Improving Automatic Recognition of Aphasic Speech with AphasiaBank

Duc Le and Emily Mower Provost

University of Michigan
Computer Science and Engineering, Ann Arbor, MI 48109, USA
{ducle, emilykmp}@umich.edu

Abstract

Automatic recognition of aphasic speech is challenging due to various speech-language impairments associated with aphasia as well as a scarcity of training data appropriate for this speaker population. AphasiaBank, a shared database of multimedia interactions primarily used by clinicians to study aphasia, offers a promising source of data for Deep Neural Network acoustic modeling. In this paper, we establish the first large-vocabulary continuous speech recognition baseline on AphasiaBank and study recognition accuracy as a function of diagnoses. We investigate several out-of-domain adaptation methods and show that AphasiaBank data can be leveraged to significantly improve the recognition rate on a smaller aphasic speech corpus. This work helps broaden the understanding of aphasic speech recognition, demonstrates the potential of AphasiaBank, and guides researchers who wish to use this database for their own work.

Index Terms: speech recognition, acoustic modeling, aphasia, AphasiaBank, out-of-domain adaptation

1. Introduction

Aphasia is a common neurological disorder that impairs a person's speech and language capabilities. It is estimated that over 1 million people in the US have aphasia and 180,000 acquire it every year[1]. Persons with aphasia (PWAs) face serious communication difficulties and frequently experience social isolation [1–4]. Speech-based technology offers PWAs many potential benefits due to its low cost and constant accessibility [5–7].

In our previous works, we investigated the feasibility of an automated intelligibility assessment system which may help PWAs improve their verbal output and long-term prognosis [6, 7]. A major roadblock in the future automation of these works is the poor performance of automatic speech recognition (ASR), which hindered the system's efficacy in real-time. ASR for aphasic speech is considerably challenging for a number of reasons. Firstly, a PWA's pronunciation can be distorted due to co-occurring motor control disorders such as apraxia of speech (AOS) and dysarthria. Secondly, language impairments may result in halting speech containing jargon and various types of paraphasia, all of which are difficult to capture with conventional ASR methods. Thirdly, the size of most aphasic speech datasets is relatively small, partly due to the difficulties involved in collecting this type of data at a large scale. This data scarcity reduces the utility of modern ASR methods such as Deep Neural Network (DNN) acoustic modeling, which typically requires a large amount of data to outperform the traditional Gaussian Mixture Model (GMM). The lack of data is further exacerbated by the high speaker variability among PWAs.

[1] http://aphasia.org/?q=content/aphasia-faq. Retrieved March 2016.

In this paper, we present a study that aims to enhance the quality of aphasic speech recognition by leveraging data from AphasiaBank, a shared multimedia database primarily used by clinicians to study aphasia [8]. We establish the first large-vocabulary continuous speech recognition (LVCSR) baseline on English AphasiaBank using DNN acoustic models. We find that appending utterance i-vectors to frame-level acoustic features results in a **3.1%** to **15.1%** relative reduction in per-speaker Phone Error Rate (PER), with more severe speakers receiving larger improvement. We investigate out-of-domain adaptation methods to adapt AphasiaBank to the University of Michigan Aphasia Program (UMAP) dataset, a smaller aphasic corpus used in our previous works for speech intelligibility assessment [6, 7]. We show that discriminative pretraining produces a **18.8%** $\pm$ **9.1%** relative reduction in per-speaker PER compared to a baseline without using AphasiaBank, with less severe PWAs benefiting more from adaptation. Our work helps further the understanding of aphasic speech recognition, provides insights into the types of speakers who would benefit from different adaptation techniques, demonstrates the potential of AphasiaBank in ASR, and suggests that real-time feedback, which relies heavily on ASR, may be feasible in certain contexts.

2. Related Work

2.1. ASR for Disordered Speech

There has been extensive work in the related field of dysarthric speech recognition [9–15]. ASR for dysarthric and disordered speech in general is faced with abnormal speech patterns, high speaker variability [16], and data scarcity [11]. Methods for alleviating these problems include speaker-dependent GMM adaptation [9, 11, 12], generation of auxiliary acoustic features used within tandem-based systems [10, 14], learning speaker-specific pronunciation [13], and speaker selection [15]. Most of these works focused on isolated word recognition and used the traditional HMM-GMM model. Our work investigates continuous ASR and the more recent HMM-DNN framework.

