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

Tanuka Bhattacharjee[1], CV Thirumala Kumar[1], Yamini Belur[2], Atchayaram Nalini[2], Ravi Yadav[2], Prasanta Kumar Ghosh[1]

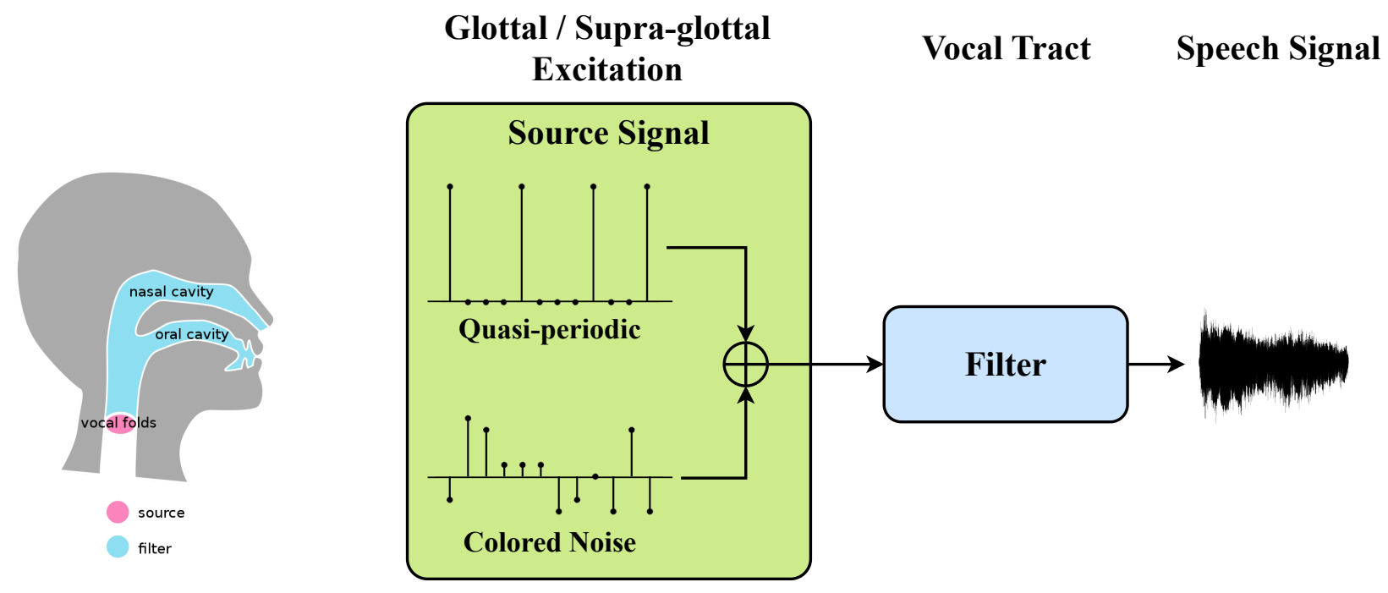
[1]SPIRE LAB, Electrical Engineering Department, Indian Institute of Science, Bengaluru, India

[2]National Institute of Mental Health and Neurosciences, Bengaluru, India

Amyotrophic Lateral Sclerosis (ALS)

- **ALS** is an **incurable** and **progressive neuro-degenerative** disease that affects **muscle movements**.
- Speech musculature, among others, get severely affected leading to **Dysarthria**.
- Speech functions including **articulation**, **phonation**, **prosody**, **respiration** and **resonance** get affected.
- **Even, elementary sustained vowel (SV) utterances get impaired.**

Source – Filter Interpretation of Vowel Production



- **Sustained vowel (SV) production** calls for
  - achieving vowel-specific source (S) and filter (F) structures
  - uniformly sustaining the structures for a prolonged duration

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

Our Objective

- We propose to capture these difficulties through **static (ST) and dynamic (DY) cues of source (S) and filter (F) components.**

