Raw speech waveform based classification of patients with ALS, Parkinson’s Disease and healthy controls using CNN-BLSTM

Jhansi Mallela¹, Aravind Illa¹, Yamini Belur², Nalini Atchayaram³, Ravi Yadav³, Prasanta Kumar Ghosh¹

¹SPIRE LAB, EE Dept., IISC, Bangalore, India
²³Dept. of SPA and NEURO., NIMHANS, Bangalore, India

INTERSPEECH 2020
Overview

1. Introduction
2. Proposed approach
3. Data collection
4. Experiments and Results
5. Conclusions
Amyotrophic Lateral Sclerosis (ALS)

- A motor neuron disorder
- Occurs due to gradual degeneration of motor neurons
- Neurons provide a communication link between the brain and voluntary muscles
- Due to the degeneration, there is a loss of muscle control

The ALS Association, "What is ALS?", May 2019.
Symptoms of ALS

- Muscle stiffness
- A hard time in holding items
- Muscle cramps
- Swallowing problems
- Speech difficulties (slurred or slowness)
Parkinson’s Disease (PD)

- A progressive brain disorder
- Occurs when Nerve cells / neurons in the brain get impaired and/or die
- These neurons produce Dopamine
- Reduced Dopamine levels in brain results in movement issues

Parkinson's Foundation, "What is Parkinson’s? "

SPIRE LAB, IISc, Bangalore
Symptoms of PD

- Tremors (hands, arms, legs)
- Stiffness in limbs and trunk
- Slowness in movements
- Difficulty in swallowing and chewing
- Speech difficulties (slurred or slowness)
Life expectancy with ALS or PD

- **ALS affected people**
  - 50% people - 3 or more years
  - 20% people - 5 or more years
  - 10% people - 10 or more years

- **PD affected people**
  - Has increased from 9.4 (1967) to 14.6 years (2016) after diagnosis

---


Introduction

Diagnosis and Treatment for ALS and PD

- Currently no specific tests can confirm of having ALS or PD \(^1\)
- Diagnosis is based on medical history (11 months) and a neurological examination
- No cure for either ALS or PD although there exists some treatment for managing their symptoms \(^1,2,3\)

Main Objective: Automated methods for detection of ALS or PD could reduce diagnosis time.

Plan for the future: To develop a mobile application that helps in early detection, assists neurologist in diagnosis and to follow the progression of the disease using speech as a biomarker.
Challenges

- Identifying speech cues that help in better detection
- Data collection from ALS and PD patients is often tedious making a large corpus development a challenging task - Automated methods require huge amount of data to train a classifier
Introduction

Speech waveforms and spectrograms of ALS, PD, and Healthy controls

Figure: Subjects repeating syllable ”pa”
Previous works studied the performance of various speech tasks in automatic classification between ALS and HC using SVM and DNN\(^1\). As seen, the spectro-temporal characteristics change depending on ALS/HC (and similarly for PD).

Literature Survey

- Performed classification tasks using MFCCs with CNN-LSTMs. Although MFCCs perform well in speech recognition, speaker verification, and several other speech related problems these hand-crafted features (MFCCs) may not be optimal for classifying ALS, PD and healthy controls.

Introduction

Goal of this work

- To study the performance of using raw speech waveform for 3 different classification tasks (ALS vs HC), (PD vs HC), and (ALS vs PD) with 4 different speech tasks
- To study the performance of each classification task in two scenarios:
  - trained and tested in a task-specific manner,
  - trained on data pooled from all tasks, and test on each task separately
- To compare the classification performance using MFCCs and the proposed approach using raw speech
Overview

1. Introduction
2. Proposed approach
3. Data collection
4. Experiments and Results
5. Conclusions
Proposed approach

Learning representations from raw waveform and classification using CNN-BLSTM network

- The work presented here makes use of learning task-specific features from the raw waveform using an end-to-end network.
- Unlike the handcrafted features (MFCCs) that are computed at frame level, representation learning using 1-D CNN might discriminate speech of ALS and PD from those of the healthy subjects in a better way.
- The learned representations are fed to the BLSTM network for classification, since LSTM networks are well-suited to capture the temporal characteristics from the time series data.
Proposed approach

Representation learning

Raw waveform → Speech frames → 1D-CNN Layer → Batch Normalization → Max pooling
Overview

1 Introduction

2 Proposed approach

3 Data collection

4 Experiments and Results

5 Conclusions
Data collection

- Speech recordings collected at the National Institute of Mental Health and Neurosciences (NIMHANS), Bengaluru, India
- Recorder: Zoom H6 X/Y recorder
- Sampling frequency: 44.1 kHz
Dataset

