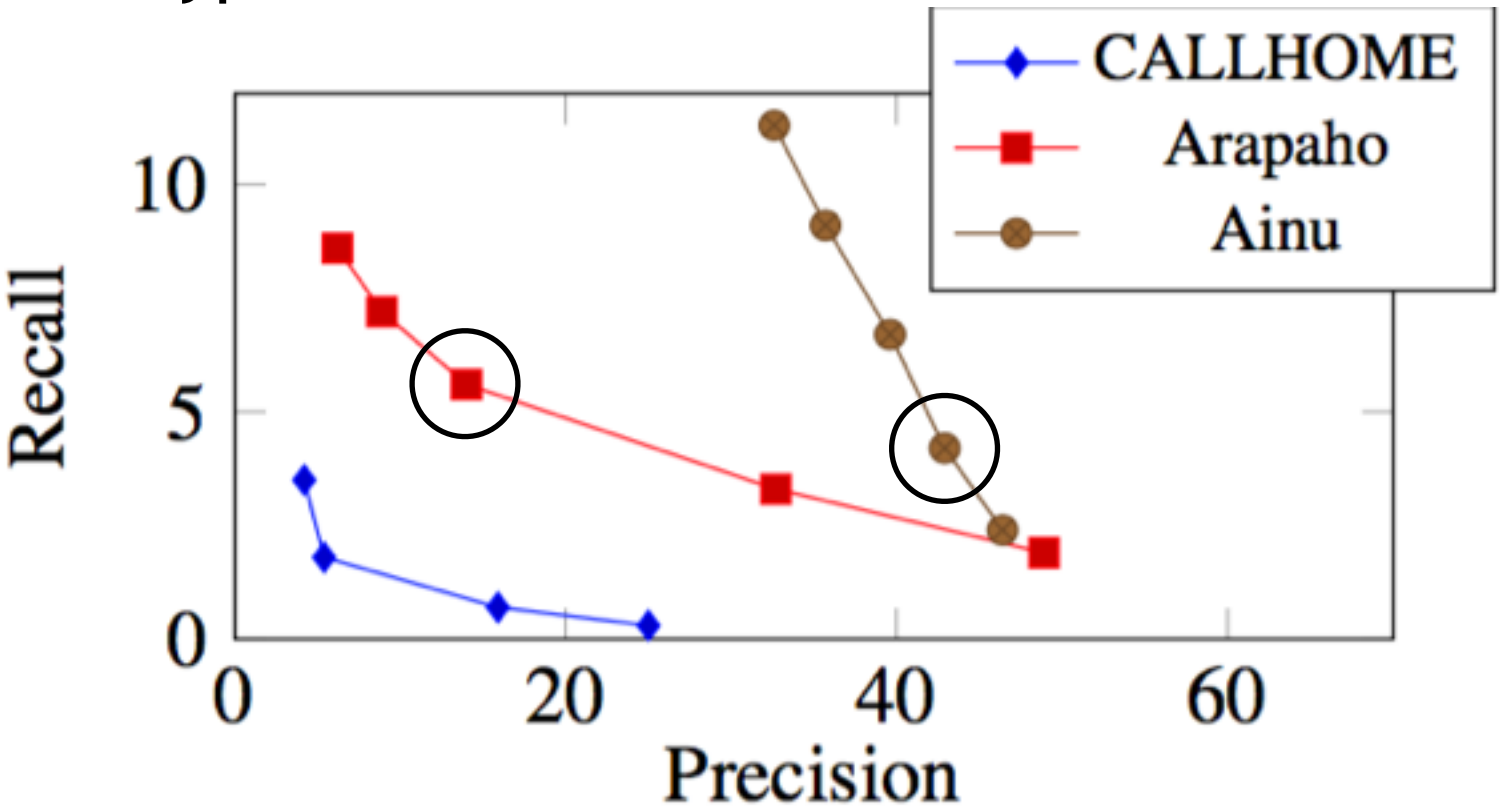
Results: Arapaho / Ainu



- Oracle recall is 48% / 64%

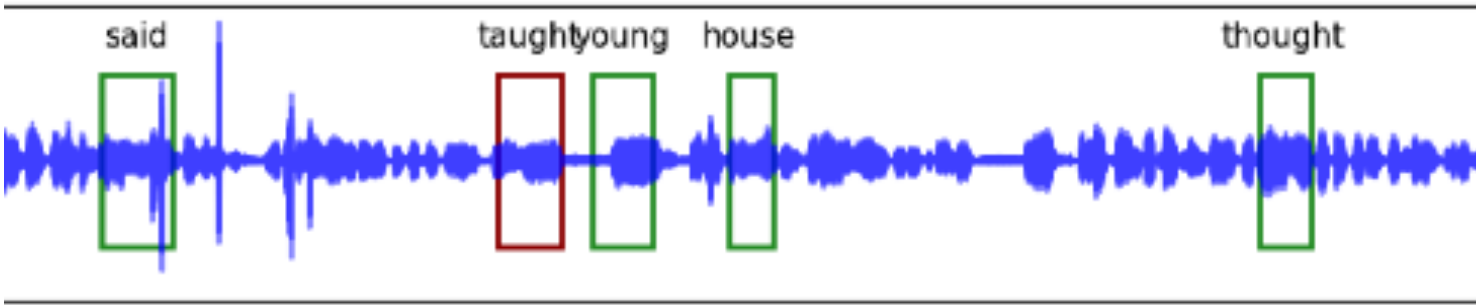
Precision/Recall tradeoff

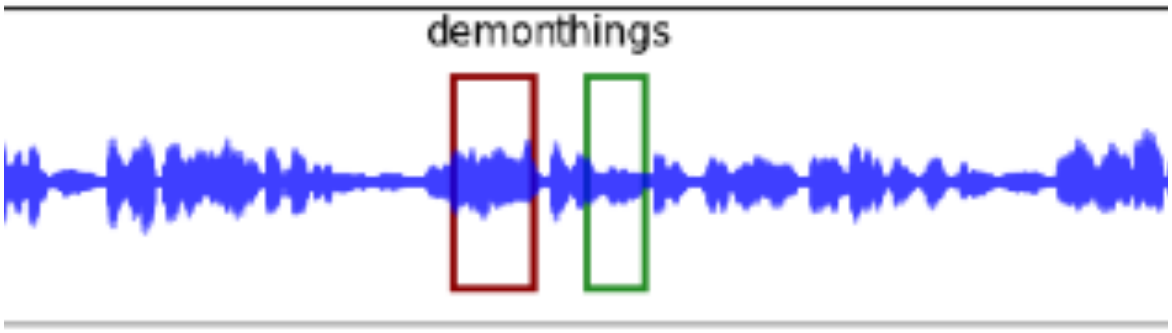
- Varying similarity threshold for matching prototypes:



Examples

Ainu #2: The Young Lad Raised by the Cat God





Conclusions

- First test of cross-lingual keyword labelling of speech data from endangered languages, using very small translated portions.
- Joint alignment/clustering outperforms pipelined system.
- Identifies a handful of terms, but speech matching problem is really hard!
- Future: consider NN approaches, improve speech feature extraction using multilingual data.