Source and Vocal Tract Cues for Speech-based Classification of Patients with Parkinson's Disease and Healthy Subjects

Tanuka Bhattacharjee[1], Jhansi Mallela[1], Yamini Belur[2], Nalini Atchayaram[3], Ravi Yadav[3], Pradeep Reddy[3], Dipanjan Gope[4], Prasanta Kumar Ghosh[1]

[1]SPIRE LAB, EE Dept., [4]ECE Dept., IISC, Bangalore, India
[2]Dept. of SPA and [3]Dept. of NEURO., NIMHANS, Bangalore, India

INTERSPEECH 2021

Overview

1 Introduction

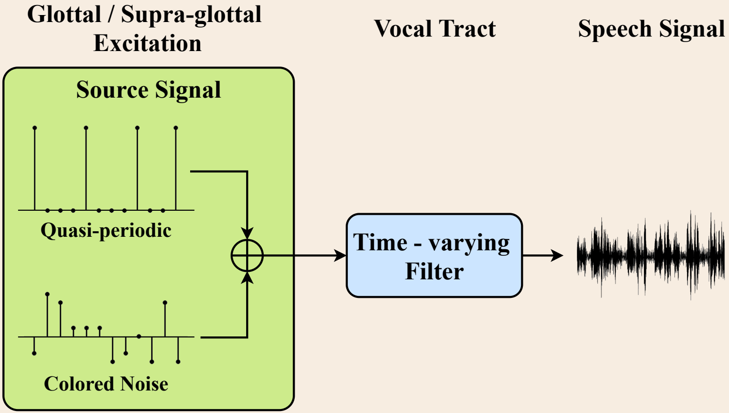
2 Dataset

3 PD vs. HC Classification

4 Experiments and Results

5 Conclusions

Source - Filter Model



G. Fant, Acoustic theory of speech production. Walter de Gruyter, no. 2, 1970.

Parkinson's Disease (PD)

- **Incurable** and **progressive neuro-degenerative** disorder affecting **muscle movements**[1]
  - Dopaminergic neurons degenerate
  - Deficit of neurotransmitter *dopamine* hampers coordinated and smooth muscular control
- Muscles responsible for speech production get affected leading to **dysarthria**[2]
  - Experienced by $\sim 90\%$ of the patients from the early stages of PD[3]

[1] https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/

[2] P. Gómez et al., "Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics," Biomedical Signal Processing and Control, vol. 52, pp. 312–320, 2019.

[3] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," Research and Reviews in Parkinsonism, vol. 9, pp. 9–16, 2019.

Effect of PD on Source and Vocal Tract

- PD impairs both **source** and **vocal tract** attributes of speech
  - **Source Impairment** - monopitch, monoloudness, low voice intensity, and reduced fundamental frequency range[1,2]
  - **Vocal Tract Impairment** - imprecise articulation, voice nasality, and increased acoustic noise[2]

[1] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," Research and Reviews in Parkinsonism, vol. 9, pp. 9–16, 2019.

[2] L. Brabenec et al., "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation," Journal of neural transmission, vol. 124, no. 3, pp. 303–334, 2017.

Our Objective

- To compare the source and vocal tract characteristics in PD patients and healthy subjects
- To analyze how the cues related to these components contribute individually and in combination toward automatic classification of individuals with PD and healthy controls (HC)

Literature Review

SPIRE LAB

Objective	Speech Features	Classifier
PD vs. HC classification	MFCC[1]	CNN-LSTM
	1D-CNN based features from raw speech[2]	BLSTM
	Auto-encoder based features from spectrogram, scalogram[3]	SVM, Softmax classifier
Classification & severity prediction of PD	MFCC, CSD, spectral dynamics, fundamental frequency variation[4]	Random Forest

[1] J. Mallela et al., "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and healthy controls with CNN-LSTM using transfer learning," in ICASSP, IEEE, pp. 6784–6788, 2020.

[2] J. Mallela et al., "Raw speech waveform based classification of patients with ALS, Parkinson's disease and healthy controls using CNN-BLSTM," in INTERSPEECH, pp. 4586–4590, 2020.

[3] B. Karan et al., "Stacked auto-encoder based time-frequency features of speech signal for Parkinson's disease prediction," in AISP, IEEE, pp. 1–4, 2020.

