Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 95 – 100

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

A Temporal Coherence Loss Function for Learning Unsupervised Acoustic Embeddings

Gabriel Synnaeve[a,b,*], Emmanuel Dupoux[b]

[a] *Facebook A.I. Research, Paris, France*

[b] *École Normale Supérieure / PSL Research University / EHESS / CNRS, France*

Abstract

We train neural networks of varying depth with a loss function which imposes the output representations to have a temporal profile which looks like that of phonemes. We show that a simple loss function which maximizes the dissimilarity between near frames and long distance frames helps to construct a speech embedding that improves phoneme discriminability, both within and across speakers, even though the loss function only uses within speaker information. However, with too deep an architecture, this loss function yields overfitting, suggesting the need for more data and/or regularization.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: unsupervised learning; speech embeddings; speech recognition; temporal coherence; zero resource speech challenge; feature extraction

1. Introduction

Deep Neural Networks (DNNs) are becoming the dominant paradigm for speech technologies, regularly breaking the state of the art obtained previously with Hidden Markov Models and signal processing systems [1,2]. However, being more powerful, DNNs are also more hungry in human annotations: commercially deployed systems require supervised training on thousands of hours of human annotated data. Yet, there are situations where human annotations are not available or two expensive to gather. Half of the human languages, for instance, have no writing system. In addition, the fact that human infants can spontaneously learn their native language through mere immersion in a linguistic environment, shows that it is theoretically possible to learn acoustic and language models with little or no human labels. It is therefore of both practical and theoretical interest to explore the so-called "zero-resource" setting [3,4,5] where linguistic structures are been learned from large amounts of unannotated data.

Here, we examine the idea that useful representations can be learned in a DNN architecture using only generic knowledge about the temporal distribution of phonetic structure. Typically, in any human language, the building blocks of words (phones) have a duration of approximatively 60-150ms [6,7]. Therefore representations with a temporal

* Corresponding author.
E-mail address: gabriel.synnaeve@gmail.com

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.035

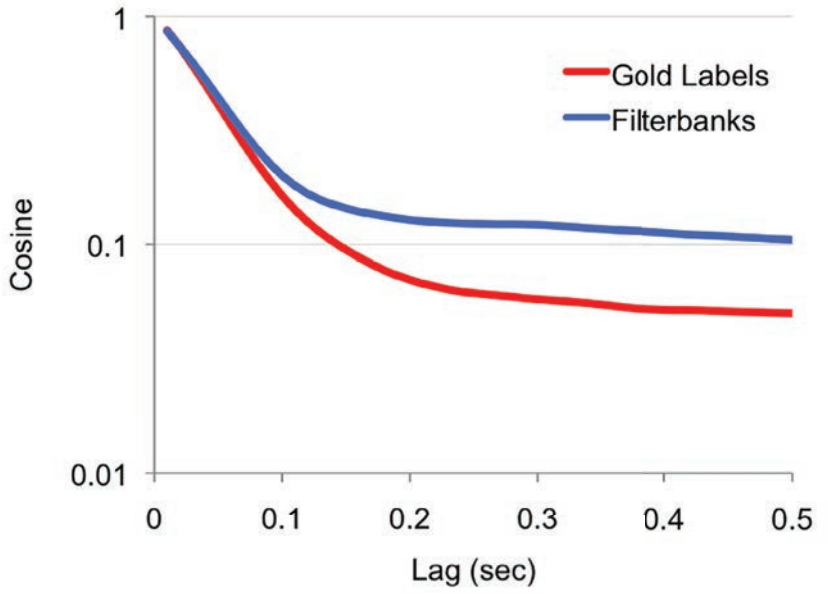

Fig. 1. Cosine similarity between different frames separated by a given time lag. In red is shown the gold labels (cosine of 1 when the labels are the same, and zero when not), in blue is shown the cosine between frames of filterbank values. The scale on the y axis is logarithmic.

