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 find matching patterns in the acoustic signal.
- The ABnet is train to minimize the distance between those matching patterns, and maximize the distance between non matching patterns.

Motivation

- The high level idea: that two randomly selected words are more distant in the acoustic space than two randomly selected phonemes. It is easier to discriminate words than phonemes.
- Our approach: extract word level information to learn phonemes.

System

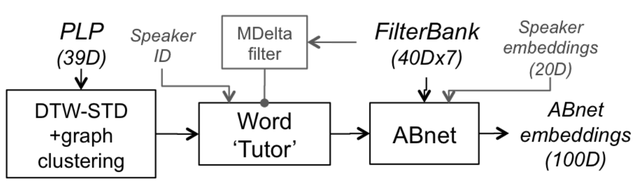

Figure: Overview of the components of our system.

Spoken term discovery

- System used as baseline for track 2[3]
- Finding patterns in an approximation of the similarity matrix.
- Computes an approximation of the similarity matrix (cosine similarity).
- Searches for diagonal patterns in that matrix.
- Those matching patterns are then filtered and clustered.

This is only a succinct description of the system and there is a lot more to it.

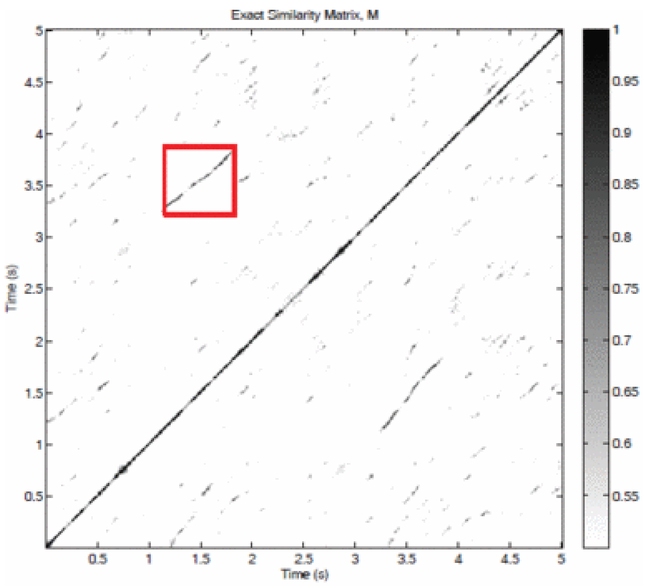

Figure: Similarity matrix of the signal.

Word tutor

The word tutor selects the words that it will feed to the ABnet, making the link between the STD system and the ABnet. Pairs of "different" words are randomly selected amongst the discovered patterns. All pairs of words (same and different) are DTW aligned.
We extract 40 dimensionnal log-energy Mel-scale filterbanks (10ms step and 25ms window size). 7 adjacent (following the "aligned" path) frames are stacked as input for the ABnet.

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.
- Input: a pair of examples (A, B) and a label Y=(same|different)
- loss function = similarity if same, distance if different.

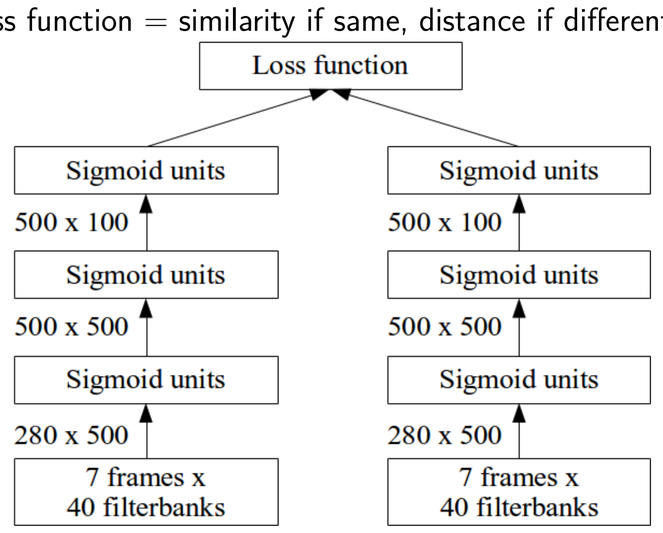

Figure: Overview of the components of our system.

- Loss function used:

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

where

$$\cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

- It learns a space where inputs with the same label are close together, and inputs with different labels are far appart.

*

Extensions
We investigate the use of temporal information, and speaker information.

Mdelta

Adding temporal information.

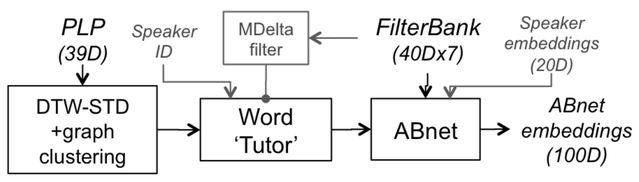

Figure: Overview of the components of our system.

*

Speaker embedding
Adding speaker information.

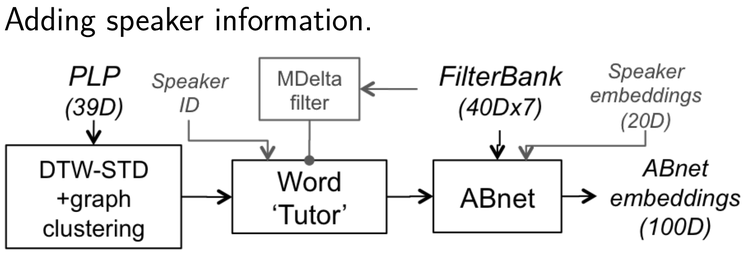

Figure: Overview of the components of our system.

Results

The speech representation is evaluated with the ABX paradigm[3].
The overall system raised a substancial improvement over baseline.

Table: Within and across speaker Minimal Pair ABX error rates for the ZeroSpeech baseline (MFCC) and topline (supervised HMM-GMM posteriorgrams), and for our systems.

	English		Xitsonga	
	Within	Across	Within	Across
Baseline (MFCC)	*15.6*	*28.1*	*19.1*	*33.8*
Topline (HMM-GMM)	*12.1*	*16.0*	*3.5*	*4.5*
STD → ABnet	**12.0**	**17.9**	**11.7**	**16.6**
STD / MDF → ABnet	12.4	18.1	12.6	18.6
STD + SpkID → ABnet	12.2	18.0	16.5	21.3

Table: Output of the spoken term discovery system. These fragments ("words") serve as input to the ABnet.

	Words	Pairs	Classes	NED	Coverage
Engl. E(1,3)	6512	4305	3149	0.219	0.163
Engl. E(2)	4334	2630	2092	0.229	0.106
Xits. E(1,3)	3582	1818	1782	0.120	0.162
Xits. E(2)	2286	1158	1138	0.105	0.106

Conclusions

The results validate the approach. Despite the low number of examples, a good speech representation can be learnt.
However, all our attempts to further improve the results by adding additional information failed.

Forthcoming Research

- Loop over the system (STD on learnt features).
- Succesfully apply MDelta to STD output to improve track 2.

[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.

[4] G. Synnaeve, "ABnet: Interspeech 2015 status," Mar. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.16411

Acknowledgements

This work really was a team effort, all authors contributed equally.
We would like to thank Aren Jensen for letting us use his spoken term discovery system pre-release, and for his technical support.