INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study

Duc Le[†], *Keli Licata*[‡], *Emily Mower Provost*[†]

University of Michigan, Ann Arbor, MI 48109, USA
[†]Computer Science and Engineering, [‡]University Center for Language and Literacy
{ducle,klicata,emilykmp}@umich.edu

Abstract

Aphasia is an acquired language disorder resulting from brain damage that can cause significant communication difficulties. Aphasic speech is often characterized by errors known as paraphasias, the analysis of which can be used to determine an appropriate course of treatment and to track an individual's recovery progress. Being able to detect paraphasias automatically has many potential clinical benefits; however, this problem has not previously been investigated in the literature. In this paper, we perform the first study on detecting phonemic and neologistic paraphasias from scripted speech samples in AphasiaBank. We propose a speech recognition system with task-specific language models to transcribe aphasic speech automatically. We investigate features based on speech duration, Goodness of Pronunciation, phone edit distance, and Dynamic Time Warping on phoneme posteriorgrams. Our results demonstrate the feasibility of automatic paraphasia detection and outline the path toward enabling this system in real-world clinical applications.
Index Terms: aphasia, paraphasia detection, pronunciation modeling, disordered speech recognition

1. Introduction

Aphasia is an acquired language disorder resulting in a loss of language skills that generally arises from focal brain damage to the left cerebral hemisphere [1]. In the US, there are approximately two million people with aphasia and more than 180,000 acquire it every year due to brain injury, most commonly from a stroke [2]. The speech-language deficits associated with aphasia impact one's ability to communicate effectively, making social interaction difficult and frustrating. Aphasia is a chronic disorder that affects the social, recreational, and vocational lives not only of the affected individuals, but also of their friends and family members. This results in feelings of social isolation, loss of autonomy, and loneliness, among others [3].

Anomia (word retrieval deficit) is the core symptom of aphasia and is present in virtually all persons with aphasia (PWAs) [4]. Those who have anomia often produce various types of paraphasias (naming errors), the most common of which are *semantic*, *phonemic*, and *neologistic*. In these three categories, respectively, the PWA may substitute the target word (e.g., *harmonica*) with a semantically related word (e.g., *flute*), a phonemically related word (e.g., *karmonica*), or a non-word (e.g., *parokada*). The type and frequency of the produced paraphasias play an important role in estimating the severity of anomia as well as determining an appropriate treatment approach [5, 6]. For example, PWAs who produce mainly semantic paraphasias may benefit from treatment approaches focusing on word meaning, while treatment approaches targeting the phonological structure of target words may be more appropriate for PWAs who produce mainly phonemic paraphasias [5, 7].

Being able to detect paraphasias automatically from a PWA's speech (e.g., through a computer-based word-finding exercise) would provide SLPs with a useful tool for both diagnostic and progress-monitoring purposes and, as such, would help guide the treatment process. Additionally, it could lead to computer-based activities for in-home practice for PWAs, thereby increasing the intensity of practice and facilitating carryover of progress from therapy to other environments. It could also serve to increase a PWA's awareness of errors and enhance self-monitoring skills and, thus, promote independence in overall communication. However, the automatic detection of paraphasias has not previously been studied in the literature.

In this work, we present a pilot study that investigates the feasibility of detecting phonemic and neologistic paraphasias automatically from aphasic speech. We demonstrate that when the target transcript is known, phonemic and neologistic paraphasias can be successfully distinguished from correctly pronounced words. We also investigate a variant of the problem in which the target transcript needs to be generated automatically. In this setup, our system is able to outperform the naïve baseline in detecting the presence of paraphasias in utterances, and achieve good correlation in estimating the rate of phonemic paraphasia production for each speaker. The results and analyses provided in this work help lay the foundation for future work targeting automatic paraphasia detection.

2. Related Work

To the best of our knowledge, no existing work has looked at paraphasia detection in aphasic speech from a technical perspective. Previous works primarily tackled utterance-level and speaker-level classification problems for therapeutic and diagnostic purposes [8–12]. Peintner et al. [8] proposed speech and language features to distinguish between three types of frontotemporal lobar degeneration, including progressive non-fluent aphasia. Fraser et al. [9] combined transcript and low-level acoustic features to classify between two subtypes of primary progressive aphasia (PPA). Le et al. tackled the problem of predicting utterance-level pronunciation, fluidity, and prosody scores given read speech samples of PWAs [10–12]. The most closely related works are those of Abad et al. [13, 14], which used keyword spotting to recognize phrases spoken by PWAs during word naming exercises. However, they did not consider fine-grained word-level labels such as paraphasias.

