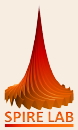
Static and Dynamic Source and Filter Cues for Classification of Amyotrophic Lateral Sclerosis Patients and Healthy Subjects

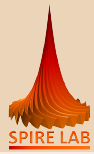
Tanuka Bhattacharjee[1], **Chowdam Venkata Thirumala Kumar**[1], Yamini Belur[2], Atchayaram Nalini[3], Ravi Yadav[3], Prasanta Kumar Ghosh[1]

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

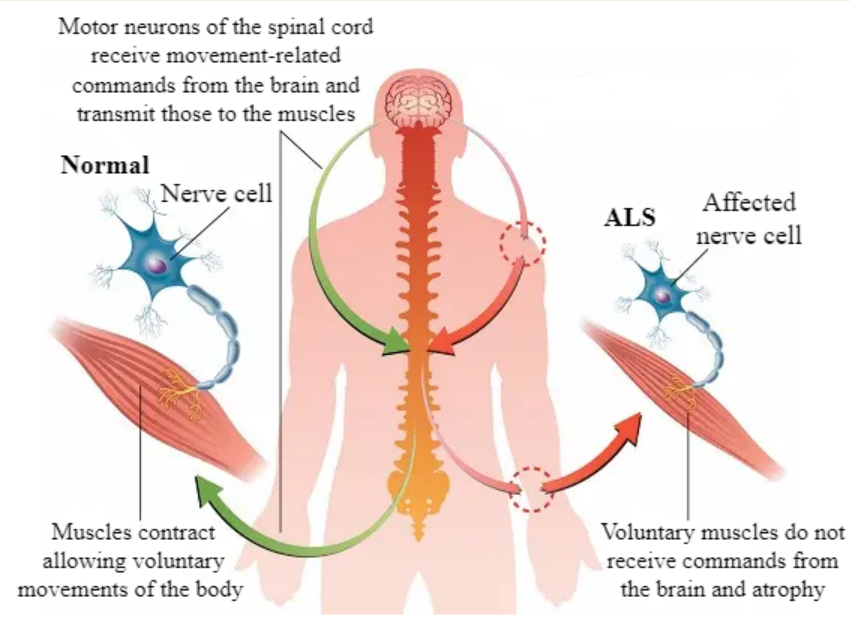


ICASSP 2023

Overview



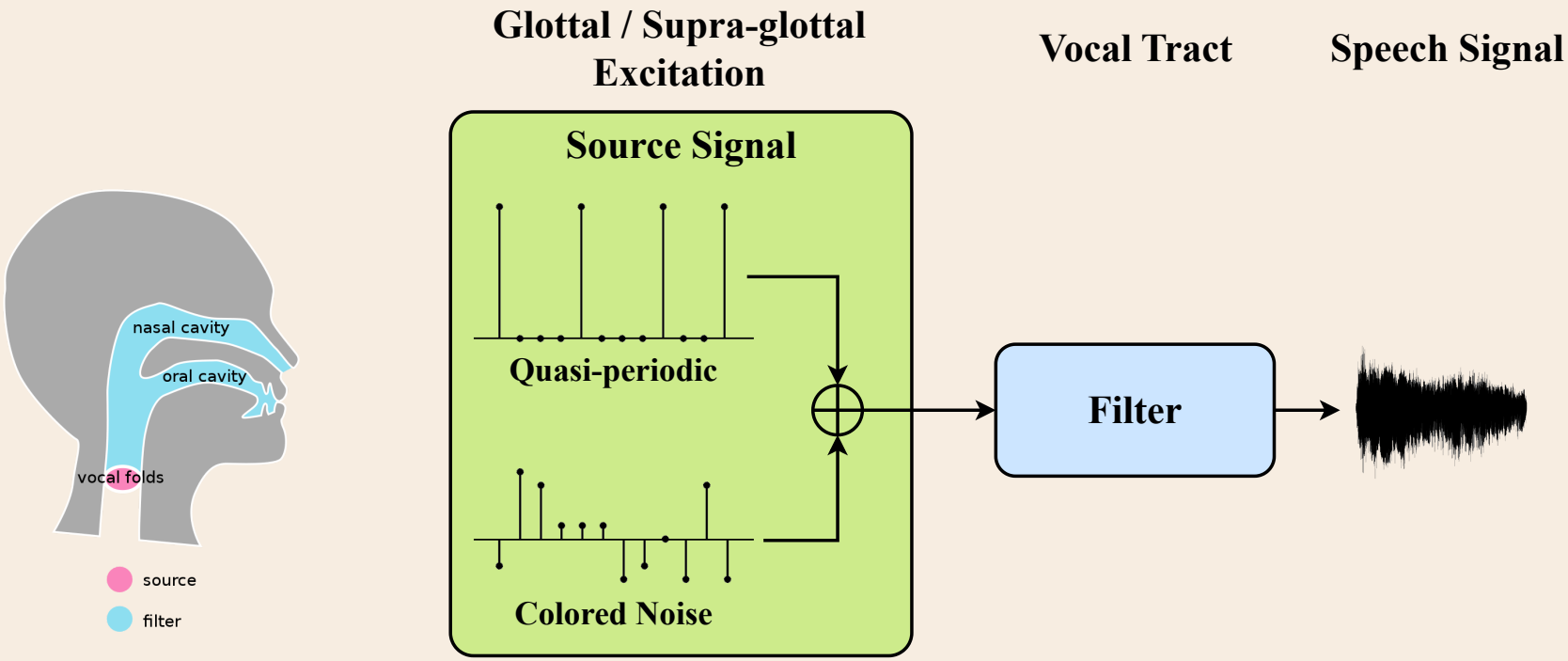
Amyotrophic Lateral Sclerosis (ALS)



