May 26th, 2020

Real-Time Speech Workload Estimation

an Honors College Thesis

Julian Fortune

Advisor: Dr. Julie A. Adams
Committee: Dr. Jamison Heard & Dr. Stefan Lee

Problem statement

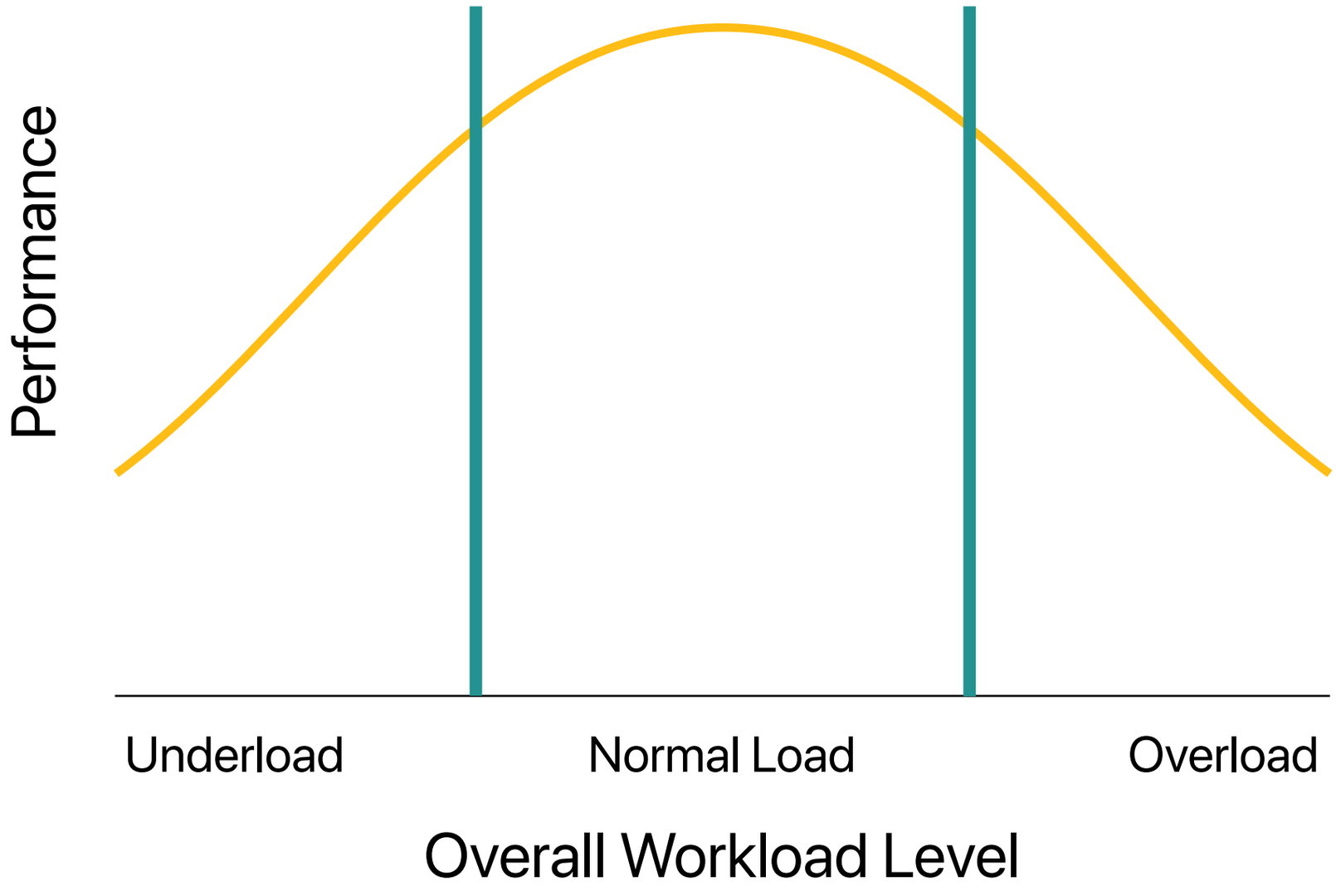
Find a means of estimating speech workload objectively in real-time.

- We want optimal performance from *human-machine teams*.
- The human-machine interface can maximize the human's performance by adapting interactions.
- Speech workload estimation is an essential component of adaptive human-machine interfaces.

https://www.dailyrecord.com/story/opinion/2017/09/01/forbes-air-traffic-control-privatize/105182818/

