Unsupervised Paraphasia Classification in Aphasic Speech

Sharan Pai[‡†], **Nikhil Sachdeva**[†], **Prince Sachdeva**[†], **Rajiv Ratn Shah**[†]
[‡] Department. of Mathematics
[†] Department of Computer Science
IIIT Delhi, India
{sharan16266, nikhil16061, prince17080, rajivratn}@iiitd.ac.in

Abstract

Aphasia is a speech and language disorder that results from brain damage, often characterized by word retrieval deficit (anomia) resulting in naming errors (paraphasia). Automatic paraphasia detection has many benefits for both treatment and diagnosis of Aphasia and its type. But supervised learning methods cant be utilized adequately as there is a lack of aphasic speech data. In this paper, we describe our novel unsupervised method, which can be implemented without the need for labeled paraphasia data. Our evaluations show that our method outperforms previous work based on supervised learning and transfer learning approaches for English. We demonstrate the utility of our method as an essential first step in developing augmentative and alternative communication (AAC) devices for patients suffering from aphasia in any language.

1 Introduction

Aphasia is a speech and language disorder commonly acquired by brain damage resulting from a stroke (Bhogal et al., 2003). Many people around the world suffer from Aphasia as there are at least 2 million patients in USA and 250,000 in Great Britain (National Aphasia Association, 2019).

Anomia, the difficulty in spoken word retrieval, is a common symptom in Aphasic speech (Laine and Martin, 2013). A majority of persons with aphasia (PWA) suffer from varying degrees of anomia (Nickels, 2002). Anomia further results in various types of Paraphasia (naming errors) which impedes the PWA's ability to carry out meaningful conversation leading to loneliness and social anxiety (Beeke et al., 2013).

There are three common types of paraphasia which occur in aphasic speech, namely *semantic, phonemic* and *neologistic* (Laine and Martin, 2013; Goodglass and Kaplan, 1972). In *semantic paraphasia*, the PWA substitutes a semantically similar word *eg.* (substituting *elbow* with *knee*). In *phonemic paraphasia*, there are various sub types involving the type of phoneme substitution such as, substituting *bat* with *lat*, inserting or deleting a phoneme (*drake* as *dake*) or phoneme movements (*candle* with *cancle*). Lastly, in *neologistic paraphasia*, the target word is substituted with a non-word (*harmonica* with *parokada*). Detecting and classifying the type of paraphasia is useful to determine the type of aphasia and which treatment to prescribe (Nickels, 2002; Friedmann et al., 2013).

Aphasia TalkBank (MacWhinney, 2007), is a large scale multi-modal online database of aphasic speech data. It contains aphasic speech data for many languages such as English, French *etc* which is primarily used by clinical researchers to study aphasia (Forbes et al., 2012). While the amount of data is sufficient for clinical researchers, there is a lack of data to implement supervised learning methods. This is true not only for a well researched language like English, but also for low-resource [1] languages like Greek, Spanish *etc*.

To counter the lack of data and to extend the proposed method for low-resource languages too, we investigate an unsupervised approach. We first consider large and available speech corpuses such as LibriSpeech (Panayotov et al., 2015) to create speech embeddings of individual words similar to (Chung et al., 2016). We then perform soft clustering using HDBSCAN on these embeddings, and classify each word by using simple rules with a cutoff hyperparameter. The whole method is end-to-end unsupervised and can be applied to any language.

In our evaluations section, we demonstrate the efficacy of our method over a naive baseline and the transfer learning method used by (Le et al., 2017)

[1] we define low-resource *wrt* amount of available aphasic speech data

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 13–19
July 5 - July 10, 2020.

for English. We hope that such an unsupervised method allows for development of AAC devices improving daily life of not only English-speaking PWA's but also PWA's in other languages.

2 Related Work

Recently, researchers have demonstrated the use of machine learning methods not only to diagnose the type of aphasia but also to rehabilitate and treat PWA's. Mainly focusing on obtaining a medical diagnosis, (Fraser et al., 2013) applied feature selection using a transcript and low-level acoustic features to classify between two sub-types of primary progressive aphasia. Likewise, (Peintner et al., 2008) used speech and language features to classify between three broad types of frontotemporal lobar degeneration, including progressive nonfluent aphasia. Further, given speech samples of PWA's, (Le et al., 2014; Le and Mower Provost, 2015; Le et al., 2016) proposed approaches for predicting the utterance-level pronunciation and prosody scores. (Abad et al., 2012, 2013) aimed to tackle the contextually similar problem through keyword spotting. It recognized target words from phrases spoken by the PWA but disregarded fine-grained word-level errors such as paraphasias.