There has been relatively little work on ASR for aphasic speech. Existing works are limited to using healthy acoustic models to recognize aphasic speech [5, 17]. Further, aphasia and dysarthria have several key differences. A PWA's verbal expression is modulated by language impairment and co-occurring motor control disorders, which often include AOS and dysarthria itself. AOS can make the speech produced by PWAs inconsistent, thus increasing the degree of intra-speaker variability. Verbal output of different PWAs may differ drastically depending on the specific aphasia type, such as fluent and non-fluent aphasia. It is unclear whether or not techniques that work for dysarthria will also translate to aphasia.

2.2. Under-Resourced ASR

Disordered speech recognition also shares important similarities with low-resource ASR due to the issue of data scarcity. Common techniques for handling this problem include deep bottleneck [18, 19] or posterior-based [20] features used within tandem-based systems [21], and discriminatively pretrained DNN acoustic model using out-of-domain data [22]. A shared theme of these methods is the use of external speech (e.g., multilingual data) for enhancing the performance of in-domain models. In the context of ASR for disordered speech, out-of-domain data usually consist of healthy speech [10, 14]. However, there is an inherent mismatch between healthy and disordered speech [11], suggesting that healthy speech data may not be the most appropriate choice for out-of-domain adaptation. We leverage aphasic speech directly as out-of-domain data in this work.

2.3. ASR with I-vectors

Another popular auxiliary feature for ASR is i-vector, which encapsulates speaker characteristics in a fixed-length representation and is commonly used in speaker verification [23, 24]. Appending i-vectors to frame-level acoustic features has been shown to improve ASR performance with DNN acoustic models [25–27]. I-vector is a promising approach for handling the high speaker variability present in disordered speech; however, its application to this type of data has been limited.

3. Data

3.1. ApahasiaBank

AphasiaBank is a shared audiovisual database containing interactions between PWAs and research investigators, and is primarily used by clinicians to study aphasia [8]. AphasiaBank is a collection of multiple sub-databases collected by different research groups under various recording conditions and elicitation protocols. For this work, we consider English sub-databases containing at least four speakers and were collected with the AphasiaBank protocol, which involves open-ended questions designed to collect verbal discourse samples from PWAs.

Our inclusion criteria selected 18 sub-databases containing 401 speakers (238 male, 163 female, age 62 $\pm$ 12). The PWA breakdown by WAB-R's Aphasia Quotient (AQ) [28] is: 43.4% mild, 32.7% moderate, 9.2% severe, 3.2% very severe, and 11.5% unknown. 63.8% and 24.9% of speakers have fluent and non-fluent aphasia, respectively; the remaining have missing diagnoses. We further discard 3.8% of utterances that have unintelligible or overlapping speech as these may fail to align properly. The final dataset contains 89.2 hours of speech, 64,748 utterances, and 458,138 instances of 11,803 unique words.

We downsample the audio to 16kHz and extract 12 MFCCs plus energy, along with delta and delta-delta coefficients. We z-normalize the features at the speaker level. Finally, we perform 4-fold partitioning to help establish an ASR baseline on AphasiaBank. We withhold 25% of speakers from each sub-database to form the test set. We further withhold 15% of training speakers from each sub-database to form a development set for parameter tuning. The test sets across these four folds form a complete partition of AphasiaBank. The per-fold training set contains approximately 56 hours of speech data.

3.2. UMAP

The UMAP dataset contains speech recordings of 17 PWAs (11 male, 6 female, age 58 $\pm$ 14) interacting with a tablet application designed for sentence building exercises. PWAs are presented with a picture stimulus and asked to verbally produce a sentence to describe the picture. Five, nine, and three speakers have mild, moderate, and severe aphasia based on WAB-R's AQ. Nine have AOS and one has dysarthria. Eight have fluent aphasia and nine have non-fluent aphasia. This dataset was used in our previous work to develop methods for automatic speech intelligibility assessment [6, 7]. A major bottleneck in these works was the reliance on human-labeled transcripts. Achieving good ASR performance on this dataset will move us closer to deploying the system for real-world usage.

The data were recorded using the tablet's built-in microphone with a 44.1 kHz sampling rate. All utterances were transcribed at the word-level with timing information by human annotators. Special events such as unintelligible words and background noise are marked with special labels. We split each utterance into continuous segments of intelligible speech, each of which contains on average 2 to 4 words. We will perform ASR evaluation on these segments. The segment-level data contains in total 2.1 hours and 12,661 instances of 1,073 unique words.