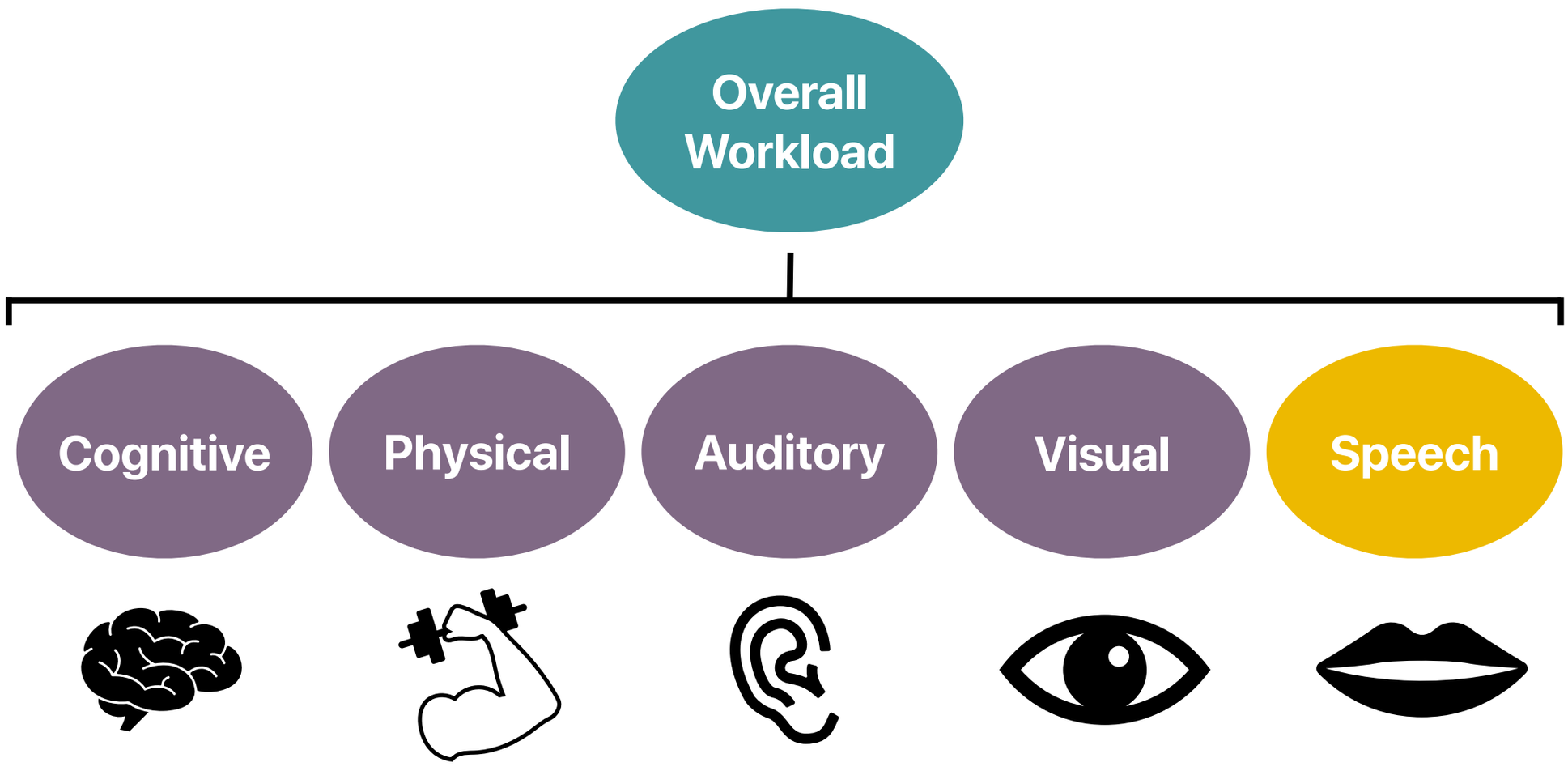
Performance and Workload

$$\text{Workload} = \frac{\text{Resources demanded}}{\text{Resources available}}$$



(Wickens, Lee, Liu, and Gordon-Becker, 2003)

Workload



(McCracken and Aldrich, 1984), (Mitchell, 2000)

Objective Speech Workload Metrics

Speech Workload Metric	Correlation	Feature Extraction Method
Intensity	Increases	Root-mean square
Intensity variation	Increases	Root-mean square
Fundamental frequency (Pitch)	Increases	Auto-correlation
Fundamental frequency variation	Increases	Auto-correlation
Speech rate	Increases	Voiced peaks
Filler utterances	Increases	Stable formants
Respiration-rate	Decreases	Physiological sensors

Existing Speech Workload Algorithms

Speech Workload Estimation for Air Traffic Control
(Luig and Sontacchi, 2010)

PHYSIOPRINT (Popovic, Stikic, Rosenthal, Klyde, and Schnell, 2015)

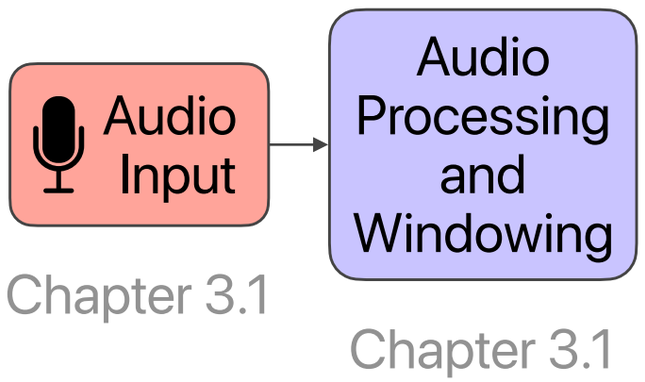
- Involved discrete classifications.
- Unable to demonstrate generalizability between individuals.
- Did not function in real-time.

Research Goal

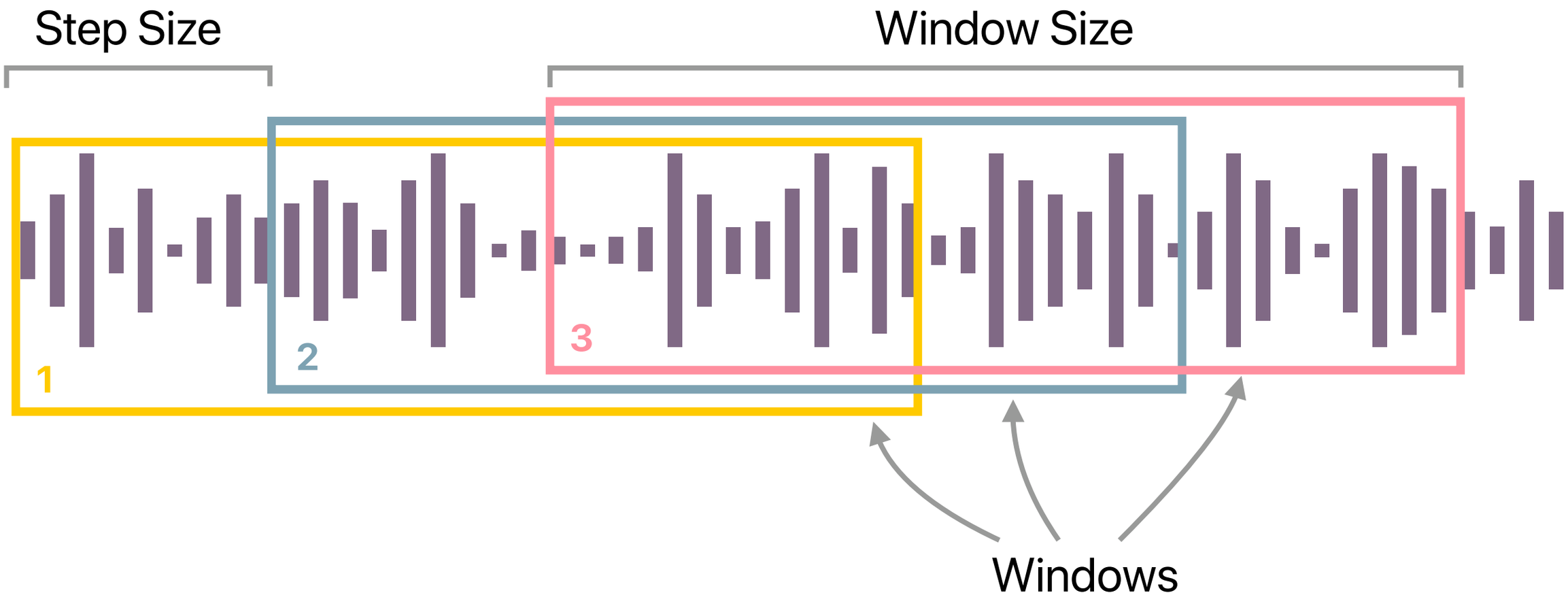
Developing an algorithm for estimating speech workload with the following attributes:

- Capable of **real-time** use.
- Calculates a **continuous** speech workload value.
- Uses **objective measures** (e.g., physiological data and audio).

