ADVERSARIAL LEARNING OF RAW SPEECH FEATURES FOR DOMAIN INVARIANT SPEECH RECOGNITION

Aditay Tripathi ⋆‡ *Aanchan Mohan* † *Saket Anand* ⋆ *Maneesh Singh* ‡

⋆ Indraprastha Institute of Information Technology, New Delhi, India.
† Synaptitude Brain Health, Vancouver, Canada.
‡ Verisk Analytics, Jersey City, USA.

ABSTRACT

Recent advances in neural network based acoustic modelling have shown significant improvements in automatic speech recognition (ASR) performance. In order for acoustic models to be able to handle large acoustic variability, large amounts of labeled data is necessary, which are often expensive to obtain. This paper explores the application of adversarial training to learn features from *raw speech* that are invariant to acoustic variability. This acoustic variability is referred to as a domain shift in this paper. The experimental study presented in this paper leverages the architecture of Domain Adversarial Neural Networks (DANNs) [1] which uses data from two different domains. The DANN is a Y-shaped network that consists of a multi-layer CNN feature extractor module that is common to a label (senone) classifier and a so-called domain classifier. The utility of DANNs is evaluated on multiple datasets with domain shifts caused due to differences in gender and speaker accents. Promising empirical results indicate the strength of adversarial training for unsupervised domain adaptation in ASR, thereby emphasizing the ability of DANNs to learn domain invariant features from raw speech.

***Index Terms*—** Unsupervised Domain Adaptation, Raw speech, ASR, Deep Learning, CNN

1. INTRODUCTION

Training neural network based acoustic models for Automatic Speech Recognition (ASR) gets challenging when limited amounts of supervised training data is available. It is expensive to obtain labeled speech data that contains sufficient variations of the different sources of acoustic variability such as speaker accent, speaker gender, speaking style, different types of background noise or the type of recording device [2]. To mitigate the effects of acoustic variability that is inherent in the speech signal, domain adaptation techniques are often used in acoustic modelling. This paper investigates the use of Domain Adversarial Neural Networks (DANNs) [1] for domain adaptation of the acoustic model from raw speech directly instead of relying on traditional log-mel features. Although MFCC or log-mel features are predominantly used in acoustic modeling, there has been significant recent interest to learn acoustic models using raw speech [3, 4, 5, 6]. For example, Sainath et.al. [7] showed that recognition performance on raw speech matches the performance on log-mel filters.

Several techniques have been proposed to mitigate the effects of acoustic variability in the speech data at test time. Feature space maximum likelihood linear regression (fMLLR) [8], Maximum Likelihood Linear Regression [8], MAP [9], Vocal Tract Length Normalization (VLTN) [10] are well-known speaker adaptation techniques used in generative acoustic models. i-Vectors [11], LHUC [12] and KL-divergence regularized DNN acoustic models[13] are the popular adaptation techniques used for discriminative acoustic models. All of these techniques require labeled data from the target domain to perform adaptation.

The success of adversarial training using DANNs for unsupervised domain adaptation in computer vision [1] suggests promising extensions to ASR applications as well. In a DANN, domain adaptation is achieved by incorporating the additional task of domain classification along with label classification. Both the domain-classifier and the label (senone) classifier share a common multi-layer CNN feature extraction module. The network is trained to *minimize* the cross-entropy loss of the label classifier and at the same time *maximize* that of the domain classifier. DANN has been used to learn domain-invariant feature representations, thus achieving unsupervised domain adaptation for acoustic models trained on log-mel features [14, 15, 16]. Variational Autoencoders (VAE) [17] have also been applied for unsupervised domain adaption. These techniques have been studied for domain adaptation for discriminative acoustic models trained on log filter-bank features. This paper presents an experimental study of unsupervised domain adaptation on discriminative acoustic models trained on raw speech by using DANNs. Unsupervised domain adaptation is used to reduce acoustic variability due to (1) speaker gender and (2) speaker accent. To study the impact of DANN unsupervised domain adaptation on acoustic variability arising due to variations in speaker gender, experimental results are presented on the TIMIT data set. Furthermore, British and American accented data from the Voxforge corpus is used to study the impact of DANN unsupervised domain adaptation on acoustic variability arising due to differing speaker accents. The experimental study in this paper shows that domain invariant features can be learned directly from raw speech with significant improvement over the baseline acoustic models trained without domain adaptation.

