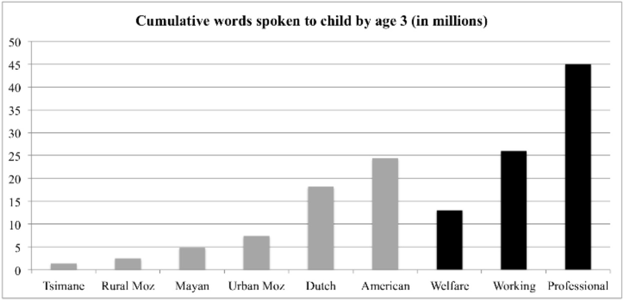
Sampling strategies in Siamese Networks for unsupervised speech representation learning

Rachid Riad, Corentin Dancette, Julien Karadayi, Neil Zeghidour, Thomas Schatz, Emmanuel Dupoux

30 juillet 2018

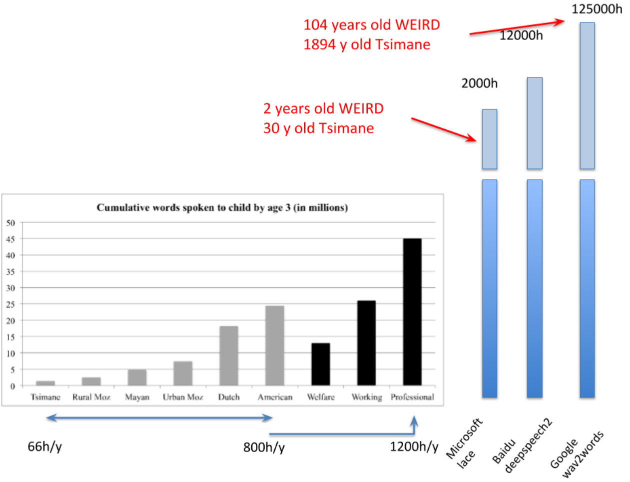
Outline

Introduction



[Cristia et al., 2017]

Introduction



[Cristia et al., 2017]

Current limits of ASR systems

Human parity for speech recognition

[Xiong et al., 2016] Microsoft claims to achieve human parity

- 2000+ hours of transcribed speech to train acoustic model
- 350M+ words to train the language model

Current limits of ASR systems

Human parity for speech recognition

[Xiong et al., 2016] Microsoft claims to achieve human parity

- 2000+ hours of transcribed speech to train acoustic model
- 350M+ words to train the language model

Too expensive and time consuming to gather these data for most languages.

Zero-resource Challenge 2015

Goal of the challenge

Unsupervised discovery of linguistic units, with two tracks

Sub-word modelling

[s] [p] [o] [w] [k]

Spoken term discovery

“magret”, “table”, “the”

Zero-resource Challenge 2015

The Buckeye dataset

- Casual conversations of English
- 12 speakers
- 5 hours of datasets, 16-30 min for each speaker

The Mboshi dataset

- Read speech of Mboshi, a Bantu language from Congo
- 24 speakers
- 2.5 hours of datasets, 2-29 min for each speaker

Zero-resource Challenge 2015

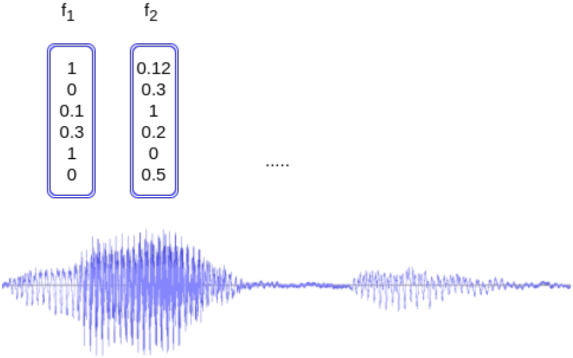
Track 1 : Sub-word modelling

- **Highlight** relevant linguistic properties : **phone** structure
- **Downplay** irelevant linguistic properties : **ID, channel, etc.**