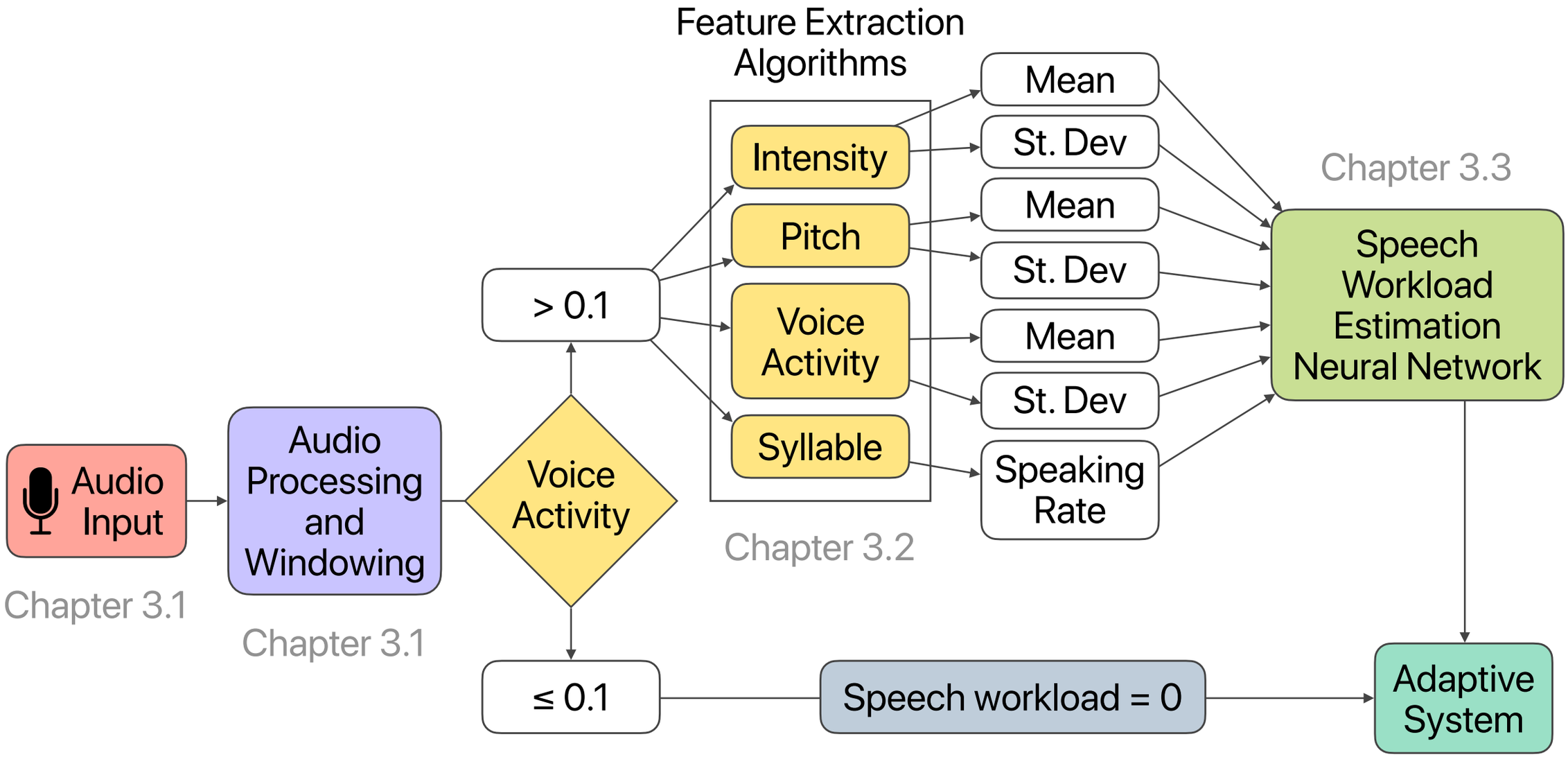
Speech Workload Estimation Algorithm



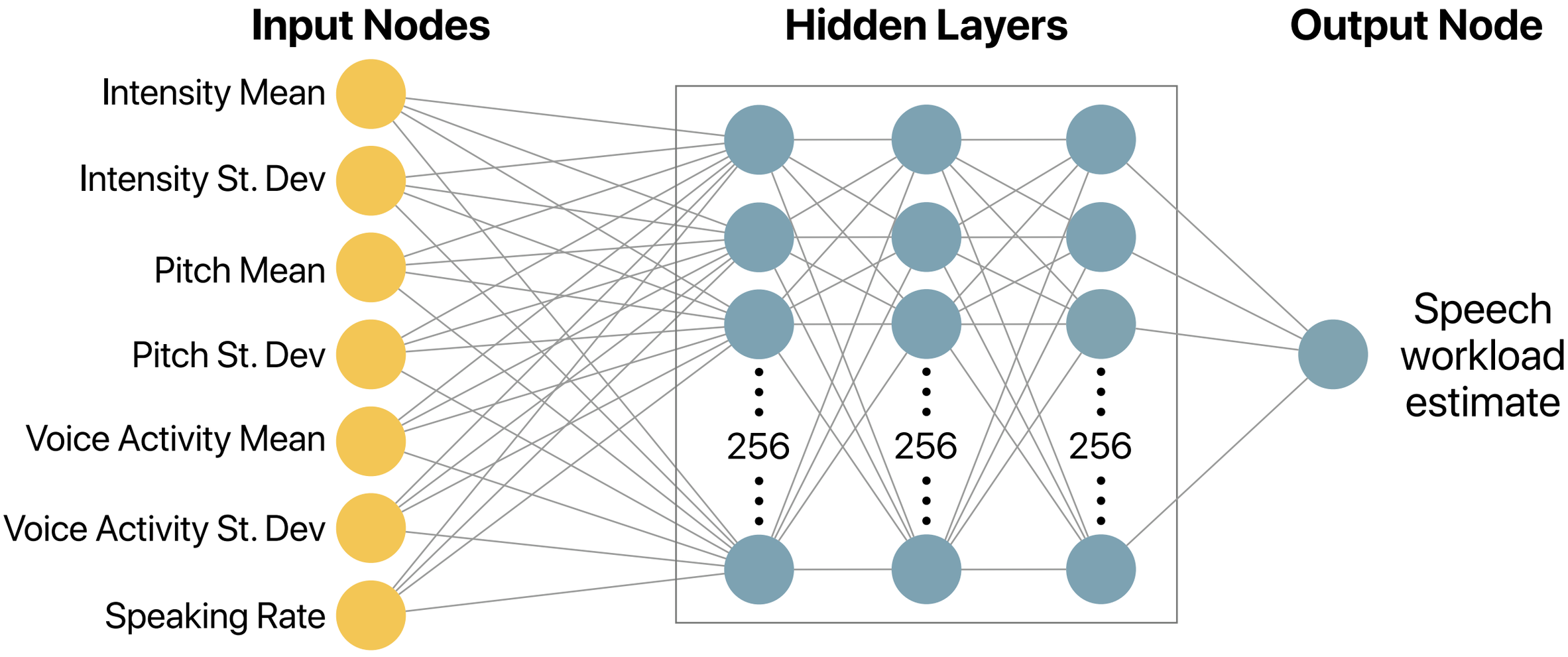
Audio Windows



Speech Workload Estimation Algorithm



Speech Workload Neural Network

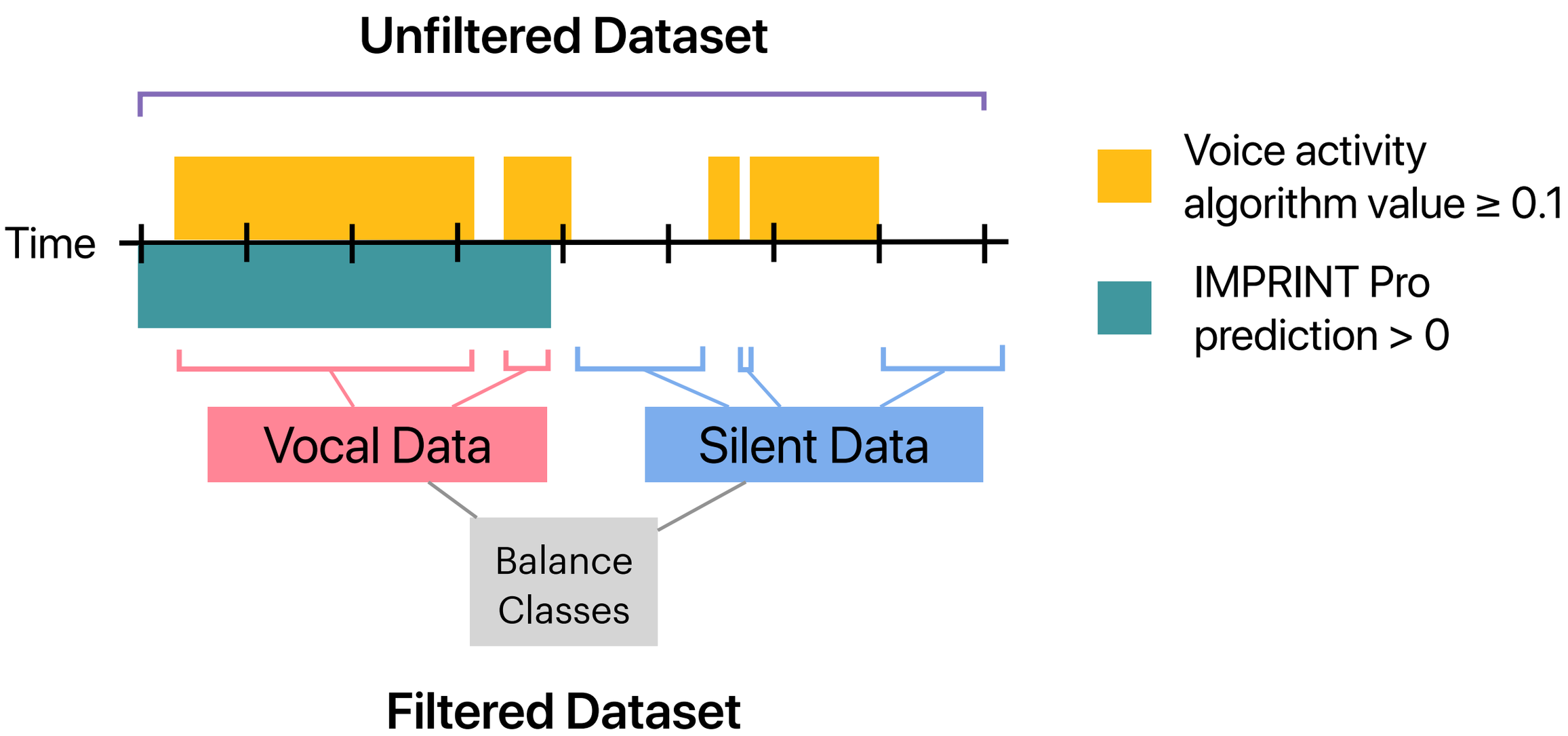


IMPRINT Pro was used for "ground-truth" labels in training and testing.

Experiments

Experiment	Research Question
Emulated Real-World Conditions	Can the speech workload estimation algorithm estimate accurately an individual's speech workload level solely using audio data?
Population Generalizability	Can the speech workload estimation algorithm estimate accurately an arbitrary, unseen individual's speech workload level solely using audio data?
Human-Robot Teaming Paradigm Generalizability	Can the speech workload estimation algorithm estimate accurately an individual's speech workload level in multiple human-robot teaming paradigms solely using audio data?
