Open access • Proceedings Article • DOI:10.1109/SLT.2014.7078558

Phonetics embedding learning with side information — Source link

Gabriel Synnaeve, Thomas Schatz, Emmanuel Dupoux

Institutions: School for Advanced Studies in the Social Sciences

Published on: 01 Dec 2014 - Spoken Language Technology Workshop

Topics: Acoustic model, Speech corpus, Voice activity detection, Speech analytics and Time delay neural network

Phonetics Embedding Learning with Side Information

Gabriel Synnaeve[1], Thomas Schatz[1,2], Emmanuel Dupoux[1]

[1]LSCP, ENS-PSL Research University/EHESS/CNRS, Paris, France

[2]SIERRA, INRIA/ENS-PSL Research University/CNRS, Paris, France

gabriel.synnaeve@gmail.com, thomas.schatz@laposte.net, emmanuel.dupoux@gmail.com

Introduction

State-of-the art speech recognition systems rely on the availability of large quantities of human-annotated signals. However, it is also of interest, both for theoretical and practical reasons, to explore the possibility of constructing speech technologies in settings where such a resource is not available.

- Unsupervised/weakly supervised acquisition of the linguistic structure happens in babies [1].
- Top-down (word-level) information can help refine phoneme categories [2]
- Infants can extract words in continuous speech before they learn the phonemes of their language [3].

Method

Related work:

- "Siamese Networks" have similar architecture [4],
- [5] used an assymetric loss function on MNIST,
- Deep semi-supervised embeddings [6].

Inputs and Dynamic Time Warping

Speech is encoded in 40 log-energy Mel-scale filter banks frames every 10ms, each computed over 25ms of speech (Hamming windowed). Separately for each of the tree sets (train, validation, test) of the standard split of the TIMIT corpus, we DTW align "long" words as in Figure 1.

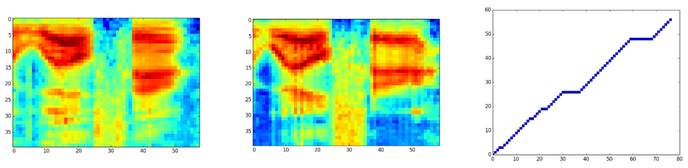

Figure: 1. the 2 left plots: filterbanks (y-axis) along frames (x-axis) for the word "welfare". Right: dynamic time warping of the left-most one to the other.

Model

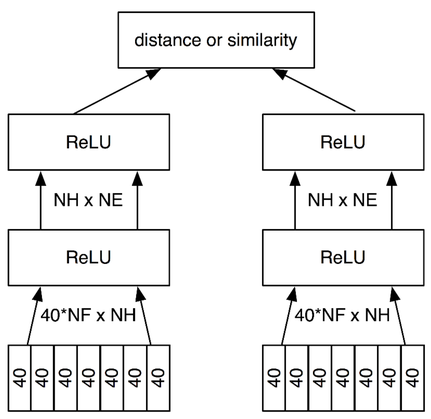

Figure: 2. AB "neural net". We feed to two copies of the same network the aligned stacked frames of a pair of words (A and B). The outputs are compared using a dissimilarity function. During training, the loss function tries to minimize the dissimilarity for "same" pairs and maximize it for "different" pairs. The loss is backpropagated in both sides of the network equivalently.

We tried several loss functions (as shown in Fig. 5), the best results were obtained with a combination of $\cos$ and $\cos^2$ similarities, as follows:

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

with

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

We built our dataset with pairs of "same" word, and we do negative sampling (pairs of "different" words) from the same dataset, with a same/difference ratio of 1.

Results

The network was trained using Adadelta (Adagrad variant of [7]) on a GPU. We trained on the training set of TIMIT (62,625 paired same words, 5.66M frames for "same" and 4.49M for "different"), with 10% held-out as validation set. The code is open source [10].

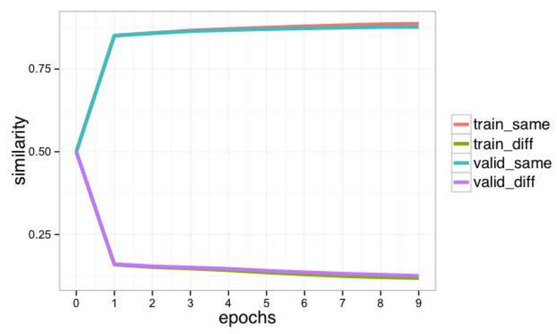

Figure: 3. Similarities ($\cos\cos^2$) for the "same" and "different" subsets of the datasets on the training set (used for learning) and on the validation set (used for early stopping).

ABX scores

We used the minimal pair ABX discrimination task [8]. Given A, B, and X, where A and B are instances of two categories, compute whether X is closer to A or a B, using the frame-wise distance $\mathrm{dist}(A, X)$ vs. $\mathrm{dist}(B, X)$ aggregated along the DTW alignment path.