Zero-resource Challenge 2015

Track 1 : Sub-word modelling

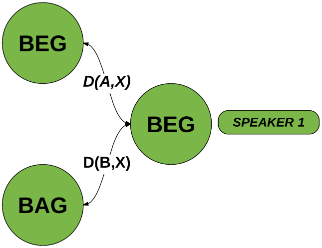
- **Highlight** relevant linguistic properties : **phone** structure
- **Downplay** irelevant linguistic properties : **ID, channel, etc.**



Zero-resource Challenge

Evaluation : ABX discriminability task [Schatz et al., 2013]

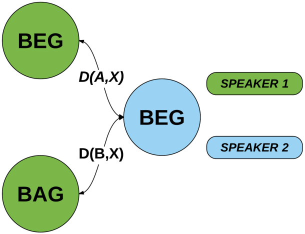
- Triplet A, B and X with A and X in the same phonetic class
- $D(A, X) < D(B, X)$ success 1, failure otherwise
- Average over all possible triplets



Zero-resource Challenge 2015

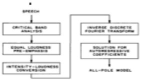
Evaluation : ABX discriminability task

- Triplet A, B and X with A and X in the same phonetic class
- $D(A, X) < D(B, X)$ success 1, failure otherwise
- Average over all possible triplets



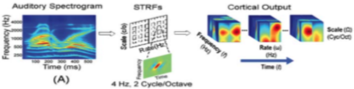
Unsupervised sub-word modelling

- Acoustic features
  PLP, RASTA



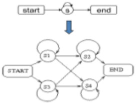
Hermanky (1990). *JASA*

- Auditory model



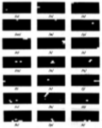
Chi, Ru, & Shamma (2005) *JASA*

- HMM state splitting



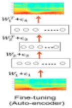
Varadarajan, Khudanpur,Dupoux, (2008)

- Kohonen's maps



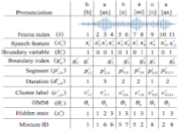
Kohonen (1988), *Computer*

- Deep autoencoders



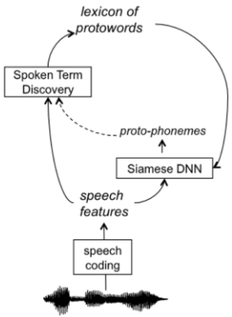
Badino, Canevari, et al (2014), *ICASSP*.

- Non Parametric Bayesian Clustering



Lee & Glass, (2012). *Proc of ACL*

Joint lexical-sublexical learning

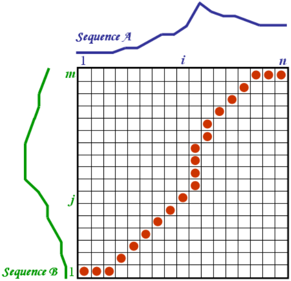


1. Discover words
2. Feed aligned frames of two words to the Siamese neural network

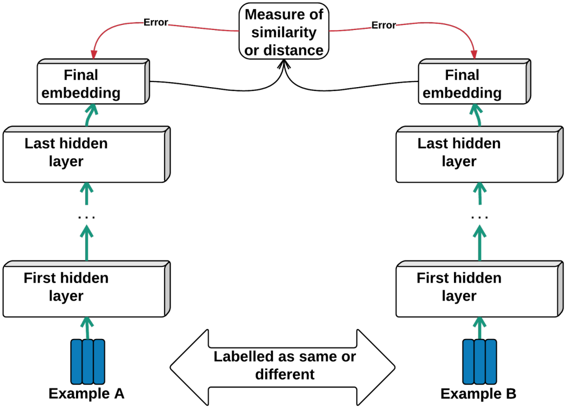
[Synnaeve et al., 2014, Thiollière et al., 2015, Dupoux, 2018]

Aligning a word pair with DTW

Dynamic Time Warping



Siamese Architecture



Loss function

Minimize distance of same discovered frames
Maximize distance of different discovered frames

Loss function

Minimize distance of same discovered frames
Maximize distance of different discovered frames

$$l_\gamma(x_1, x_2, y) = \begin{cases} -\cos(e(x_1), e(x_2)), & \text{if } y = 1 \\ \max(0, \cos(e(x_1), e(x_2)) - \gamma), & \text{otherwise} \end{cases}$$

Sampling Schema for pairs

How do we select training data for siamese network ?