Task Environment Generalizability	Can the speech workload estimation algorithm estimate accurately an arbitrary, unseen individual's speech workload level in multiple task environments solely using audio data?
Real-Time Window Size	What is the optimal window size for real-time speech workload estimation?
Physiological Data and Filler Utterances	Does adding respiration rate and filler utterance features to the base set of features improve speech workload estimation accuracy?

Filtered and Unfiltered Datasets



Emulated Real-World Conditions Experiment

Hypotheses

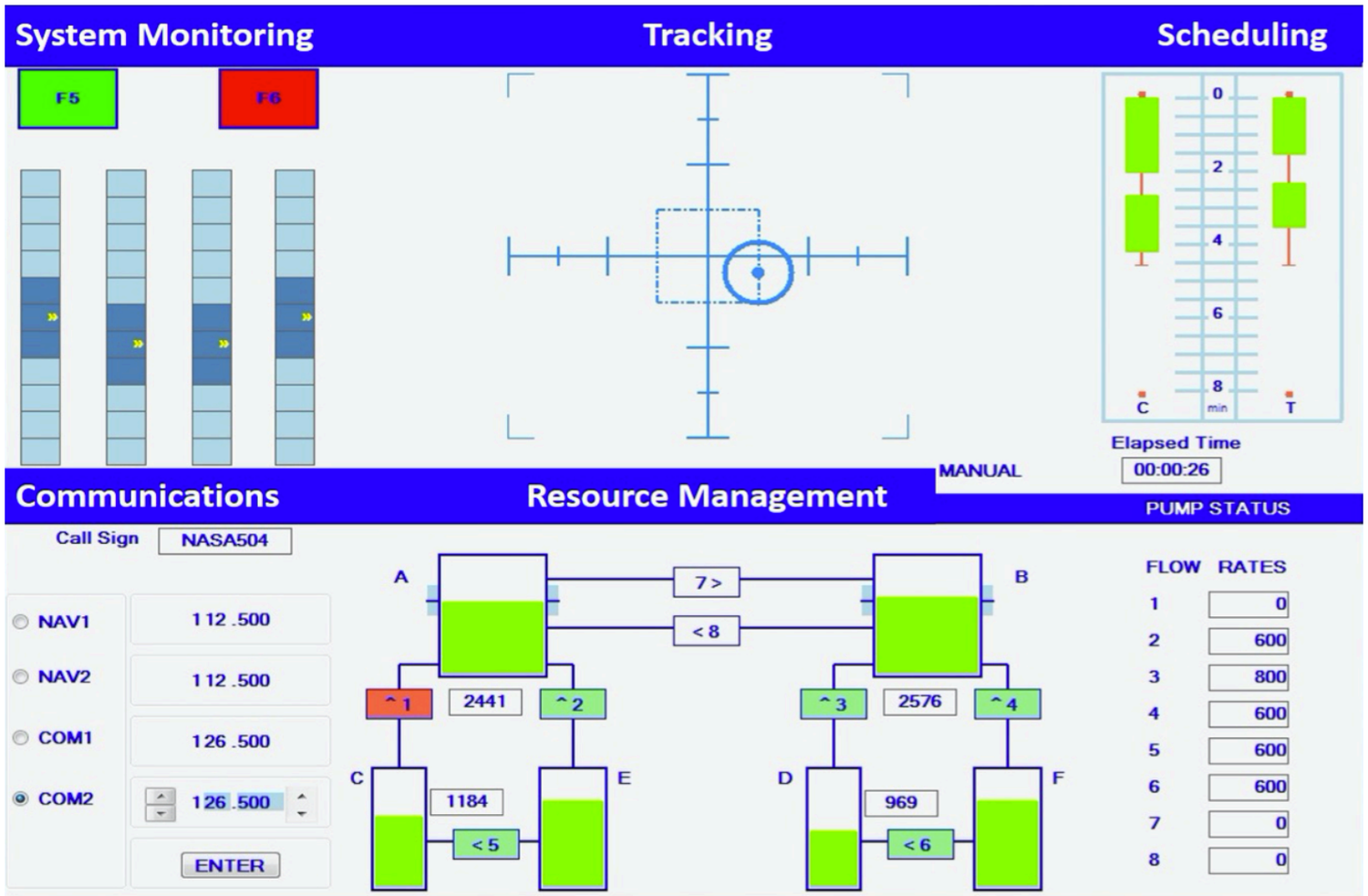
H_1: The speech workload estimate and the IMPRINT Pro model **means** will be within **one standard deviation** of each other.