We apply an identical feature extraction pipeline used in AphasiaBank. ASR evaluation will be done through leave-one-speaker-out cross-validation, which results in 17 folds where data from one speaker are withheld for testing and the rest are used for training. We further withhold 15% of utterances from each training speaker to form a development set. The size of the per-fold training set ranges from 1.7 to 2 hours.

4. Experimental Setup

4.1. Intra-Database Speech Recognition

In this section, we outline our experiments for intra-database speech recognition, which will result in a speaker-independent cross-validated PER for each database. We consider two classes of methods, one based on the traditional context-dependent tied-state triphone HMM-GMM model, and one based on the more modern hybrid HMM-DNN system [29, 30]. We train two versions of HMM-DNN, one with and one without i-vectors in the input features. Details about this experiment are summarized in Table 1. We used Kaldi [31] for HMM-GMM modeling and i-vector extraction, and Theano [32] for DNN training. Additional data for replicating this work, such as fold selection, transcription, and audio segmentation, are available online[2]. For the remainder of this section, we will elaborate on the hyperparameter choice, learning schedule, and i-vector extraction.

4.1.1. Hyperparameter Selection

HMM-GMM and HMM-DNN both require a number of hand-picked hyperparameters, such as the number of Gaussians and tied-states for the former, and the DNN architecture and training recipe for the latter. Hyperparameters for AphasiaBank were selected based on the average PER achieved on the development set across all four cross-validation folds. On the other hand, hyperparameters for UMAP were selected using an oracle method that optimizes for test PER. Doing so helps us obtain the strongest UMAP baseline to compare against out-of-domain adaptation techniques described in later sections.

4.1.2. Learning Schedule

Learning schedule refers to the adjustment of learning rate after each DNN stochastic gradient descent epoch. We find that

[2] http://www.umich.edu/ ducle/IS16appendix

	AphasiaBank	UMAP
HMM-GMM	Context-dependent tied-state triphone trained with Maximum Likelihood estimation.	
	Parameters:	
	25,000 Gaussians; approx. 3,000 tied states.	8,000 Gaussians; 700-800 tied states.
HMM-DNN	Randomly initialized DNN (5 hidden layers, 1024 units per layer, sigmoid activation) trained with stochastic gradient descent using 27-frame context windows, HMM-GMM alignments, and Cross-Entropy objective.	
	Training without i-vectors: exponential-decay learning schedule (0.4 initial learning rate, 0.05% threshold).	
	No regularization.	2×10^{-5} L2 regularization weight.
	Training with i-vectors: step-decay learning schedule (learning rates: 0.4 initial, 0.01 minimum).	
	10^{-5} L2 regularization weight.	2×10^{-5} L2 regularization weight.
i-vectors	***Universal Background Model (UBM)***: 1024 Gaussians trained on 9-frame spliced MFCCs, followed by 40-dimensional Linear Discriminant Analysis (LDA) with senones as class labels. Only voiced frames are used.	
	Type of i-vector:	
	32-dimensional utterance-level.	32-dimensional session-level.
Decoding	Continuous phone loop using trigram phone-level language model with backoff.	

Table 1: Training and decoding methods for intra-database ASR. See text for description of learning schedule and i-vector type.

different learning schedules must be used for models with and without i-vectors to achieve optimal results.

Exponential-decay: This schedule first trains the network at a fixed initial learning rate (e.g., 0.4). Once the change in frame-level error on the development set drops below a threshold (e.g., 0.05% absolute), we halve the learning rate after every epoch. The training process terminates once the change in development error once again drops below the threshold. We find that this schedule is appropriate for models without using i-vectors, possibly because it finishes faster and avoids overfitting the network to the training set, which is easier to do without having additional features to model.

Step-decay: This schedule is similar to the one used in [29]. Instead of halving the learning rate after every epoch, it halves the learning rate and restores previous network weights whenever the development error does not improve. The training process terminates once the learning rate drops below a minimum value (e.g., 0.01). We find that this schedule is appropriate for less stable and more slowly converging learning process, such as when i-vectors are used.