- **Incurable** and **progressive neuro-degenerative** disease affecting **muscle movements**[1]
  - Motor neurons degenerate
- Speech musculature get severely affected leading to **Dysarthria**
  - Affects articulation, phonation, prosody, respiration and resonance[2]
  - **Even, elementary Sustained Vowel (SV) utterances get impaired**

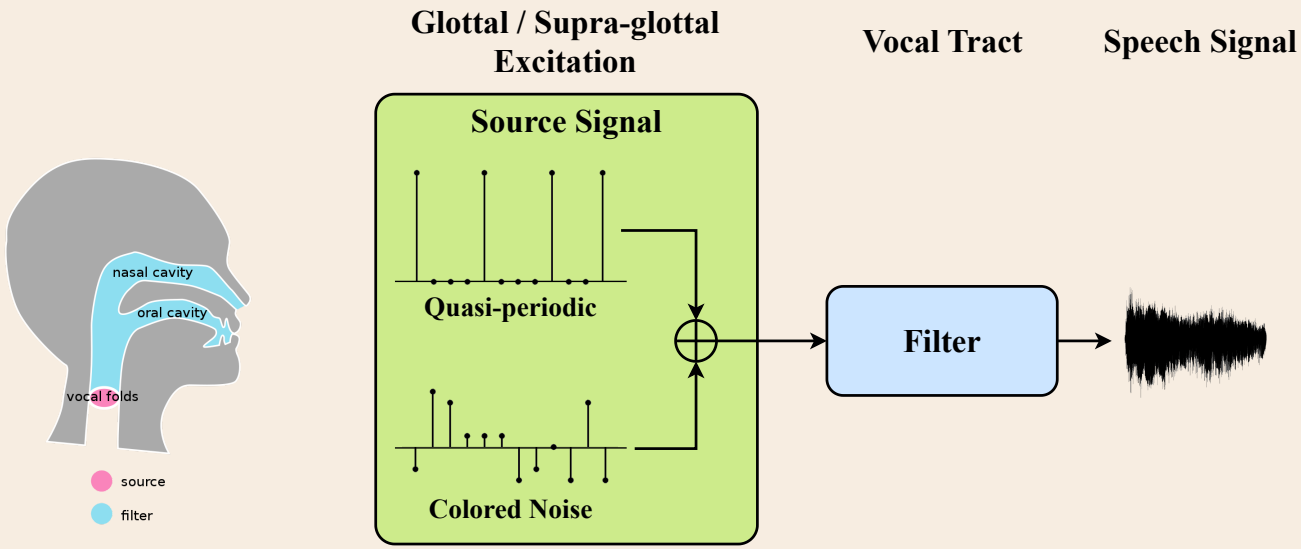
1. https://www.als.org/understanding-als/what-is-als/
2. Lavoisier Leite and Ana Carolina Constantini, "Dysarthria and quality of life in patients with Amyotrophic Lateral Sclerosis," Revista CEFAC, vol. 19, pp. 664–673, 2017.