Deep learning methods to detect paraphasia was first demonstrated in (Le et al., 2017). It worked around the notion of mispronunciation detection, adopting the methods of (Lee et al., 2013; Lee and Glass, 2013), which used Dynamic Time Warping (DTW) features to provide a quantitative comparison of word and phone-level pronunciations between native and non-native speakers. Similarly, (Le et al., 2017) has used DTW and other acoustic features like Phone Edit distance and Goodness of Pronunciation, to distinguish between target transcripts and paraphasias. Consequently, it has also used Automatic Speech Recognition (ASR) techniques to generate the target transcripts from the paraphasias automatically. In the end, all of these proposed methods require target transcripts for their core functioning.

To the best of our knowledge, no existing work provides an unsupervised approach to detecting and classifying paraphasia from aphasic speech. In this paper, we explore a realistic scenario where we have access only to the free form discussion with PWA's.

3 Method

Aphasic speech data can be collected in mainly two ways: as a free form discussion between a PWA and an interviewer or a PWA reading a set of provided scripts. While a PWA reading from scripts is conducive to supervised learning methods, it is rarely the case in real life. Hence, our goal is to perform paraphasia detection and classification in the wild *i.e.* without any target scripts. Another motivation for classification in the wild is the lack of labeled English aphasic speech data. Further, the available speech data has a class imbalance (phonemic and neologistic paraphasias account for 12.0 and 6.4 percent respectively). Low-resource languages such as Hindi, Greek *etc.* have a serious lack of aphasia speech data and almost non-existent labeled speech data. Using transfer-learning approaches similar to (Le et al., 2017), would not allow extending it to such low-resource languages. Hence, it was necessary to investigate unsupervised approaches for paraphasia classification. In this section, we outline our proposed unsupervised method which consists of first creating speech embeddings of non-aphasic speech data and then performing soft clustering to further classify the type of paraphasia detected.

3.1 Speech Embedding

In order to classify phonemic and neologistic paraphasia, capturing phoneme placement in a word is necessary.

Previous work, used features such as Goodness of Pronunciation and Phoneme Edit-Distance to do the same. Hence, we adopt speech embeddings which focus on phoneme pronunciation.

In particular, we use the Audio-Word2Vec embeddings outlined in (Chung et al., 2016) as they have demonstrated good performance in distinguishing utterances that have large (>3) phoneme sequence edit distance and grouping utterances with low phoneme sequence edit distance (0 to 2). These speech embeddings are created in an unsupervised fashion. Each word utterance is passed through a sequence-to-sequence encoder and reconstructed via a decoder. This process preserves the acoustic information in the embedding.

(Chung et al., 2016) further demonstrated that sequential phoneme structure is preserved in the vector space. This property can be exploited using density based clustering, the next step of our proposed method.

Classifying semantic paraphasia requires different approaches which cannot be encompassed in methods used to classify phonemic and neologistic paraphasia and hence is left as future work.

3.2 Probing Tasks

Unsupervised word embeddings can be improved further and geared specifically for aphasic speech, but in order to understand what these embeddings are capturing it is important to probe them. Taking inspiration from (Conneau et al., 2018), we create probing tasks specifically for paraphasia. Probing tasks are simple classification tasks for embeddings. We detail three probing tasks specifically for phonemic and neologistic paraphasia.

1. *Phoneme-Movement*: Phonemic paraphasia is often characterized with phoneme movement, usually involving a shift in the position of one or two phonemes. In this binary classification task, the embeddings are used to determine if a phoneme shift took place or not.

2. *Phoneme-Add/Delete*: The addition or deletion of a phoneme is seen in phonemic paraphasia. We use the generated embeddings to determine if the word utterance has a phoneme addition/deletion or is unchanged.

3. *In-Dictionary*: In this task, we check if the embeddings can classify if the word is in the language's dictionary or not. Neologistic paraphasia occurs when PWA's substitute target words with non-words.

These three probing tasks, while not exhaustive, can be used to determine how well the speech embeddings can perform for paraphasia detection.

3.3 Density based Clustering

As our method is unsupervised, we do not have access to whether each word utterance is a paraphasia (further what type) or not. To classify each utterance, we use techniques similar to anomaly detection.