4.1.3. I-vector Extraction

We set the i-vector dimension to 32, based on validation results on one training fold and the system described in [27]. Following [26], we extract utterance-level i-vectors for AphasiaBank. However, we find that this type of i-vector does not work well on UMAP, possibly because the utterances in the latter are too short. We instead use session i-vectors, which are extracted from all speech data produced by the PWA in one single recording session. There are 125 sessions, each containing 1 minute of speech on average. Refer to Table 1 for more information.

4.2. Adapting AphasiaBank to UMAP

We consider two methods in this work to use AphasiaBank data to improve recognition results on UMAP.

merged: In this method, we merge the full AphasiaBank corpus' training and development set with the UMAP counterparts, and train a new DNN using the same recipe and architecture described in Table 1. This method allows the network to directly model UMAP data while also modeling the large amount of speech present in AphasiaBank. A potential disadvantage of this method is that it might not model UMAP data extensively since UMAP contributes only a relatively small fraction of training data. It is also more computationally expensive.

dpAB: We investigate **d**iscriminative **p**retraining with **A**phasia**B**ank data inspired by the work of Thomas et al. for low-resource ASR [22]. The authors in [22] proposed retraining only the softmax layer while keeping the lower layers fixed. However, we find that retraining the entire AphasiaBank DNN on the UMAP training set, using the step-decay learning schedule and no regularization, yields better results. This suggests that the high-level representation learned by AphasiaBank DNN does not transfer directly to UMAP data. This indicates a large mismatch between the two datasets, and further suggests that methods which aim to constrain the shift in parameters from the original model by inserting additional layers on top of a fixed network [33] or regularizing the change in output distribution [34] may have limited efficacy. However, speaker adaptation on the same dataset, which does not suffer from such data mismatch, may benefit greatly from these approaches.

We also considered using deep bottleneck features (DBNFs) generated by AphasiaBank DNN in a tandem-based system. However, our preliminary experiments were not able to outperform the HMM-GMM baseline. Again, this may be due to the high level of mismatch between AphasiaBank and UMAP. As a result, we do not consider DBNFs in this paper.

5. Results and Discussion

5.1. AphasiaBank Phone Error Rate

Table 2 summarizes the mean and standard deviation of speaker-level PERs on AphasiaBank, where the speakers are grouped by the level of severity defined by WAB-R's AQ.

We first turn attention to the relatively high PERs achieved on this dataset. This may be caused by the abnormal speech patterns associated with aphasia that are difficult to capture with

Severity	No i-vectors	With i-vectors
mild	48.95 ± 11.55	47.41 ± 10.46
moderate	57.04 ± 13.22	52.79 ± 10.37
severe	65.44 ± 18.65	61.00 ± 13.20
v. severe	89.27 ± 29.14	75.81 ± 18.65
unknown	60.36 ± 29.75	54.35 ± 18.64

Table 2: AphasiaBank per-speaker PER, grouped by severity.

conventional ASR techniques. Speech data in AphasiaBank were recorded using video cameras situated far away from the speaker. This far-field recording condition is known to significantly reduce recognition performance. Two observations can be made from these results. One, if we want to apply ASR technology to help improve the well-being of PWAs, it is crucial to constrain the recognition problem in some way, such as restricting the vocabulary or task grammar. Aphasic speech may be too challenging for unconstrained LVCSR to achieve an acceptable recognition accuracy. Two, it is important to realize that ASR is only a precursor and not an end goal for speech-based technology aimed toward PWAs. It will be interesting to investigate tasks that can be performed reasonably well given imperfect ASR output. This will help us better understand what kind of ASR-dependent technology is feasible for aphasic speech.

These results also show that both the mean and standard deviation of per-speaker PERs tend to increase as a PWA's aphasia becomes more severe on the WAB-R's AQ scale. This is a useful observation as it shows that AQ, despite being a measure of general language skills and not of speech itself, can be a reasonable estimate for the effectiveness of ASR. Being able to predict how well an ASR system will work for a speaker using readily available information such as AQ may help the system adapt to that speaker more quickly and effectively. A natural extension of this observation is to use a speaker's severity level directly as input to the DNN, such as encoding it as a one-hot vector. However, our preliminary experiments indicate that this approach does not yield additional improvement on top of i-vectors. We will explore different methods to augment acoustic modeling with PWAs' diagnoses in future work.