H_2: The **Pearson's correlation coefficient** between the IMPRINT Pro model's predictions and speech workload estimates will be **positive** and **significant**.

H_3: The **root-mean-square error (RMSE)** between the estimate and Imprint Pro model will be **less than 5% overall**, even though the underload, normal load, and overload workload conditions may be different.

Supervisory Evaluation

- NASA MATB-II
- First day: three trials at each workload level: underload, normal load, and overload.
- Second day: one trial where the workload condition transitioned every 5 minutes.



(Heard, 2019)

The Correlation Between the Algorithm's Estimates and the IMPRINT Pro Model Predictions

Dataset	Condition	Coefficient
Unfiltered	Underload	0.144**
	Normal Load	0.046**
	Overload	0.008
	All	0.088**
Filtered	Underload	1.000**
	Normal Load	1.000**
	Overload	1.000**
	All	0.929**

Note: ** represents $p < 0.0001$ and * represents $p < 0.05$.

Conclusion

- H_1 was fully supported.
- **H_2** was partially supported.
- H_3 was not supported.

Lessons Learned

- IMPRINT Pro was not sufficient for modeling the human's actual speech workload.
- The speech workload algorithm accurately estimated speech workload.

Human-Robot Teaming Paradigm Generalizability Experiment

Hypotheses

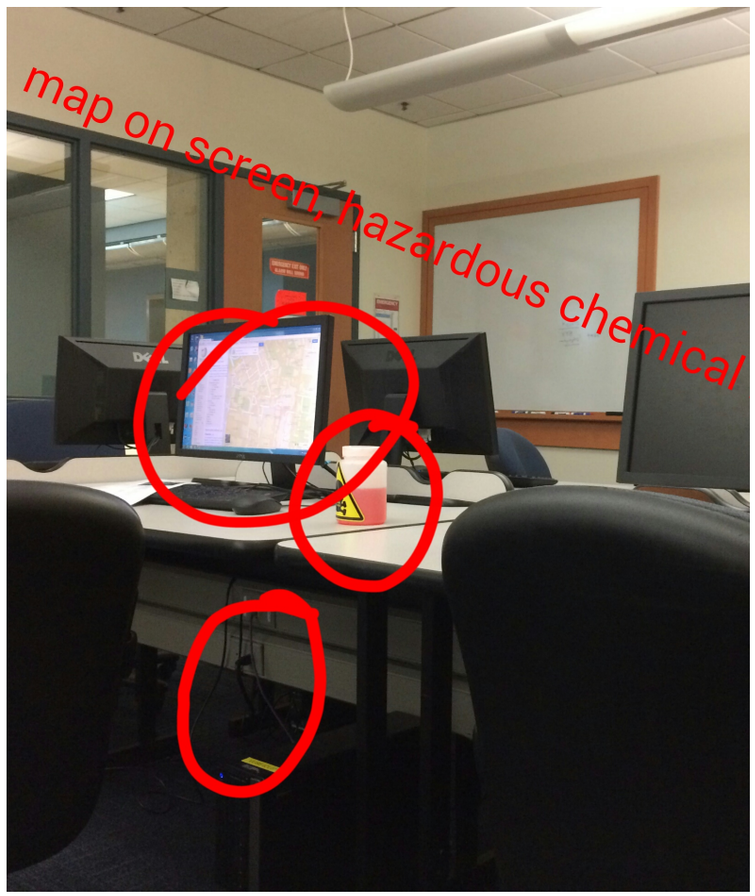
H_7: The speech workload estimate and the IMPRINT Pro model **means** will be within **one standard deviation** of each other.

H_8: The **Pearson's correlation coefficient** between the IMPRINT Pro model's predictions and speech workload estimates will be **positive** and **significant**.

H_9: The **root-mean-square error (RMSE)** between the estimate and Imprint Pro model will be **less than 5% overall**, even though the underload, normal load, and overload workload conditions may be different.

Peer-Based Evaluation

- Search for suspicious objects and collect samples in multiple environments aided by a robot assistant.
- Four non-concurrent tasks, each randomly assigned a workload condition: Low or High.