Firstly, the embeddings generated for each word, represent only non-paraphasia words. This is because the dataset used to create these embeddings consists of only correct words utterances. We cluster these non-paraphasia embeddings into distinct clusters where the members of each cluster are embeddings of the same word. We use individual words as centroids rather than phoneme based centroids. This is because, phoneme based centroid choices such as monophones, senones *etc.* creates a surjective mapping from embeddings to centroids (*eg.* both words *cat* and *hat* contain the same phoneme *ae*, hence both words will be assigned to the same centroid), whereas word based centroids has a bijective mapping.

Secondly, we use HDBSCAN (McInnes et al., 2017) to perform density based clustering as it allows for cluster densities of varying size. The two most influential parameters namely, minimum cluster size and minimum samples are chosen so as to produce number of clusters equal to the vocabulary size of the dataset.

Lastly, we exploit the soft clustering property of HDBSCAN to detect paraphasias. We use simple rule based methods to perform classification. When a word utterance is correct *i.e* it is not a paraphasia, the top 1 cluster probability should be high, as the embedding should have a core distance of 0. Hence if the utterance satisfies top_1 probability $\geq \alpha$ then it is classified as a correct word. We use $\alpha = 0.75$ in our experiments.

Now, if a word utterance is phonemic paraphasia, HDBSCAN returns near similar cluster membership probabilities for 2 to 3 clusters (*eg. lat* will be clustered close to correct words *bat*, *late etc.*)

$$top_1 - top_2 \leq \beta \tag{1}$$

If a word utterance satisfies equation 1 then we can classify it as a phonemic paraphasia. We use $\beta = 0.2$ in our experiments.

For a neologistic paraphasia, the cluster membership probabilities are evenly low, as the word utterance is a non-word and was never seen by HDBSCAN while clustering. Hence, a utterance that satisfies

$$\sum^{k} top_i \leq \gamma$$

is classified as a neologistic paraphasia. In our experiments $k = 5$ and $\gamma = 0.5$

This clustering based method does not violate the unsupervised nature of the proposed goal. Our reasoning is validated by the empirical evaluations performed in further sections.

4 Evaluation

In order to validate the claims made in the previous section, we perform the following evaluations. For a fair comparison, we use the same test dataset used in (Le et al., 2017), and perform further analysis

on our soft clustering approach. In this section, we detail the experimental setup used including the model structure and hyperparameters, the metrics and the baselines used to compare and finally expand on the results of our method.

4.1 Data

We use two speech datasets, one to create word utterance embeddings and perform HDBSCAN clustering and another to test our method.

As detailed in (Chung et al., 2016), we used the LibriSpeech corpus (Panayotov et al., 2015) to create audio-word2vec embeddings. We have used the *train-clean-100* subset to train the Seq2Seq autoencoder and a combination of *dev-clean* and *test-clean* subsets to perform density based soft clustering. MFCC's of 13 feature-coefficient were used as input to the models.

For our test dataset we used speech data from Aphasia TalkBank (MacWhinney, 2007), specifically, the *Scripts* section of the English section. *Scripts* contains recordings of PWA's reading a script, with each word utterance conveniently labeled as [*p:n] and [*n:k] for phonemic and neologistic paraphasia. (Le et al., 2017) uses the *Fridriksson* subset consisting of 12 PWA's reading 4 predefined scripts each, allowing (Le et al., 2017) to use supervised learning to classify paraphasia as they have access to the target word. We used this same subset, for our experiments to remain consistent.

4.2 Analysis

In this section we provide empirical evidence to substantiate our intuition while building our unsupervised method.

4.2.1 Probing Tasks

The three probing tasks are used to determine how well the unsupervised embeddings are performing on specific tasks. We examine three different types of embedding methods. First is the original setup (Chung et al., 2016) utilized, an Sequence-to-Sequence autoencoder with both the RNN Encoder and Decoder consisting of one hidden layer of 100 LSTM units was used. The networks were trained with SGD without momentum with a fixed learning rate of 0.3 and for 500 epochs. Secondly we improve upon the autoencoder architecture by using 2 instead of 1 hidden layer of 100 bidirectional LSTM units. (Chung et al., 2016) noticed that the embeddings favoured phonemes towards the end of the word, this problem is alleviated by using bidirectional LSTM. The networks were trained with Adam with a learning rate of 0.01 and for 500 epochs.

Method	**Ph-Move**	**Ph-Add/Del**	**In-Dict**
Audio-word2vec	68%	**81%**	76%
Bi-LSTM	**73%**	77%	**83%**

Table 1: Performance of embedding generation methods on probing tasks reported as averaged accuracy values.