Vowel Production - Source-Filter Model



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

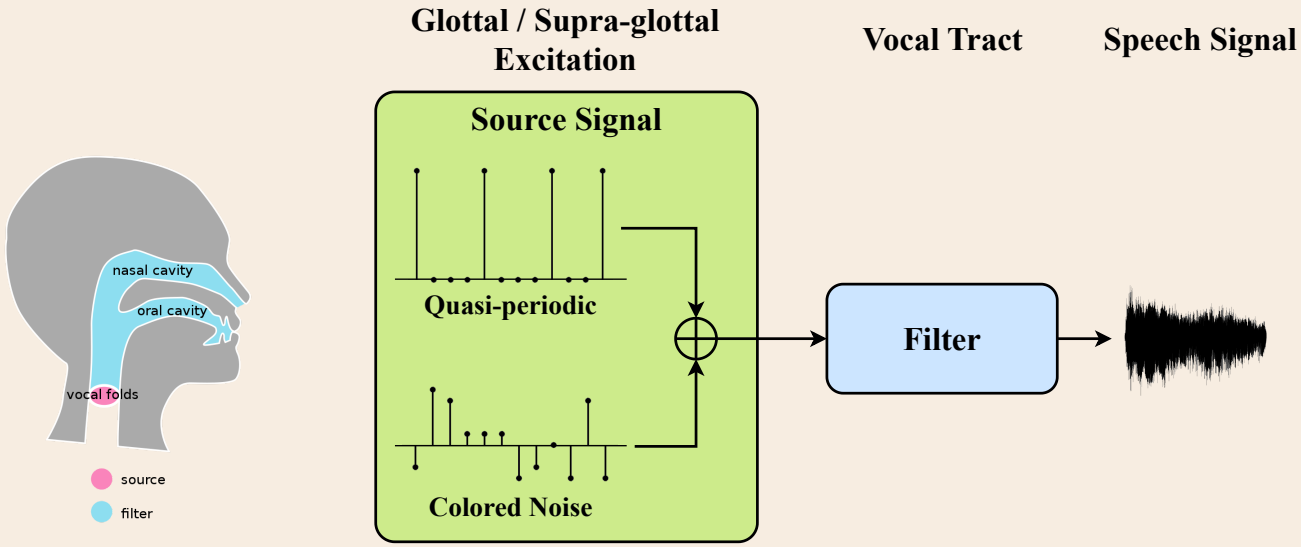
Sustaining a Vowel



Sustained Vowel (SV) production calls for

- achieving vowel-specific source (S) and filter (F) configurations
- uniformly sustaining the configurations for a prolonged duration