profile that is either much smaller or much larger are likely to be not very useful for the purpose of word recognition. The typical duration of phonemes is illustrated in Figure 1, where we have plotted the average cosine similarity between short stretches of speech (frames) separated by different lags. As one can see, the similarity is a decreasing function of lag. On a "gold" phoneme representation (each 10ms frame is represented by a binary N-dimensional vector, where each dimension codes for one of the phoneme classes), the average cosine can be interpreted as the propensity for two frames separated by a given lag to belong to the same phoneme class. When the same plot is done on filterbank representations (each frame is composed of 40 Mel-frequency spectral coefficients over an Hamming window of 25ms, with a step size of 10ms), the same general curve is obtained, but the drop is less steep than for 'gold' labels. This is due to the fact that filterbank representations encode information that change less quickly than phoneme (e.g. information relevant to talker identity), and therefore display more long distance similarity than abstract, talker invariant labels do.

In this paper, we therefore propose a loss function for training DNNs that minimizes the difference between embeddings similarities at short lags (where frames are likely to belong to the same phoneme), and maximizes (shatters them) at long lags (where frames are likely to belong to different phonemes). We implement this idea within a siamese network architecture following our previous work on ABnets[8].

2. Related Work

The idea of using the temporal structure to evaluate speech representations has been proposed in[9]. In this study, the M-measure, defined as a between-frame correlation between posteriogramms of a speech recognizer can be used to predict the word error rate. In a more recent paper[10], the M-measure is modified to become the M-delta, which amounts to the difference in M-measure between a long and short lag correlogram. Models of the temporal structure of phonemes have also be used in systems that try to learn acoustic models from raw speech. As shown in[11], unsupervised clustering applied to continuous speech results in units smaller to phonemes (more akin to phone states). Other studies[12,13] improved on this idea by imposing three-state HMM structure to phonemes, the clustering being done using a Chinese restaurant process. Similarly,[14] model clusterized (unsupervised) phoneme states by splitting word chunks into phoneme-size units using the average duration of the phonemes.

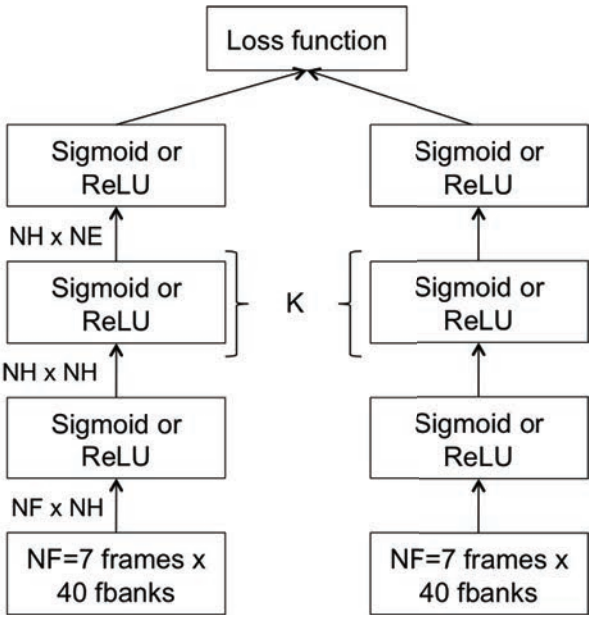

Fig. 2. Architecture of the ABnets used in the present study.

With regard to the model, siamese networks[15] is an architecture in which two copies of the same network are fed with different inputs and trained with an asymmetric loss function which tries to distinguish inputs belonging to same or different classes. It has been used in[16] for learning invariants in images. Other relevant papers that used (and refined) a maximum margin ranking loss include learning an embedding of both images and text[17]. Siamese networks have also been used to predict both phonetic- and speaker-related information in[18].

3. Model

We use a siamese neural architecture as shown in Figure 2. The number of units in each hidden layer was NH=500, the number of units in the output embedding layer was NE=100, and we varied the number of hidden layers to be 1, 3, and 5 (K=0, 2, and 4, respectively). We used the sigmoid non linearity. As input we used 40 log-compressed filterbanks (MFSC, Mel frequency spectral coefficients), stacked for 7 frames.

