A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling

Roland Thiollière, Ewan Dunbar*, Gabriel Synnaeve**
*Maarten Versteegh, Emmanuel Dupoux**

Ecole Normale Supérieure / PSL Research University / EHESS / CNRS, France

rolthiolliere@gmail.com, emd@umd.edu, gabrielsynnaeve@gmail.com
maartenversteegh@gmail.com, emmanuel.dupoux@gmail.com

10.21437/Interspeech.2015-640

Abstract

We report on an architecture for the unsupervised discovery of talker-invariant subword embeddings. It is made out of two components: a dynamic-time warping based spoken term discovery (STD) system and a Siamese deep neural network (DNN). The STD system clusters word-sized repeated fragments in the acoustic streams while the DNN is trained to minimize the distance between time aligned frames of tokens of the same cluster, and maximize the distance between tokens of different clusters. We use additional side information regarding the average duration of phonemic units, as well as talker identity tags. For evaluation we use the datasets and metrics of the Zero Resource Speech Challenge. The model shows improvement over the baseline in subword unit modeling.

Index Terms: zero resource speech challenge, feature extraction, deep learning

1. Introduction

The automatic discovery of linguistic units from the raw speech stream may seem to be a daunting task from a scientific and technological point of view. Yet, the fact that infants spontaneously converge on what amounts to a functional speech recognizer (language-tuned acoustic and language models) within a year or so, by mere immersion in a linguistic community [1, 2], indicates that it is not an impossible one. Infants do not construct a speech recognizer by being fed hours of speech paired with phone labels. However, their task is not totally unsupervised either. Apart from the fact that infant speech input is accompanied by multimodal signals that may contain relevant side information (visual context, social signals etc.), the speech signal itself is produced by an adult linguistic system which has a particular universal structure that the infant learner could exploit. Here, we will exploit one such source of information; the fact that speech contains hierarchically organized levels of structure: utterances are made of words, and words are made of phonemes.

In particular, our paper rests on two critical assumptions regarding words and phonemes: (1) word types are more separated in acoustic space than phoneme classes, (2) pairs of word tokens share most (or all) of their phonemes if they belong to the same class, and differ in most of their phonemes if not. At a high level, this suggests the following learning strategy: first, discover word-like units from the signal using spoken term discovery (assumption 1), then, use the discovered word-like units to construct a word 'tutor' that trains phoneme-like representations (assumption 2). The tutor works by aligning each pair of word-like tokens and declaring that the aligned frames represent the same "phonemes" (abstract speech feature representations) when the word-like tokens are in the same class, and different ones when they are in different classes. We learn these using a deep neural network.

Previous work in the psycholinguistic literature has shown that lexical information can help in discovering phoneme identity. Using a Bayesian model, Feldman et al. [3] demonstrated the feasibility of this principle in a toy model with simple Gaussian data. Others (cf. [4, 5]) demonstrated using transcribed speech that such top down information can help to cluster allophones into phonemes. Synnaeve et al. [6, 7] showed that, using oracle word labels, a deep neural network can be trained to yield a phonetic embedding whose performance on cross-speaker phoneme discrimination improves substantially on the raw input features. This architecture, ABNET, is described below. Finally, Jansen et al. [8], using an architecture on denoising auto-encoders, found that both oracle word labels and speech fragments discovered automatically can yield good speech features. Here we test whether the ABNET architecture can be coupled with a spoken term discovery system, discovering words to help discover phonemes.

In addition to exploiting the relationships between phonemes and words, our paper explores two other potential sources of side information. The first one is that speech not only conveys linguistic information, but also information pertaining to speaker identity. Here, we assume that speaker identity is accessible to infants, either through an analysis of the speech signal alone (as in speaker diarization) or using multimodal input. With a DNN architecture, a common technique is to concatenate a speaker embedding with the input features during the training of the network [9, 10]. We will explore the fact that having access to speaker identity information is useful, since it provides information about what information to *ignore* or *normalize for* in the speech signal. The second source of information could be an innate knowledge of the temporal properties of phoneme units in human languages. Typically, phonemes have a duration around 70ms; one could use this source of knowledge to either discard sections of speech that are corrupted and therefore do not obey this structure, or to discard potentially flawed feature representations for speech. Several papers have used the difference in autocorrelation between adjacent and distant frames as a measure of the quality of a speech recognizer [11, 12], but, to our knowledge, not to aid in weakly supervised learning.

