# Dialect Adaptation of Deep Neural Network for Phoneme Recognition

DT2119 - Speech and Speaker Recognition

KTH - Royal Institute of Technology

Antoine Broyelle
broyelle@kth.se

Guillaume Coppens
coppens@kth.se

June 7, 2017

**Abstract**

This report summarize Adaptive Learning (AL) techniques for deep neural networks (DNN). The main goal is to generate a new model based on a trained model in order to specify it for a given scenarios. In this report, we will first explain the problem of adaptive learning. Then we will detail different architectures that we studied. We will then show with the experiments done on the TIMIT dataset that AL with DNNs does not significantly increase the accuracy of speech recognition. Finally we will discuss why it has not worked.

## 1 Introduction

One big problem in speech technology and speech recognition is to get samples which are relevant for the task we want to achieve. Generally it is not the case when an application is deployed, moreover the environments or the users may change over time.

In speech recognition, and more generally in machine learning, we are looking for good models with great generalization in order to get good performance and to deal with the difference between the database and real cases. The first naive idea to improve the model is to increase the amount of training data. However large dataset are generally owned by big companies and may not work so well [1].

Adaptive Learning is a great way to face this problem. The main idea behind AL is to introduce a transformation of the data or the model in order to make them match. As an example, in speech recognition, the preprocessing for feature adaptation could be vocal tract length normalization[2]. In this article we will focus on the transformation of the model, in our case neural networks. The DNN adaptation techniques can be classified in three categories[3] : linear transformation, conservative training, and subspace method. We decided in this project to focus on linear transformation. A general introduction to linear adaptation model for neural network in given in Section 2

For the experiment, we use the TIMIT dataset[4] and performed a dialect adaptation on the phoneme recognizer. A general model is trained on 4 dialects and 4 specified model are generated for 4 other dialects. Both the general model and the specified ones will be obtains under supervised learning. The general model is trained with a hybrid method combining DNN and Hidden Markov Model (HMM).

# 2 Method

## 2.1 Global Description

Deep Neural Networks have become the state-of-the-art in many fields and speech recognition is not an exception. The hybrid system DNN-HMM is known to outperform HMM [5]. With this idea in mind, We train an HMM per phoneme using 3 states and GMM emissions. We than use the Viterbi algorithm to provide the DNN training targets. The work of the DNN is to estimate the probability of an HMM state given an observation. Using a simple Bayes rule, we obtain the probability of an observation given a state in order to perform the HMM decoding. [5]

For training the GMM-HMM, MFCC with it $1^{st}$ and $2^{nd}$ order derivatives. The MFCC features allow to get a low number of uncorrelated features (in general 13). The same features have been used as a DNN input.

## 2.2 DNN Adaptation

With DNN, given an already trained network, a linear adaptation is simply made by adding a linear layer to the network. This linear layer can be added everywhere in the network architecture as long as it respect size constraints. During the specialization the weights of the original DNN are kept fixed.

Our approach is to apply a linear transformation to either the input features (Linear Input Networks) [6], or to the output vector (Linear Output Networks [7]). When a new hidden layer added, it also introduces a new set of parameters which can be apply before (denoted as b) or after (denoted as a) the new layer (before and after denomination are based on the forward pass of the network). We will only focus on the extreme cases, i.e. a feature linear regression of the input LIN(b) and an output linear regression LON(b).

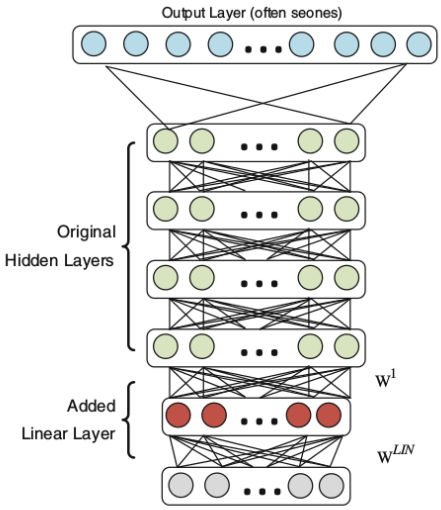


**Figure 1:** Linear Input Networks - LIN *illustration from [3]*

The Linear Input Networks Fig.1 allows to map the speaker dependent features to the average speaker features via a linear transformation $W^{LIN}, b^{LIN}$, where $W^{LIN} \in \mathbb{R}^{N_0 \times N_0}, b^{LIN} \in \mathbb{R}^{N_0 \times 1}$, with $N_0$ the number of input features.

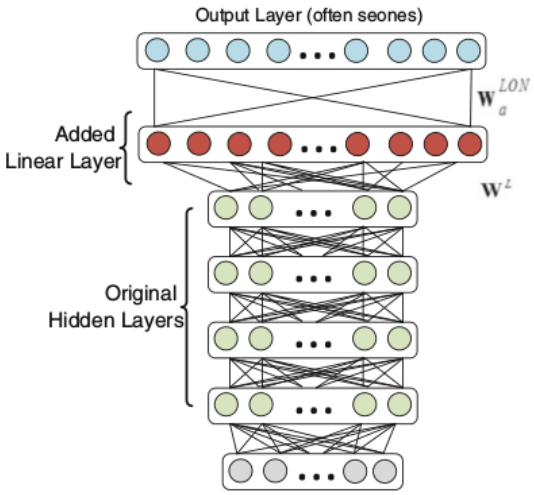


**Figure 2:** Linear Output Networks - LON *illustration from [3]*

The LON architecture works as a linear regression with the DNN as the feature extractor.

# 3 TIMIT's Dataset

## 3.1 Data Description

The experiences have been done on the TIMIT corpus. This data set contains 6300 sentences, spoken from 630 speakers from 8 major dialect regions of the United States. The dialect label is one of the major interest for this dataset for this project.