The loss function (that we call M-delta) is based on the idea of maximizing the similarity between frames that belong to the same phoneme and minimizing it for frames belonging to different phonemes. Here we use, for two inputs A and B a simple asymmetric same/different coscos^2 loss (see[8] for a comparison of other similarities and metrics, also on speech acoustics tasks) in the embedding space (noted Y):

$$\ell_{\text{coscos2}}(A,B) = \begin{cases} (1-\cos(Y_A,Y_B))/2 & \text{if same} \\ \cos^2(Y_A,Y_B) & \text{if different} \end{cases}$$

with

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

Over a whole batch, for a given frame A, we sample four "different" frames $C_1..C_4$, for one pairing "same" frame B. Here, "same" corresponds to frames with a lag of 1, and 'different' to frames with lags of 15, 20, 25, and 30 frames. The overall loss boils down to:

$$\ell(A,B,C_{1..4}) = \frac{1-\cos(Y_A,Y_B)}{2} + \sum_{i=1}^{4} \cos^2(Y_A,Y_{C_i})$$

The network was trained using stochastic gradient descent with Adadelta[19] and early stopping on a held-out validation set, on batches of 100 samples (500 in total with 1 same and 4 different), using the Theano python package[20].

For comparison, we used a fully supervised architecture that gives good performance both in phones classification and ABX tasks[18]. It is constructed as follows: 11 stacked filterbanks, 4 hidden layers of 2400 rectified linear units, and 46 (time aligned) phones as outputs of the logistic regression (with a 37.9% frame-wise classification accuracy). It is trained with dropout[21], and early stopping on the validation set.

4. Evaluation

In order to test for the adequacy of the "phoneme" embeddings that are discovered, we use the minimal pairs ABX evaluation metric[22]. This metric enables to evaluate the quality of a speech representation for the purpose of word recognition, without necessitating the training of a phoneme classifier. It focuses on the minimal distinctions that need to be done to recognize words (i.e., the distinction between one-phoneme neighbors, as in 'bed' and 'bad'), with the idea that if a representation fails to capture these differences, then this is not a good representation. The only information needed is a vector representation of the speech for each frame, plus a frame-wise distance metric. In the ABX task, the system is asked to compute the distance between A and X, B and X, and pick the closest one. For instance A=bed, B=bad, X=bed. This task is increasingly used in the evaluation of speech features and unsupervised systems[23,4].

Here, we setup three versions of this task (see Table 1). The first one is a *within talker* phone discrimination, where A, B and X are spoken by the same talker. The second one is a *between talker* phone discrimination, in which A and B are spoken by the same talker, and X by a different one. The third one is a *between phone* talker discrimination, in which A and B are the same syllable spoken by different talkers but X is a different syllable.

For the three tasks we use as frame-wise distance the symmetrized Kullback-Leibler divergence. For more details about this evaluation, see[8].

Table 1. Example of stimuli in our three ABX tasks.

Task	A	B	X
phone within talker	bed_{T1}	bad_{T1}	bed_{T1}
phone between talker	bed_{T1}	bad_{T1}	bed_{T2}
talker between phone	bed_{T1}	bed_{T2}	bad_{T2}

5. Results

All the experiments were conducted on about 1/3rd (12 speakers) of the Buckeye (spontaneous speech) corpus[†]. Our code is available online[‡].

We present in Figure 5 the ABX score for the three tasks for the M-delta trained siamese networks varying in number of hidden layers and for comparison, the same tasks run on the raw input features, and on the final hidden layer of a fully supervised model.

The result show that, compared to the raw filterbank features baseline, the three M-delta trained networks improve the scores on the phoneme discrimination tasks, and degrade the performance on the speaker discrimination task. This degradation was expected, since the speaker identity remains constant throughout a sentence. Therefore, forcing the M-delta loss to be maximal effectively induces the DNN to remove these constant speaker-specific features. This is confirmed through the improvement on the across-speaker phone discrimination score over the baseline. Interestingly, the M-delta networks also improves within-speaker discrimination, showing that the loss function is removing constant (or slow moving) features beyond speaker features.

