Joint Learning of Speaker and Phonetic Similarities with Siamese Networks

Neil Zeghidour\textsuperscript{1,2}, Gabriel Synnaeve\textsuperscript{1}, Nicolas Usunier\textsuperscript{1}, Emmanuel Dupoux\textsuperscript{2,3}

\textsuperscript{1}Facebook AI Research, Paris
\textsuperscript{2}Ecole Normale Supérieure, Paris
\textsuperscript{3}Ecole des Hautes Etudes en Sciences Sociales, Paris
{neilz,gab,usunier}@fb.com, emmanuel.dupoux@gmail.com

Abstract

Recent work has demonstrated, on small datasets, the feasibility of jointly learning specialized speaker and phone embeddings, in a weakly supervised siamese DNN architecture using word and speaker identity as side information. Here, we scale up these architectures to the 360 hours of the Librispeech corpus by implementing a sampling method to efficiently select pairs of words from the dataset and improving the loss function. We also compare the standard siamese networks fed with same (AA) or different (AB) pairs, to a 'triamese' network fed with AAB triplets. We use ABX discrimination tasks to evaluate the discriminability and invariance properties of the obtained joined embeddings, and compare these results with mono-embeddings architectures. We find that the joined embeddings architectures succeed in effectively disentangling speaker from phoneme information, with around 10% errorsLibri121matchingtasks and embeddings (speaker ask on speaker embeddings, and phone (ask)-255(on)-256(phone)-255(embedding))-255(and)-255(near)-256(chance)-255 mismatched
of frames, \( y_{\text{phn}} \in \{0, 1\} \) is 1 if \( x \) and \( x' \) are phonetically similar and \( y_{\text{spk}} \in \{0, 1\} \) is 1 if \( x \) and \( x' \) are said by the same speaker. Given \( x \), the network outputs a phonetic embedding \( e_{\text{phn}}(x) \in \mathbb{R}^d \) and a speaker embedding \( e_{\text{spk}}(x) \in \mathbb{R}^d \), the same architecture and parameters are used for \( e_{\text{phn}}(x) \) and \( e_{\text{spk}}(x) \), except for the last layer.

Siamese networks are trained using a loss function defined on pairs which, given any two embeddings \( e, e' \in \mathbb{R}^d \) and a label \( y \in \{0, 1\} \), enforces that \( e \) should be close to \( e' \) if \( y = 1 \), while the two embeddings should be far away if \( y = 0 \). The similarity between embeddings is measured by their cosine:

\[
\cos(e, e') = \frac{e \cdot e'}{|e||e'|}. 
\]

The pairwise loss function we propose is

\[
\ell_\gamma(e, e', y) = \begin{cases} 
- \cos(e, e') & \text{if } y = 1 \\
\max(0, \cos(e, e') - \gamma) & \text{if } y = 0 
\end{cases},
\]

where \( \gamma \) is a margin hyperparameter. The loss of the multi-output network is then

\[
L(x, x', y_{\text{phn}}, y_{\text{spk}}) = \ell(e_{\text{phn}}(x), e_{\text{phn}}(x'), y_{\text{phn}}) + \ell(e_{\text{spk}}(x), e_{\text{spk}}(x'), y_{\text{spk}}).
\]

We also experimented with single output networks, which learn only either \( e_{\text{phn}} \) or \( e_{\text{spk}} \).

2.3. Triamese network

The triamese network uses a triplet-based loss function [8, 9, 6]. The model has the same architecture as before, but now the data takes the form \( (x_1, x_2, x_3) \) where \( x_1, x_2 \) are input stacks with similar phonetic content from two different speakers, and \( x_3 \) is a stack from two different words said by the same speaker.