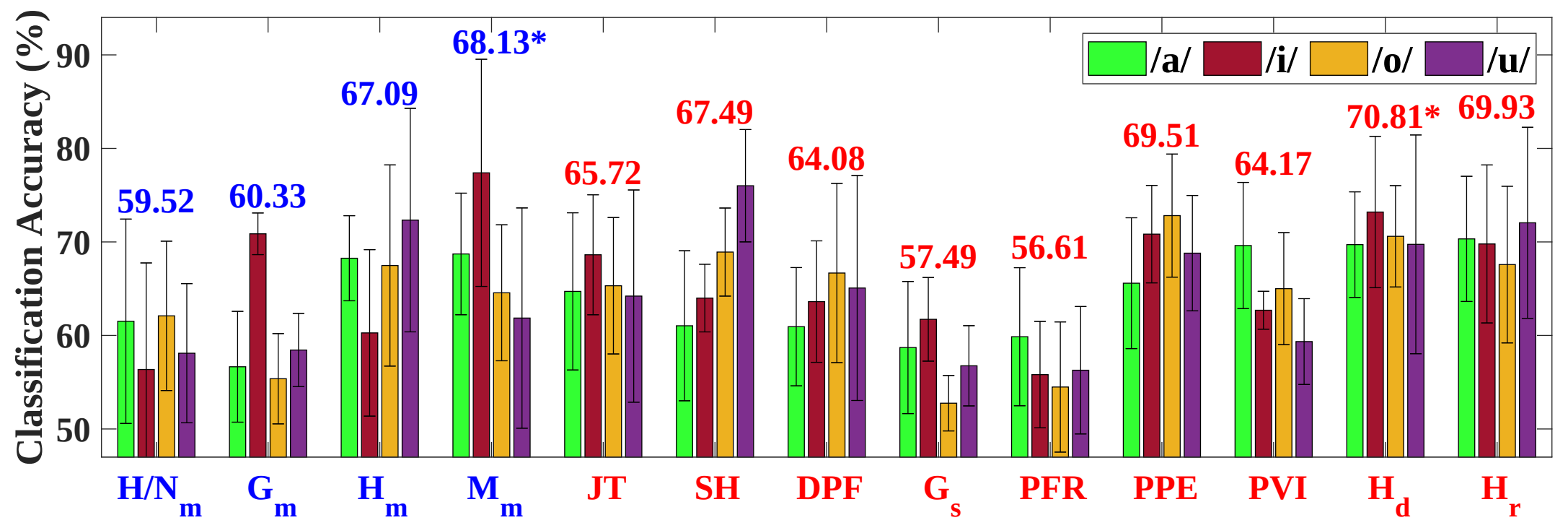
		Description	Potential reason	Clinical sign	Acoustic cues
Source (S)	Static (ST)	Unusual average characteristics of source excitation	Impaired respiratory and laryngeal function	Weakened or strained voice, hoarseness	Mean HNR, average loudness
	Dynamic (DY)	Unusual temporal variations in source excitation	Impaired laryngeal control	Difficulties in controlling pitch	Jitter, pitch period entropy
Filter (F)	Static (ST)	Impaired vocal tract configuration	Restricted articulatory mobility	Poor articulation	Mean spectral envelope, mean log-area ratio
	Dynamic (DY)	Unusual temporal fluctuations in vocal tract configuration	Articulatory muscle weakening	Irregular articulation	Temporal variations in spectral envelope

- **We aim to analyze the relative discriminative capabilities of source-static (S-ST), source-dynamic (S-DY), filter-static (F-ST) and filter-dynamic (F-DY) cues for SV-based ALS vs. healthy control (HC) classification.**

Experimental Details

- **Dataset**
  - **Place of data collection:** NIMHANS, Bengaluru, India
  - **Subjects:** 80 ALS (50M, 30F), 80 HC (62M, 18F) (Every subject gave an informed consent.)
  - **Speech task:** Sustained utterances of /a/, /i/, /o/ and /u/
  - **Total #utterances:** 858 (ALS), 842 (HC)
  - **Mean (SD) of utterance duration (sec):** 4.05 (2.29) (ALS), 5.71 (1.98) (HC)
  - **Recording device:** Zoom H6 with XYH-6 capsule (at 44.1 kHz sampling frequency)
- **Validation Protocol:** 5-fold cross-validation at subject level
- **Classifier:** Linear discriminant analysis (LDA)

Choice of Static and Dynamic Cues



Feature set (64D) adopted from Ref 2:
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

Mean (SD) of ALS vs. HC classification accuracies obtained using 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; here, * indicates the features having the highest average accuracy over all vowels among each of ST and DY groups

- **M_m** and **H_d** perform the best among ST and DY group respectively.
  - **Chosen as the representative ST and DY cues**
- The most stable articulatory configuration is expected to be attained during the middle portion of an SV.
  - **M_m** and **H_d** are derived from the middle 1 sec of the utterances - **M_m^1** and **H_d^1**

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

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

- M_m^1 and H_d^1 perform statistically similar 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.