[4] T. Khan et al., "Assessing Parkinson's disease severity using speech analysis in non-native speakers," Computer Speech Language, vol. 61, p. 101047, 2020.

Overview

1 Introduction

2 Dataset

3 PD vs. HC Classification

4 Experiments and Results

5 Conclusions

Dataset Description

- All speech data were collected at **National Institute of Mental Health and Neurosciences (NIMHANS)**, Bangalore, India

Condition	#Male	#Female	#Subjects	Age range (years)
PD	45	14	59	35 - 79
HC	44	16	60	22 - 53
Total	**89**	**30**	**119**	**22 - 79**

- Subjects had **six** different **native languages** - Bengali, Hindi, Kannada, Odiya, Tamil, and Telugu
- PD subjects had dysarthria severity in the range of 0 - 2 as per the UPDRS-III scale[1]

[1] D. J. Gelb et al., "Diagnostic criteria for Parkinson's disease," Archives of Neurology, vol. 56, no. 1, pp. 33–39, 1999.

Dataset Description

Speech Task	Duration (hours)
Image description (IMAG)	~ 12.83
Diadochokinetic Rate (DIDK)	~ 4.65
Spontaneous speech (SPON)	~ 5.62

- IMAG and SPON tasks were performed in the subjects' native language
- **Audio Recorder:** Zoom H6 with XYH-6 stereo microphone capsule
- **Sampling frequency:** 44.1 kHz (downsampled to 16 kHz)

Overview

1 Introduction

2 Dataset

3 PD vs. HC Classification

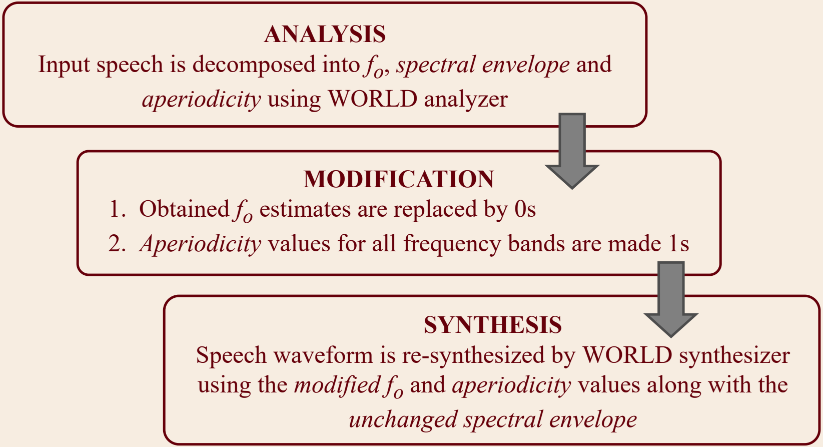
4 Experiments and Results

5 Conclusions

Source and Vocal Tract Features

Speech Component	Feature
Source	Fundamental frequency (f_o)
Vocal Tract	Voicing-removed MFCC (vrMFCC) [MFCC computed after voicing removal]
Source + Vocal Tract	MFCC

Voicing Removal Procedure



ANALYSIS

Input speech is decomposed into f_o, *spectral envelope* and *aperiodicity* using WORLD analyzer

MODIFICATION

1. Obtained f_o estimates are replaced by 0s
2. *Aperiodicity* values for all frequency bands are made 1s

SYNTHESIS

Speech waveform is re-synthesized by WORLD synthesizer using the *modified* f_o and *aperiodicity* values along with the *unchanged spectral envelope*

M. Morise et al., "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE Transactions on Information and Systems, vol. 99, no. 7, pp. 1877–1884, 2016.

Feature Extraction

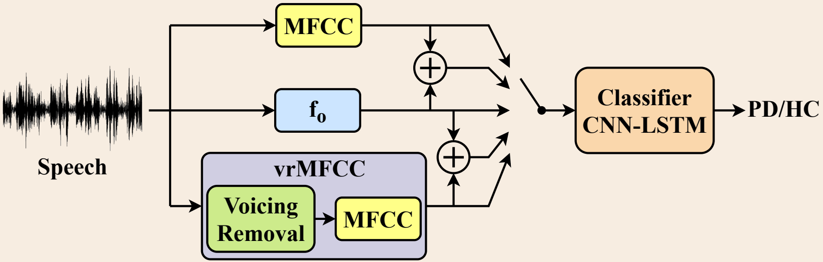
	$\mathbf{f_o}$	**MFCC**
Algorithm/Toolkit	SWIPE[1]	KALDI[2]
Dimension	3 ($1\ f_o + 1\ \Delta f_o + 1\ \Delta^2 f_o$)	39 (13 MFCC $+ 13\ \Delta$MFCC $+ 13\ \Delta^2$MFCC)
Temporal Setting	extracted every 10 ms	20 ms frame length, 10 ms overlap

