# Joint Learning of Speaker and Phonetic Similarities with Siamese Networks

*Neil Zeghidour*[1,2], *Gabriel Synnaeve*[1], *Nicolas Usunier*[1], *Emmanuel Dupoux*[2,3]

[1]Facebook AI Research, Paris
[2]Ecole Normale Supérieure, Paris
[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% errors for the matching tasks and embeddings (speaker task on speaker embeddings, and phone task on phone embedding) and near chance for the mismatched task. Furthermore, the results carry over in out-of-domain datasets, even beating the best results obtained with similar weakly supervised techniques.

## 1. Introduction

A current problem in machine learning is to learn representations that are invariant with respect to a set of transformations [1, 2]. This is especially relevant in speech, where the acoustic signal carries simultaneously linguistic, speaker-specific and channel information. It would therefore be of great interest to be able to untangle these representations, i.e., to derive representations that are selective to one dimension and invariant to the others.

Typically, selectivity is achieved through a supervised classification task. For instance, to learn speaker and channel invariant linguistic representations, one trains a phone classifier on a corpus that contains enough variation in all of these dimensions. Vice versa, to learn speaker representations, one trains a speaker classifier. Previous work has explored the possibility to use weak supervision to achieve invariant representation: [3] used a siamese architecture [4] where two otherwise identical copies of the same deep neural network (DNN) were presented with pairs of words that were either phonetically the same or different. The only information provided to the network was whether the words were the same or not, and a contrastive loss function on the output layer tried to maximize the similarity of the two representations when the words were the same and maximize the dissimilarity when they were different. The results showed that this technique reached the same level of invariance than supervised techniques.

Further work explored the possibility of disentangling phoneme and speaker information within the same network [5]. Here, two output layers were defined, one dedicated to phone representations (phone embedding) while the other was dedicated to speaker representations (speaker embedding). The input pairs of words were either phonetically the same or different, and that were spoken either by the same speaker or different speakers. A contrastive loss function on each of the embeddings tried to emphasize one dimension while ignoring the other, and vice versa. The results ran on a small corpus were encouraging, as the dual training network performed comparably to two single training networks separately, thereby saving computation time for identical results. However, the representations were far from achieving complete phoneme / speaker disentanglement in either regimes.

Here, we expand on this previous work in three ways. First, we scale up the architecture to deal with a considerably larger dataset. Secondly, we improve on the loss function by adding a margin and investigating a triplet-based loss function as in [6]. Finally we present out-of-domain experiments that show the selectivity and invariance properties on different languages.

## 2. Model

### 2.1. Weak supervision for sub-word units

We represent the speech signal using log compressed mel filterbanks frames (Mel Filterbanks Spectral Coefficients, MFSC) with a window size of 25ms and a shift of 10ms. The networks learn phonetic and/or speaker embeddings of sub-words units, provided an input defined as a stack of 7 or 15 of successive filterbank frames. Instead of forcing the network to encode the input into specific sub-word units (phonemes, diphones, triphones, phonetic features), we use the weakly supervised technique of [3] which only specifies whether the inputs are the same or different in terms of phonemes and speakers, and let the network figure out by itself what are the most appropriate sub-word units, provided they show the right level of invariance.

For training, we use annotation at the word level, and in terms of speaker identifiers. For pairs of identical words (same or different speakers), we first realign them at the frame level using Dynamic Time Warping (DTW) [7]. Sliding windows of stacked frames are then presented to the two entries of the siamese network. Dissimilar pairs are simply aligned along the diagonal.

### 2.2. Siamese network

The multi-output siamese architecture is trained using labeled pairs $(x, x', y^{phn}, y^{spk})$ where $x$ and $x'$ are two input stacks of frames, $y^{phn} \in \{0,1\}$ is 1 if $x$ and $x'$ are phonetically similar and $y^{spk} \in \{0,1\}$ is 1 if $x$ and $x'$ are said by the same speaker. Given $x$, the network outputs a phonetic embedding $\mathbf{e}^{phn}(x) \in \mathbb{R}^d$ and a speaker embedding $\mathbf{e}^{spk}(x) \in \mathbb{R}^d$; the same architecture and parameters are used for $\mathbf{e}^{phn}(x)$ and $\mathbf{e}^{spk}(x)$, except for the last layer.