A triplet loss enforces constraints on relative similarities between pairs. For phonetic embeddings \( e_{\text{phn}} \), the units from the same word but different speakers \((x_1, x_2')\) should be more similar than the units from different words but the same speaker \((x_1', x_2)\). The rule is inverted for speaker embeddings. Formally, the triplet loss is defined for any three embeddings \( e, e', e'' \) as

\[
\tilde{\ell}_\gamma(e, e', e'') = \max\left(0, \gamma - \cos(e, e') + \cos(e, e'')\right).
\]

In the final model, we may have different margin parameters \( \gamma_{\text{phn}} \) and \( \gamma_{\text{spk}} \) for phonetic and speaker embeddings respectively. The losses for each embeddings are then

\[
\tilde{\ell}_{\gamma_{\text{phn}}}(x_1', x_2, x_3) = \tilde{\ell}_{\gamma_{\text{phn}}}(e_{\text{phn}}(x_1'), e_{\text{phn}}(x_2), e_{\text{phn}}(x_3)) \quad \text{and} \quad \tilde{\ell}_{\gamma_{\text{spk}}}(x_1, x_2', x_3) = \tilde{\ell}_{\gamma_{\text{spk}}}(e_{\text{spk}}(x_1), e_{\text{spk}}(x_2'), e_{\text{spk}}(x_3)).
\]

For the multi-output network, the final loss is \( \tilde{\ell}_{\gamma_{\text{phn}}} + \tilde{\ell}_{\gamma_{\text{spk}}} \).

A multi-output triamese is shown in Fig. 1.

3. Experiments

3.1. Experimental Setup

The neural networks are trained on the 360 hours of read speech (920 speakers) constituting the train\_clean\_360 subset of the Librispeech dataset [10]. We obtained the speech fragments for each word of the dataset by force-aligning a state-of-the-art HMM-DNN [10] with transcription at the phone level, and then segmenting the speech at word boundaries.

After preliminary experiments, we focused on a deep neural net architecture with four hidden layers with 1000 units and a final embedding layer of size \( d = 100 \). A ReLU nonlinearity [11] is applied at each layer (ReLUs exhibited similar performances). We used Adadelta [12] with interpolation parameter 0.9 and epsilon \( 10^{-6} \) to train the siamese architecture, whereas plain stochastic gradient descent (SGD) seemed to perform slightly better for the triamese model. The learning rate for SGD starts at 0.01 and is halved when the error on the development set stops to decrease (with a minimum of \( 10^{-6} \)). The margin parameters \( (\gamma, \gamma_{\text{phn}}, \gamma_{\text{spk}}) \), the weight decay, and the number of frames in an input stack were respectively chosen among \( \{0.15, 0.5, 0.85\}, \{0, 0.001\} \) and \( \{7, 15\} \). The dev-clean split of the dataset is used for early stopping and hyperparameter selection.

3.2. Evaluation metrics and datasets

We evaluate the selectivity and invariance properties of the embeddings learned by the system with ABX discrimination tasks [13, 14]. An ABX task is performed on three utterances \( A, B, \) and \( X \), with \( A \) and \( B \) belonging to different classes and \( X \) matching the category of either \( A \) or \( B \). Let us assume for the sake of example that \( X \) matches the class of \( A \). If \( D(A, X) > D(B, X) \), with \( D \) some distance function, then the error is 1 (failure), else it is 0 (success). By averaging the error over all relevant \( A, B \) and \( X \) that can be found in the data, we can evaluate the discriminability of the class on which \( A \) and \( B \) differ, from 0% to 50% (chance level) in the representation space where the tasks are performed.

In our experiments, \( A, B \) and \( X \) are triphones that may only differ by their central phoneme. When evaluating phonetic discriminability, \( A \) and \( B \) share the same speaker while their central phoneme is different, and \( X \) matches \( A \) on its phonetic content but is pronounced by a different speaker. Hence, this phonetic discriminability task is performed across speakers.
which makes it a harder task than if A, B and X shared the same speaker. Switching B and X provides a speaker discriminability task across phonemes. Table 1 shows examples for both tasks.

Precisely, each triphone is represented as a stack of frames in the embedding space (each embedding is considered to be time-aligned with the central frame of the input stack), and the distance between triphones is computed as the sum of the cosine distances between aligned frames after DTW. An ABX task is then performed per triplet and we show the average error over all triplets that can be found in the data.