The remainder of the paper is organized as follows. Section 2 reviews related work done in the area of domain-invariant feature learning using domain adversarial neural networks. Section 3 describes the adversarial training for unsupervised domain adaptation for ASR trained on raw speech and Section 4 presents the experimental setup, description of input features and a description of the configuration of the acoustic model. Section 5 presents the experimental results obtained followed by a discussion in Section 6 and conclusion in Section 7.

2. RELATION TO PRIOR WORK

Domain adaptation using adversarial training was first introduced by Ganin et.al [1] for domain adaptation in computer vision. It has since been used for noise invariant feature learning in ASR using supervised labels and filter-bank features [14, 15]. Sun et.al. [16]

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

used adversarial training for unsupervised domain adaptation for robust speech recognition using filter bank features and used the WSJ and Librispeech speech corpora as data from different domains. All of these studies used filter-bank features to learn the domain invariant features for ASR. In this study, domain invariant features are learned directly from raw speech to mitigate acoustic variabilities due to speaker gender and accent which often adversely affect ASR performance [18, 19].

3. DOMAIN ADAPTATION USING RAW SPEECH FEATURES

This section explains unsupervised domain adaptation using adversarial training on raw speech features. Consider a classification problem with the input feature vector space X and $Y = \{0, 1, 2, ..., L-1\}$ as the set of labels in the output space. Let $S(x, y)$ and $T(x, y)$ be *unknown* joint distributions defined over $X \times Y$, referred to as the source and target distributions respectively. The unsupervised domain adaptation algorithm requires input as the *labeled source domain data*, sampled from $S(x, y)$ and *unlabeled target domain data*, sampled from the marginal distribution $T(x)$ i.e.

$$\{(x_i, y_i)\}_{i=0}^{n} \sim S(x, y);\ \{(x_i)\}_{i=n+1}^{n+n'=N} \sim T(x),$$

where $N = n+n'$ is the total number of input samples. As opposed to the class labels, which are assumed only for the source domain data, the *binary* domain labels ($d_i = \{0, 1\}$) are defined as

$$d_i = \begin{cases} 0 \text{ for } x_i \sim S(x, y) \\ 1 \text{ for } x_i \sim T(x). \end{cases} \tag{1}$$

and are assumed to be known for each sample. The neural network architecture is as shown in Fig. 1 and comprises of three modules: the feature extractor, label classifier and the domain classifier. The feature extractor is a multi-layer Convolutional Neural Network (CNN) that takes as input x_i, while its output is available to the label and domain classifiers, which predict the labels y_i and d_i respectively. At training time, the label classifier's loss is only computed over labeled samples from $S(x, y)$, whereas the domain classifier's loss is computed over both, labeled samples from $S(x, y)$ and unlabeled samples from $T(x)$. The feature extractor and the two classifier modules are described in detail in Sections 3.1 and 3.2.

3.1. Feature Extractor

The feature extractor, G_f, shown in Fig. 1 is a multi-layer CNN and takes the raw speech input vector x_i and generates a d-dimensional feature vector $f_i \in R^d$ i.e.,

$$f_i = G_f(x_i; \Theta_f), \tag{2}$$

where Θ_f are the parameters of the feature extractor i.e. weights and biases of the convolutional layers. The input vector x_i can be from the source distribution $S(x, y)$ or the target distribution $T(x)$. The 1-d convolution operation in the convolutional layer in the network is defined as below

$$f_i^{m,c,1} = \sigma\Big(\sum_{j=m}^{m+k-1} \theta_f^{j-m,c,1} \cdot x_i^j\Big), \tag{3}$$

Eq. (3) gives feature vector output at index m from the first layer convolution operation on input feature vector x_i, $\theta_f^{c,1}$ denotes the k-dimensional vector of weights and biases of the first convolutional layer and c^{th} convolutional filter. The function $\sigma(\cdot)$ is a non-linear activation function like the sigmoid or ReLU.