Sampling Schema for pairs

How do we select training data for siamese network ?

"the" "at"

Sampling Schema for pairs

How do we select training data for siamese network ?

"the"

"at"

"magret"

"table"

Sampling Schema for pairs

How do we select training data for siamese network ?

Previously in Siamese networks... [Thiollière et al., 2015]

- Randomly choosing two words in the dataset
- Balancing same / different pairs
- Balancing same / different speakers pairs

Sampling Schema for pairs

How do we select training data for siamese network ?

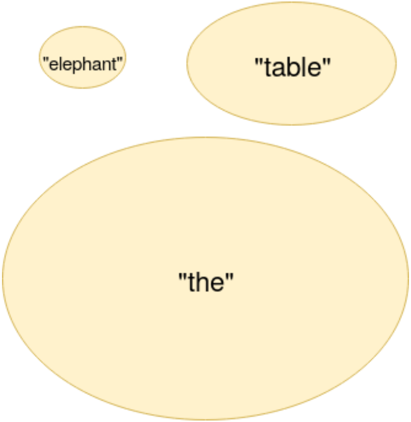
Sampling Schema for pairs

How do we select training data for siamese network ?

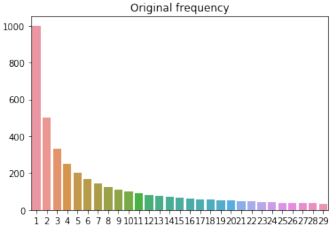
Parameters

- ϕ Distribution choice to sample words
- P_w^- Same versus Different word Ratio
- P_s^- Same versus Different Speaker Ratio

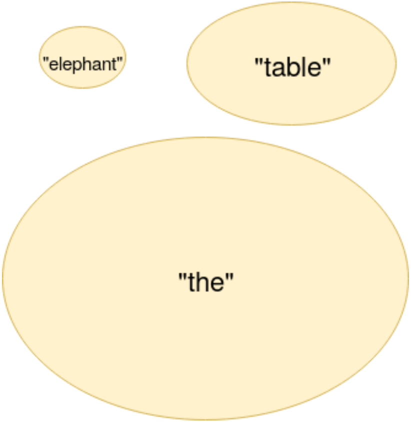
Word frequencies



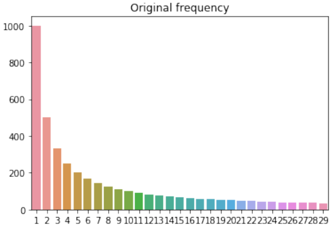
Distribution follows an empirical power law : **Zipf's Law**. [Zipf, 1949]



Word frequencies



Distribution follows an empirical power law : **Zipf's Law**. [Zipf, 1949]



$$f_w \propto \frac{1}{r_w^{\alpha}}, \alpha \approx 1$$

Sampling compression function

Let say a word w as a number of occurrences n_w.
We define the sampling compression function Φ such as :

Sampling compression function

$$\mathbb{P}(w) = \frac{\phi(n_w)}{\sum\limits_{\forall w'} \phi(n_{w'})} \tag{1}$$

Sampling compression function

Let say a word w as a number of occurrences n_w.
We define the sampling compression function Φ such as :

Sampling compression function

$$\mathbb{P}(w) = \frac{\phi(n_w)}{\sum_{\forall w'} \phi(n_{w'})} \tag{1}$$

[Mikolov et al., 2013, Levy et al., 2015]