(1) Photo search task

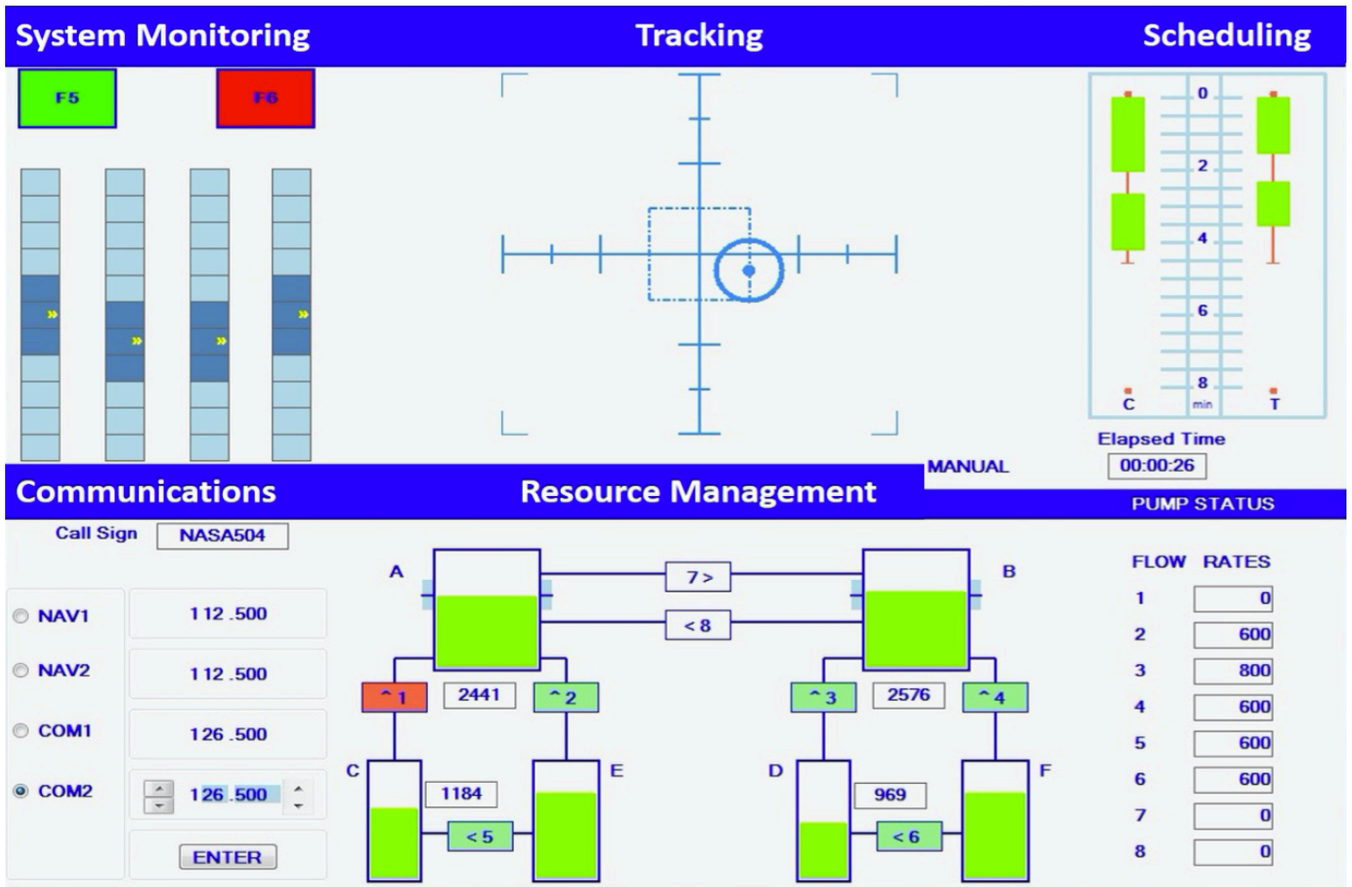
(2) Item search task

(3) Solid contaminant sampling task &
(4) Liquid contaminant sampling task

(Harriott, Zhang, and Adams, 2013)

Methodology

- The analysis was performed via two-fold cross-validation.
  - Both Supervisory Evaluation days.
  - Peer-Based Evaluation.



The Correlation Between the Algorithm's Estimates and the IMPRINT Pro Model Predictions

Dataset	Condition	Coefficient
Unfiltered	Underload	0.117**
	Normal Load	0.005
	Overload	-0.015*
	All	0.073**
Filtered	Underload	0.987**
	Normal Load	0.989**
	Overload	0.983**
	All	0.937**

Note: ** represents $p < 0.0001$ and * represents $p < 0.05$.

Conclusion

- H_7 was fully supported.
- **H_8** was partially supported.
- H_9 was not supported.

Lessons Learned

- The algorithm is robust to changes in human-machine teaming paradigm.

Real-Time Window Size Experiment

Hypotheses

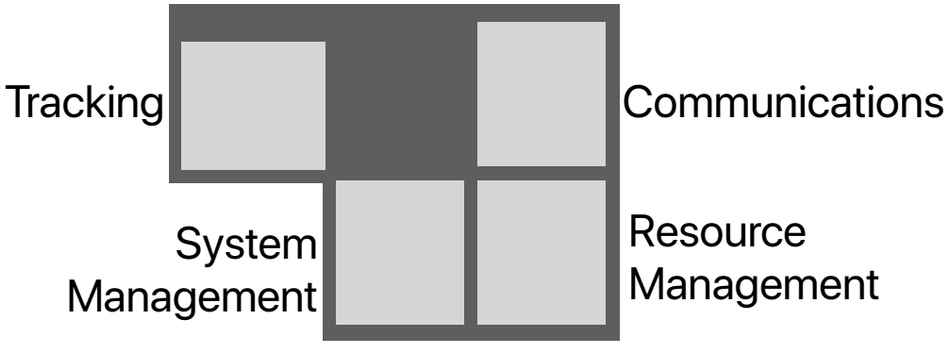
H_{13}: The **correlation** between the algorithm's estimates and the IMPRINT Pro speech workload predictions will **increase** as the window size increases.

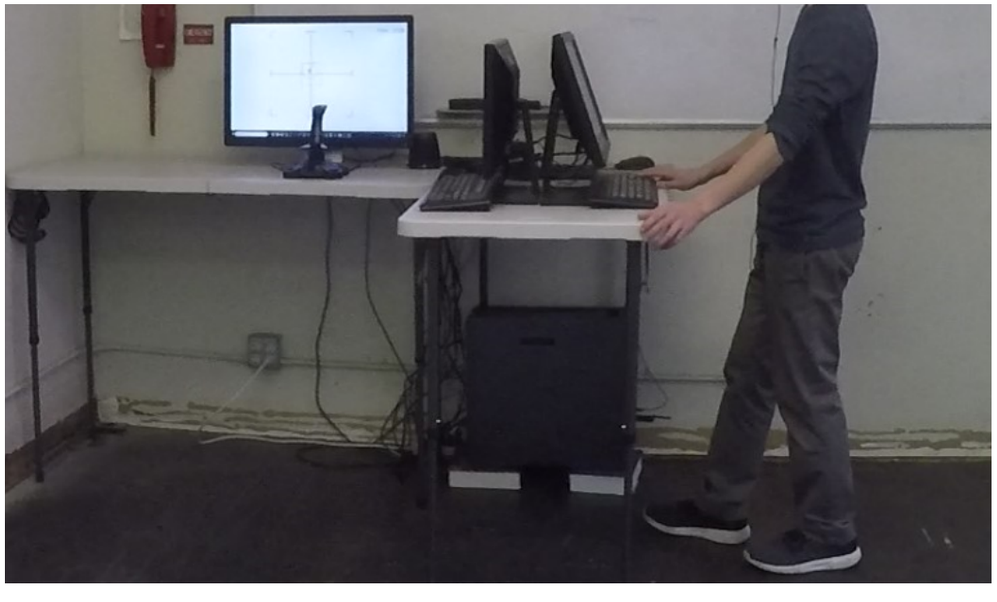
H_{14}: The **RMSE** of the algorithm's estimates, when compared to the IMPRINT Pro speech workload predictions, will **decrease** as the window size increases.

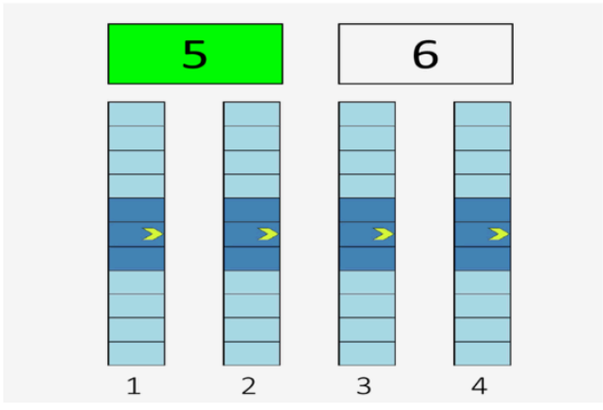
H_{15}: The **time** required to calculate the features will **increase** as the window size increases, but will **remain less than 1s**.

Real-Time Evaluation

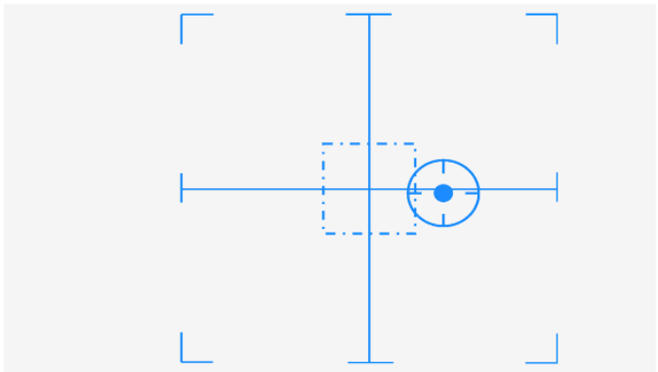
- Physically separated NASA MATB-II.
- One 52.5-min trial: seven consecutive 7.5-min workload conditions.