As seen in table 1, the bi-directional LSTM version of audio-word2vec performs better and hence going further we use this setup for creating word utterance embeddings.

4.2.2 Soft Clustering

We empirically demonstrate that the word embedding clusters behave similar to the format outlined in the Methods section. We use (McInnes et al., 2017) implementation of HDBSCAN in our experiments.

First we report the HDBSCAN cluster membership scores for correct, phonemic and neologistic paraphasias in Table 2. The paraphasia are transcribed in CHAT transcription format.

Word	Top 1	Top 2	Top 3
Correct Words			
weather	.882	.073	.032
hot	.821	.072	.053
rarely	.764	.213	.014
Phonemic Paraphasia			
u@u (to)	.537	.419	.065
duz@u (choose)	.501	.324	.171
fpł@u (spring)	.461	.253	.258
Neologistic Paraphasia			
ziz@u (easy)	.277	.102	.156
muz@u (use)	.196	.162	.153
zt@u (vast)	.234	.142	.077

Table 2: Top k cluster membership probability scores for correct, phonemic and neologistic paraphasia. Correct word for corresponding paraphasia is included in parenthesis

The cluster membership probabilities, align with the choice of cutoff rules used in the Methods section. Phonemic paraphasia is usually assigned a membership score split across two or three clusters.

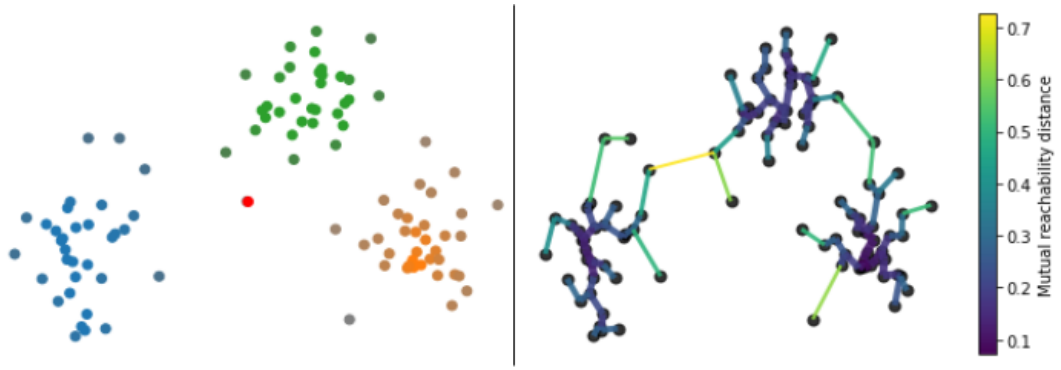

Figure 1: (a) TSNE projections of phonemic paraphasia (in red) with top 1,2 and 3 clusters. The darker the color the higher the cluster membership probability. (b) Minimum spanning tree based on mutual reachability scores

This is true because of the phoneme movement, addition or deletion property leaving rest of the word unaffected, causing confusion so as to which cluster the utterance belongs to. TSNE projection of a sample phonemic paraphasia with its top 1,2 and 3 clusters is displayed in Figure 1. The minimum spanning tree of the clusters also displays the confusion in allocating cluster membership to the phonemic error. Similarly neologistic paraphasia, has uniformly low cluster membership scores, as the utterance is never seen by HDBSCAN as it is a non-word.

A very small set of word utterances ($\leq$ 20) satisfied the condition for both phonemic and neologistic paraphasia *eg.*(top 1,2 and 3 probabilities were .32 .11 and .09) These utterances were classified as phonemic due to the higher value of top 1 than the average neologistic paraphasia.

4.3 Results

As noted by (Le et al., 2017), it is necessary to classify if the word is correct in addition to phonemic or neologistic for future ASR and AAC system development. We report the averaged F1 score on three binary classification schemes, namely *C-pn* (correct vs. phonemic or neologistic), *C-p* (correct vs. phonemic) and *C-n* (correct vs. neologistic)

As baselines, we compare with a naive baseline which classifies all words as correct (the majority class) and the DBLSTM-RNN acoustic model by (Le et al., 2017). It is necessary to note that the DBLSTM-RNN was trained on supervised data using transfer learning methods.

Our method demonstrates results in table 3 which are comparable to the supervised learning method. It outperforms the other baselines for *C-pn* and *C-p*.

Method	**C-pn**	**C-p**	**C-n**
Majority Baseline	.442	.461	.484
(Le et al., 2017)	.704	.632	**.761**
Ours	**.761**	**.683**	.728

Table 3: Paraphasia detection and further classification reported as averaged F1 scores.