3.2. Label and Domain Classifiers

The feature vector f_i, which is extracted from G_f, is mapped to class label $y_i = G_y(f_i; \Theta_y)$ by the label classifier G_y and to domain label $d_i = G_d(f_i; \Theta_d)$ by a domain classifier G_d as shown in Fig. 1. Both the label classifier as well as domain classifier are multi-layer feed-forward neural networks with parameters collectively denoted as Θ_f and Θ_d respectively. The unsupervised domain adaptation is achieved by training the network to minimize the cross-entropy based label classification loss on the labeled source domain data and at the same time *maximize* the cross-entropy domain classification loss on the supervised source domain data and unsupervised target domain data. The classification losses is the cross-entropy costs. The total loss is given by

$$E(\Theta_f, \Theta_y, \Theta_d) = \sum_{i=1..N, d_i=0} L_y(G_y(G_f(x_i; \Theta_f); \Theta_y), y_i) - \lambda \sum_{i=1..N} L_d(G_d(G_f(x_i; \Theta_f); \Theta_d), d_i). \tag{4}$$

The parameter λ is a hyper-parameter that weighs the relative contribution of the two costs. To simplify the above equations, these are written in the compressed form as below

$$E(\Theta_f, \Theta_y, \Theta_d) = \sum_{i=1..N, d_i=0} L_y^i(\Theta_f, \Theta_y) - \lambda \sum_{i=1..N} L_d^i(\Theta_f, \Theta_d). \tag{5}$$

The label classifier tries to minimize the label classification loss $L_y^i(\Theta_f, \Theta_y)$ on the data from source distribution $S(x, y)$, therefore the parameters of both feature extractor (Θ_f) and label predictor(Θ_y) are optimized. This ensures that the features f_i are discriminative enough to perform good prediction on samples from the source domain. At the same time the extracted features should be invariant to the shift in domain. In order to obtain domain invariant features, the parameters of feature extractor Θ_f are optimized to maximize the domain classification loss $L_y(\Theta_f, \Theta_d)$ while at the same time domain classifier Θ_d tries to classify the input features. In other words, the domain classifier of the trained network should not be able to correctly predict the domain labels of the features coming from the feature extractor.

The desired parameters $\hat{\Theta_f}, \hat{\Theta_y}, \hat{\Theta_d}$ give the saddle point at the training and are estimated as:

$$(\hat{\Theta_f}, \hat{\Theta_y}) = arg \min_{\Theta_f, \Theta_y} E(\Theta_f, \Theta_y, \hat{\Theta_d}) \tag{6}$$

$$\hat{\Theta_d} = arg \max_{\Theta_d} E(\hat{\Theta_f}, \hat{\Theta_y}, \Theta_d). \tag{7}$$

The model can be optimized by the standard stochastic gradient descent (SGD) based approaches. The parameter updates during the SGD becomes

$$\Theta_f \leftarrow \Theta_f - \mu\Big\{\frac{\partial L_y^i}{\partial \Theta_f} - \lambda \frac{\partial L_d^i}{\partial \Theta_f}\Big\} \tag{8}$$

$$\Theta_y \leftarrow \Theta_y - \mu \frac{\partial L_Y^i}{\partial \Theta_y} \tag{9}$$

$$\Theta_d \leftarrow \Theta_d - \mu \frac{\partial L_d^i}{\partial \Theta_d}. \tag{10}$$

where, μ is the learning rate. Eq. 8-10 can be implemented in some form of SGD by using a special Gradient Reversal Layer (GRL) at the end of feature extractor and at the beginning of domain classifier as shown in fig. 1. During the backward propagation, GRL reverses