Sustaining a Vowel

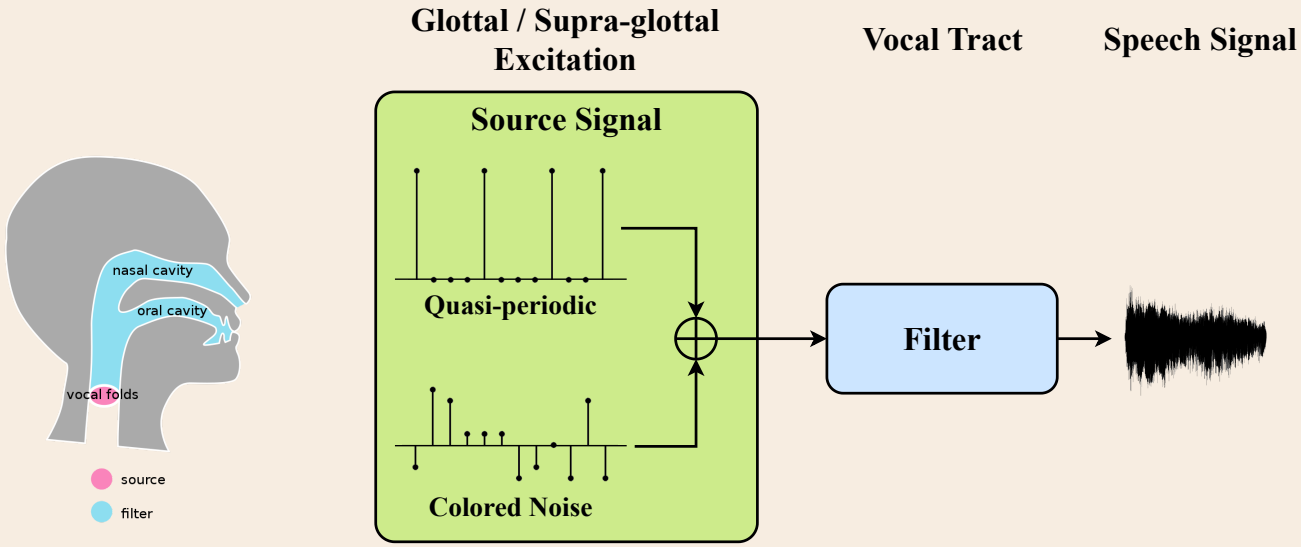


Sustained Vowel (SV) production calls for

- achieving vowel-specific source (S) and filter (F) configurations
- uniformly sustaining the configurations for a prolonged duration

Due to restricted muscular control, ALS patients might face difficulties in accomplishing either/both of these goals.

Sustaining a Vowel



Sustained Vowel (SV) production calls for

- achieving vowel-specific source (S) and filter (F) structures - **Static cues (ST)**
- uniformly sustaining the structures for a prolonged duration - **Dynamic cues (DY)**

Due to restricted muscular control, ALS patients might face difficulties in accomplishing either/both of these goals.

Our Objective

- To **analyze** the relative **discriminative capabilities** of the following cues for SV-based **ALS vs. healthy control (HC) classification**:
  - source-static (S-ST)
  - source-dynamic (S-DY)
  - filter-static (F-ST)
  - filter-dynamic (F-DY)

Literature

Authors [Ref]	Vowels	Features	Classifiers
Suhas et al. [1]	/a/, /i/, /o/, /u/, /æ/	Log-mel spectrogram	2D-CNN
Vashkevich et al. [2]	/a/, /i/	Spectral, noise, perturbation & F_0 contour based features; e.g. MFCC, HNR, jitter, pitch period entropy etc.	LDA
Tena et al. [3]	/a/, /e/, /i/, /o/, /u/	Phonatory-subsystem & time-frequency features	SVM, LDA, RF, logistic regression, dense NN
Mallela et al. [4]	/a/, /i/, /o/, /u/, /æ/	1D-CNN based features from raw speech	BLSTM

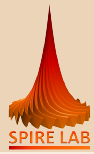
1. BN Suhas et al., "Speech task based automatic classification of ALS and Parkinson's disease and their severity using log mel spectrograms," in SPCOM, IEEE, pp. 1–5, 2020.

2. M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," Biomedical Signal Processing and Control, vol. 65, pp. 102350, 2021.

3. A. Tena et al., "Detecting bulbar involvement in patients with Amyotrophic Lateral Sclerosis based on phonatory and time-frequency features," Sensors, vol. 22, no. 3, pp. 1137, 2022.

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

Overview



Dataset Description

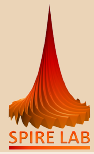
- **Place of data collection:**
  - National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, India
- **Speech task:**
  - Sustained utterances of /a/, /i/, /o/ and /u/
  - 1-3 utterances per vowel per subject

Table: Subject and utterance details

Condition	#M:#F	Age range (years)	#Utterances	Mean (SD) of utterance duration (sec)
ALS	50:30	28 - 77	858	4.05 (2.29)
HC	62:18	22 - 65	842	5.71 (1.98)

- Data were arranged in 5-fold cross-validation setup with disjoint subjects in the 5 groups.