[†] http://buckeyecorpus.osu.edu
[‡] https://github.com/SnippyHolloW/abnet

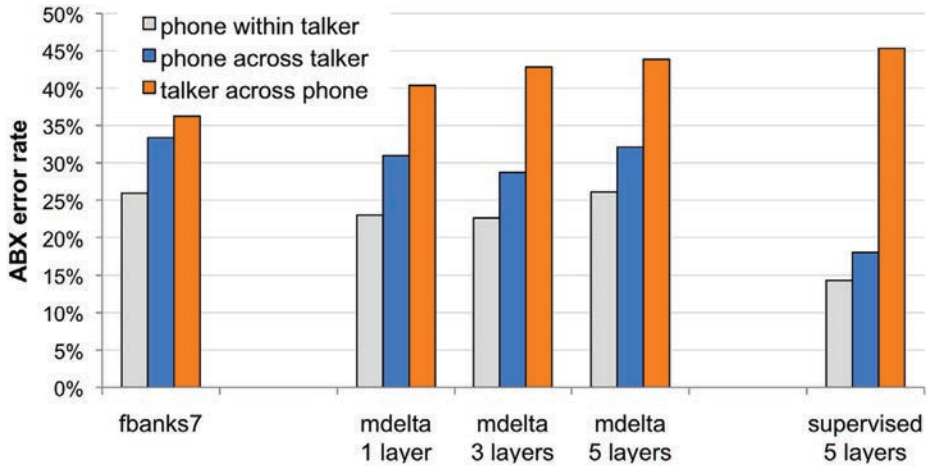

Fig. 3. ABX error rates on the raw filterbanks, on the output layer of the M-delta models and on the on last hidden layer of a supervised DNN on three ABX tasks: two of phone discrimination, and one of speaker discrimination.

Two other aspects of the results are worth noting: First, the performance does not increase monotonically with the number of hidden layers. In fact, the optimal performance (for the phoneme discrimination tasks) is found with 3 hidden layers, and degrades somewhat with 5 hidden layers, whereas the training error is lower with more layers (not shown here). This is perhaps due to the fact that the unregularized M-delta loss is not constraining enough for a large DNN to learn without overfitting. Second, while the performance is better than baseline, the M-delta loss does not reach that of the supervised topline. Interestingly, though, the topline performance on speaker discrimination is not that different from that of the M-delta networks with 3 and 5 hidden units, suggesting that we have reached some kind of optimum, at least in terms of removing talker effects.

6. Conclusion

We have demonstrated that extremely general temporal information (the M-delta loss) can help in learning a representation that is more speaker invariant than the original spectrographic representation. Note that this was achieved through a very generic kind of information: we only try to maximize the difference in similarity between very close frames (lag of 1) and distant frames (more than 15 frames). We did not try to incorporate more specific temporal information, such as fitting the red curve in Figure 1, or by taking into account phoneme-specific temporal information. Yet, we obtained substantial improvement (3-4% absolute, 13-14% relative) in error rate on a phone discrimination task both within and across talkers. In parallel, our learner representation lost the ability to perform speaker discrimination (with an increase in error rate up to 8% absolute, 21% relative).

In brief, the M-delta loss enables to learn a representation of speech sounds that is more speaker invariant than the filterbank representation. It is, however, far from achieving perfect invariance compared to supervised losses. The gain in performance achieved with the M-delta loss is only about a third of the improvement obtained by supervised learning. It remains to be seen whether the M-delta loss could be combined with some other loss, such as the same-different loss[8,5], or an auto-encoder reconstruction loss[24] as a way to improve the phoneme discrimination score without supervised labels.

Finally, not covered in this study is a thorough examination of the optimum parameters for the M-delta loss: ratio of positive to negative examples, network architecture, etc. It also remains to be seen whether performance increases with the size of the dataset, or whether some asymptote is reached after a critical amount of data.

7. Acknowledgment

This project is supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, the Region Ile de France (DIM cerveau et pense), and an AWS in Education Research Grant award. We thank Nicolas Usunier for proof-reading.