In an oracle setting where we have access to a PWA's target transcript, automatic paraphasia detection shares certain similarities with mispronunciation detection, an extensively studied problem in the literature. The task in both cases is to classify each word in the transcript as either correct or containing errors. We adopt techniques proposed by Lee et al. [15–17], which compared a non-native speaker's word- and phone-level pronunciations against those of a native speaker, using Dynamic Time Warping (DTW) features extracted on phoneme posterior-

http://dx.doi.org/10.21437/Interspeech.2017-626

Table 1: *Example AphasiaBank transcripts.*

Target	I have aphasia
P1	I have the aphasia
P2	have æfeziəˈ [: aphasia] [* n:k]
P3	I have vəfeɜəˈ [: aphasia] [* p:n]

Table 2: *Dataset summary.*

Speaker	**Utts**	**Words**	**Phonemic**	**Neologistic**
P1	85	787	90	72
P2	108	879	108	66
P3	109	1060	113	46
P4	88	767	108	75
P5	67	652	101	36
P6	37	262	28	61
P7	103	1118	67	18
P8	104	1076	117	24
P9	93	901	146	53
P10	6	47	2	4
P11	67	607	136	112
P12	123	1154	101	32
Total	**990**	**9310**	**1117**	**599**

grams. However, PWAs often do not produce the correct target due to their speech-language impairments. Consequently, target transcriptions may not be available, and reference utterances do not always exist, making it difficult to apply techniques from mispronunciation detection. In this paper, we investigate the oracle use case where target transcripts are available, as well as a more realistic scenario in which automatic speech recognition (ASR) is used to generate the transcripts automatically.

3. Data

AphasiaBank is a large-scale audiovisual dataset primarily used by clinical researchers to study aphasia [18, 19]. It contains a number of sub-datasets collected by different research groups under various recording conditions and elicitation protocols. We focus on the *Fridriksson* sub-dataset of the *Scripts* portion of English AphasiaBank, which contains recordings of 12 PWAs reading from four predefined scripts (*advocacy*, *eggs*, *vast*, and *weather*). The other *Scripts* sub-dataset, *Adler*, consists of six high-functioning PWAs and very few instances of paraphasias. We therefore exclude it from this study.

Each utterance in this set was transcribed verbatim with word-level error codings in concordance with the CHAT transcription format [20]. Word-level error codes include semantic, phonemic, and neologistic paraphasias, each of which is accompanied by a target word. Table 1 shows example transcripts of three PWAs reading the prompt "*I have aphasia*." P1 produced the target without any paraphasia, but added an extra "*the*." P2 and P3 produced neologistic and phonemic paraphasias, respectively, for the target word "*aphasia*." The actual pronunciation was transcribed in IPA format (ending with ˈ).

We target phonemic and neologistic paraphasias in this work. Detecting semantic paraphasias requires a different approach and will be addressed in future work. Table 2 summarizes the 12 speakers in the dataset, along with the utterance and word count, as well as the number of phonemic and neologistic paraphasias. In total, phonemic and neologistic paraphasias account for 12.0% and 6.4% of the words, respectively.

All experiments in this paper are performed with leave-one-speaker-out cross-validation in order to assess the system's performance on unseen speakers. We further withhold 10% of utterances from each training speaker to form a development set.

4. Paraphasia Detection

4.1. With Known Target Transcripts

We first want to determine if it is possible to separate phonemic and neologistic paraphasias from correct words. We define the target transcript of an utterance as the original transcript in which all phonemic and neologistic paraphasias are replaced with their corresponding targets. Thus, the target transcripts in Table 1 will be: "*I have the aphasia*" (P1), "*have aphasia*" (P2), and "*I have aphasia*" (P3). We assume that we have access to the target transcripts. The goal is then to label each word according to one of the following binary classification schemes:

- *C–pn*: correct (C) vs. phonemic or neologistic (pn).
- *C–p*: correct (C) vs. phonemic (p).
- *C–n*: correct (C) vs. neologistic (n).

where correct words are defined as those without any error code. We exclude words that do not fall under any labeling category (e.g., semantic paraphasias), as well as audible background noise, breath sounds, fillers, and laughters.

