Hybrid Spoken Term Discovery-ABnet System
Finding words to learn segments

Roland Thiollière*, Ewan Dunbar*, Gabriel Synnaeve*†, Maarten Versteegh*, Emmanuel Dupoux*

*LSCP, École Normale Supérieure / EHESS / CNRS, Paris, France
†now at Facebook AI Research
rolthiolliere@gmail.com, emd@umd.edu, gabrielsynnaeve@gmail.com, maartenversteegh@gmail.com, emmanuel.dupoux@gmail.com

Introduction

- This system is the combination of two architectures: a **spoken term discovery**[1] (STD) and a **siamese neural network**[2] (ABnet).
- The STD system finds matching patterns in the acoustic signal.
- The ABnet is trained to minimize the distance between those matching patterns, and maximize the distance between non matching patterns.

Motivation

The high level idea: It is easier to discriminate words than phonemes → **use Track 2 to help with Track 1**.

System

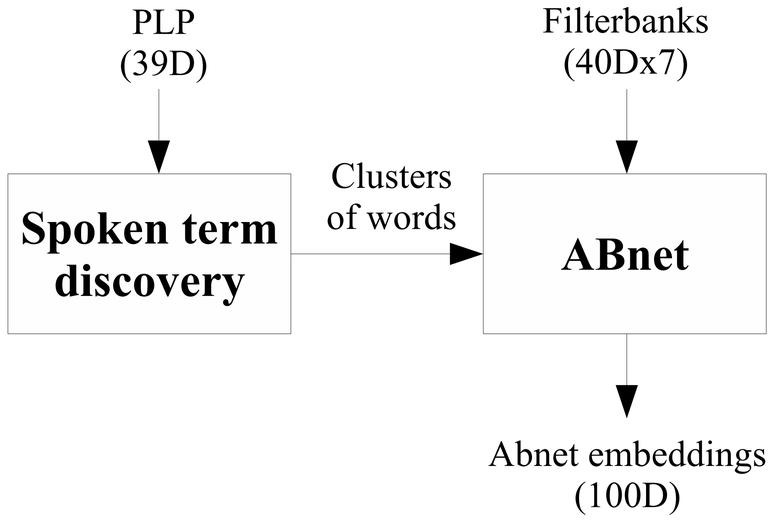

Figure: Overview of the components of our system.

Spoken term discovery

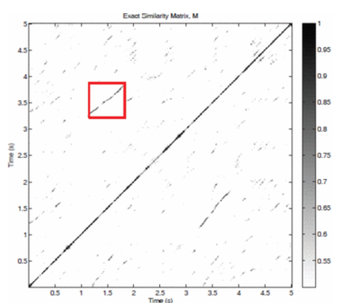

Figure: Similarity matrix of the signal. Source: [1]

- System used as baseline for track 2 [3].
- Computes an approximation of the similarity matrix (cosine similarity).
- Searches for diagonal patterns
- Filters and cluster those patterns

See [1] for more.

ABnet

- The ABnet is a siamese neural network architecture: the weights in the second branch are the duplicate of the weights in the first branch.
- It learns a space where inputs with the same label are close together, and inputs with different labels are far apart.
- Input: a pair of examples (A, B) and a label Y=(same|different), here 7 stacked frames of 40 dimensionnal filterbanks.
- loss function = similarity if same, distance if different.

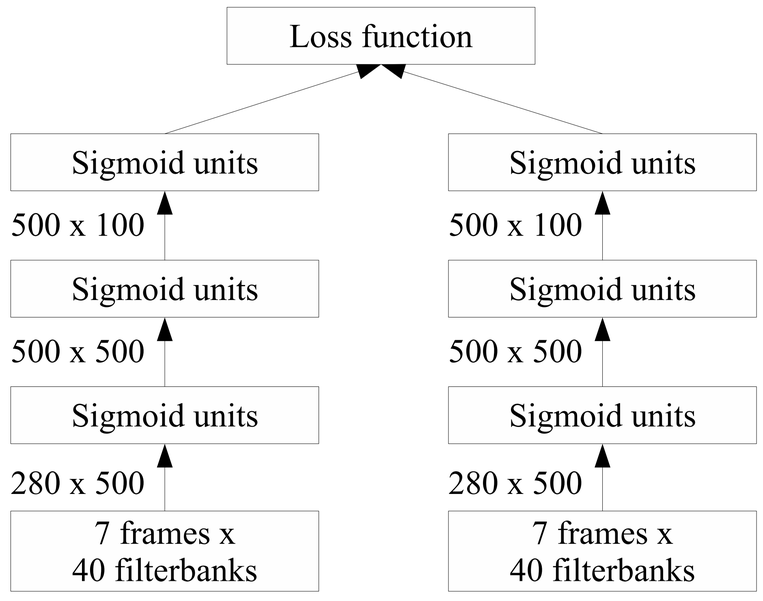

Figure: Overview of the components of our system.

- Loss function used:

$$\mathcal{L}(A, B) = \begin{cases} (1 - \cos(Y_A, Y_B))/2 & \text{if same} \\ \cos^2(Y_A, Y_B) & \text{if different} \end{cases}$$

- Pairs of "different" words are randomly selected amongst the discovered patterns. Pairs of same words are DTW aligned.
- Pairs of same/different words → pairs of same/different frames

Results

Table: Evaluation of the STD system. These fragments serve as input to the ABnet.

	Words	Pairs	Classes	NED	Coverage
English	6512	4305	3149	0.219	0.163
Xitsonga	3582	1818	1782	0.120	0.162

- The speech representation is evaluated with the ABX paradigm (see [3]).
- Two tasks: discriminate phones within speakers, and discriminate phones across speakers.

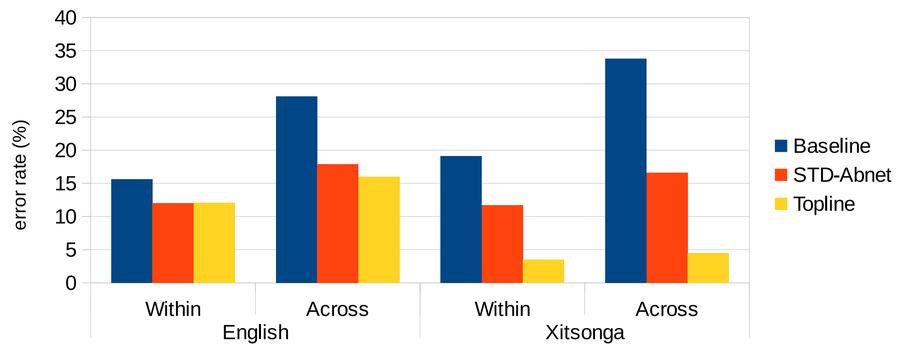

Figure: Within and across speaker Minimal Pair ABX error rates.

Conclusions

- Despite the low number of examples, a good speech representation can be learnt.
- Close to topline in english.
- Forthcoming research: Loop over the system (STD on learnt features).

References & Acknowledgements

[1] A. Jansen and B. van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proceedings of ASRU*, 2011.

[2] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *IEEE SLT*, 2014.

[3] M. Versteegh, R. Thiolliere, T. Schatz, X. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge," in *Submitted to Interspeech*, 2015.

We would like to thank Aren Jansen for letting us use his spoken term discovery system pre-release, and for his technical support.