A DEEP SCATTERING SPECTRUM - DEEP SIAMESE NETWORK PIPELINE FOR UNSUPERVISED ACOUSTIC MODELING

Neil Zeghidour[1] *Gabriel Synnaeve*[2] *Maarten Versteegh*[1*] *Emmanuel Dupoux*[1]

[1]École Normale Supérieure / PSL Research University / EHESS / CNRS, France
[2] Facebook A.I. Research, Paris, France

ABSTRACT

Recent work has explored deep architectures for learning acoustic features in an unsupervised or weakly-supervised way for phone recognition. Here we investigate the role of the input features, and in particular we test whether standard mel-scaled filterbanks could be replaced by inherently richer representations, such as derived from an analytic scattering spectrum. We use a Siamese network using lexical side information similar to a well-performing architecture used in the Zero Resource Speech Challenge (2015), and show a substantial improvement when the filterbanks are replaced by scattering features, even though these features yield similar performance when tested without training. This shows that unsupervised and weakly-supervised architectures can benefit from richer features than the traditional ones.

Index Terms— speech recognition, scattering transform, siamese network, ABnet, ABX

1. INTRODUCTION

In the task of learning an acoustic model, unsupervised systems are on the rise [1], however the gap between these systems and the supervised ones remains considerable.

In supervised ASR, standard spectral features such as mel-filterbanks or MFCCs are often used to represent the acoustic signal. The emphasis is on improving the supervised classifier, since a good classifier can compensate for flaws in the representation. An example of this is the elimination of noisy and uninformative features. However, in an unsupervised setting, we lack the guidance of the labels to help a system learn to select features, scale them relative to their importance in the classification task, and extract useful information. Unfortunately, unsupervised algorithms are very sensitive to these drawbacks, hence learning a good acoustic representation is the key to success for unsupervised speech recognition.

In previous work, several architectures have been explored to learn an acoustic model in an unsupervised or weakly-supervised way [2, 3, 4, 5]. However, regardless of the sophistication of these algorithms, their performance remains inherently limited by the amount of information available in their input representation.

In this study we investigate the importance of choosing an appropriate input representation to learn an acoustic model in an unsupervised or weakly-supervised way. We focus on one particular model, the ABnet [2, 5], a Siamese network that learns a phone representation from pairs of words, and we study the effect of switching its input representation from mel-filterbanks to the scattering spectrum [6].

Information is lost in the mel-fiterbanks computation process, mainly when averaging the spectrogram over the filters, and this loss of information puts an upper bound on the learning potential of classifiers that are built on these features. The untransformed waveform is the richest representation possible, and some supervised systems are able to learn phonetic classes directly from it [7]. However, in an unsupervised or weakly-supervised setting, it would be extremely difficult to extract information directly from the waveform. Here, we strike a middle ground by replacing the filterbanks by a deep scattering spectrum, a representation that has many of the desirable properties of standard spectral features, i.e. they are stable and can be efficiently exploited by classifiers, while retaining more information than filterbanks.

The current study will show that both in a weakly supervised setting with gold word-level annotations on the TIMIT (American English) corpus and in a purely unsupervised setting on the Buckeye (American English) and NCHLT (Xitsonga) corpora, combining the scattering spectrum and the ABnet significantly improves the learned representation, in terms of its ABX error [8, 9], a score that characterizes the discriminability of phone classes in the embedding space. This improvement holds in comparison to an ABnet trained on standard spectral features, and even in comparison to supervised systems.

2. METHODS

2.1. Scattering Transform

For a signal x we define the following wavelet transform Wx as a convolution with a low-pass filter ϕ and higher frequency

*Currently at TextKernel B.V.

complex analytic wavelets ψ_{λ_1}:

$$Wx = (x \star \phi(t), x \star \psi_{\lambda_1}(t))_{t\in\mathbb{R},\lambda_1\in\Lambda_1} \quad (1)$$

We apply a modulus operator to the wavelets coefficients to remove complex phase and extract envelopes at different resolutions:

$$|W|x = \Big(x \star \phi(t)\,,\, |x \star \psi_{\lambda_1}(t)|\Big)_{t\in\mathbb{R},\lambda_1\in\Lambda_1} \quad (2)$$