Finally, we note the effectiveness of i-vectors for aphasic speech recognition. Models that use i-vectors in the input features experience a reduction in both the mean and standard deviation of per-speaker PER. While the relative improvement for speakers with mild aphasia is relatively small (3.1%), the improvement is more noticeable for those with moderate to severe (6.8% – 7.5%), and especially very severe aphasia (15.1%). Christensen et al. noted that although their out-of-domain adaptation technique is quite effective, speakers with more severe dysarthria tend to benefit less from adaptation [14]. Our results suggest a complementary method for improving the recognition rate of the more severe population.

5.2. UMAP Phone Error Rate

Figure 1 shows the PERs for different speakers in the UMAP dataset using the HMM-GMM baseline model. The PERs range from 20.8% to 71.2% (mean 39.7%, std. deviation 11.1%). We will estimate the effectiveness of different adaptation methods based on the resulting change in PER for each speaker. These are summarized in Table 3.

The first two rows, *AB-DNN* and *UMAP-DNN*, refer to

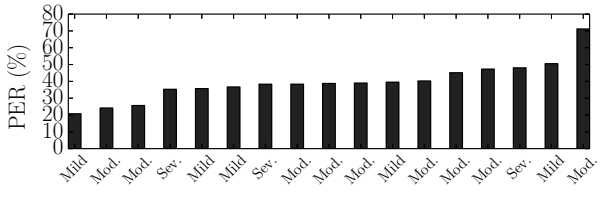

Figure 1: UMAP per-speaker PER using HMM-GMM baseline trained only on UMAP data. x-axis denotes AQ severity level.

Model	No i-vectors	With i-vectors
AB-DNN	4.8 ± 15.1	3.4 ± 15.5
UMAP-DNN	-1.0 ± 7.6	2.9 ± 9.0
merged	-14.7 ± 9.3	**-16.6 ± 8.9**
dpAB	**-18.8 ± 9.1**	-15.9 ± 7.6

Table 3: Relative change (%) in UMAP per-speaker PER over HMM-GMM baseline. A negative value means reduced PER. *AB-DNN* and *UMAP-DNN* are DNNs trained only on AphasiaBank and UMAP data, respectively.

DNN acoustic models trained only on AphasiaBank and UMAP data, respectively. Compared to the baseline, the resulting PERs for both models improve for some speakers and worsen for others, and there is no clear advantage to using either model. The fact that *UMAP-DNN* was not able to outperform the HMM-GMM baseline reinforces the data scarcity problem in aphasic speech recognition. Looking at individual speakers, *AB-DNN* tends to work better for those who are similar to the typical speakers in AphasiaBank, namely those with mild and fluent aphasia. On the other hand, there is no obvious pattern as to which type of speaker benefits from the *UMAP-DNN* model.

Of the two adaptation methods, the best result (**18.8% ± 9.1%** relative improvement) is achieved with *dpAB*, which uses UMAP data to finetune a DNN that was discriminatively pretrained on AphasiaBank. Speakers with mild severity receive the largest improvement (22.5% ± 7.5%), while those with fluent and non-fluent aphasia experience a similar degree of PER reduction (19.2% ± 9.3% vs. 18.3% ± 9.0%). On the other hand, the *merged* adaptation method provides more benefit to those with fluent aphasia, resulting in 18.7% ± 6.1% relative improvement compared to 14.7% ± 10.5% for non-fluent.

Finally, we analyze the effect of i-vectors on adaptation. Using i-vectors resulted in better performance for *AB-DNN* and *merged*, but worse performance for *UMAP-DNN* and *dpAB*. The common theme among the two methods that were not able to take advantage of i-vectors is that only UMAP i-vectors were used for DNN training. On the other hand, using UMAP i-vectors directly in testing (*AB-DNN*) or training them jointly with AphasiaBank i-vectors (*merged*) proved beneficial. A possible explanation is that the 125 UMAP session i-vectors are too few in number and too dissimilar for the network to take advantage of in a speaker-independent setup. Additional work is needed to leverage i-vectors in limited-data situations.

6. Conclusion and Future Work

In this work, we establish the first LVCSR baseline on AphasiaBank, and show that AphasiaBank data can be leveraged to improve the recognition rate on a smaller aphasic speech corpus by a large margin through discriminative pretraining. Our analysis suggests that discriminative pretraining provides more benefit to PWAs with lower severity, while i-vector-based adaptation benefits those with higher severity. However, more work is needed to combine the benefit of both approaches.