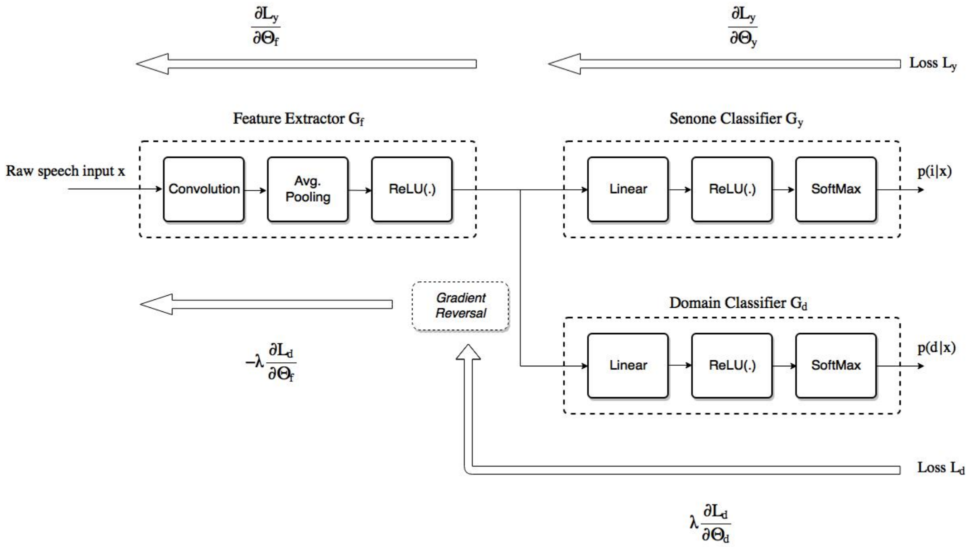

Fig. 1. The feature extractor extracts discriminative features directly from raw speech. It consists of several stages of convolution/pooling/ReLU. Senone classifier along with the feature extractor forms a standard architecture. Domain classifier is also connected to feature extractor through a gradient reversal. Reversal of gradient force the feature extractor to learn domain invariant features.

the sign of gradients, multiply them with λ and pass it onto the subsequent layer, while in forward propagation GRL acts as an identity transform.

At the test time, domain classifier and GRL are discarded. The data samples are passed through the feature extractor and label classifier to get the predictions.

4. EXPERIMENTAL SETUP

This section explains the setup used for domain adaptation experiments. Sec. 4.1 describes the dataset used for the experiments, sec. 4.2 details the input feature preparation and sec. 4.3 discusses the model architecture and training details.

4.1. Dataset Description

TIMIT [20] and Voxforge [21] datasets are used to perform domain adaptation experiments. For TIMIT speech corpus domain adaptation is performed by taking male speech as source domain and female speech corpus as target domain. For the Voxforge corpus domain adaptation is performed by taking American accent and British accent as source domain and target domain respectively and vice-versa.

For TIMIT speech corpus male and female speakers are separated into source domain and target domain datasets. TIMIT is a read speech corpus in which a speaker reads a prompt in front of the microphone. It consists of a total of 6300 sentences, 10 sentences spoken by each of the 630 speakers for 8 major dialect regions of the United States of America. It consists a total of 3,696 training utterances sampled at 16kHz, excluding all SA utterances because they create a bias in the dataset. The training set consists of 438 male speakers and 192 female speakers. The core test set is used to report the results. It consists of 16 male speakers and 8 female speakers from all of the 8 dialect regions.

For the Voxforge dataset American accent speech and British accent speech are taken as two separate domains. Voxforge is a multi-accent speech dataset with 5 sec speech samples sampled at 16 KHz. Speech samples are recorded by users with their own microphones therefore quality varies significantly among samples. Voxforge corpus has 64 hrs of American accent speech and 13.5 hrs of British accent speech totaling to 83 hrs of speech. Results are reported on 400 utterances each for both the accents. Alignments are obtained by using HMM-GMM acoustic model trained using Kaldi [22].

4.2. Input Features

Raw speech features are obtained by using a rectangular window of size 10 ms on raw speech with a frame shift of 10 ms. A context of 31 frames is added to windowed speech features to get a total of 310 ms of context dependent raw speech features. These context dependent raw speech features are mean and variance normalized to obtain final features.

4.3. Model Description

Feature extractor is a 2 layer convolutional neural network. The first convolutional layer has filter size of 64 with 256 feature maps along with the step size of 31. The second convolutional layer has filter size of 15 with 128 feature maps and step size of 1. After each convolutional layer a avg-pool layer is used with pooling size of 2 and a ReLU activation unit. Both the senone classifier and domain classifier are 4 layer and 6 layer fully connected neural networks with ReLU activation unit and hidden unit size of 1024 and 2048 for TIMIT and Voxforge respectively.