Siamese networks are trained using a loss function defined on pairs which, given any two embeddings $\mathbf{e}$, $\mathbf{e}'$ in $\mathbb{R}^d$ and a label $y \in \{0,1\}$, enforces that $\mathbf{e}$ should be close to $\mathbf{e}'$ if $y = 1$, while the two embeddings should be far away if $y = 0$. The similarity between embeddings is measured by their cosine $\cos(\mathbf{e}, \mathbf{e}') = \frac{\mathbf{e}.\mathbf{e}'}{||\mathbf{e}||_2||\mathbf{e}'||_2}$. The pairwise loss function we propose is

$$\ell\gamma(\mathbf{e}, \mathbf{e}', y) = \begin{cases} -\cos\left(\mathbf{e}, \mathbf{e}'\right) & \text{if } y = 1 \\ \max(0, \cos(\mathbf{e}, \mathbf{e}') - \gamma) & \text{if } y = 0 \end{cases},$$

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

$$\begin{aligned} L(x, x', y^{phn}, y^{spk}) = & \quad \ell(\mathbf{e}^{phn}(x), \mathbf{e}^{phn}(x'), y^{phn}) \\ & + \ \ell(\mathbf{e}^{spk}(x), \mathbf{e}^{spk}(x'), y^{spk}) . \end{aligned}$$

We also experimented with single output networks, which learn only either $\mathbf{e}^{phn}$ or $\mathbf{e}^{spk}$.

### 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 $(x_1^1, x_1^2, x_2^1)$ where $(x_1^1, x_1^2)$ are input stacks with similar phonetic content from two different speakers, and $(x_1^1, x_2^1)$ are stacks from two different words said by the same speaker.

A triplet loss enforces constraints on relative similarities between pairs. For phonetic embeddings $\mathbf{e}^{phn}$, the units from the same word but different speakers $(x_1^1, x_1^2)$ should be more similar than the units from different words but the same speaker $(x_1^1, x_2^1)$. The rule is inverted for speaker embeddings. Formally, the triplet loss is defined for any three embeddings $\mathbf{e}, \mathbf{e}', \mathbf{e}''$ as

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

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

$$\tilde{\ell}^{phn}(x_1^1, x_1^2, x_2^1) = \tilde{\ell}_{\gamma^{phn}}\left(\mathbf{e}^{phn}(x_1^1), \mathbf{e}^{phn}(x_1^2), \mathbf{e}^{phn}(x_2^1)\right) ,$$

$$\tilde{\ell}^{spk}(x_1^1, x_1^2, x_2^1) = \tilde{\ell}_{\gamma^{spk}}\left(\mathbf{e}^{spk}(x_1^1), \mathbf{e}^{spk}(x_2^1), \mathbf{e}^{spk}(x_1^2)\right) .$$

For the multi-output network, the final loss is $\tilde{\ell}^{phn} + \tilde{\ell}^{spk}$. 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

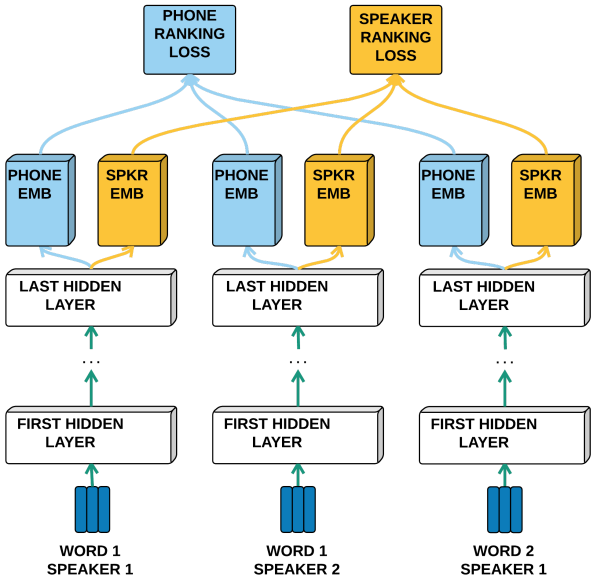


Figure 1: A multi-ouput triamese network. All parameters of each of the three branches at a given depth are shared.