Metric: although the focus of this work is to detect phonemic and/or neologistic paraphasias, we argue that detecting correctly produced words is equally important. We therefore utilize the average F1 score across classes for evaluation.

Baseline: no baseline currently exists as this is the first work to tackle paraphasia detection. We adopt a simple approach that labels every word as correct (i.e., the majority class).

4.2. Without Known Target Transcripts

The target transcripts will not be available in advance for many real-world applications. We propose to transcribe test utterances automatically with ASR to overcome this limitation. Given the hypothesized transcripts, we can utilize the same classification models in Section 4.1 to obtain predicted word labels.

We consider three types of evaluation metrics that measure the system's performance at the word, utterance, and speaker level. These metrics will help determine the system's applicability under different levels of analyses.

Word-Level Metric: the ideal paraphasia detection system should simultaneously generate the correct target transcripts and label each word accurately. We encode this idea by augmenting the hypothesized and reference target transcripts with corresponding word labels. Under the *C–pn* classification scheme, the augmented reference transcripts in Table 1 will be: "*I/C have/C the/C aphasia/C*" (P1), "*have/C aphasia/pn*" (P2), and "*I/C have/C aphasia/pn*" (P3). Given an augmented hypothesized transcript, its Word Error Rate (WER) compared to the reference captures both transcription and word labeling errors. We henceforth refer to this metric as augmented WER (AWER).

Utterance-Level Metric: aphasic speech is known to be difficult to recognize [21], thus achieving good AWER may be challenging. Instead of providing detailed word-level predictions, the system can simply output whether or not a given utterance contains paraphasias, i.e., a binary prediction problem. We again adopt average F1 as the evaluation metric.

Speaker-Level Metric: using the same reasoning, the system can be modified to estimate the rate of paraphasia pro-

duction for a given speaker, which helps indicate anomia severity. We evaluate this task by computing the Pearson correlation coefficient (r) between the predicted and actual paraphasia occurrence rate per minute for all speakers in the dataset.

5. Methods

5.1. Acoustic Modeling

Given the small size of the dataset, we adopt an out-of-domain training approach, motivated by previous work in disordered speech recognition [21, 22]. We first train an acoustic model on the core AphasiaBank dataset, which contains approximately 126 hours of spontaneous speech elicited through the AphasiaBank protocol. We then adapt (retrain) the model on each training fold in our dataset. We refer to these two models as the out-of-domain (OOD) and in-domain (ID) models, respectively.

We utilize a multi-task deep bidirectional long-short term memory recurrent neural network (DBLSTM-RNN) to predict both the correct senone and monophone labels for each frame. DBLSTM-RNN acoustic models have been shown to achieve state-of-the-art results on various ASR benchmarks [23–25], while training on senones and monophones jointly is known to improve performance [26–28]. In addition, the monophone output of the network represents a distribution over phonemes, also referred to as phoneme posteriorgrams. They can be viewed as a compact representation of each speech frame.

Input Features: we use Kaldi [29] to extract 40-dimensional log Mel filterbank coefficients, using a 25ms window and 10ms frame shift. We perform per-speaker z-normalization and augment each feature frame with five left and right neighbors, resulting in 440 dimensions per frame.

Model Architecture: our multi-task DBLSTM-RNN consists of four hidden BLSTM layers, each with 1200 units (600 for forward, 600 for backward). The senone and monophone output layers contain 4550 and 46 units, respectively.

OOD Training: we train the network with the Adam optimizer [30], full Backpropagation Through Time, Cross Entropy (CE) loss, 0.4 dropout, and an initial learning rate of 0.001. We perform early stopping using the development frame error rate and an exponential-decay learning schedule [21].