The weights are initialized in Glorot [23] fashion. The model is trained with SGD with momentum [24]. Learning rate is selected during the training using formula $\mu_p = \frac{\mu_0}{(1+\alpha*p)^\beta}$, where p increases linearly from 0 to 1 as training progresses, $\mu_0 = 0.01$, $\alpha = 10$, and $\beta = 0.75$. A momentum of 0.9 is also used. The adaptation parameter λ is initialized at 0 and is gradually changed to 1 according to the formula $\lambda_p = \frac{2}{1+\exp(-\gamma*p)} - 1$, where γ is set to 10 as suggested in [1]. Domain labels are switched 10% of the times to stabilize the adversarial training.

5. RESULTS

This section presents the results of experiments performed to evaluate the impact of adversarial training for unsupervised domain adaptation. The experiments specifically study the acoustic variabilities like speaker gender and accent using TIMIT and Voxforge speech corpus respectively. Sec. 5.1 presents the domain adaptation experiments for speaker gender variability. Due to insufficient labeled female speech data in TIMIT corpus domain adaptation, experiments

Labeled source data	Unlabeled target data	Test data	NN	DANN
Male + Female		Male	21.25	
Male + Female		Female	23.21	
Male	Female	Male	24.63	25.37
Male	Female	Female	37.20	**32.26**

Table 1. % PER for for acoustic model trained on supervised data from source domain and unsupervised data from target domain for TIMIT corpus taking male speech as source and female speech as target.

are performed only for male speech as the source domain and female speech as target domain. Sec. 5.2 discusses the experiments performed to mitigate speaker accent variability using American and British accent speech data in Voxforge. Experiments are performed by taking American accent as source domain and British accent as target domain and vice versa. Additional experiments are also performed by training the acoustic model on the labeled data from both the domains which work as the lower limit for the achievable WER. In Table 1 and 2, DANN represents the domain adapted acoustic model using labeled data from the source domain and unlabeled data from the target domain and NN represents the acoustic model trained on the labeled data from the source domain only. We also trained a fully connected DNN using MFCC features, with two layers for the feature extractor and three each for the senone and domain classifiers. Each fully connected layer comprised of 1024 nodes and the total number of parameters is similar to that of the model described in Sec. 4.3.

5.1. Domain adaptation for male and female speech domains in TIMIT corpus

The first two rows in Table 1 list the PER results for the acoustic model trained on labeled data from both the domains with no domain adaptation. This acoustic model gives the best results and is the lower limit for the PER. Rows 3 and 4 give the acoustic model trained on labeled data from the male speakers and adapted using unlabeled data from female speakers. Row 3 indicates the effect of domain adaptation on the performance on data from source domain which is male speech in this case. Row number 4 gives the PER for the unadapted and adapted acoustic models for data from target domain which is female speech in this case. With male and female speech as source and target domains respectively, and using MFCC features as input, the PER dropped from 33.825% to 31.375% upon applying domain adaptation, indicating a higher absolute accuracy, but a poorer relative improvement compared to raw speech.

5.2. Domain adaptation for American and British accents domains in Voxforge corpus

Rows 1 and 2 in Table 2 are the WER values for the acoustic model trained on labeled data from both the domains and without any domain adaptation. These values correspond to lower limits for the WER for both the domains. Row number 3 and 4 represents the effect of domain adaptation on the performance of acoustic model on the data from source domain which is American and British respectively. Rows number 5 and 6 gives the WER for target domain data on unadapted and adapted acoustic models. With American and British speech as source and target domains respectively, and using MFCC features, the WER dropped from 24.19% to 23.73% upon applying domain adaptation. These results indicate a poorer absolute accuracy as well as relative improvement compared to the raw speech experiments reported in Table 2.

Labeled source data	Unlabeled target data	Test data	NN	DANN
American + British		American	10.87	
American + British		British	15.01	
American	British	American	11.50	16.53
British	American	British	18.41	19.62
American.	British	British	28.11	**23.10**
British	American	American	23.37	**23.16**

Table 2. % WER for acoustic models trained on supervised data from source domain and unsupervised data from target domain for Voxforge dataset taking American and British accents as two different acoustic domains.