$S_0x = x \star \phi(t)$ is locally invariant to translation thanks to the time averaging ϕ. This time-averaging loses the high frequency information, which is retrieved in the wavelet modulus coefficients $|x \star \psi_{\lambda_1}|$. However, these wavelet modulus coefficients are not invariant to translation, and as for S_0 a local translation invariance is obtained by a time averaging, which defines a first layer of scattering coefficients:

$$S_1x(t,\lambda_1) = |x \star \psi_{\lambda_1}| * \phi(t) \quad (3)$$

It is shown in [6] that if the wavelets ψ_{λ_1} have the same frequency resolution as the standard mel-filters, then the S_1x coefficients approximate the mel-filterbanks coefficients. Unlike with the mel-filterbanks computation process, we here have a strategy to recover the lost information, by passing the wavelet modulus coefficients $|x \star \psi_{\lambda_1}|$ through a bank of higher frequency wavelets ψ_{λ_2}:

$$|W_2|\,|x\star\psi_{\lambda_1}| = \Big(|x\star\psi_{\lambda_1}|\star\phi\,,\, ||x\star\psi_{\lambda_1}|\star\psi_{\lambda_2}|\Big)_{\lambda_2\in\Lambda_2} \quad (4)$$

This second layer of wavelet modulus coefficients is still not invariant to translation, hence we average these coefficients with a low-pass filter ϕ to derive a second layer of scattering coefficients:

$$S_2x(t,\lambda_1,\lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \quad (5)$$

Repeating these successive steps of computing invariant features and retrieving lost information leads to the scattering spectrum, as seen in Fig. 1, however speech signals are almost entirely characterized by the first two layers of the spectrum, that is why a two layers spectrum is typically used for speech representation. It is shown in [6] that this representation is invariant to translations and stable to deformations, while keeping more information than the mel-filterbanks coefficients.

2.2. ABnet

Siamese networks were first introduced for written signatures verification [10]. The main intuition behind these architectures is that given an abstract notion of similarity on the data we can use pairwise relations between samples to learn a representation where the distance between the embeddings of objects will reflect the abstract similarity between these objects.

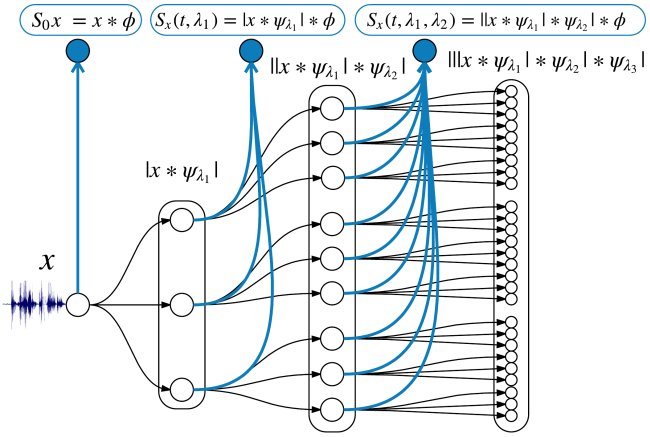

Fig. 1. A scattering spectrum with two layers.

In other words, we want to learn a mapping $\mu(X)$ such that for a certain similarity metric D in the embedding space we have $D(\mu(X_1), \mu(X_2))$ small if X_1 and X_2 are *same* objects, and large if they are *different*. Because of this architecture, the supervision required by a Siamese network consists of pairs of samples that are labeled *same* or *different*, rather than labels for individual samples.

An ABnet is a particular case of Siamese neural network. It uses pairs of words to learn a representation of phones. Once words are paired, the feature frames that constitute them are aligned with Dynamic Time Warping (DTW) [11] and make the pairs of samples that are fed to the ABnet. The motivation for using such lexical feedback comes from the fact that the lexicon is typically quite sparse in phonetic space. As a result, two randomly selected words will mismatch in most of their phonemes. This makes lexical clustering an easier task than phoneme clustering. In addition, experiments and computational models in psycholinguistics have shown that lexical information can help refine phonetic categories [12, 13].