* These authors contributed equally to this work.

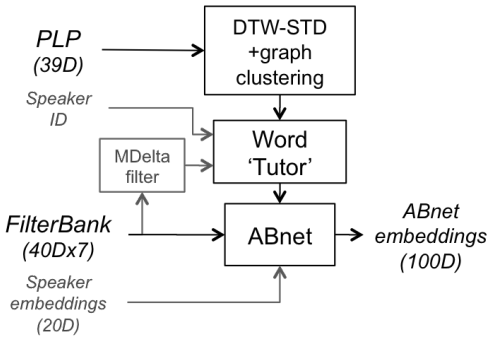

Figure 1: Overview of the components of our system.

2. System

The general organization of our system is displayed in Figure 1. Each of the components is described in the sections that follow.

2.1. Spoken term discovery

We use the general framework of spoken term discovery systems proposed in [13]. These systems search for repeated acoustic patterns using dynamic time warping (DTW) across the entire frame-level similarity matrix. Typically, repeated patterns show up as diagonal stretches of high acoustic similarity, which are then segmented out and clustered together into classes. This process is computationally intensive and several techniques exist to speed up the pattern discovery. Here, as in [14], we use random vector projections followed by bit quantization to make the process computationally tractable. The algorithm outputs pairs of matched fragments together with their DTW score (the average similarity along the matching path). Further post processing involves recomputing the DTW scores using the real cosine function on the original input features, and applying connected component clustering of the found fragments, as in [13], to construct classes. In this paper, we used the implementation of [14], with a similarity threshold of .5, a DTW threshold of .89, and a connected component threshold of .98. This is the system used to generate the baseline of the Zero Resource Speech Challenge Track 2.

2.2. Word Tutor

The input to the ABNET system is based on a set of patterns grouped into classes, discovered by the STD system. To obtain inputs for ABNET, we take the pairs of matching (same-class) patterns and convert them into collections of pairs of matching fixed-width fragments, aligned by DTW with frame-level cosine distance on the feature representations. We do the same for a subset of all pairs of mismatching (different-class) patterns, except we just align them along the diagonal (truncating the longest pattern). We use a subset, first, in order to have the same number of matching as mismatching word pairs, and second, to balance the matching and mismatching word pairs to have both same-speaker pairs and different-speaker pairs. (The ratio of same to different speaker was left free to vary depending on the input fragments.) This avoids a statistical bias in which the information about matching versus mismatching phonetic content is accidentally correlated with same versus different speaker (which would turn the task into an easier speaker identity task).

2.3. ABnet

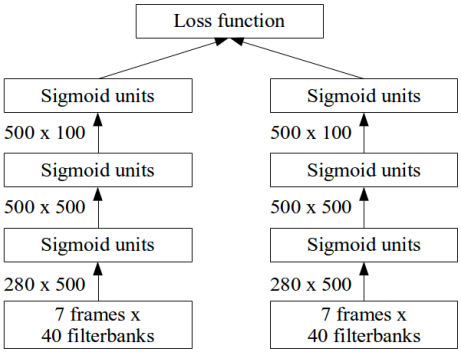

Figure 2: The Siamese neural network.

This part of the system learns a phonetic embedding, i.e., a vector representation of speech sounds. It is based on [6]. We use a Siamese network architecture [15] (see Figure 2) in which we stack 7 frames of features in the input layer (a center frame, and 3 context frames on each side), followed by 2 layers of 500 units, and a final output layer of 100 units. The activation functions are sigmoid functions. Two identical copies of the same network are fed by the features of the members of each pair, A and B. These are then forward-propagated in the ABNET, where we finally use an asymmetric loss as in [16] for learning invariants in images, with a margin as in [17]. In the embedding, Y, we get:

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

where

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

Over a whole batch, as we supply negative samples at a positive to negative ratio of 1:1, this reduces to a loss of:

$$\mathcal{L}(A,B,C) = \frac{1-\cos(Y_A,Y_B)}{2} + \cos^2(Y_A,Y_C)$$

We refer to this loss function as COSCOS2. Roughly speaking, this loss function is at its minimum when the vector representations are collinear for matching frames and orthogonal for mismatching frames. This COSCOS2 loss was shown to perform well in [6]. The network is initialized with random weights and then trained on 50% of the data. The rest of the data is used as a heldout validation set, which is tested to stop training in order to prevent the network from overfitting. The networks were trained for a maximum of 500 epochs by mini-batch stochastic gradient descent (using Adadelta [18]) on an Nvidia K20 Tesla GPU. The ABNET code uses the Theano library[19, 20], and is freely available [21]. This neural network outputs 100-dimensional vectors, on which the pairwise distance is best characterised by the symmetrized Kullback-Leibler (KL) divergence.

2.4. Phoneme duration

To make use of prior information about phoneme duration, we apply an additional filter on the results of the STD system to

remove patterns that do not clearly signal phoneme duration in the acoustic signal. To do this, we calculate a version of the M-delta measure of [12]. This calculation begins with a reference curve, derived independently from a phone-labelled corpus, (in this case the TIMIT corpus [22]), giving the overall probability of a phone label being repeated at various time lags. At short lags, (e.g., one or two frames away), the probability is relatively high, and the shape of the curve gives an indication of phoneme duration.

The M-delta measure quantifies how well this repetition probability curve is reflected in a measure of acoustic divergence at the same lags. In previous work [12], the divergence used was the M-measure [11], a time-averaged KL-divergence, but we instead use $d = 1 - \cos(x, y)$, where x and y are feature vectors representing individual frames. The measure is the difference between two regression coefficients

$$M_\Delta = \mu_{\text{across}} - \mu_{\text{within}}$$

which are the coefficients of the overdetermined system

$$d = (1 - p) \cdot \mu_{\text{across}} + p \cdot \mu_{\text{within}}$$

where p is the phone label repetition probability. For each frame, we can compute M_Δ with either positive or negative lag, since the curve will necessarily be the same in both directions. We will apply it to intervals; we take the M-delta measure of an interval to be the median value of all calculated frame M_Δ values within that interval, both negative and positive lags. We exclude values where the frame is within 500ms of the edge of the interval (the right edge for positive-lag M_Δ, the left edge for negative-lag). This is acceptable, because we never apply the filter to intervals shorter than 500ms.

To filter the intervals returned by STD on the grounds that they do not adequately preserve information about phoneme duration, we set a cutoff for the M-delta measure of an interval, and discard intervals that fall below that cutoff.

2.5. Speaker ID and embedding

The speaker identity for each file was provided as part of the datasets. This information is used to balance the same and different pairs in the Word Tutor. We also explore the possibility of using a speaker embedding as side information, fed into the ABNET. To derive the speaker embedding, we used the same ABNET architecture and tutor, with the same cost function, but computed over the categories of same and different speakers, instead of same or different words in order to optimize speaker discrimination. This architecture and the results are described in [7].

3. Experiments

3.1. Datasets

We evaluate our system on two datasets prepared in the context of the ZeroSpeech 2015 Challenge [23]. The first dataset is a subset of the Buckeye Corpus of conversational American English[24]. The subset consists of around 5 hours of speech of 12 speakers. The second dataset was selected from the NCHLT Speech Corpus of South African Languages [25]. The selected subset consists of read speech in Xitsonga from 24 speakers for a total of roughly 2.5 hours. Voice activity detection information was provided, as well as speaker identity. No annotation information, whether orthographic or phonemic, was made available and our evaluation relies solely on the tools made available by the challenge.

Table 1: Output of the spoken term discovery system. These fragments ("words") serve as input to the ABNET. E(1,3) refers to Experiment 1 and 3, described in sections 3.3 and 3.5, respectively. E(2) refers to Experiment 2, described in section 3.4.

	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

We extract speech features from these data in two ways, for different parts of our system. First, for the spoken term discovery system, 13-dimensional perceptual linear prediction (PLP) features were extracted along with first and second derivatives. Second, the input to the ABNET consists of stacks of 40 Mel-scaled filters with logarithmic amplitude compression, computed over 25ms windows with a 10ms window shift. Both sets of features were mean-variance normalized file by file, skipping the areas indicated as non-speech (by the voice activity information provided with the dataset) in the computation of the mean and variance.