6. DISCUSSION

As can be seen from rows 3 and 4 of Table 1, the acoustic variability due to speaker gender results in a 12.57 % absolute increase in PER when the acoustic model (NN) trained on male speech (source domain) is tested on female speech (target domain). The effect of applying DANN improves the absolute PER performance on the target domain by nearly 5%. Similarly, domain shifts due to speaker accent cause the target domain performance deteriorate significantly. Table 2 reports an absolute degradation of 16.61% when the acoustic model is trained on American speech (source domain) and tested on British speech (target domain). Upon applying DANN using the unlabeled British speech, the absolute WER drops by nearly 5%. General trends that appear from our experimental analysis point out that models undergoing adversarial training based unsupervised domain adaptation, improve in performance on the target domain data as opposed to their unadapted counterparts. This improvement however, comes at a cost of drop in source domain accuracy. This observation is not unexpected as the feature extractor module perhaps learns to ignore some domain-specific features in the pursuit of learning invariant representations.

7. CONCLUSION

The paper proposes unsupervised domain invariant features learning directly from raw speech using domain adversarial neural networks. The senone classification model is modified for domain adaptation by using an additional domain classifier and modifying the loss such that the network learns features from raw speech that are sufficiently discriminative for the senone classifier and invariant enough to fool the domain classifier. This experimental study also shows that there is a significant acoustic variability present in the speech signal due to speaker gender and accent which adversely affects the performance of models trained in a domain agnostic setting. The performance loss due to this variability can be alleviated by adversarial training based domain adaptation using unlabeled target domain data. Moreover, the analysis suggests that raw speech features along with a convolutional neural network based feature extractor may be more amenable to an advesarial approach to domain adaptation as opposed to handcrafted features like MFCC, particularly when large amounts of data is available. The evaluation of the proposed approach is done using two benchmarking datasets: TIMIT for gender based domain shift and Voxforge corpus for using American and British accents for the domain shift.

8. ACKNOWLEDGMENT

While the first author was at IIIT-Delhi, this work was partially supported by the Infosys Center for Artificial Intelligence, IIIT-Delhi.

9. REFERENCES

[1] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[2] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al., "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10, pp. 763–786, 2007.

[3] Dimitri Palaz, Ronan Collobert, et al., "Analysis of CNN-based speech recognition system using raw speech as input," Tech. Rep., Idiap, 2015.

[4] Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4295–4299.

[5] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks.," in *INTERSPEECH*, Frdric Bimbot, Christophe Cerisara, Ccile Fougeron, Guillaume Gravier, Lori Lamel, Franois Pellegrino, and Pascal Perrier, Eds. 2013, pp. 1766–1770, ISCA.

[6] Pegah Ghahremani, Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Acoustic modelling from the signal domain using CNNs.," in *INTERSPEECH*, 2016, pp. 3434–3438.

[7] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Mark Gales and Steve Young, "The application of hidden Markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[9] Yoo Rhee Oh and Hong Kook Kim, "MLLR/MAP adaptation using pronunciation variation for non-native speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 216–221.

[10] Puming Zhan and Alex Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," Tech. Rep., Carnegie-Mellon Univ Pittsburgh PA School Of Computer Science, 1997.

[11] Vishwa Gupta, Patrick Kenny, Pierre Ouellet, and Themos Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6334–6338.

[12] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[13] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[14] Yusuke Shinohara, "Adversarial Multi-Task learning of deep neural networks for robust speech recognition.," in *INTERSPEECH*, 2016, pp. 2369–2372.

[15] Dmitriy Serdyuk, Kartik Audhkhasi, Philémon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, "Invariant representations for noisy speech recognition," *arXiv preprint arXiv:1612.01928*, 2016.

[16] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, 2017.

[17] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," *arXiv preprint arXiv:1707.06265*, 2017.

[18] Chao Huang, Tao Chen, and Eric Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2, pp. 141–153, 2004.

[19] Rachael Tatman and Conner Kasten, "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions," *Proc. Interspeech 2017*, pp. 934–938, 2017.

[20] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[21] Ken MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on ASRU*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[23] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks.," in *Aistats*, 2010, vol. 9, pp. 249–256.

[24] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.