ID Adaptation: we adapt the OOD network to the smaller training set using the same strategies as in OOD training, with two modifications. Firstly, we modify the loss function to also minimize the Kullback-Leibler divergence (KLD) between the ID and OOD model outputs. This has been shown to be an effective regularization technique [31]. Secondly, we employ the step-decay learning schedule [21] with a 0.00005 minimum learning rate. We select the KLD weight (0.25 or 0.5) and dropout rate (0.4 or 0.6) based on the development frame error.

5.2. Feature Extraction

The ID acoustic model obtained from the previous step can be used to detect word and phone boundaries via forced alignment with the target transcripts. In addition, the phoneme posteriorgrams produced by the model provide a compact representation of word and phone segments. Given this information, our objective is to extract features for each word that can help separate phonemic/neologistic paraphasias from correct words. Our features can be grouped into the following sets.

Goodness of Pronunciation (GOP): GOP is a widely used metric for assessing pronunciation, first proposed by Witt and Young [32]. It has also been used successfully in our previous work to estimate aphasic speech quality [10, 12]. GOP involves calculating the difference between the average acoustic log-likelihood of a force-aligned word-level segment and that of an unconstrained phone loop. The closer this number is to 0, the more likely that the pronunciation of this word is correct. We extract the GOP as well as the raw forced alignment score for each word. All calculations are performed on our DBLSTM-RNN's phoneme posteriorgram output.

Phone Edit Distance (DIST): both phonemic and neologistic paraphasias involve deviations between the spoken and correct phone sequences. The spoken phone sequence can be estimated from an unconstrained phone loop over the word segment, and the correct phone sequence can be obtained from forced alignment results on the target transcript. For each sequence pair, we extract the raw edit distance, edit distance normalized by alignment length, as well as the number of insertions, deletions, and substitutions normalized by alignment length.

Dynamic Time Warping (DTW): the underlying assumption behind these features is that the phoneme posteriorgrams of phonemic and neologistic paraphasias are different from those of correct words. Given a *candidate* word, we can find *references* of this word in the ID training set that are marked as correctly produced, along with their phoneme posteriorgrams. Following Lee et al. [15–17], we compare posteriorgram pairs using DTW, where the distance between two frames c_i and r_j is defined as their inner product distance:

$$D(c_i, r_j) = -\log(c_i \cdot r_j) \tag{1}$$

We extract the following features for each candidate-reference posteriorgram pair: raw DTW distance, DTW distance normalized by aligned path length, and length of the longest horizontal/vertical aligned segment normalized by aligned path length. We extract the mean, median, lower and upper quartile, and standard deviation of each feature group to produce word-level features. We extract a similar set of features for all candidate-reference phone pairs within the word, given that they might provide complementary information. If a candidate word has fewer than three references, we use the average features of all correct words in the training set. This accounts for 6–7% of all candidate words across the 12 training folds.

Duration Measures (DUR): these features are also inspired by Lee et al. [15–17] and extracted similarly to DTW. However, we compare the differences in durations instead of posteriorgrams. For each candidate-reference word/phone pair, we extract the ratio between their durations, and the difference in duration normalized by the candidate and reference durations.

As a final post-processing step, we z-normalize all features using statistics computed from correct words in the training set.

5.3. Automatic Transcription

Automatic transcription of test utterances can be performed by combining our DBLSTM-RNN acoustic model with a language model (LM) for decoding. We experiment with two LM types in this work. Firstly, we use a trigram model estimated on the ID training and development set. We refer to this model as the *global* LM. Secondly, we take advantage of the fact that utterances in the dataset are limited to four predefined scripts with different vocabulary and sentence structures. Therefore, it may be beneficial to use a trigram model estimated on the portion of the training and development set corresponding to the same script as the current test utterance. We refer to this as the *task-specific* LM. In both cases, the LM weight and word insertion penalty are chosen based on the development WER.

Table 3: *WER with different language and acoustic model types.*

	Global LM	**Task LM**
OOD AM	65.82	60.97
ID AM	47.68	**45.11**

Table 3 lists the test WERs for different acoustic and language model combinations. As expected, the best performance is obtained with an in-domain acoustic model and task-specific language model. We will use the hypothesized transcripts produced by this system for all relevant experiments.