- Number of subjects used in this work:
  - 60 ALS (30 Male, 30 Female)
  - 60 PD (34 Male, 26 Female)
  - 60 healthy control (HC) (30 Male, 30 Female)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Gender</th>
<th>Count</th>
<th>Age Range (Avg) in years</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS</td>
<td>M</td>
<td>30</td>
<td>33 - 76 (58.60)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>38 - 75 (56.02)</td>
</tr>
<tr>
<td>PD</td>
<td>M</td>
<td>34</td>
<td>34 - 78 (58.22)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>26</td>
<td>36 - 74 (56.99)</td>
</tr>
<tr>
<td>HC</td>
<td>M</td>
<td>30</td>
<td>26 - 68 (44.21)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>30</td>
<td>31 - 65 (46.93)</td>
</tr>
</tbody>
</table>

Table: Subject count and age range for each condition - gender pair
Number of subjects used in this work: 60 ALS, 60 PD, and 60 HC

Speech tasks: considering all subjects
- Spontaneous speech (SPON) - 21 hours
- Diadochokinetic rate (DIDK) - 22.42 hours
- Sustained phoneme production (PHON) - 25.84 hours
- Image Description (IMAG) - 25.22 hours
Overview

1. Introduction
2. Proposed approach
3. Data collection
4. Experiments and Results
5. Conclusions
Experimental setup for proposed approach

- Proposed approach: CNN-BLSTM network using raw waveform as input
  - 5 fold cross validation (each fold consists of 12 ALS, 12 PD, 12 HC)
  - Features: Raw speech waveform downsampled to 8kHz (framed with window length of 20ms & shift of 10ms).
  - No. conv. filters or ‘feature maps’ = 256/128/64 (for which the val. loss is minimal)
  - Activation function : ReLUlog (softmax @ output)
  - Kernel : A 120 × 1 (represented by 1D Conv)
  - No. BLSTM layers : 3 (tanh as activation function)
  - No. units for each BLSTM layer : 150
Experiments and Results

Experimental setup for baseline

- Baseline: BLSTM network using MFCCs as input
  - 5 fold cross validation (each fold consists of 12 ALS, 12 PD, 12 HC)
  - Features: MFCC (computed with window length of 20ms & shift of 10ms)
  - No. BLSTM layers: 3 (tanh as activation function)
  - No. units for each BLSTM layer: 150
  - Activation function: softmax @ output
Evaluation metrics

- **Classification accuracy**
- **p-value from sign rank test**\(^1\) : performed using the five folds’ classification accuracies of the baseline and the proposed approach where each test fold is again split into three sub folds since a minimum of five variables are required for this test.

---

Experiments

1. 2 class ALS vs HC
2. 2 class PD vs HC
3. 2 class ALS vs PD
Experiments

Different speech tasks

- **Task-specific**
  - Trained with IMAG
  - Tested with IMAG

- Trained with PHON
  - Tested with PHON

- Trained with DIDK
  - Tested with DIDK

- Trained with SPON
  - Tested with SPON

**Pooled model**
(trained with all speech tasks combinedly and tested task independently)
ALS vs HC classification performance

<table>
<thead>
<tr>
<th>Speech tasks</th>
<th>IMAG</th>
<th>PHON</th>
<th>DIDK</th>
<th>SPON</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task-specific model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>90.14 (4.85)</td>
<td>87.67 (1.10)</td>
<td>93.63 (3.35)</td>
<td>96.51 (3.37)</td>
</tr>
<tr>
<td>Raw speech</td>
<td>97.31 (1.75)</td>
<td>89.47 (3.82)</td>
<td>96.39 (1.94)</td>
<td>96.16 (2.80)</td>
</tr>
<tr>
<td><strong>Pooled model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC</td>
<td>93.08 (1.2)</td>
<td>85.56 (5.0)</td>
<td>95.88 (2.2)</td>
<td>98.69 (1.9)</td>
</tr>
<tr>
<td>Raw speech</td>
<td>98.02 (1.75)</td>
<td>92.34 (1.43)</td>
<td>96.73 (2.45)</td>
<td>97.86 (2.38)</td>
</tr>
</tbody>
</table>

Table: Average classification accuracy (SD in brackets) of the proposed approach and baseline (MFCC). Blue entries indicate the cases where the proposed approach performs significantly (p < 0.05) better than MFCC.
Experiments

1. 2 class ALS vs HC
2. 2 class PD vs HC
3. 2 class ALS vs PD
## Experiments and Results