- The f_o estimates for unvoiced/silence regions are replaced by 0s
- Utterance-level Z-score normalization is applied to each feature dimension independently

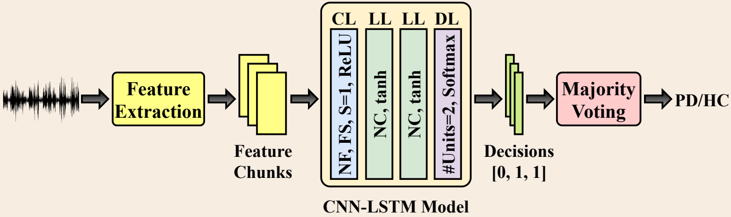
[1] A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” The Journal of the Acoustical Society of America, vol. 124, no. 3, pp. 1638–1652, 2008.

[2] D. Povey et al., “The Kaldi speech recognition toolkit,” in Workshop on automatic speech recognition and understanding, IEEE Signal Processing Society, 2011.

Classification Scheme



Classifier Configuration



CL: 1D-CNN layer
LL: LSTM layer
DL: Dense layer
NF: #filters
FS: Filter size
S: Stride
NC: #cells in LL

Feature Set	NF	FS	NC	#param	FLOPs
$\mathbf{f_o}$	18	20	64	55500	175.48k
MFCC / vrMFCC	5	20	64	54979	174.46k
$\mathbf{f_o}$**+MFCC / **$\mathbf{f_o}$**+vrMFCC**	5	20	64	55279	175.06k

J. Mallela et al., "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and healthy controls with CNN-LSTM using transfer learning," in ICASSP, IEEE, pp. 6784–6788, 2020.

Noise Conditions

- **Noise:**
  - Additive White Gaussian Noise (AWGN)
  - High-Frequency Channel Noise (HF)[1]
  - Pink Noise[1]
  - Babble Noise[1]
- **SNR:**
  - 0, 5, 10 and 20 dB
- **Train-Test Settings:**
  - **Matched:** Noise and SNR of the data used in training and testing the classifier are matched
  - **Mismatched:** Classifier trained with clean data is used to test both clean and noisy test samples

[1] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech communication, vol. 12, no. 3, pp. 247–251, 1993.

Overview

1 Introduction

2 Dataset

3 PD vs. HC Classification

4 Experiments and Results

5 Conclusions

Evaluation Protocol

- **Validation Scheme:**
  - 5-fold cross-validation
    - Each fold contains almost equal number of subjects from PD and HC classes
    - Similar distributions of age, gender, language and dysarthria severity are maintained across folds
- **Evaluation Metrics:**
  - Classification accuracy
  - Wilcoxon signed rank test[1] at 10% significance level

[1] RF Woolson, "Wilcoxon signed-rank test," Wiley encyclopedia of clinical trials, pp. 1–3, 2007.

Source (f_o) or Vocal Tract (vrMFCC)?

Table: Mean classification accuracies in % (SD in bracket); here blue colour indicates superiority at 10% significance level

Feature Set	Speech Task			
	IMAG	DIDK	SPON	Overall
f_o	74.19 (4.67)	75.33 (2.86)	88.12 (4.44)	79.21
vrMFCC	83.17 (3.56)	76.45 (4.31)	83.26 (3.41)	80.96

- Relative contributions of source and vocal tract cues toward PD vs. HC classification vary with the speech tasks at hand

Are They Complementary?

SPIRE LAB

Table: Mean classification accuracies in % (SD in bracket); # indicates that MFCC outperforms vrMFCC at 10% significance level; * and △ indicate that f_o+vrMFCC outperforms f_o & vrMFCC, respectively, at 10% significance level

Feature Set	Speech Task			
	IMAG	DIDK	SPON	Overall
f_o	74.19 (4.67)	75.33 (2.86)	88.12 (4.44)	79.21
MFCC	85.30 (4.92)	81.23 (2.40)#	88.04 (2.84)#	84.86
vrMFCC	83.17 (3.56)	76.45 (4.31)	83.26 (3.41)	80.96
f_o+vrMFCC	84.74 (3.69)*	83.42 (1.29)*△	90.36 (4.03)△	86.17

- Source and vocal tract cues complement each other in all tasks

Does Fusion of f_o and MFCC Help?