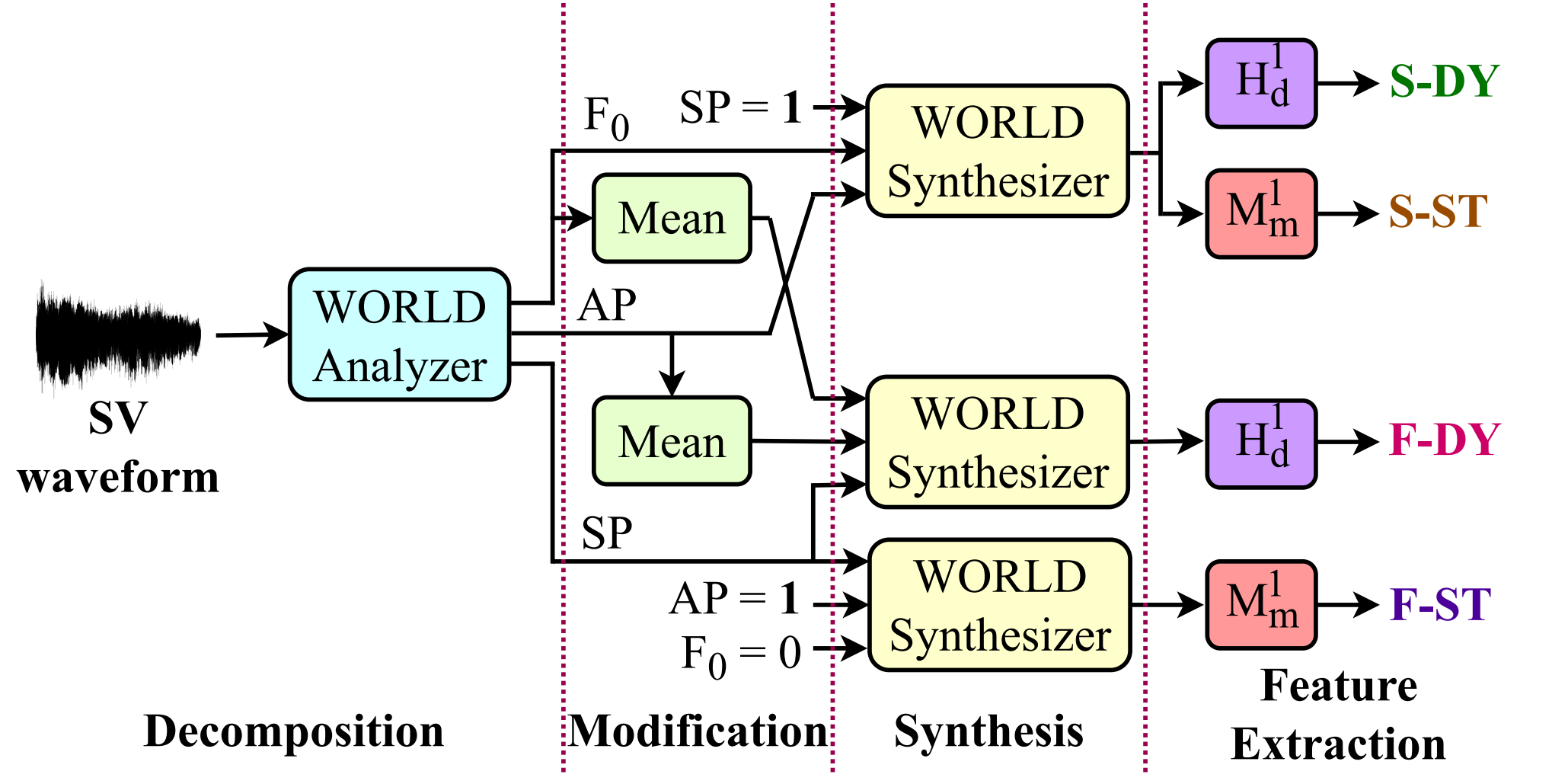
- **Comparison with baseline**

Mean (SD) of ALS vs. HC classification accuracies in % obtained using M_m^1 + H_d^1 cues of SVs as compared to baseline feature sets

Features	Vowels			
	/a/	/i/	/o/	/u/
M_m^1 + H_d^1	70.80 (5.20)	79.37 (9.70)	74.28 (7.29)	71.62 (8.29)
Baseline-64D[2] (from entire utterance)	73.76 (8.36)	81.00 (5.63)	73.22 (6.33)	73.24 (3.28)
Baseline-64D[2] (from middle 1.5 sec)	73.85 (5.09)	80.74 (4.97)	70.81 (9.78)	71.36 (6.87)
MFCC + CNN-LSTM[3]	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.