The ABnet is composed of two copies of the same network, each copy being fed with one of the elements of a pair of input samples. These identical networks project the samples into the embedding space, through several hidden layers. A measure of similarity or distance is then computed between the two pairs depending on their relation label (*same* or *different*), and the error is propagated evenly in the two copies. An ABnet can be seen in Fig. 2.

3. ABX EVALUATION

The standard way of evaluating features is to train a supervised classifier and compare the classification performance to the performance we get with a similar classifier trained on other features. However, supervised classifiers can compensate for properties of the features that would constitute con-

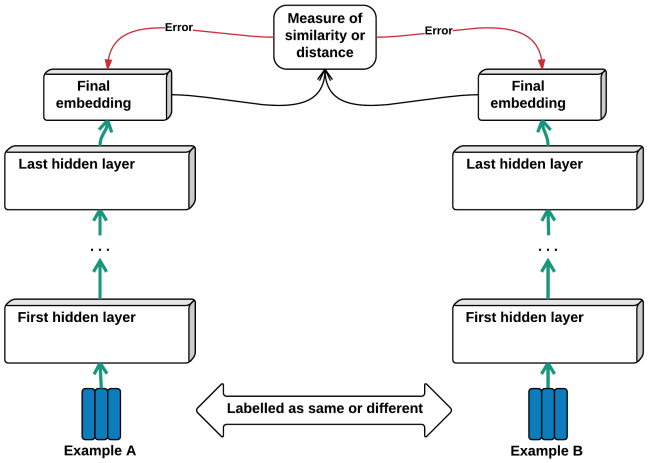

Fig. 2. The ABnet architecture.

siderable flaws in an unsupervised setting (e.g. poor scaling, uninformative dimensions). Hence, supervised classification performance obtained on features is not a reliable indicator of the performance of these features in an unsupervised setting. Rather, we evaluate a different property of the representation, its phonetic discriminability, i.e. how well phonetic classes are separated in the embedding space, since well separated classes lead clustering algorithms to discover meaningful clusters. This evaluation is done by computing an ABX score [8, 9].

An ABX task consists in presenting three stimuli A, B and X, with A and B belonging to different categories and X matching the category of either A or B, let us assume in this example that X belongs to the same category as A. Distances $D(A, X)$ and $D(B, X)$ are computed in the embedding space and compared. If $D(A, X) < D(B, X)$ then the score is 1 (success), else it is 0 (failure).

The experiments in this paper are evaluated with a particular type of ABX task, adapted for speech, the triphone minimal-pair ABX task. A minimal pair is a pair of sounds, composed each of three phonemes, that only differ by their central phoneme ("beg" vs "bag"). In section 4.1 we perform this task across-speaker (A: "bag" by speaker 1, B: "beg" by speaker 1, X: "bag" by speaker 2). Since this task is particularly hard, a representation that yields a good ABX score on such a task can be considered a good representation for phone recognition. In section 4.2 we also perform this task within-speaker (A: "bag" by speaker 1, B: "beg" by speaker 1, X: "bag" by speaker 1). A global score is obtained by averaging ABX errors over all relevant triplets that can be found in the corpus. We obtain an error between 0% and 50% (the chance level), a low error characterizing a representation in which categories are well separated from each other.

4. EXPERIMENTS

In the following experiments, we use a two-layers scattering spectrum over 16ms windows, with normalized second order coefficients and log-frequency scattering [6]. The features are computed with the ScatNet toolbox [6].

4.1. Weakly-supervised phone representation learning on TIMIT

The TIMIT dataset [14] is a corpus of clean read speech containing a set of 10 sentences read by 630 speakers of eight major dialects of American English. All the words of more than 5 characters that are repeated in the corpus are extracted and matched as pairs of *same*. This yields 62,625 pairs of same words represented as time bounding-boxes in the signal. We extract the scattering features within these boxes and align them with DTW, yielding 6.77M feature frames. We then sample the same number of *different* pairs, which are sampled randomly. Even if there is a risk of "false negatives" i.e. labeling frames as *different* while they actually have the same phonetic content, this probability is relatively low due to the distribution of the 39 phonemes inside the English language. The pairs of *different* objects are not aligned with DTW but just aligned on the shortest one, since DTW looks for matching acoustic units and would thus increase the risk of getting false negatives.