Overview



Choice of Static and Dynamic Cues

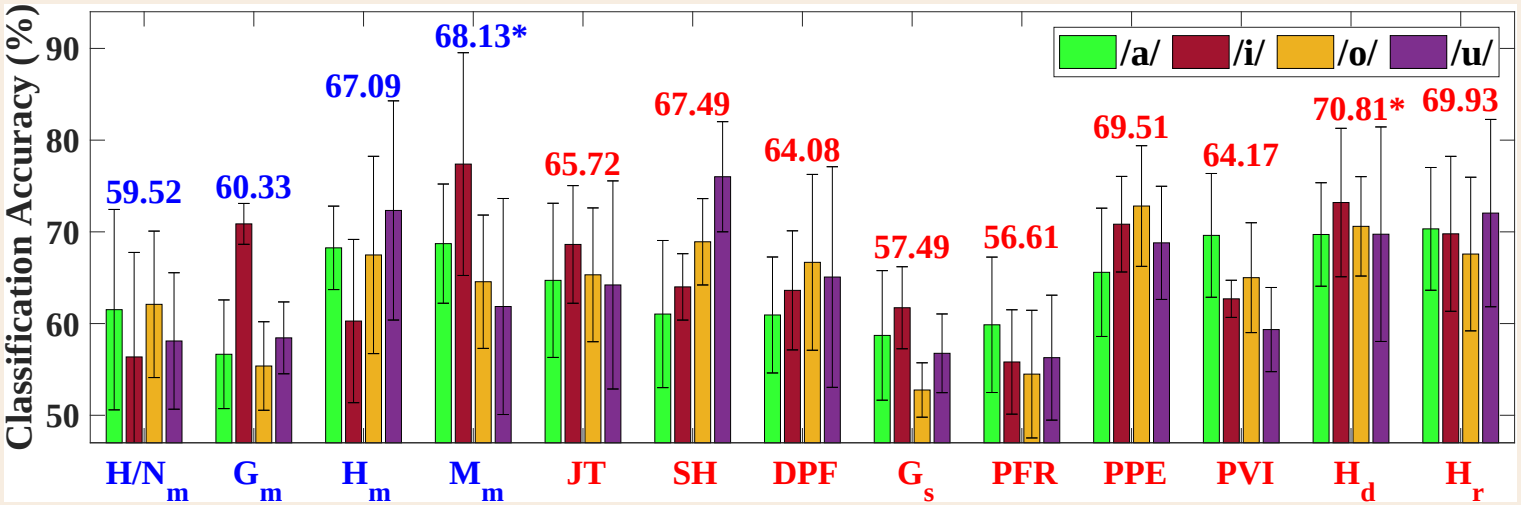


Figure: Mean ALS vs. HC classification accuracies (SD in error bar) obtained using LDA with different ST (blue) and DY (red) cues extracted from complete durations of SVs; accuracies averaged over all vowels are shown on top of each group of bars; * indicates the features having the highest average accuracy over all vowels among each of ST and DY groups

H/N_m: mean harmonic-to-noise ratio, **G_m:** mean glottal-to noise excitation ratio, **H_m:** mean spectral amplitudes over time at the first 8 harmonic frequencies, **M_m:** mean MFCC, **JT**: jitter, **SH**: shimmer, **DPF**: directional perturbation factor, **G_s**: SD of glottal-to noise excitation ratio, **PFR**: phonatory frequency range, **PPE**: pitch period entropy, **PVI**: pathological vibrato index, **H_d:** SD of spectral amplitudes over time at the first 8 harmonic frequencies, **H_r:** inverse of the sum of absolute values of H_m and H_d

- **M_m** and **H_d** perform the best among ST and DY group respectively.
  - selected as the **representative ST and DY cues**

M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," Biomedical Signal Processing and Control, vol. 65, pp. 102350, 2021.

Choice of Static and Dynamic Cues

Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of SVs

Features	Vowels			
	/a/	/i/	/o/	/u/
$\mathbf{M_m}$	68.72 (6.50)	77.39 (12.14)	64.56 (7.27)	61.86 (11.78)
$\mathbf{H_d}$	69.71 (5.64)	73.20 (8.09)	70.60 (5.42)	69.74 (11.70)
$\mathbf{M_m^1}$	62.24 (7.35)	75.75 (10.92)	64.12 (7.41)	58.80 (6.55)
$\mathbf{H_d^1}$	73.92 (3.20)	71.69 (4.50)	75.57 (2.44)	68.49 (3.28)

Here, M_m^1 and H_d^1 refer to M_m and H_d computed from the middle 1sec of an utterance.

- M_m^1 and H_d^1 perform statistically similar (as per Wilcoxon signed-rank test at 1% significance level) to M_m and H_d respectively.
- In most cases, SD of accuracies are lower for M_m^1 and H_d^1 than M_m and H_d respectively.