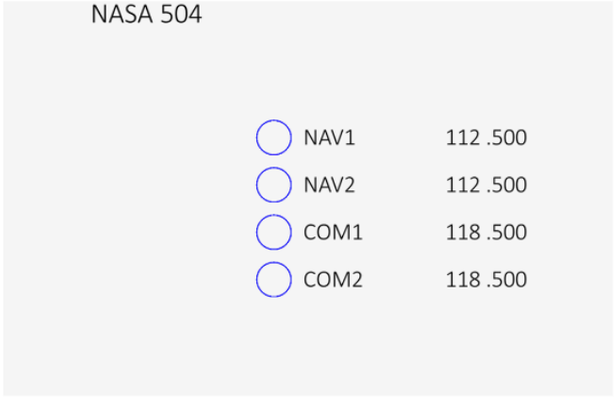




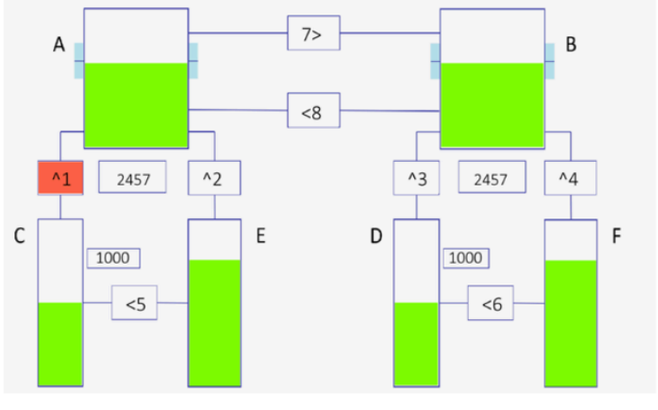
(a) System Management (SYSMAN).



(b) Tracking.



(c) Communications (COMM).



(d) Resource Management (RMAN).

(Heard, 2019)

Methodology

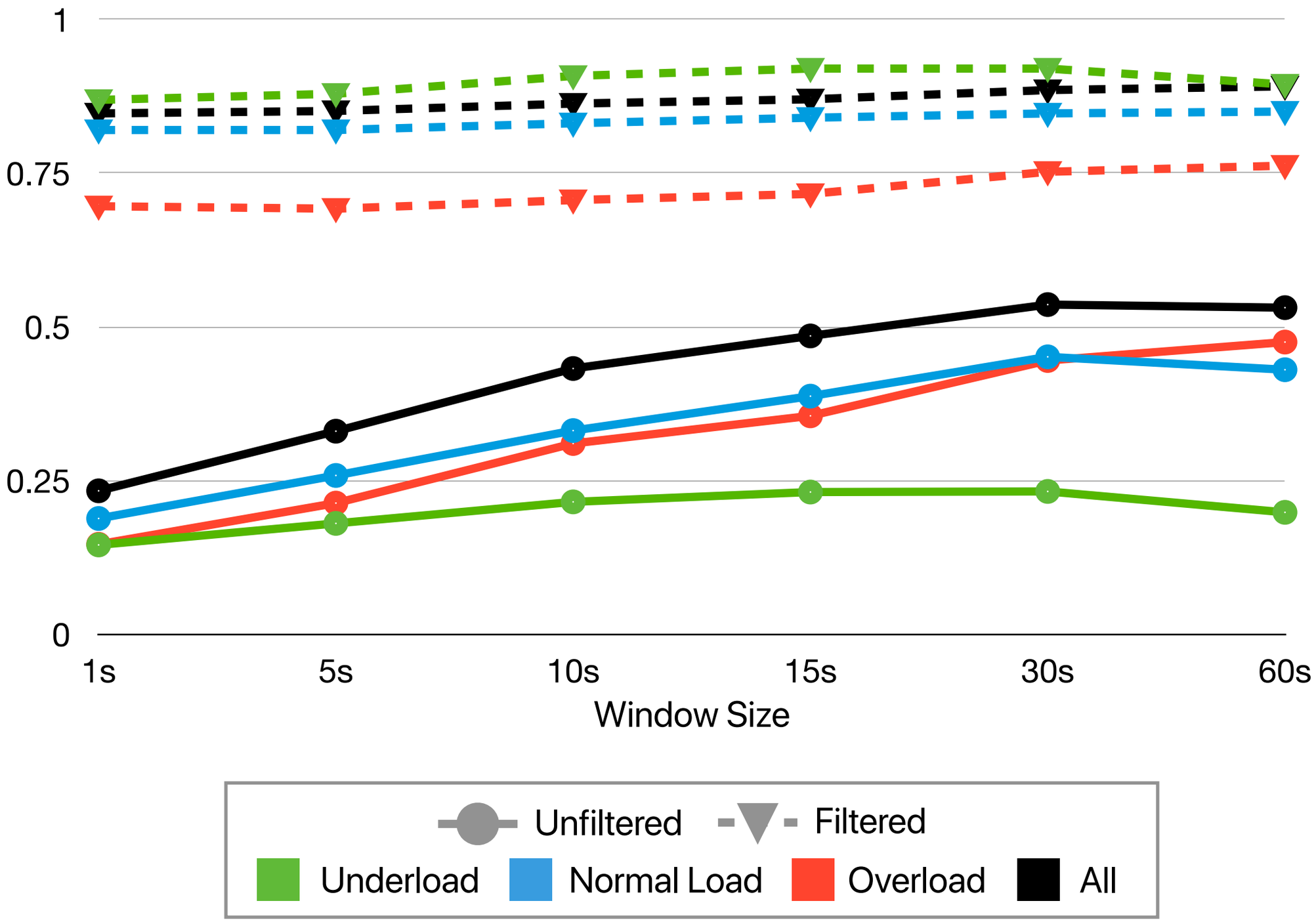
- The investigated window sizes were 1s, 5s, 10s, 15s, 30s, and 60s.
- For each window size:
  - Run-time was recorded for all four features.
  - Estimation accuracy was assessed by leave-one-participant-out cross-validation.
- Employed data from the Real-time Evaluation.

The Correlation Between the Algorithm's Estimates and the IMPRINT Pro Model Predictions by Window Size

Dataset	Condition	Window Size					
		1s	5s	10s	15s	30s	60s
Unfiltered	Underload	0.145**	0.18**	0.215**	0.231**	0.232**	0.198**
	Normal Load	0.188**	0.258**	0.331**	0.387**	0.451**	0.43**
	Overload	0.146**	0.213**	0.31**	0.355**	0.445**	0.475**
	All	0.233**	0.33**	0.432**	0.485**	0.536**	0.531**
Filtered	Underload	0.869**	0.879**	0.908**	0.92**	0.92**	0.894**
	Normal Load	0.82**	0.82**	0.831**	0.84**	0.847**	0.85**
	Overload	0.696**	0.692**	0.706**	0.716**	0.752**	0.762**
	All	0.847**	0.851**	0.863**	0.87**	0.885**	0.891**

Note: ** represents $p < 0.0001$ and * represents $p < 0.05$.
Color intensity corresponds to the correlation's magnitude.
Green cells support the hypothesis (monotonic increase), red cells do not.