References

1. Mohamed, A., Dahl, G., Hinton, G.. Deep belief networks for phone recognition. In: *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*. 2009.
2. Dahl, G.E., Sainath, T.N., Hinton, G.E.. Improving deep neural networks for lvcsr using rectified linear units and dropout. ICASSP; 2013.
3. Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., et al. A summary of the 2012 JH CLSP workshop on zero resource speech technologies and models of early language acquisition. In: *Proceedings of ICASSP 2013*. 2013.
4. Versteegh, M., Thiollière, R., Schatz, T., Xuan-Nga, C., Anguera, X., Jansen, A., et al. The zero resource speech challenge 2015. In: *INTERSPEECH-2015*. 2015.
5. Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., Dupoux, E.. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
6. Umeda, N.. Vowel duration in american english. *The Journal of the Acoustical Society of America* 1975;**58**(2):434–445.
7. Umeda, N.. Consonant duration in american english. *The Journal of the Acoustical Society of America* 1977;**61**(3):846–858.
8. Synnaeve, G., Schatz, T., Dupoux, E.. Phonetics embedding learning with side information. In: *IEEE Spoken Language Technology Workshop (SLT 2014)*. IEEE; 2014, doi:10.1109/slt.2014.7078558.
9. Hermansky, H., Variani, E., Peddinti, V.. Mean temporal distance: predicting asr error from temporal properties of speech signal. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE; 2013, p. 7423–7426.
10. Hermansky, H., Burget, L., Cohen, J., Dupoux, E., Feldman, N., Godfrey, J., et al. Towards machines that know when they do not know: Summary of work done at 2014 frederick jelinek memorial workshop in prague. In: *ICASSP-2015 (IEEE International Conference on Acoustics Speech and Signal Processing)*. 2015.
11. Varadarajan, B., Khudanpur, S., Dupoux, E.. Unsupervised learning of acoustic subword units. In: *Proceedings of ACL-08: HLT*. 2008, p. 165–168.
12. Lee, C.y., Glass, J.. A nonparametric Bayesian approach to acoustic model discovery. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. 2012, p. 40–49.
13. Siu, M.h., Gish, H., Chan, A., Belfield, W., Lowe, S.. Unsupervized training of an HMM-based self-organizing recognizer with applications to topic classification and keyword discovery. *Computer Speech & Language* 2013;**preprint**.
14. Jansen, A., Church, K.. Towards unsupervised training of speaker independent acoustic models. In: *Proceedings of INTERSPEECH*. 2011, p. 1693–1696.
15. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., et al. Signature verification using a siamese time delay neural network. *Internat Journ of Pattern Recog and Artific Intell* 1993;**7**(04):669–688.
16. Hadsell, R., Chopra, S., LeCun, Y.. Dimensionality reduction by learning an invariant mapping. In: *Computer vision and pattern recognition, 2006 IEEE computer society conference on*; vol. 2. IEEE; 2006, p. 1735–1742.
17. Weston, J., Bengio, S., Usunier, N.. Wsabie: Scaling up to large vocabulary image annotation. In: *IJCAI*; vol. 11. 2011, p. 2764–2770.
18. Synnaeve, G., Dupoux, E.. Weakly supervised multi-embeddings learning of acoustic models. In: *ICLR*. 2014.
19. Zeiler, M.D.. Adadelta: An adaptive learning rate method. *arXiv preprint:12125701* 2012.
20. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., et al. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop; 2012.
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 2014;**15**(1):1929–1958.
22. Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hynek, H., Dupoux, E.. Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In: *INTERSPEECH-2013*. 2013, p. 1781–1785.
23. Schatz, T., Peddinti, V., Cao, X.N., Bach, F., Hermansky, H., Dupoux, E.. Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise. In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.
24. Deng, L., Seltzer, M.L., Yu, D., Acero, A., Mohamed, A.R., Hinton, G.E.. Binary coding of speech spectrograms using a deep auto-encoder. In: *Interspeech*. Citeseer; 2010, p. 1692–1695.