On Speech Datasets in Machine Learning for Healthcare

Jekaterina Novikova
Heriot-Watt University

On Speech Datasets in Machine Learning for Healthcare

Jekaterina Novikova and Aparna Balagopalan
Winterlight Labs
{jekaterina, aparna}@winterlightlabs.com

Abstract

Multi-language speech datasets are scarce and often have small sample sizes in the medical domain. We address this problem by employing the cross-linguistic transfer methods and by collecting the large longitudinal dataset of impaired speech. The cross-lingual method demonstrates improvements in Aphasia detection over unilingual baselines, and the early results on the newly collected dataset show the promise to achieve a strong baseline in Alzheimer's disease detection.

Index Terms: cognitive impairment, speech classification, computational linguistics

1 Introduction

Machine learning has great potential in detecting cognitive, mental and functional health disorders from speech, as acoustic properties of speech and corresponding patterns in language are modified by a wide variety of health-related effects [1, 7]. However, machine learning models strongly rely on availability of datasets of an appropriate quality and size. This is particularly a problem in the Machine Learning for Healthcare domain, where standards for state-of-art are high, due to the high cost of errors, as well as for ensuring fair and unbiased decisions for all individuals [10]. A common issue in healthcare is that large datasets developed in commercial settings are not accessible to academia, while smaller clinical datasets that are considered standard in academic research often do not permit commercial usage. This leads to both difficulties in academic machine learning research and restrictions in deploying validated models in commercial applications [5].

Various solutions have been proposed for mitigating this problem in the domain of cognitive impairment detection. Few examples of such methods include: creating novel sources of data [3], developing semi-supervised algorithms to utilize a large amount of unlabeled data for Alzheimer's disease detection [12, 4], using pre-trained embeddings to enrich NN-models on smaller speech datasets for impairment detection [11], domain adaptation from low-resource to resource-rich domains [8]. In our work, we address the problem of availability of speech datasets in two ways: 1) we employ the cross-linguistic transfer methods, and 2) we are collecting the large dataset of impaired speech and developing a strong baseline for it.

2 Cross-Language Speech Impairment Detection

In our recent work [2], we studied cross-linguistic transfer of aphasia detection models trained on English speech from a multi-lingual dataset of healthy and aphasic speech, AphasiaBank [9]. This is investigated to translate developments made in the resource-rich English language to other languages. We adapt linguistic features from speech in different languages to English using domain adaptation with Optimal Transport (OT), using a large *unaligned* multi-lingual dataset. OT involves minimizing the cost of moving samples from source to target Probability Distributions Functions (PDFs), here PDFs of linguistic features in each language (see Figure 1).

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

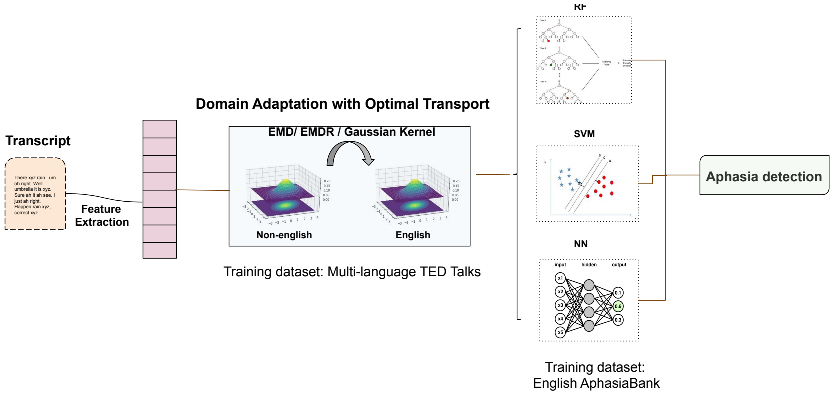
Transcript
There xyz rain...um oh right. Well umbrella it is xyz. Sure ah it ah see. I just ah right. Happen rain xyz, correct xyz.
Feature Extraction
Domain Adaptation with Optimal Transport
EMD/ EMDR / Gaussian Kernel
Non-english
English
Training dataset: Multi-language TED Talks
RF
Majority Vote
Random Forest's decision
SVM
NN
input
hidden
output
x1
x2
x3
x4
x5
0.1
0.6
0.3
Aphasia detection
Training dataset:
English AphasiaBank

Figure 1: Pipeline for processing a speech transcript from a non-English language. Cross-lingual representations are obtained with multiple Optimal Transport algorithms, and Aphasia detection classification performed with different ML models.