Sampling compression functions

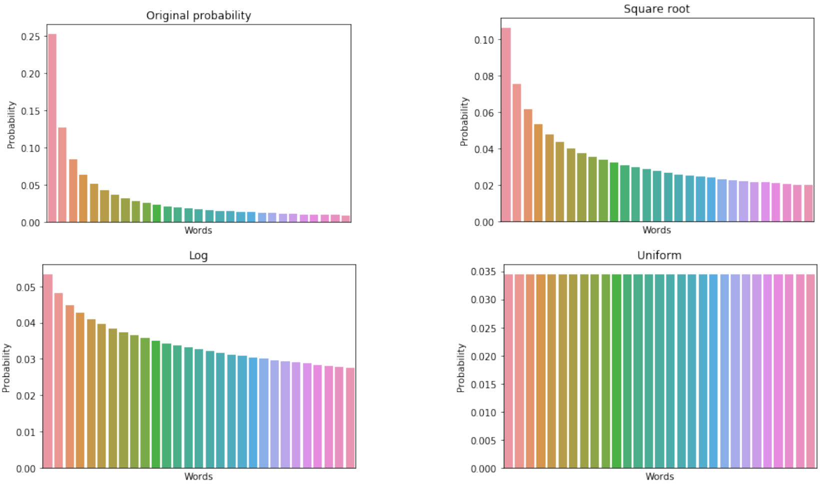
We evaluated 5 different sampling compression functions

Sampling functions Φ

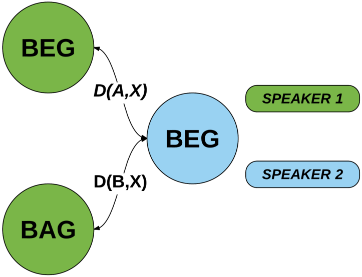
- $n \rightarrow n$
- $n \rightarrow \sqrt{n}$
- $n \rightarrow \sqrt[3]{n}$
- $n \rightarrow \log(1 + n)$
- $n \rightarrow 1$

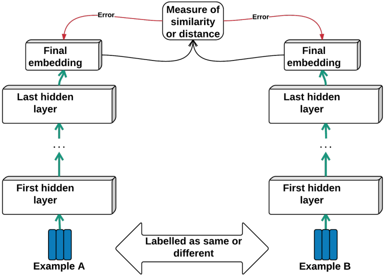
The objective is to balance the effect of Zipf's law on the sampling.

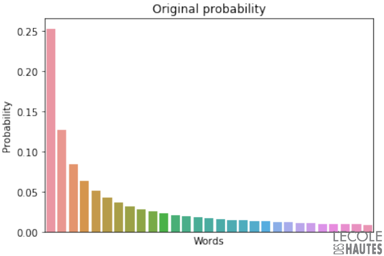
Sampling compression function



Summary of methods







Measure sampling contribution

Weakly-supervised setting

Lexical of real words as weak labels

Training

- 12 speakers from the Buckeye corpus, 5 hours
- Split in 4 different sizes

Evaluation on test set

- 2 other speakers from the Buckeye corpus

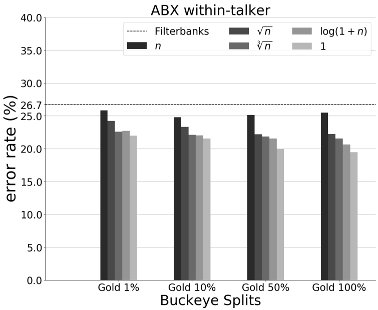
Weakly supervised setting

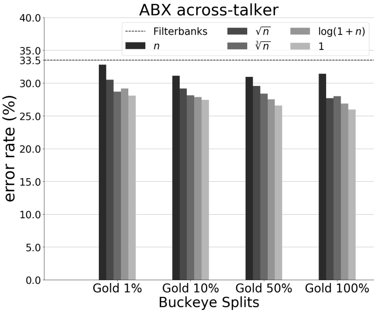
Table – Statistics for the 4 Buckeye splits used for the weakly supervised training, the duration in minutes expressed the total amount of speech for training

	Duration	#tokens	#words	#possible pairs
1%	3.0 min	1006	355	$\sim 5.10^5$
10%	29.9 min	7189	1297	$\sim 2.10^7$
50%	149.5 min	34912	3112	$\sim 6.10^8$
100%	299.1 min	69543	4538	$\geq 2.10^9$

Influence of Φ on results

Experience using gold words of buckeye.