Comparison with Baseline

Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of SVs as compared to baseline feature sets

Features	Vowels			
	/a/	/i/	/o/	/u/
$\mathbf{M_m^1 + H_d^1}$	70.80 (5.20)	79.37 (9.70)	74.28 (7.29)	71.62 (8.29)
Baseline-64D[a] (from entire utterance)	73.76 (8.36)	81.00 (5.63)	73.22 (6.33)	73.24 (3.28)
Baseline-64D[a] (from middle 1.5 sec)	73.85 (5.09)	80.74 (4.97)	70.81 (9.78)	71.36 (6.87)
MFCC + CNN-LSTM[b]	77.82 (6.12)	68.62 (5.13)	74.19 (4.80)	64.96 (8.87)

- $M_m^1 + H_d^1$ can achieve classification accuracies comparable to the baselines.

a. M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," Biomedical Signal Processing and Control, vol. 65, pp. 102350, 2021.

b. 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.

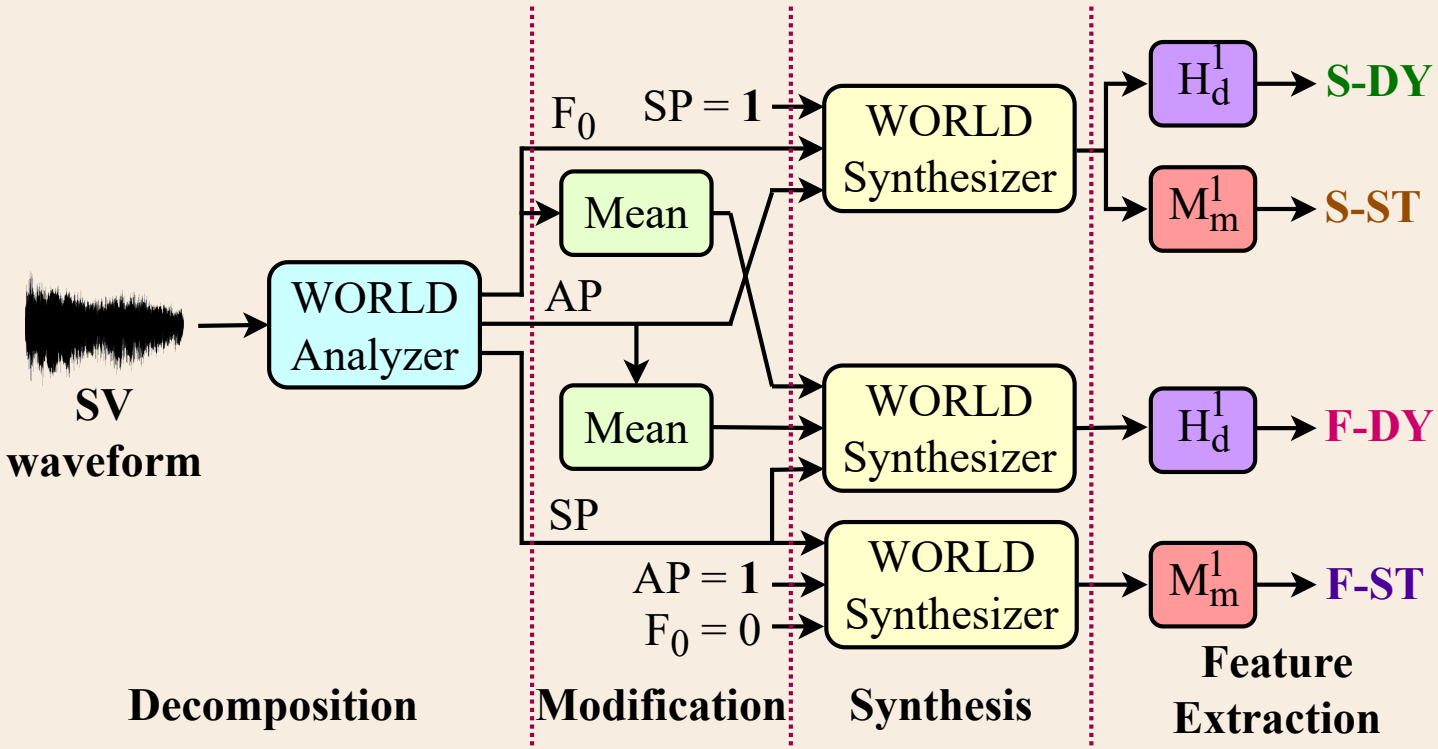
Static vs. Dynamic

Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using M_m^1 and H_d^1; here, * indicates that H_d^1 outperforms M_m^1 as per Wilcoxon signed-rank test at 1% significance level

Features	Vowels			
	/a/	/i/	/o/	/u/
$\mathbf{M_m^1}$	62.24 (7.35)	75.75 (10.92)	64.12 (7.41)	58.80 (6.55)
$\mathbf{H_d^1}$	73.92 (3.20)*	71.69 (4.50)	75.57 (2.44)*	68.49 (3.28)

- For /a/, /o/ and /u/, H_d^1 (DY) outperforms M_m^1 (ST).
  - ALS and HC differ in the extent of variations in the target S-F configuration over the course of an utterance.
- For /i/, M_m^1 (ST) achieves higher average classification accuracy than H_d^1 (DY).
  - Gross S-F configurations differ predominantly between ALS and HC subjects.