We utilize out-of-domain, unpaired, single-speaker, healthy speech data for training multiple OT domain adaptation systems. We learn mappings from other languages to English and detect aphasia from linguistic characteristics of speech, and show that OT domain adaptation improves aphasia detection over unilingual baselines for French (6% increased F1) and Mandarin (5% increased F1). Further, we show that adding aphasic data to the domain adaptation system significantly increases performance for both French and Mandarin, increasing the F1 scores further (10% and 8% increase in F1 scores for French and Mandarin, respectively, over unilingual baselines).

3 Collecting the Dataset for Alzheimer's Disease Detection

Currently, there are no publicly available speech datasets with samples many years before diagnosis of individuals who later developed Alzheimer's Disease (AD). We are collecting such a dataset for early detection of AD from publicly available speech of celebrities to leverage new sources of data. This *Famous People* dataset currently consists of 326 spontaneous speech samples of 30 celebrities (e.g. Ronald Reagan and Woody Allen; 18 healthy and 12 diagnosed with AD) over the period of 1956 to 2018, spanning periods from early adulthood to older age. Audio is being collected online from publicly-available recordings, such as press conferences, interviews etc., and transcribed.

We plan to use the collected data for developing machine learning models that recognize symptoms from linguistic and acoustic characteristics of speech [7]. Since the dataset is currently still relatively small, we refrain from training on it. Instead, we develop models on available normative datasets, and test on our dataset. However, a significant problem we face is that while the *Famous People* dataset consists of unstructured conversations, other datasets consist of participants performing several structured speech tasks (such as describing a picture shown to them, recalling a paragraph from a book etc.). As a result, the distribution of linguistic features would be different, though common patterns might exist. In order to pick up task-independent voice patterns, we use first order meta-learning [6]. With this approach, an NN-based classifier samples from and is trained on a batch of different speech tasks for cognitive impairment detection, and is tested on the Famous People dataset. Early results on *Famous People* show an accuracy of 73.33%, sensitivity of 71.25% and specificity of 75% at subject-level, where the dataset includes samples from 1-5 years before diagnosis and healthy data.

4 Conclusions and Future Work

Availability of datasets of an appropriate quality and size is essential in the ML for Healthcare domain, due to the high cost of errors, as well as to ensure fair decisions for all individuals [10]. In our work, we utilise cross-lingual transfer methods to address this problem, as well as collecting the new longitudinal dataset of impaired speech. While our work shows promising results, the more permanent solution would help the community of ML in Healthcare enormously, involving the agreement upon privacy and de-identification regulations to pathological and normative speech datasets in order to drive innovation.

References

[1] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[2] Aparna Balagopalan, Jekaterina Novikova, Matthew McDermott, Bret Nestor, Tristan Naumann, and Marzyeh Ghassemi. Cross-Language Aphasia Detection using Optimal Transport Domain Adaptation. In *Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2019*, 2019.

[3] Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech. In *Proceedings of the Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*, 2018.

[4] L Bull, K Worden, G Manson, and N Dervilis. Active learning for semi-supervised structural health monitoring. *Journal of Sound and Vibration*, 437:373–388, 2018.

[5] Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. *Gigascience*, 5(1):12, 2016.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[7] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

[8] Bai Li, Yi-Te Hsu, and Frank Rudzicz. Detecting dementia in mandarin chinese using transfer learning from a parallel corpus. *arXiv preprint arXiv:1903.00933*, 2019.

[9] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

[10] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 2018.

[11] Leandro B dos Santos, Edilson A Corrêa Jr, Osvaldo N Oliveira Jr, Diego R Amancio, Letícia L Mansur, and Sandra M Aluísio. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. *arXiv preprint arXiv:1704.08088*, 2017.

[12] Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. Semi-supervised Classification by Reaching Consensus Among Modalities. In *Proceedings of the Interpretability and Robustness in Audio, Speech, and Language (IRASL) Workshop at NeurIPS 2018*, 2018.