The dialect regions are :

- dr1: New England
- dr2: Northern
- dr3: North Midland
- dr4: South Midland
- dr5: Southern
- dr6: New York City
- dr7: Western
- dr8: Army Brat (moved around)

| Dialect Region(dr) | Male | Female | Total |
|---|---|---|---|
| 1 | 31 (63%) | 18 (27%) | 49 (8%) |
| 2 | 71 (70%) | 31 (30%) | 102 (16%) |
| 3 | 79 (67%) | 23 (23%) | 102 (16%) |
| 4 | 69 (69%) | 31 (31%) | 100 (16%) |
| 5 | 62 (63%) | 36 (37%) | 98 (16%) |
| 6 | 30 (65%) | 16 (35%) | 46 (7%) |
| 7 | 74 (74%) | 26 (26%) | 100 (16%) |
| 8 | 22 (67%) | 11 (33%) | 33 (5%) |
| 8 | 438 (70%) | 192 (30%) | 630 (100%) |

**Table 1:** Dialect distribution of speakers in the database

In the experiments we have done linear adaptation from a general model trained with dialects dr1 to dr4. We have adapted the general model for four different networks with respectively the dialect dr5 to dr8.

We have split each of theses training sets with the proportions 80%, 20% to create a training set and a validation set. Training and validation set follow the same gender distribution. We also paid attention to include a speaker only in one dataset.

## 3.2 Preprocessing

Stress markers were removed and only the pronunciation of nouns were kept in case verb or adjectives differ.

As it often done, we mapped the phonemes set (61 components) to a smaller one (only 39)[8]. As it said in the HTK [9] documentation for HLEd scripts :

> The aim of this mapping is to delete all glottal stops, replace all closures preceding a voiced stop by a generic voiced closure (vcl), all closures preceding an unvoiced stop by a generic unvoiced closure (cl) and the different types of silence to a single generic silence (sil)

| From | To |
|---|---|
| aa,ao | aa |
| ah, ax, ax-h | ah |
| er, axr | er |
| hh, hv | hh |
| ih, ix | ih |
| l, el | l |
| m, em | m |
| n, en, nx | n |
| ng, eng | ng |
| sh, zh | sh |
| uw, ux | uw |
| pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi | sil |
| q | (discarded) |

**Table 2:** Mapping of TIMIT's phonemes (from 61 to 39)

One have normalized each data set according to the mean and standard deviation along each feature of the global training set.

# 4 Experiments

The general model is trained with training set from dialects dr1 to dr4. We have generated four specified networks for dialects dr5 to dr8. We used an hybrid system which combining HMM and DNN [5]. They both use MFCC and its two first derivatives as their inputs. The HMM was trained with GMMs of 16 gaussians. Three states were used to represent each phoneme. The DNN has 4 hidden layers and each of them contain 1024 nodes. The output of the DNN is a probability distribution of HMM states given an observation.

# 5 Results

The performances of the global model are summarized in Table 3. The accuracy is given in terms of HMM state recognition and phoneme recognition. Table 4 show results with the adapted models.

| Test Set | dr{1-4} | dr5 | dr6 | dr7 | dr8 |
|---|---|---|---|---|---|
| state acc. | 58.56 | 55.35 | 58.67 | 58.53 | 58.58 |
| phon. acc. | 63.84 | 60.77 | 64.21 | 63.73 | 63.83 |

**Table 3:** accuracy of the global model

| Model | LIN | | | |
|---|---|---|---|---|
| Dialect | dr5 | dr6 | dr7 | dr8 |
| state acc | 55.21 | 57.84 | 58.81 | 57.88 |
| phon. acc | 60.71 | 63.42 | 64.09 | 63.37 |
| Model | LON | | | |
| Dialect | dr5 | dr6 | dr7 | dr8 |
| state acc | 51.42 | 51.99 | 54.97 | 51.18 |
| phon. acc | 57.96 | 58.56 | 61.08 | 57.95 |

**Table 4:** Accuracy for specified models

We do no get expected such results. Indeed LIN and LON has worse accuracy than the global model. That might be explained that the data set is not adapted for adaptive learning as dialect are quite close. However, we observe that the LIN model is in general better than the LON.

# 6 Discussion and Conclusions

In this context both LIN and LON do not show significant improvements for phoneme recognition. Indeed the global model is already good on unseen dialects. The dataset might not be adapted for adaptive learning. For further exploration on AL with DNN we want to try to fine tune all the network instead of freezing the weights of the general model. It appears to perform better according to [10].

Also, the added layer was just a regression. Adding a normal layer (fully connected + normalization + activation) might help.

# References


[1] Markus Forsberg. Why is speech recognition difficult? 2 2003.

[2] H. Gish E. Eide. A parametric approach to vocal tract length normalization.

[3] Dong Yu and Li Deng. Automatic speech recognition: A deep learning approach. chapter 11. Springer Publishing Company, Incorporated, 2014.

[4] Victor Zue, Stephanie Seneff, and James Glass. Speech database development at mit: Timit and beyond. *Speech Communication*, 9(4):351 – 356, 1990.

[5] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 312–317. IEEE, 2013.

[6] Jan Trmal, Jan Zelinka, and Luděk Müller. Adaptation of a feedforward artificial neural network using a linear transform. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue*, TSD'10, pages 423–430, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 366–369. IEEE, 2012.

[8] Hsiao-Wuen Hon Kai-Fu Lee. Speaker-independent phone recognition using hidden markov models. IEEE Transactions on accoustics, speech and signal processing(37):1641–1648, 1989.

[9] Editing label files @ONLINE, 5.

[10] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.