Fig. 3 shows ABX errors on the across-speaker task. The distances used for the ABX tasks are the cosine distance for the raw features, and the symmetric KL-divergence for all trained models. "Shallow" models have one hidden layer while "Deep" models have three. Even though raw scattering features do not yield a better ABX error than mel-filterbanks, their use as an input representation leads to a substantial improvement after training an ABnet, with a best error of 9.8% against 11.8% for the best ABnet trained on mel-filterbanks. Our best Scattering-ABnet model even gives a better ABX score than the HMM-GMM posteriorgrams (11%), very close from the output of the deep supervised network (9.6%). In fact, changing the input representation of the Shallow ABnet from mel-filterbanks to scattering coefficients has an impact on the ABX error (from 12.4% to 10.2%) that is 3.7 times higher than adding hidden layers to get a Deep ABnet (from 12.4% to 11.8%).

4.2. Unsupervised phone representation learning on Buckeye and NCHLT

In this experiment we run our model under the conditions of the Zero Resource Speech Challenge 2015 [15]. One of the tasks in this challenge was unsupervised acoustic modeling. The challenge provided two data sets (a subset of the Buckeye Corpus of conversational English [16] and a subset of the NCHLT corpus of Xitsonga [17]) and baseline and topline ABX scores for both data sets, both for within- and

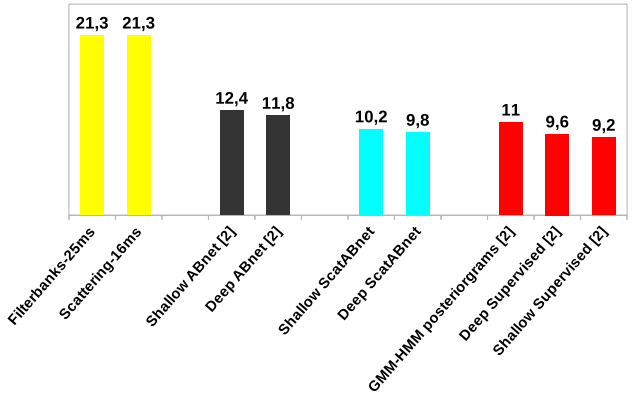

Fig. 3. From left to right: Across-speaker ABX error on TIMIT (as percentages), measured on raw features (yellow bars), best ABnet models trained on mel-filterbanks (purple bars), best ABnet models trained on scattering spectrum (blue bars), and outputs of three supervised systems (red bars).

across-speakers. The baselines provided by the challenge are MFCC's, the toplines GMM-HMM posteriorgrams. In the spirit of the challenge, we extract the pairs of speech segments used for training our model in an unsupervised manner. That is, rather than taking matching words from the gold transcription, we extract them from the signal by an unsupervised spoken term discovery (STD) algorithm [18]. This algorithm discovered 3149 pairs of similar segments of speech from the English corpus and 1782 pairs from the Xitsonga corpus, 50% being used for training and 50% for early stopping. These pairs form the *same* input to the ABnet. The *different* input is composed of randomly matched segments of speech. Here, the "Deep" ScatABnet architecture consists of 2 layers of 500 nodes, with a sigmoid activation function, exactly as in a previously published study using an ABnet with filterbanks [19]. The "Shallow" one has only one hidden layer.

We compare our model against the challenge baselines (MFCC) and toplines (supervised HMM-GMM posteriorgrams) and also against the best performing system submitted to the challenge [20], a DPGMM system that takes as input talker-normalized MFCC's. In Table 1, we can see that the both ABnet variants perform better than the baseline. For English within-speaker, both systems actually outperform the supervised topline. ScatABnet has lower error scores than the FbanksABnet on all conditions except Xitsonga within-speaker. The table further shows that ScatABnet is competitive with the state of the art system of [20], in one case, across-speaker for Xitsonga, producing the lowest ABX error. These performances are remarkable given that the number of pairs on which the ABnet is trained is much lower than for the TIMIT. This low number of pairs can also explain why here a shallow architecture with fewer parameters to learn gives a higher performance than a deep one.