While, a tighter set of cutoff hyperparameters can be used to classify the paraphasias as the AAC devices and systems gets further personalized. Our choice of hyperparameters is purposely kept generalized so as to accommodate various PWA speakers. We also believe a better embedding method will allow for better scores even with our general cutoff hyperparameters, especially neologistic paraphasia as it will be further from any word cluster.

5 Conclusion

The work presented in this paper is heavily inspired by (Le et al., 2017), but differs and improves it in the following ways. We provide a completely unsupervised method which outperforms previous work in paraphasia classification and detection. While we maintain that our method can be used for all languages, irrespective of aphasic speech data, due to time constraints we could include only English in our evaluations. We lay the ground-work for paraphasia classification in low-resource languages allowing for development of ASR and AAC systems for not only English-speaking PWA's but also PWA's in developing nations. Our future work will target demonstrating the method on other languages. We also hope to address semantic paraphasia in future work and create, deploy AAC systems building on the method proposed in this paper.

Acknowledgement

Sharan Pai is partly supported by Mantle Labs and MIDAS Lab, IIIT Delhi. Nikhil and Prince Sachdeva are partly supported by MIDAS lab, IIIT Delhi. Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant (ECR/2018/002776) by SERB, Government of India. We would like to thank RAs Srimoyee Chaudhury and Sakshi Labhane for their help in earlier versions of the study.

References

Alberto Abad, Anna Pompili, Angela Costa, and Isabel Trancoso. 2012. Automatic word naming recognition for treatment and assessment of aphasia. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2.

Alberto Abad, Anna Pompili, Angela Costa, Isabel Trancoso, Jos Fonseca, Gabriela Leal, Luisa Farrajota, and Isabel Martins. 2013. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech Language*, 27:12351248.

Suzanne Beeke, Firle Beckley, Wendy Best, Fiona Johnson, Susan Edwards, and Jane Maxim. 2013. Extended turn construction and test question sequences in the conversations of three speakers with agrammatic aphasia. *Clinical linguistics & phonetics*, 27(10-11):784–804.

Sanjit K Bhogal, Robert W Teasell, Norine C Foley, and Mark R Speechley. 2003. Rehabilitation of aphasia: more is better. *Topics in Stroke Rehabilitation*, 10(2):66–76.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *arXiv preprint arXiv:1603.00982*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Margaret M Forbes, Davida Fromm, and Brian MacWhinney. 2012. Aphasiabank: A resource for clinicians. In *Seminars in speech and language*, volume 33, pages 217–222. Thieme Medical Publishers.

Kathleen Fraser, Frank Rudzicz, and Elizabeth Rochon. 2013. Using text and acoustic features to diagnose progressive aphasia and its subtypes.

Naama Friedmann, Michal Biran, and Dror Dotan. 2013. Lexical retrieval and its breakdown in aphasia and developmental language impairment. *The Cambridge handbook of biolinguistics*, pages 350–374.

Harold Goodglass and Edith Kaplan. 1972. *The assessment of aphasia and related disorders*. Lea & Febiger.

Matti Laine and Nadine Martin. 2013. *Anomia: Theoretical and clinical aspects*. Psychology Press.

Duc Le, Keli Licata, Elizabeth Mercado, Carol Persad, and Emily Mower Provost. 2014. Automatic analysis of speech quality for aphasia treatment. pages 4853–4857.

Duc Le, Keli Licata, Carol Persad, and Emily Mower Provost. 2016. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24:1–1.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic paraphasia detection from aphasic speech: A preliminary study. In *Interspeech*, pages 294–298.

Duc Le and Emily Mower Provost. 2015. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation.

Ann Lee and James R. Glass. 2013. Pronunciation assessment via a comparison-based system. In *SLaTE*.

Ann Lee, Yaodong Zhang, and James Glass. 2013. Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. pages 8227–8231.

Brian MacWhinney. 2007. The talkbank project. In *Creating and digitizing language corpora*, pages 163–180. Springer.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

National Aphasia Association. 2019. Aphasia statistics. [Online; accessed 30-January-2020].

Lyndsey Nickels. 2002. Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10-11):935–979.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Bart Peintner, William Jarrold, Dimitra Vergyri, Colleen Richey, Maria Luisa Gorno-Tempini, and

Jennifer Ogar. 2008. Learning diagnostic models using speech and language measures. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2008:4648–51.