a final embedding layer of size $d = 100$. A RReLU non-linearity [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 ($\gamma$, $\gamma^{phn}$ and $\gamma^{spk}$), 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,

| Discriminability | A | B | X |
|---|---|---|---|
| Phonetic | /beg/ $sp_1$ | /bag/ $sp_1$ | /beg/ $sp_2$ |
| Speaker | /beg/ $sp_1$ | /beg/ $sp_2$ | /bag/ $sp_1$ |

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

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 zerospeech 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).

| | | phone embed. | | speaker embed. | |
|---|---|---|---|---|---|
| model | task | phn | spk | phn | spk |
| MFSC7 | - | 24.5 | 32.9 | 24.5 | 32.9 |
| Sia7 | single | 10.9 | 46.0 | 46.4 | 23.9 |
| | double | 10.5 | 45.9 | 45.4 | 9.3 |
| Sia15 | single | 9.7 | 47.1 | 45.8 | 12.4 |
| | double | 10.2 | 46.6 | 45.3 | 8.7 |
| Tri7 | single | 10.0 | 46.0 | 45.0 | 10.0 |
| | double | 11.5 | 45.0 | 45.7 | 9.4 |
| Tri15 | single | 9.8 | 46.9 | 44.6 | 9.4 |
| | double | 10.7 | 46.2 | 44.7 | 8.1 |

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.

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

| model | phone embed. phn | spk | speaker embed. phn | spk |
|---|---|---|---|---|
| Language: English (TIMIT) [15] | | | | |
| MFSC7 | 20.5 | 39.7 | 20.5 | 39.7 |
| DNN supervised[3] | 9.2 | | | |
| Best ScatABnet[17] | 9.8 | | | |
| Tri15 (double) | 10.3 | 47.9 | 43.4 | 14.2 |
| Tri15 (single) | 9.2 | 48.7 | | |
| Language: Xitsonga (NCHLT) [18] | | | | |
| MFSC7 | 30.1 | 25.8 | 30.1 | 25.8 |
| Best ScatABnet[17] | 15.8 | | | |
| Tri15 (double) | 15.4 | 41.6 | 44.7 | 14.3 |
| Tri15 (single) | 15.5 | 42.6 | | |

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 filterbank baseline, and ScatABnet 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].

and modal voiced consonants which is totally absent in English. Third, these out-of-domain models happen to beat 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, we showed that it was possible to learn these two orthogonal embeddings within the *same* network (ie, a network that carried out the two tasks using the same connections, except the last layer), thereby demonstrating effective disentanglement of phoneme and speaker information. Finally, we showed that these disentangling networks could generalize their performance in out-of-domain datasets (a different English dataset, and a different, under-resourced, language), even beating the state of the art in these languages.

In detail, the double-output networks differed somewhat from the single-output ones. In particular, whereas the speaker task benefited consistently from the joint training, this was not the case for the phone task. This asymmetry may be related to the observed finding that speaker ID systems based on i-vectors improve their performance if they use phone embeddings as inputs as opposed to raw MFCC [19, 20]. Reciprocally, the benefit of speaker normalization in state of the art DNN-based ASR has been more elusive. Further work is necessary to understand this asymmetry and to further improve representation disentanglement.

# 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

[1] F. Anselmi, L. Rosasco, and T. Poggio, "On invariance and selectivity in representation learning," *arXiv preprint arXiv:1503.05938*, 2015.

[2] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1872–1886, 2013.

[3] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetics embedding learning with side information," in *IEEE Spoken Language Technology Workshop*. IEEE, 2014.

[4] 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," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[5] G. Synnaeve and E. Dupoux, "Weakly supervised multi-embeddings learning of acoustic models," *ICLR Workshop: arXiv:1412.6645*, 2014.

[6] H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," *arXiv preprint arXiv:1510.01032*, 2015.

[7] 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.

[8] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *The Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[9] L. Van Der Maaten and K. Weinberger, "Stochastic triplet embedding," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[11] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

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

[13] 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.

[14] 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.

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

[16] M. Versteegh, R. Thiolliere, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," in *Proc. of INTERSPEECH*, 2015.

[17] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum-deep siamese network pipeline for unsupervised acoustic modeling," in *ICASSP*, 2016.

[18] 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.

[19] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," *Proc. Odyssey-14, Joensuu, Finland*, 2014.

[20] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014, pp. 293–298.