Model	English		Xitsonga	
	within	across	within	across
Baseline (MFCC)	15.6	28.1	19.1	33.8
Topline (Supervised)	12.1	16.0	3.5	4.5
FbanksABnet [19]	12.0	17.9	11.7	16.6
Deep ScatABnet	11.3	17.1	12.5	16.2
Shallow ScatABnet	11.0	17.0	12.0	**15.8**
DPGMM [20]	**10.8**	**16.3**	**9.6**	17.2

Table 1. ABX error (as percentages) on the ZeroSpeech 2015 datasets (English, Xitsonga) for the ABX within- and across-speaker tasks. The best scores for each condition are in **bold**.

5. CONCLUSION AND FUTURE WORK

This study confirms that the input representation of deep architectures has a substantial impact on the performance of the pipeline for acoustic modeling. The experiments on TIMIT in section 4.1 show that switching from standard spectral features to the scattering spectrum yields a substantial gain (about 17% in relative error rate), a higher gain than switching from a shallow to a deep network. This shows that in acoustic representation learning, putting more emphasis on the input representation might give a larger performance increase than improving the learning architecture. These results suggest that deep architectures that are trained on standard spectral features are not exploited to their full potential, as previously shown in a supervised setting in [21].

For future work we will use the output of our system as input to the spoken term discovery system, thus bringing it full circle. Since our system has improved the representation at the phone level, we expect the spoken term discovery system to find more and better matching pairs than with its original input features.

Another way of improving the system would be to go beyond spectral features: since using a representation with more information releases the potential of the learning algorithm, we might try to train the ABnet directly on the raw signal. As said before, this is currently done in supervised systems, but it still remains very hard to do in an unsupervised setting. Limiting the search space by imposing strong structural constraints on the architecture (e.g. convolutional lower layers) could make this objective possible.

6. ACKNOWLEDGEMENTS

NZ, MV, and ED's research was funded by the European Research Council (ERC-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG) and the Fondation de France. It was also supported by ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC.

7. REFERENCES

[1] A. Jansen, E. Dupoux, S. Goldwater, M. Johnson, S. Khudanpur, K. Church, N. Feldman, H. Hermansky, F. Metze, R. Rose, 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*.

[2] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 106–111.

[3] L. Badino, A. Mereta, and Lorenzo Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *Proceedings of Interspeech*, 2015.

[4] D. Renshaw, H. Kamper, A. Jansen, and Sharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proceedings of Interspeech*, 2015.

[5] Gabriel Synnaeve and Emmanuel Dupoux, "Weakly supervised multi-embeddings learning of acoustic models," in *ICLR*, 2014.

[6] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *Signal Processing, IEEE Transactions on*, vol. 62, no. 16, pp. 4114–4128, 2014.

[7] D. Palaz, R. Collobert, et al., "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of Interspeech*, 2015, number EPFL-CONF-210029.

[8] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "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*, 2013, pp. 1–5.

[9] T. Schatz, V. Peddinti, X-N. Cao, F.Bach, H.Hermansky, and E.Dupoux, "Evaluating speech features with the minimal-pair abx task (ii): Resistance to noise," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[12] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "Learning phonetic categories by learning a lexicon," in *Proceedings of the 31st annual conference of the cognitive science society*, 2009, pp. 2208–2213.

[13] Abdellah Fourtassi and Emmanuel Dupoux, "A rudimentary lexicon and semantics help bootstrap phoneme acquisition," *CoNLL-2014*, p. 191, 2014.

[14] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall, "The darpa speech recognition research database: specifications and status," in *Proc. DARPA Workshop on speech recognition*, 1986, pp. 93–99.

[15] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, "The zero resource speech challenge 2015," in *Proc. of Interspeech*, 2015.

[16] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," www.buckeyecorpus.osu.edu, 2007.

[17] N.J. de Vries, M.H. Davel, J. Badenhorst, W.D. Basson, F. de Wet, E. Barnard, and A. de Waal, "A smartphone-based asr data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[18] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 401–406.

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

[20] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "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*, 2015.

[21] V. Peddinti, T. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep scattering spectrum with deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 210–214.