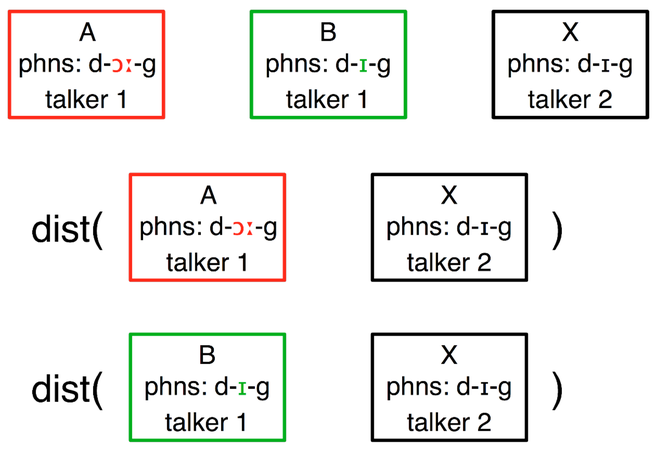

Figure: 4. "Phones accross talker" ABX task

We used the cosine similarity as well as the symmetric Kullback-Leibler divergence as distances in our ABX evaluations shown in Figure 5.

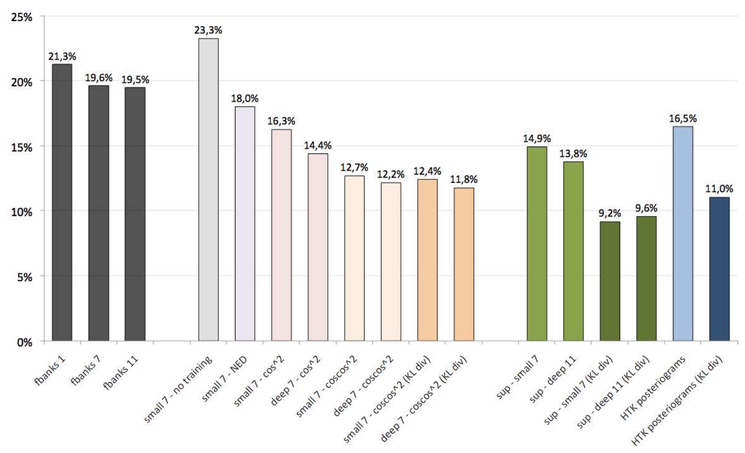

Figure: 5. Average phone discrimination error-rate (ABX task, averaged over talkers and triphone contexts)

Qualitative evaluation

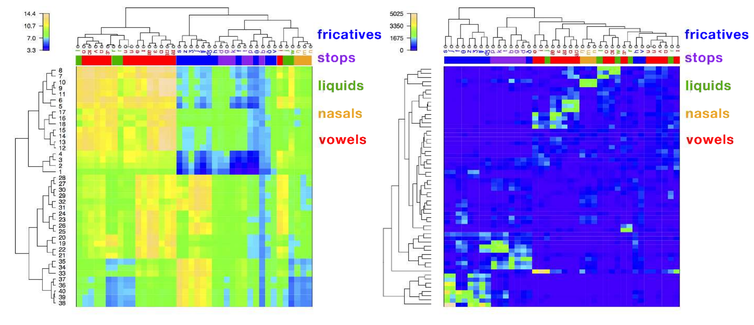

Figure: 6. Bi-clustering of the mean activations of (left) the 40 filterbank features (y-axis) and (right) the most activated 58 embedding features (y-axis), with the phonetic input (x-axis).

Summary and conclusions

It is possible to learn an efficient acoustic model using only word-level same/different annotations. No need for detailed phonetic labeling: it is only necessary to have stretches of speech that are known to the same. An acoustic model is obtained by training shallow and deep neural networks, using an architecture and a cost function well-adapted to the nature of the provided information.

ABX scores for the phones accross talker condition:

- our best: 11.8% ABX error rate
- best raw speech features: 19.6%
- best neural network matched for number of weights: 9.2%, HMM-GMM: 11%

For low-resources speech technology, it provides a practical way to learn an efficient acoustic model. For studies of language acquisition, it lends plausibility to the hypothesis that simple measures of similarity between word-size units of speech signal constitute one of the sources of information that are used by infants when learning the phonetic categories of their language.

To obtain full unsupervision, future work will use spoken terms discovery [9] to seed the inventory of same/different words.

References

[1] Werker, J.F. and Tees, R.C., Influences on infant speech processing: Toward a new synthesis, *Annual review of psychology*, 1999

[2] Swingley, Daniel, Statistical clustering and the contents of the infant vocabulary, *Cognitive psychology*, 2005

[3] Jusczyk, P. and Aslin, R.N., Infants' detection of sound patterns of words in fluent speech, *Cognit. Psychol.*, 1995

[4] Bromley, J. et al., Signature verification using a "Siamese" time delay neural network, *IJPRA*, 1993

[5] Hadsell, R. et al., Dimensionality reduction by learning an invariant mapping, *CVPR*, 2006

[6] Weston, J. et al., Deep learning via semi-supervised embedding. *Neural Networks: Tricks of the Trade* 2012

[7] Zeiler, M.D., ADADELTA: An adaptive learning rate method, *arXiv* 2012

[8] Schatz T. et al., Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline, *INTERSPEECH*, 2013

[9] Park, A. S., and Glass, J. R. Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1), 186–197, 2008.

[10] https://github.com/SnippyHolloW/abnet

NIPS Deep Learning workshop (2014, Montréal, Canada)