Table: Mean classification accuracies in % (SD in bracket); * and △ indicate that f_o+MFCC/vrMFCC outperforms f_o & MFCC/vrMFCC, respectively, at 10% significance level

Feature	Speech Task			
Set	IMAG	DIDK	SPON	Overall
f_o	74.19 (4.67)	75.33 (2.86)	88.12 (4.44)	79.21
MFCC	85.30 (4.92)	81.23 (2.40)	88.04 (2.84)	84.86
f_o+MFCC	88.65 (4.21)*	83.28 (4.09)*	91.91 (1.31)*$^{\triangle}$	87.95
f_o+vrMFCC	84.74 (3.69)*	83.42 (1.29)*$^{\triangle}$	90.36 (4.03)$^{\triangle}$	86.17

- Source information encoded in MFCC and f_o are different and complementary
- PD vs. HC classification accuracy benefits from f_o+MFCC fusion
- f_o+MFCC outperforms f_o+vrMFCC

Effect of Noise: Matched Train - Test

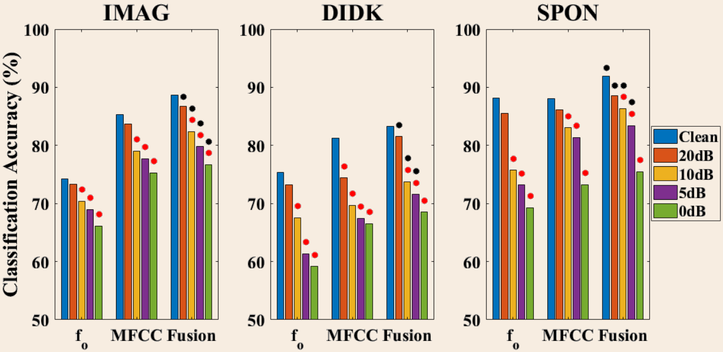


Figure: Mean classification accuracy over AWGN, HF, pink, and babble noise; here ● indicates drop in accuracy w.r.t. clean case at 10% significance level and ● marks the feature set which outperforms the other two at 10% significance level for a particular SNR

- Source and vocal tract features are complementary in the matched train-test noisy conditions

Effect of Noise: Mismatched Train - Test

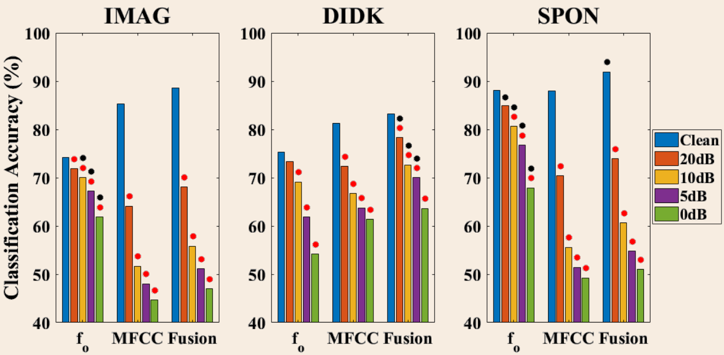


Figure: Mean classification accuracy over AWGN, HF, pink, and babble noise; here ● indicates drop in accuracy w.r.t. clean case at 10% significance level and ● marks the feature set which outperforms the other two at 10% significance level for a particular SNR

- f_o is highly robust against unseen noise and SNR conditions

Overview

SPIRE LAB

1 Introduction

2 Dataset

3 PD vs. HC Classification

4 Experiments and Results

5 Conclusions

Key Takeaways

- Relative merits of source and vocal tract cues vary in different speech tasks
- However, the two components complement each other consistently
- Among all the feature sets considered, f_o+MFCC is found to attain the highest classification accuracy under both clean and matched train-test conditions
- Robustness against unseen noise is predominantly observed in the case of source features encoded in f_o

Future Work

- To perform similar analysis using other source cues like glottal flow
- To assess the dysarthria severity using source and vocal tract cues

Acknowledgement

- 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 @ `spirelab.ee@iisc.ac.in`