5.4. Paraphasia Classification

We consider three standard classification algorithms implemented in scikit-learn [33], decision trees (DT), logistic regression (LR), and support vector machines (SVM). Hyperparameters are selected based on the average F1 score on the development set. Test predictions are aggregated across all 12 training folds. We report the test performance on this aggregated test set.

6. Results and Discussion

6.1. Paraphasia Detection With Known Transcripts

Paraphasia classification results from known transcripts using different feature sets and labeling schemes, measured in average F1 scores, are summarized in Table 4. We show results from the classifier that yields the best overall test performance.

All of our systems are able to outperform the naïve baseline, demonstrating that it is feasible to automatically separate correctly produced words and phonemic/neologistic paraphasias. In particular, neologistic paraphasias (*C–n*) are easier to detect than phonemic paraphasias (*C–p*) under this problem setup. This is consistent with the clinical definitions of these two paraphasia types. Because neologistic paraphasias, by definition, involve more deviations from the sounds in the target word, they are better characterized by our proposed features.

In all three labeling schemes (*C–pn*, *C–p*, and *C–n*), the best performance is obtained by using all features, with DTW generating the best individual results. This demonstrates the utility of the phoneme posteriorgram representation produced by our multi-task DBLSTM-RNN acoustic model. A potential method to further exploit phoneme posteriorgrams is to use them as features in whole-word acoustic modeling, which may lead to better discrimination than template matching techniques. GOP features traditionally perform favorably compared to DTW for mispronunciation detection [16], but not so in our work. A possible way to improve GOP performance in this task is to extract phone-level GOP scores alongside word-level features. Likewise, duration-based features (DUR) may benefit from established measures in rhythm analysis, such as Pairwise Variability Error [34]. We will explore these ideas in future work.

Finally, we observe that different feature sets benefit from different classification algorithms. Logistic regression and SVM work well with primarily continuous features such as GOP, DTW, and DUR. By contrast, decision tree yields better performance on DIST, whose features are largely discrete.

6.2. Paraphasia Detection Without Known Transcripts

We are interested in how the best (bolded) models in Table 4 perform when target transcripts for test utterances are generated automatically with ASR. Table 5 lists the results at the word, utterance, and speaker level, as described in Section 4.2.

Table 4: *Paraphasia detection results with known target transcripts, measured in average F1. The best performing classifiers are indicated in parentheses.*

	C–pn	**C–p**	**C–n**
Baseline	.442	.461	.484
GOP	.615 (SVM)	.560 (LR)	.590 (SVM)
DIST	.619 (DT)	.556 (DT)	.662 (DT)
DTW	.699 (SVM)	.611 (LR)	.746 (LR)
DUR	.628 (LR)	.556 (DT)	.652 (LR)
All Feats.	**.704 (LR)**	**.632 (LR)**	**.761 (LR)**

Table 5: *Paraphasia detection results without known target transcripts. Naïve baseline performance is in parentheses.*

	C–pn	**C–p**	**C–n**
Word [AWER]	53.46 (53.39)	54.18 (51.48)	47.84 (47.18)
Utterance [Avg. F1]	.594 (.412)	.611 (.373)	.604 (.404)
Speaker [r]	.479 (N/A)	.749* (N/A)	.057 (N/A)

*statistically significant ($p \approx 0.005$, 2-tailed test)

For word-level, the goal of the system is to simultaneously recognize and label each word accurately. However, our system is unable to outperform the naïve baseline in terms of AWER. As previously discussed, this is challenging because aphasic speech poses significant problems for ASR, and it is difficult to obtain reliable word-level predictions without accurate target transcripts. This suggests that aphasic speech ASR performance must be improved before paraphasias can be detected reliably at the word level without known transcripts.

Meanwhile, utterance-level results, which involve detecting the presence of paraphasias in an utterance, appear more promising. Our system outperforms the naïve baseline in all three classification schemes, suggesting that although word-level predictions may be unreliable, meaningful clinically-relevant information can still be extracted at a coarser level of analysis.