3.2. Evaluation

We use the Track 1 (subword unit discovery) evaluation toolkit of the ZeroSpeech challenge to evaluate our system. This evaluation consists of an aggregate minimal pair ABX discrimination score run on all of the triplets of phones of the given dataset. The basic idea behind it is to measure the discriminability between two sound categories A and B by measuring the error rate in deciding whether a speech token X is closer to a token from category A or category B in terms of their DTW distances. The aggregate error rate indicates the separability in the embedding space of the phoneme classes, both within-speaker and across-speaker. For more details on the rationale behind this metric, see [23].

3.3. Experiment 1: Spoken Term Discovery to ABNET

The spoken term discovery described in section 2.1 was applied to both the English and Xitsonga datasets. We describe the results, which will serve as input to the ABNET, in Table 1. The normalized edit distance (NED) was computed on the phoneme sequences corresponding to the discovered fragments. The coverage indicates the proportion of the dataset that is covered by the discovered fragments. The table indicates that we get comparable output for both languages. Note that we did not optimize over these scores and only supply this table to illustrate that the input to the ABNET was obtained in an unsupervised manner and contains a substantial amount of noise.

As described above in section 2.2, the output of the spoken term discovery system is piped into the ABNET. Filterbank stacks corresponding to these discovered pairs of fragments were passed as "same" input to the ABNET, while pairs of fragments coming from different classes, and therefore not corresponding to discovered pairs, served as "different" input.

3.4. Experiment 2: M-delta based filtering (MDF)

We compute an M-delta measure for each STD interval, as described in section 2.4. Then, for each corpus, we discard all the intervals falling in the lower quartile for the M-delta measure, and any intervals left in singleton classes. The result of filtering

is summarized in Table 1. We then pass the result as the set of discovered words to ABNET as before.

3.5. Experiment 3: Adding a speaker embedding

In this experiment we add information about speaker identity to the model. As described in section 2.5, an ABNET is trained on speaker discrimination. It outputs a speaker embedding in 100 dimensions. We average those embeddings for each speaker and do an Independent Component Analysis (ICA) to reduce the number of dimensions to 20. We then train a new ABNET on word discrimination as described in section 2.3, which uses as input the concatenation of the features of a fragment and the speaker embedding associated with that fragment, resulting in an input vector of dimension $7 \times 40 + 20 = 300$. In this way, we provide the system with speaker identification, which could help with cross speaker word discrimination.

4. Results

The Track 1 ABX evaluation is shown in Table 2. The results show a marked improvement over the MFCC baseline. In fact, on the English dataset, the STD $\rightarrow$ ABNET system comes close to the supervised HMM-GMM topline, and beats it in the within-talker condition. On the Xitsonga dataset, the results are midway between the baseline and the topline. This is probably due to the fact that the Xitsonga STD system outputs half as many pairs as compared to the English STD system. This may be due to the particular nature of the Xitsonga corpus, which is composed of read speech made for the purpose of training speech technology systems. In contrast, the English dataset is composed of conversations and presumably contains more repeated materials. As expected, the across-speaker comparisons are more difficult than within-speaker, but the difference between these two conditions is reduced both in relative and absolute terms, compared to the baseline. This indicates that the STD-based tutoring succeeded in helping to construct a more speaker-invariant representation, despite the fact that the word-like pairs that were extracted were themselves mostly within speaker (94% and 95% respectively).

The results on the optional M-delta filtering and speaker side information do not show a marked improvement over the base STD $\rightarrow$ ABnet system. On the contrary, the results were somewhat less good. This rather unsatisfying result has to be moderated by two considerations. Regarding the M-Delta filter, its net effect is to reduce the number of available word pairs that can be used to tutor the ABNET. For the English dataset the result of this filtering turned out to be of worse quality than without, see Table 1. The smaller number of pairs constitutes an unfavorable situation for the strategy outlined in this study, but this may change with a larger dataset. The talker side information necessitates a more complex network. This may hamper the otherwise beneficial effect of talker side information that has been noted in other studies [9, 10].