Static & Dynamic Source & Filter Cues



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

Mean (SD) of ALS vs. HC classification accuracies in % obtained using ST and DY cues of S and F components of SVs

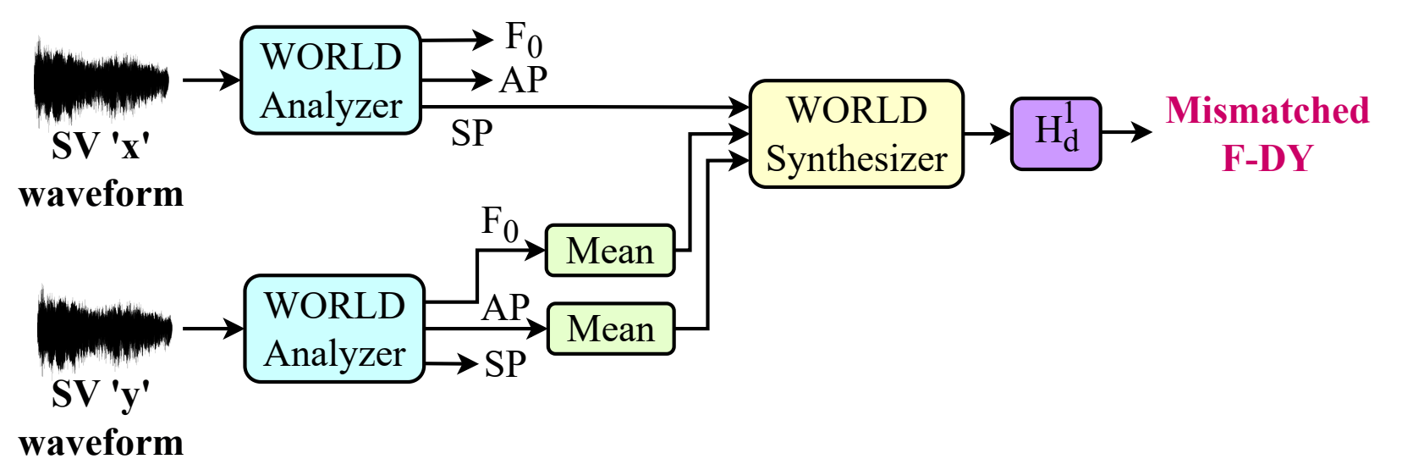
Features	Vowels			
	/a/	/i/	/o/	/u/
S-ST	55.27 (2.82)	61.85 (7.83)	56.32 (5.33)	55.82 (8.26)
S-DY	62.11 (2.68)	57.90 (5.86)	60.00 (4.59)	57.18 (5.16)
F-ST	60.25 (6.57)	**76.66 (12.90)**	64.27 (6.55)	63.51 (6.60)
F-DY	**66.29 (8.43)**	68.86 (1.91)	**73.03 (3.49)**	**70.27 (5.27)**

- 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/.
- 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.

Effect of Harmonic Locations

- For computing H_d^1, speech spectrum is sampled at the harmonic frequencies.
- For F-DY computation, harmonics are kept constant throughout the SV.
- We analyze if the locations of the harmonics (though constant) contribute towards the discriminative capabilities of the feature.

F-DY computation from mismatched utterances



Mean (SD) of ALS vs. HC classification accuracies in % obtained using F-DY cues of mismatched utterances

		F_0 + aperiodicity			
		/a/	/i/	/o/	/u/
spectral envelope	/a/	-	66.85 (6.03)	75.13 (3.85)	76.93 (2.82)
	/i/	73.08 (2.49)	-	69.57 (6.47)	70.75 (3.23)
	/o/	74.37 (4.38)	66.55 (3.73)	-	73.07 (4.22)
	/u/	71.22 (4.70)	69.70 (4.43)	74.40 (6.49)	-

- F_0 and aperiodicity of /a/, /o/ and /u/ when used with the spectral envelope of any vowel lead to mostly similar classification performances.
- F_0 and aperiodicity of /i/ always have inferior performance.
- Locations of the harmonics, or the frequencies at which the spectrum is sampled, are indeed important for capturing F-DY attributes.

Conclusion

- **Key Takeaways:**
  - Different cues capture predominant discriminative information in case of different vowels.
  - F-DY cues achieve the highest mean classification accuracy for 3 out of 4 vowels at hand.
  - 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

References

1. Gunnar Fant, Acoustic theory of speech production, Walter de Gruyter, 1970.
2. Maxim Vashkevich and Yu Rushkevich, “Classification of ALS patients based on acoustic analysis of sustained vowel phonations,” Biomedical Signal Processing and Control, vol. 65, pp. 102350, 2021.
3. Jhansi Mallela, Aravind Illa, BN Suhas, Sathvik Udupa, Yamini Belur, Nalini Atchayaram, Ravi Yadav, Pradeep Reddy, Dipanjan Gope, and Prasanta Kumar Ghosh, “Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson’s disease and healthy controls with CNN-LSTM using transfer learning,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 6784–6788.
4. Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “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.

Acknowledgement: We thank Navaneetha G and Agniv Chatterjee for their valuable assistance in data preparation. We also thank the Department of Science and Technology (DST), Govt. of India for supporting this work.