We plan to extend this work in two major directions. Firstly, we are interested in the extent to which an improved ASR model can replace human-labeled transcripts in our system for automatic quantification of aphasic speech intelligibility [6, 7]. Secondly, we will investigate more fine-grained adaptation methods based on diagnoses and other speaker properties. Given the high speaker variability present in aphasic speech, more highly personalized models may result in further gain.

7. References

[1] A. Basso, *Aphasia and Its Therapy*. Oxford University Press, 2003.

[2] G. A. Davis, *Aphasiology: Disorders and Clinical Practice*, 2nd ed. Pearson, 2006.

[3] L. R. Cherney, A. S. Halper, A. L. Holland, and R. Cole, "Computerized Script Training for Aphasia: Preliminary Results," *American Journal of Speech-Language Pathology*, vol. 17, no. 1, pp. 19–34, Feb 2008.

[4] N. Helm-Estabrooks, M. L. Albert, and M. Nicholas, *Manual of Aphasia and Aphasia Therapy*, 3rd ed. Pro-Ed, 2013.

[5] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Computer Speech and Language*, vol. 27, no. 6, pp. 1235–1248, 2013.

[6] D. Le, K. Licata, E. Mercado, C. Persad, and E. Mower Provost, "Automatic Analysis of Speech Quality for Aphasia Treatment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[7] D. Le and E. M. Provost, "Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation," in *Proc. of the 15th Annual Conference of the ISCA (INTERSPEECH)*, Singapore, 2014.

[8] B. Macwhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for Studying Discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[9] H. V. Sharma and M. Hasegawa-Johnson, "State-transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition," in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Los Angeles, CA, USA, 2010, pp. 72–79.

[10] M. Aniol, "Tandem features for dysarthric speech recognition," Master's thesis, Edinburgh University, United Kingdom, 2012.

[11] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc. of the 13th Annual Conference of the ISCA (INTERSPEECH)*, Portland, OR, USA, 2012.

[12] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*, vol. 27, no. 6, pp. 1147 – 1162, 2013.

[13] H. Christensen, P. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech," in *Proc. of the 14th Annual Conference of the ISCA (INTERSPEECH)*, Lyon, France, 2013.

[14] H. Christensen, M. B. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Proc. of the 14th Annual Conference of the ISCA (INTERSPEECH)*, Lyon, France, 2013.

[15] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic Selection of Speakers for Improved Acoustic Modelling: Recognition of Disordered Speech with Sparse Data," in *IEEE Workshop on Spoken Language Technology (SLT)*, South Lake Tahoe, NV, USA, 2014.

[16] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal, "Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker," *Expert Systems with Applications*, vol. 42, no. 8, pp. 3924 – 3932, 2015.

[17] T. Lee, Y. Liu, P. Huang, J. Chien, W. Lam, Y. Yeung, T. Law, K. Lee, A. Kong, and S. Law, "Automatic Speech Recognition for Acoustical Analysis and Assessment of Cantonese Pathological Voice and Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[18] D. Yu and M. Seltzer, "Improved Bottleneck Features Using Pretrained Deep Neural Networks," in *Proc. of the 12th Annual Conference of the ISCA (INTERSPEECH)*, Florence, Italy, 2011.

[19] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 3377–3381.

[20] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *IEEE Workshop on Spoken Language Technology (SLT)*, Miami, FL, USA, 2012.

[21] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, pp. 1630–1635.

[22] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 6704–6708.

[23] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, 2011.

[24] O. Glembek, L. Burget, P. Matejka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 4516–4519.

[25] 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)*, Olomouc, Czech Republic, 2013, pp. 55–59.

[26] A. Senior and I. Lopez-Moreno, "Improving DNN Speaker Independence with I-vector Inputs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[27] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi, "Robust i-vector based adaptation of DNN acoustic model for speech recognition." in *Proc. of the 16th Annual Conference of the ISCA (INTERSPEECH)*, Dresden, Germany, 2015, pp. 2877–2881.

[28] A. Kertesz, *The Western Aphasia Battery - Revised*. Texas: Harcourt Assessments, 2006.

[29] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[30] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, 2011.

[32] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016.

[33] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc. of the 11th Annual Conference of the ISCA (INTERSPEECH)*, Chiba, Japan, 2010.

[34] H. S. G. L. D. Yu, K. Yao and F. Seide, "KL-Divergence Regularized Deep Neural Network Adaptation For Improved Large Vocabulary Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.