We also observe positive results for estimating the paraphasia production frequency of a particular speaker, which can be tied to anomia severity. Specifically, we obtain a statistically significant Pearson correlation coefficient of 0.749 ($p \approx 0.005$, 2-tailed test) for estimating the rate of phonemic paraphasia production. However, there is virtually no correlation for neologistic paraphasias. We hypothesize that while neologistic paraphasias are easy to classify from known transcripts, they are difficult to detect in a free-form setting because our ASR system fails to recognize them. This again calls for further improvement in aphasic speech recognition.

7. Conclusion and Future Work

In this paper, we presented the first study on detecting phonemic and neologistic paraphasias automatically from aphasic speech, utilizing techniques from ASR and mispronunciation detection. We demonstrated the feasibility of detecting paraphasias from known target transcripts. We showed the utility of utterance- and speaker-level analysis when target transcripts are generated automatically with ASR. For future work, we will investigate additional feature extraction methods, experiment with ways to further improve and analyze aphasic speech recognition performance, and tackle semantic paraphasia detection.

8. References

[1] S. K. Bhogal, R. W. Teasell, N. C. Foley, and M. R. Speechley, "Rehabilitation of aphasia: more is better," *Topics in Stroke Rehabilitation*, vol. 10, no. 2, pp. 66–76, 2003.

[2] N. A. Association, "Aphasia," http://www.aphasia.org/, 2016, accessed: 2016-11-12.

[3] N. Simons-Mackie, A. Raymer, E. Armstrong, A. Holland, and L. Cherney, "Communication partner training in aphasia: A systematic review," *Archives of Physical Medicine and Rehabilitation*, vol. 91, no. 12, pp. 1814–1837, December 2010.

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

[5] L. Nickels, "Therapy for naming disorders: Revisiting, revising, and reviewing," *Aphasiology*, vol. 16, no. 10-11, pp. 935–979, 2002.

[6] N. Friedmann, M. Biran, and D. Dotan, *Lexical retrieval and its breakdown in aphasia and developmental language impairment*, ser. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2013.

[7] K. Makin, B. McDonald, L. Nickels, C. Taylor, and M. Moses, "The facilitation of word production in aphasia: What can it do for the clinician," *Acquiring Knowledge in Speech, Language and Hearing*, vol. 6, no. 90-92, 2004.

[8] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. G. Tempini, and J. Ogar, "Learning Diagnostic Models Using Speech and Language Measures," in *Proc of the 30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, Canada, 2008.

[9] K. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes," in *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013.

[10] 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.

[11] 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 International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.

[12] D. Le, K. Licata, C. Persad, and E. Mower Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE Transactions on Audio, Speech, and Language*, vol. 24, pp. 2187–2199, 2016.

[13] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[14] 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.

[15] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2012, pp. 382–387.

[16] A. Lee, Y. Zhang, and J. R. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 8227–8231.

[17] A. Lee and J. R. Glass, "Pronunciation assessment via a comparison-based system," in *ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, Grenoble, France, 2013, pp. 122–126.

[18] M. M. Forbes, D. Fromm, and B. MacWhinney, "Aphasiabank: A resource for clinicians," in *Seminars in Speech and Language*, vol. 33, no. 3. NIH Public Access, 2012, p. 217.

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

[20] AphasiaBank, *Error Coding*, Accessed: 2015-11-13. [Online]. Available: http://talkbank.org/APhasiaBank/transcribe/errors.doc

[21] D. Le and E. Mower Provost, "Improving Automatic Recognition of Aphasic Speech with AphasiaBank," in *Interspeech*, San Francisco, USA, 2016.

[22] 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 International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013.

[23] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, Vancouver, BC, Canada, 2013.

[24] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*, Singapore, 2014.

[25] H. Sak, A. W. Senior, K. Rao, and F. Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," *CoRR*, vol. abs/1507.06947, 2015.

[26] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *ICASSP*, Vancouver, BC, Canada, 2013.

[27] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," in *ICASSP*, Brisbane, Australia, 2015.

[28] P. Bell and S. Renals, "Complementary tasks for context-dependent deep neural network acoustic models," in *INTERSPEECH*, Dresden, Germany, 2015.

[29] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[31] D. Yu, K. Yao, H. Su, G. Li, 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, Canada, May 2013.

[32] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 23, pp. 95 – 108, 2000.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing Suprasegmental English Through Parroting," in *Proc. of Speech Prosody*, Chicago, IL, USA, 2010.