### PD vs HC classification performance

<table>
<thead>
<tr>
<th>Task-specific model</th>
<th>Speech tasks</th>
<th>IMAG</th>
<th>PHON</th>
<th>DIDK</th>
<th>SPON</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>84.91 (2.13)</td>
<td>65.39 (2.92)</td>
<td>81.89 (5.32)</td>
<td>90.14 (2.92)</td>
<td></td>
</tr>
<tr>
<td>Raw speech</td>
<td>95.01 (2.17)</td>
<td>73.57 (5.05)</td>
<td>89.51 (2.20)</td>
<td>95.71 (4.10)</td>
<td></td>
</tr>
<tr>
<td>Pooled model</td>
<td>Speech tasks</td>
<td>IMAG</td>
<td>PHON</td>
<td>DIDK</td>
<td>SPON</td>
</tr>
<tr>
<td>MFCC</td>
<td>91.51 (2.06)</td>
<td>70.25 (8.93)</td>
<td>89.87 (3.48)</td>
<td>95.34 (2.73)</td>
<td></td>
</tr>
<tr>
<td>Raw speech</td>
<td>97.28 (2.01)</td>
<td>82.18 (9.86)</td>
<td>93.08 (4.92)</td>
<td>97.41 (3.88)</td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Average classification accuracy (SD in brackets) of the proposed approach and baseline (MFCC). Blue entries indicate the cases where the proposed approach performs significantly ($p < 0.05$) better than MFCC.
Experiments

1. 2 class ALS vs HC
2. 2 class PD vs HC
3. 2 class ALS vs PD
### Experiments and Results

#### ALS vs PD classification performance

<table>
<thead>
<tr>
<th>Task-specific model</th>
<th>Speech tasks</th>
<th>IMAG</th>
<th>PHON</th>
<th>DIDK</th>
<th>SPON</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>71.88</td>
<td>72.72</td>
<td>79.73</td>
<td>68.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.08)</td>
<td>(2.91)</td>
<td>(3.90)</td>
<td>(10.06)</td>
<td></td>
</tr>
<tr>
<td>Raw speech</td>
<td>76.87</td>
<td>72.67</td>
<td>76.43</td>
<td>78.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(8.20)</td>
<td>(5.76)</td>
<td>(2.75)</td>
<td>(7.84)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pooled model</th>
<th>Speech tasks</th>
<th>IMAG</th>
<th>PHON</th>
<th>DIDK</th>
<th>SPON</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>74.56</td>
<td>66.54</td>
<td>78.71</td>
<td>73.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.83)</td>
<td>(3.81)</td>
<td>(2.08)</td>
<td>(10.09)</td>
<td></td>
</tr>
<tr>
<td>Raw speech</td>
<td>78.28</td>
<td>69.33</td>
<td>82.09</td>
<td>82.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.32)</td>
<td>(4.67)</td>
<td>(3.16)</td>
<td>(7.84)</td>
<td></td>
</tr>
</tbody>
</table>

Table: Average classification accuracy (SD in brackets) of the proposed approach and baseline (MFCC). Blue entries indicate the cases where the proposed approach performs significantly ($p < 0.05$) better than MFCC.
Illustration of learned CNN filters

Figure: Magnitude response of 256 filters for ALS/PD, PD/HC, and ALS/HC
1-D CNN output

Figure: Illustration of /pa/ sequence spoken by ALS and PD patients using (a) speech waveform, (b) 1-D CNN output, and (c) spectrogram
Overview

1. Introduction
2. Proposed approach
3. Data collection
4. Experiments and Results
5. Conclusions
Conclusions

- Task-specific+Raw wav outperforms Task-specific+MFCC across all classification tasks (ALS vs HC, PD vs HC, and ALS vs PD).
- Pooled+Raw wav in turn performs better than Task-specific+Raw wav or Pooled+MFCC in all classification tasks.
- From the analysis of the learned CNN filter response, it is revealed that the filters are low pass in nature and the center frequencies lie below 800Hz, 500Hz, and 400Hz for ALS/PD, PD/HC, and ALS/HC, respectively.
Key Takeaways

1. A comparison of the proposed approach with baseline acoustic features (MFCC) revealed that the proposed approach significantly performs better than baseline.

2. The proposed approach allows the classifier to extract features related to speech rate cues by enhancing vowels in low-frequency regions and help in accurate classification.
Future work

1. To investigate the proposed approach of classification for severity estimation of ALS and PD patients
2. To investigate the proposed approach incorporating the attention mechanism which enables learning high quality features
Acknowledgement

- Authors thank Ms Renuka and Ms Hima Jyothi for their technical assistance in generating plots.
- Authors thank the Department of Science and Technology, Govt. of India for their support in this work.
THANK YOU

Have Questions/Suggestions?
Write to us at spirelab.ee@iisc.ac.in