Extracting Static & Dynamic Source & Filter Cues



F_0: fundamental frequency, AP: aperiodicity,
SP: spectral envelope, **1**: matrix with all entries as 1

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.

Relative Performance

Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of S and F components of SVs; here # and † indicate respectively that F-ST significantly outperforms S-ST and F-DY significantly outperforms S-DY as per Wilcoxon signed-rank test at 1% significance level

Features	**Vowels**			
	/a/	**/i/**	**/o/**	**/u/**
S-ST	55.27 (2.82)	61.85 (7.83)	56.32 (5.33)	55.82 (8.26)
S-DY	62.11 (2.68)	57.90 (5.86)	60.00 (4.59)	57.18 (5.16)
F-ST	60.25 (6.57)	76.66 (12.90)$^{\#}$	64.27 (6.55)	63.51 (6.60)
F-DY	66.29 (8.43)	68.86 (1.91)†	73.03 (3.49)†	70.27 (5.27)†

- Source cues (ST/DY) are not the primary discriminators.

Relative Performance

Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of S and F components of SVs; here # and † indicate respectively that F-ST significantly outperforms S-ST and F-DY significantly outperforms S-DY as per Wilcoxon signed-rank test at 1% significance level

Features	**Vowels**			
	/a/	**/i/**	**/o/**	**/u/**
S-ST	55.27 (2.82)	61.85 (7.83)	56.32 (5.33)	55.82 (8.26)
S-DY	62.11 (2.68)	57.90 (5.86)	60.00 (4.59)	57.18 (5.16)
F-ST	60.25 (6.57)	76.66 (12.90)$^{\#}$	64.27 (6.55)	63.51 (6.60)
F-DY	66.29 (8.43)	68.86 (1.91)†	73.03 (3.49)†	70.27 (5.27)†

- For /a/, /o/ and /u/, the F-DY attributes contribute the most.
  - Holding the target vocal tract shape for long appears to be the primary challenge for the ALS patients in case of /a/, /o/ and /u/.

Relative Performance

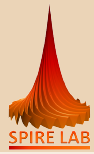
Table: Mean ALS vs. HC classification accuracies in % (SD in bracket) obtained using representative ST and DY cues of S and F components of SVs; here # and † indicate respectively that F-ST significantly outperforms S-ST and F-DY significantly outperforms S-DY as per Wilcoxon signed-rank test at 1% significance level

Features	**Vowels**			
	/a/	**/i/**	**/o/**	**/u/**
S-ST	55.27 (2.82)	61.85 (7.83)	56.32 (5.33)	55.82 (8.26)
S-DY	62.11 (2.68)	57.90 (5.86)	60.00 (4.59)	57.18 (5.16)
F-ST	60.25 (6.57)	76.66 (12.90)$^{\#}$	64.27 (6.55)	63.51 (6.60)
F-DY	66.29 (8.43)	68.86 (1.91)†	73.03 (3.49)†	70.27 (5.27)†

- For /i/, the F-ST cues achieve the highest mean classification accuracy.
  - ALS patients seem to face difficulties in forming the front closed vocal tract structure of /i/, possibly due to the impaired tongue height control.

Overview

Key Takeaways



- Different cues capture predominant discriminative information in case of different vowels.
- F-DY cues achieve the highest mean classification accuracy in case of 3 out of 4 vowels under consideration.
- Achieving the vocal tract configuration involving proximal placement of the tongue and palate, specific to the front close vowel /i/, seems to get difficult for the patients having ALS-induced dysarthria.
- Maintaining a constant vocal tract shape seems to become the primary hurdle in the cases of the other three vowels - /a/, /o/ and /u/.

Future Work

- To combine cues from different vowels for ALS vs. HC classification
- To analyze the effect of increasing dysarthria severity on the ST and DY cues under consideration

THANK YOU

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