### 3.2.1. In-domain evaluation

Evaluations on the Librispeech dataset are computed on the test-clean subset. We use annotations at the phoneme level from the forced alignment to extract all relevant triplets from the test set. We then subsample randomly 10% of the triplets to get 600k ABX triplets for the evaluation, from 40 speakers.

### 3.2.2. Out-of-domain evaluation

In order to evaluate the robustness of the learned representation across datasets and languages, we also performed two sets of out-of-domain experiments. First, we evaluated our embeddings on the training set of the TIMIT dataset [15], a corpus of clean read speech containing 10 sentences read by 630 speakers of 8 major dialects of American English. We extracted all triplets from the train set of the standard train/dev/test split. We then subsample randomly 10% of these triplets, and obtain 1.87m ABX triplets total, with 462 speakers.

We also evaluated out-of-domain performance across languages by evaluating our embeddings on the Xitsonga dialect, subset of the NCHLT corpus. This corpus was used in the zero-shot 2015 challenge [16], for unsupervised discovery of phone embeddings, and we will compare our method to the best in-domain unsupervised system. The corpus used for evaluation contains 240k ABX triplets, for 24 speakers.

### 3.3. Results

We present the ABX error rate of phone and/or speaker embeddings, each one on both phone across speaker and speaker across phone task. For the phone embedding, lower ABX error rates are better on the phone across speaker task (high selectivity), but a score close to 50% is better for the speaker across phone task because it means high invariance. Conversely, better speaker embeddings have lower speaker across phone error rate. For each size of input stack (7 or 15), the chosen hyperparameters for the siamese networks are $\gamma = 0.5$ and a weight decay of 0 and for the triamese we use $\gamma_{phn} = 0.85$, $\gamma_{spk} = 0.5$ and a weight decay of 0.001.

#### 3.3.1. In-domain results

The results on the test set of Librispeech are presented in Table 2. As a baseline, we also present the results of stacks of 7 MFSC frames (input features), which was shown to give good results on TIMIT [3]. The main results are clear. For all networks, the phone and speaker tasks show high selectivity on their matched embeddings with an error rate around 10% (best score, respectively of 9.7% and 8.7%). At the same time, the scores on the mismatched embeddings (phonetic embedding for a speaker task and speaker embedding for a phonetic task) are within 5% of the chance level. This means that the embeddings have learned not only to be selective for the relevant dimension, but also to ignore the irrelevant one. This contrasts with the MFSC7 input representations that encode both dimensions. Moreover, even though comparisons are limited because the datasets are different, we achieve here a level of disentanglement that was not obtained in [5], in which phonetic embeddings had phonetic discriminability close to the raw MFSC (30.4% and 34.1% error respectively), and were less speaker-invariant than MFSC (30.8% and 38% error respectively).

In addition, we can see that the double embedding architectures do roughly as well as the single ones, even though the former have to share most of the network’s weights for the two competing tasks. The speaker embedding (tested on the speaker task) seems to consistently benefit from the double training regime compared to a network trained only on a single task, these gains ranging from 0.6% to 14.5% (absolute). The phone tasks, in contrast, are less consistently affected, some architectures showing a small gain and most others a small cost.

#### 3.3.2. Out-of-domain results

The results on TIMIT and Xitsonga are shown in Table 3. For reference, we show the same MFSC baselines, a supervised phone classifier (DNN from [3]) on TIMIT, and the previous best weakly-supervised trained (in-domain) siamese neural networks on these datasets (using scattering features [17]). These results are remarkable for several reasons: First, models trained on Librispeech generalize properly to TIMIT (no dataset overfit), both for tasks on phones and on speakers, i.e. they keep very good level of selectivity and invariance compared to the training dataset, even though they are tested on 462 speakers (40 for the in-domain evaluation). Second, models trained on English (Librispeech) generalize to Xitsonga, a language typologically unrelated to English, containing a large array of consonants (54) including some click consonants and a contrast between breathy

<table>
<thead>
<tr>
<th>Discriminability</th>
<th>A</th>
<th>B</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonetic</td>
<td>/beg/</td>
<td>/sp1</td>
<td>/beg/</td>
</tr>
<tr>
<td>Speaker</td>
<td>/beg/</td>
<td>/sp1</td>
<td>/beg/</td>
</tr>
</tbody>
</table>

Table 1: Examples of A, B and X for both phonetic and speaker discriminability tasks. ”sp_i” stands for speaker number i.