The Correlation Between the Algorithm's Estimates and the IMPRINT Pro Model Predictions by Window Size



Run-Time of Feature Extraction by Window Size

Feature	Window Size					
	1s	**5s**	**10s**	**15s**	**30s**	**60s**
Intensity	.001 (.00)	.007 (.00)	.013 (.00)	.019 (.00)	.038 (.00)	.069 (.01)
Pitch	.051 (.02)	.246 (.08)	.490 (.15)	.734 (.22)	1.46 (.44)	2.84 (.88)
Voice Activity	.004 (.00)	.024 (.00)	.049 (.00)	.074 (.00)	.149 (.00)	.286 (.02)
Speech Rate	.004 (.00)	.024 (.00)	.047 (.00)	.070 (.00)	.139 (.00)	.258 (.03)
All Features	.061 (.02)	.301 (.08)	.599 (.15)	.897 (.22)	1.78 (.44)	3.45 (.88)

Mean (St. Dev.) run-time in seconds.

Conclusion

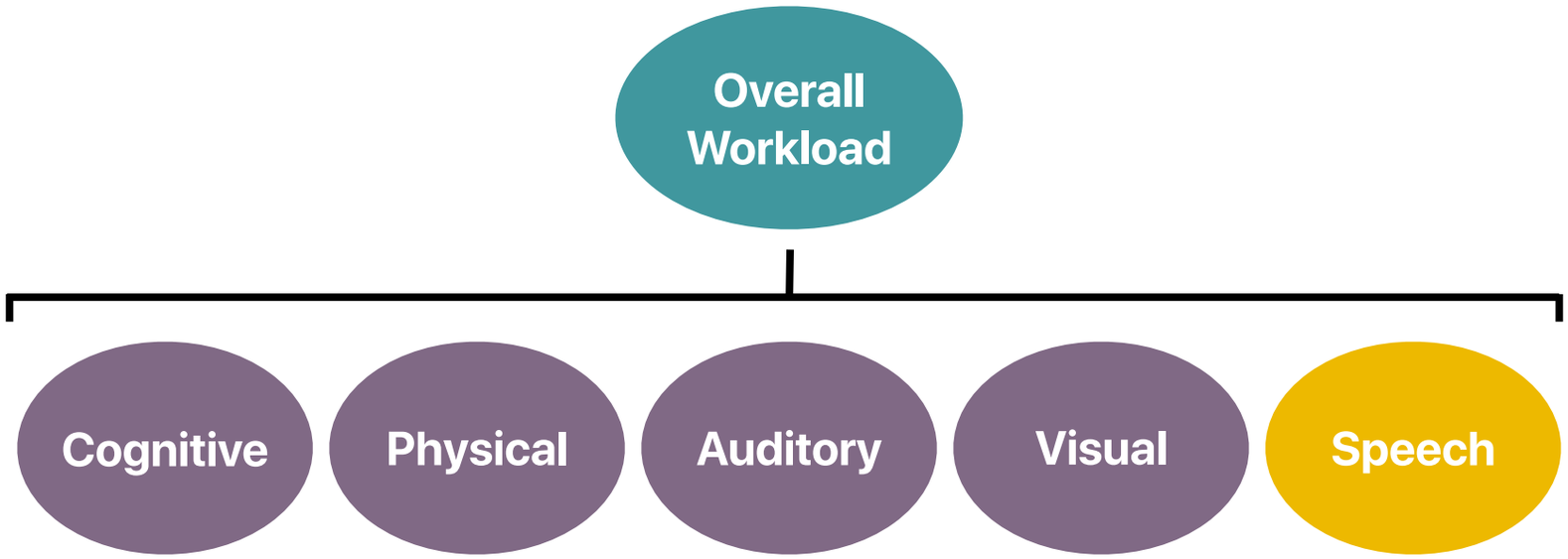
- **H_{13}** was partially supported.
- **H_{14}** was partially supported.
- **H_{15}** was not supported.

Lessons Learned

- Overall, a window size of 15 seconds is the most feasible size for real-time applications.
- A window size of 30 seconds is the most reliable for offline speech workload estimation.

Primary Contributions

- Design, development, and validation of a speech workload algorithm.



- The algorithm is invariant to:
  - Individual.
  - Human-machine teaming paradigm.
  - Task environment (stationary vs. non-stationary).
- The algorithm operates in real-time.

Secondary Contributions

- Effective for post-hoc assessment.
- Identification of appropriate window sizes for real-time and offline use.
- Analysis of adding physiological metrics and filler utterances.

Thank You

The floor is open for questions

Publications

J. Fortune, J. Heard, and J. A. Adams, "Speech workload estimation for human-machine interaction," 2020. Submitted to the Human Factors and Ergonomics Society Annual Meeting.

J. Heard, J. Fortune, and J. A. Adams, "SAHRTA: A supervisory-based adaptive human-robot teaming architecture." IEEE Conference on Cognitive and Computational Aspects of Situation Management, 2020. arXiv:2003.05823.

J. Heard, J. Fortune, and J. A. Adams, "Speech workload estimation for human-machine interaction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 277–281, 2019.

Acknowledgments

This work was partially supported by NASA Cooperative Agreement No. NNX16AB24A and DoD Contract Number W81XWH-17-C-0252 from the CDMRP Defense Medical Research and Development Program.

Appendix

The RMSE of the Algorithm's Estimates Compared to the IMPRINT Pro Model Predictions from the Emulated Real-World Conditions Experiment

Dataset	Condition	RMSE	Percent Error
Unfiltered	Underload	1.238	242.238%
	Normal Load	1.819	358.997%
	Overload	2.494	248.383%
	All	1.859	279.272%
Filtered	Underload	1.295	154.95%
	Normal Load	0.005	0.388%
	Overload	0.006	0.212%
	All	0.812	48.568%

The RMSE of the Algorithm's Estimates Compared to the IMPRINT Pro Model Predictions from the Human-Robot Teaming Paradigm Generalizability Experiment

Dataset	Condition	RMSE	Percent Error
Unfiltered	Underload	2.330	216.129%
	Normal Load	1.992	209.718%
	Overload	2.155	213.553%
	All	0.701	41.234%
Filtered	Underload	1.022	80.246%
	Normal Load	0.892	60.885%
	Overload	2.330	216.129%
	All	1.992	209.718%

The RMSE of the Algorithm's Estimates Compared to the IMPRINT Pro Model Predictions by Window Size from the Real-Time Window Size Experiment

Dataset	Condition	Window Size					
		1s	**5s**	**10s**	**15s**	**30s**	**60s**
Unfiltered	Underload	0.472	0.487	0.478	0.482	0.495	0.502
	Normal Load	1.151	1.113	1.067	1.027	0.980	1.005
	Overload	1.793	1.656	1.490	1.405	1.350	1.364
	All	1.271	1.199	1.109	1.058	1.018	1.034
Filtered	Underload	0.285	0.270	0.222	0.204	0.182	0.184
	Normal Load	0.586	0.600	0.590	0.578	0.566	0.570
	Overload	0.754	0.735	0.714	0.697	0.665	0.661
	All	0.583	0.581	0.564	0.551	0.530	0.531