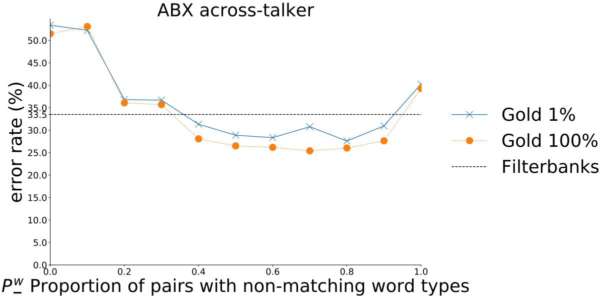
It is beneficial to mitigate the effect on zipf's law on sampling.

Influence of the proportion of different-types pairs

What is the proportion P_w^- of negative pairs of words should we sample ?

Influence of the proportion of different-types pairs

What is the proportion P_w^- of negative pairs of words should we sample ?

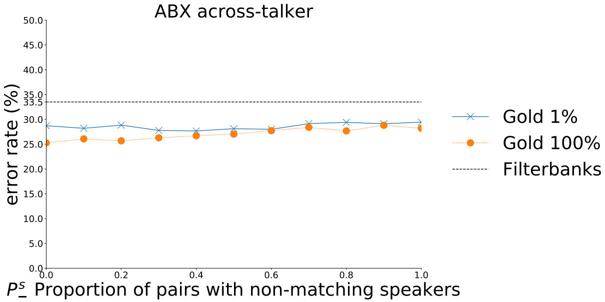


Influence of the proportion of different-speaker pairs

What is the proportion P_s^- of pairs of words with different speakers we should sample ?

Influence of the proportion of different-speaker pairs

What is the proportion P_s^- of pairs of words with different speakers we should sample ?



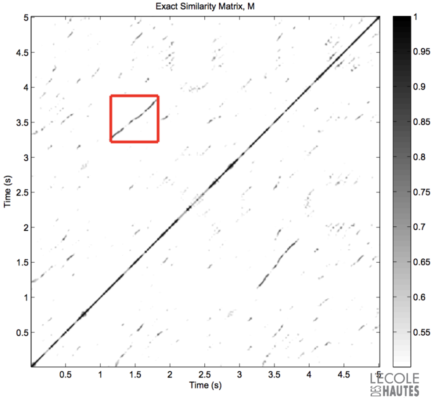
Parameters for best performance on weakly-supervised setting

1. $\phi : n \rightarrow 1$: **remove** the influence of word frequency
2. $P_w^- = 0.7$: **preference** for negative pairs (different words)
3. $P_s^- = 0$: **only** same-speaker pair

Unsupervised Spoken Term Discovery

[Jansen and Van Durme, 2011]

Discover words instead of using the labels of words.

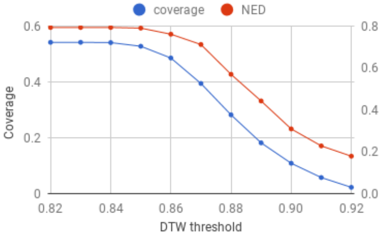


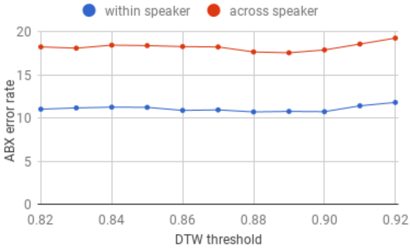
Test on zerospeech 2015 challenge

Table – ABX discriminability results for the ZeroSpeech2015 datasets. The best error rates for each conditions for siamese architectures are in **bold**. The best error rates for each conditions overall are underlined.