5. Discussion

This paper described a hybrid approach to unsupervised optimization of feature representations for the ABX discrimination task. First, pairs of intervals of high acoustic similarity were extracted using a spoken term discovery system. Next, an ABNET was trained to construct a speech feature embedding that reflects this information. We showed that, with this architecture, we can construct a feature representation that does extremely well on the minimal-pair ABX task. For both the English and the Xitsonga datasets, our representation improves on the baseline by a significant margin. In the case of the English dataset our system even beats the topline, which was constructed using a supervised system. The addition of information about the speaker or phoneme duration was shown not to improve the results of the base system.

Table 2: 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 $\rightarrow$ ABNET	**12.0**	**17.9**	**11.7**	**16.6**
STD / MDF $\rightarrow$ ABNET	12.4	18.1	12.6	18.6
STD + SpkID $\rightarrow$ ABNET	12.2	18.0	16.5	21.3

This shows that "lexical" information, even when it has been extracted automatically, and is hence of very low quality and quantity compared to the gold lexicon, can indeed help in establishing robust phonetic representations. Further work needs to be done to establish whether the resulting embeddings can in turn help word discovery. The results we found are encouraging in this respect, since our embeddings improve the most on across-speaker discrimination, compared to the MFCC baseline. However, for a synergistic situation to emerge it would be necessary to construct a pair of systems—a spoken term detection system and a phonetic learner—finely tuned to work together.

6. Acknowledgements

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. The authors thank Aren Jansen for his help with the spoken term discovery system used in this paper.

7. References

[1] P. W. Jusczyk, *The discovery of spoken language*. Cambridge, Mass.: MIT Press, 1997.

[2] P. K. Kuhl, "A new view of language acquisition," *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 850–11 857, 2000.

[3] N. Feldman, T. Griffiths, S. Goldwater, and J. Morgan, "A role for the developing lexicon in phonetic category acquisition," *Psychological review*, vol. 120, no. 4, pp. 751–778, 2013.

[4] A. Martin, S. Peperkamp, and E. Dupoux, "Learning phonemens with a proto-lexicon," *Cognitive Science*, in press.

[5] A. Fourtassi, T. Schatz, B. Varadarajan, and E. Dupoux, "Exploring the relative role of bottom-up and top-down information in phoneme learning," in *Proceedings of the 52nd Annual Meeting of the ACL*, vol. 2. Association for Computational Linguistics, 2014, pp. 1–6.

[6] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *IEEE SLT*, 2014.

[7] G. Synnaeve and E. Dupoux, "Weakly supervised multi-embeddings learning of acoustic models," *CoRR*, vol.

abs/1412.6645, 2014. [Online]. Available: http://arxiv.org/abs/1412.6645

[8] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proceedings of ICASSP*, 2015.

[9] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[10] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014.

[11] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: predicting asr error from temporal properties of speech signal," in *Proceedings ICASSP*, 2013, pp. 7423–7426.

[12] T. Ogawa, S. Mallidi, E. Dupoux, J. Cohen, N. Feldman, and H. Hermansky, "M-delta measure for accuracy prediction and its application to multistream-based unsupervised adaptation," in *Proceedings of ICASSP*, 2015.

[13] A. Park and J. Glass, "Unsupervised pattern discovery in speech," in *Transactions of ASLP*, vol. 16, no. 1, 2008, pp. 186–197.

[14] A. Jansen and B. van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proceedings of ASRU*, 2011.

[15] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Internat. Journ. of Pattern Recog. and Artific. Intell.*, vol. 7, no. 04, pp. 669–688, 1993.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 1735–1742.

[17] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *IJCAI*, vol. 11, 2011, pp. 2764–2770.

[18] M. D. Zeiler, "Adadelta: An adaptive learning rate method," *arXiv preprint:1212.5701*, 2012.

[19] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010, oral Presentation.

[20] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

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

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, 1993, lDC93S1.

[23] 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.

[24] M. 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.

[25] E. Barnard, M. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The nchlt speech corpus of the south african languages," in *SLTU 2014*, 2014, pp. 194–200.