<table>
<thead>
<tr>
<th>model</th>
<th>task</th>
<th>phone embed.</th>
<th>speaker embed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFSC7</td>
<td>-</td>
<td>24.5</td>
<td>23.9</td>
</tr>
<tr>
<td>Sia7</td>
<td>single</td>
<td>10.9</td>
<td>24.5</td>
</tr>
<tr>
<td>Sia15</td>
<td>double</td>
<td>10.5</td>
<td>24.5</td>
</tr>
<tr>
<td>Tri7</td>
<td>single</td>
<td>9.7</td>
<td>10.5</td>
</tr>
<tr>
<td>Tri15</td>
<td>double</td>
<td>11.5</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Table 2: ABX error rates on Librispeech. The evaluation tasks are either ABX on phones across speakers (phn) or ABX on speakers across phones (spk). MFSC7 is a no training baseline where 7 stacked filterbanks are used as both phone and speaker embeddings."Sia" is for siamese, "Tri" for triamese networks, followed by the number of frames in an input stack. “single” means that the phonetic and speaker embeddings were trained separately in single output networks, whereas “double” refers to a multi-training multi-output network.
and modal voiced consonants which is totally absent in English. Third, these out-of-domain models happen to be the previous in-domain state of the art (trained with the same general architecture). On TIMIT the single output triamese network trained on pairs of words has a phone across speaker ABX (9.2%) which is equivalent to the in-domain supervised phone classifier DNN.

4. Conclusion

We have demonstrated that a siamese or triamese architecture, together with a weak supervision using only same-different information regarding word and speaker identity can learn embeddings that are very selective in one dimension and invariant in the other: indeed, our best embeddings showed around a 10% error rate in one task and near chance in the other. Moreover, in the other: indeed, our best embeddings showed around a 10% error rate in one task and near chance in the other. Furthermore, our best embeddings showed around a 10% error rate in one task and near chance in the other. Moreover, we showed that it is possible to learn these two orthogonal embeddings within the same network (ie, a network that carried together with a weak supervision using only same-different information). On TIMIT, the DNN “topline” is the output of a supervised neural network trained as a phone classifier on the TIMIT train set [3].

5. Acknowledgments

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 pensée).

6. References


<table>
<thead>
<tr>
<th>model</th>
<th>phone embed.</th>
<th>speaker embed.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>phn</td>
<td>spk</td>
</tr>
<tr>
<td>Language: English (TIMIT) [15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFSC7</td>
<td>20.5</td>
<td>39.7</td>
</tr>
<tr>
<td>DNN supervised[3]</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>Best ScatABN[17]</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>Tri15 (double)</td>
<td>10.3</td>
<td>47.9</td>
</tr>
<tr>
<td>Tri15 (single)</td>
<td>9.2</td>
<td>48.7</td>
</tr>
<tr>
<td>Language: Xitsonga (NCHLT) [18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFSC7</td>
<td>30.1</td>
<td>25.8</td>
</tr>
<tr>
<td>Best ScatABN[17]</td>
<td>15.8</td>
<td></td>
</tr>
<tr>
<td>Tri15 (double)</td>
<td>15.4</td>
<td>41.6</td>
</tr>
<tr>
<td>Tri15 (single)</td>
<td>15.5</td>
<td>42.6</td>
</tr>
</tbody>
</table>

Table 3: Out-of-domain ABX results on a different dataset in English (TIMIT) and on a different language, Xitsonga. The results for Tri15 are obtained from extracting output embeddings from a (single or double) Triamese neural network previously trained on Librispeech. MFSC7 is an untrained stacked filter-bank baseline, and ScatABN is the state-of-the-art model for a weakly supervised siamese architecture trained on the TIMIT or an unsupervised architecture trained on the Xitsonga dataset (respectively) using scattering coefficients as input features [17]. For TIMIT, the DNN “topline” is the output of a supervised neural network trained as a phone classifier on the TIMIT train set [3].