Models	English		Xitsonga	
	within	across	within	across
baseline (MFCC)	15.6	28.1	19.1	33.8
supervised topline (HMM-GMM)	12.1	16.0	04.5	03.5
Our ABnet with $P^{w}_{-} = 0.7, P^{s}_{-} = 0, \phi : n \rightarrow 1$	**10.4**	17.2	**9.4**	**15.2**
CAE [Renshaw et al., 2015]	13.5	21.1	11.9	19.3
ABnet [Thiollière et al., 2015]	12.0	17.9	11.7	16.6
ScatABnet [Zeghidour et al., 2016]	11.0	**17**	12.0	15.8
DPGMM [Chen et al., 2015]	10.8	16.3	9.6	17.2
DPGMM+PLP+bestLDA+DPGMM [Heck et al., 2016]	10.6	16.0	8.0	12.6

Spoken term discovery threshold





Conclusions

- The Sampling component plays a major part in the process of training Siamese Network, especially the compression function choice.

Conclusions

- The Sampling component plays a major part in the process of training Siamese Network, especially the compression function choice.
- Transfer findings from weakly-supervised learning to unsupervised does not yield as much improvements.

Conclusions

- The Sampling component plays a major part in the process of training Siamese Network, especially the compression function choice.
- Transfer findings from weakly-supervised learning to unsupervised does not yield as much improvements.
- Optimal trade-off for the sub-word modelling performances between the amount and quality of discovered words.

Conclusions

- The Sampling component plays a major part in the process of training Siamese Network, especially the compression function choice.
- Transfer findings from weakly-supervised learning to unsupervised does not yield as much improvements.
- Optimal trade-off for the sub-word modelling performances between the amount and quality of discovered words.

Code in Python, pytorch

Code available online https://github.com/bootphon/abnet3

Next steps

Understand why using different-speaker pairs doesn't improve results

Fully unsupervised loop between Spoken Term Discovery and ABNet

Learn fixed-size representation of words and co-training with sub-word discriminative loss

Thank You!

References I

Chen, H., Leung, C.-C., Xie, L., Ma, B., and Li, H. (2015).
Parallel inference of dirichlet process gaussian mixture models for unsupervised acoustic modeling : A feasibility study.
In *Sixteenth Annual Conference of the International Speech Communication Association*.

Cristia, A., Dupoux, E., Gurven, M., and Stieglitz, J. (2017).
Child-directed speech is infrequent in a forager-farmer population : A time allocation study.
Child development.

Dupoux, E. (2018).
Cognitive science in the era of artificial intelligence : A roadmap for reverse-engineering the infant language-learner.
Cognition, 173 :43–59.

Heck, M., Sakti, S., and Nakamura, S. (2016).
Unsupervised linear discriminant analysis for supporting dpgmm clustering in the zero resource scenario.
Procedia Computer Science, 81 :73–79.

Jansen, A. and Van Durme, B. (2011).
Efficient spoken term discovery using randomized algorithms.
In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 401–406. IEEE.

Levy, O., Goldberg, Y., and Dagan, I. (2015).
Improving distributional similarity with lessons learned from word embeddings.
Transactions of the Association for Computational Linguistics, 3 :211–225.

References II

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pages 3111–3119.

Renshaw, D., Kamper, H., Jansen, A., and Goldwater, S. (2015).
A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge.
In *Sixteenth Annual Conference of the International Speech Communication Association*.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. (2013).
Evaluating speech features with the minimal-pair abx task : Analysis of the classical mfc/plp pipeline.
In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Synnaeve, G., Schatz, T., and Dupoux, E. (2014).
Phonetics embedding learning with side information.
In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 106–111. IEEE.

Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., and Dupoux, E. (2015).
A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.
In *Sixteenth Annual Conference of the International Speech Communication Association*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. (2016).
Achieving human parity in conversational speech recognition.
arXiv preprint arXiv :1610.05256.

References III

Zeghidour, N., Synnaeve, G., Versteegh, M., and Dupoux, E. (2016).
A deep scattering spectrum - deep siamese network pipeline for unsupervised acoustic modeling.
In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4965–4969. IEEE.

Zipf, G. K. (1949).
Human behaviour and the principle of least effort. an introduction to human ecology, hafner reprint, new york, 1972.

Zipf's law for the Buckeye dataset