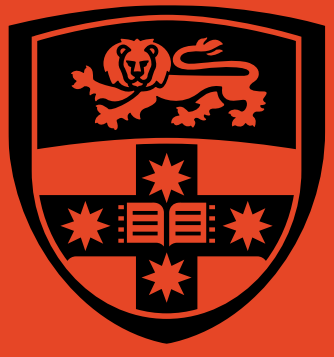




Multi-Modal Deep Learning for Voice Disorder Classification

Changbao Ren¹, Brian Stasak², Cate Madill², Craig Jin¹¹ School of Electrical and Computer Engineering, Faculty of Engineering, NSW Australia² Voice Research Laboratory, Sydney School of Health Sciences, Faculty of Medicine and Health, NSW Australia

Introduction

Automatic acoustic classification of speech disorders presents numerous challenges. Differences in recording environments, speaker demographics (e.g., age, gender), lifestyle (e.g., drinking, smoking), illness types, and metadata annotation standards across different voice corpora can reduce the generalization effectiveness of deep learning models.

Aim

To build a multi-modal automatic voice disorder classification system using different datasets with stronger generalization ability, adapting to more to patients' background and voice disorder categories.

This study uses the Feature-Wise Linear Modulation (FiLM) algorithm^[1] with combined acoustic and demographic/lifestyle features to improve automatic voice disorder classification. This novel approach employs a self-supervised model to address the scarcity of specific voice disorder types in the FEMH dataset^[2]. Further, the VOICED dataset^[3] is utilized to validate model cross-dataset generalization ability.

Materials & Methods

2 Datasets:

Far Eastern Memorial Hospital Voice (FEMH)^[2]
VOice ICar fEDerico II (VOICED)^[3]
2000 and 208 patients; up to 18 disorder types

4 Diagnostic Disorder Classes:

Muscle Tension, Organic Inflammatory,
Organic Neuromuscular, Organic Structural

Self-Supervised Learn Encoders:

Wav2vec, wav2vec 2.0^[4], HuBERT WavLM
Base, UniSpeech-SAT, data2vec

Model Training Parameters:

Self-supervised learning (SSL) encoders^[5],
class-weighted cross-entropy, model selection
validation, 10 independent random seeds

Performance Metric: Balanced Accuracy (%)

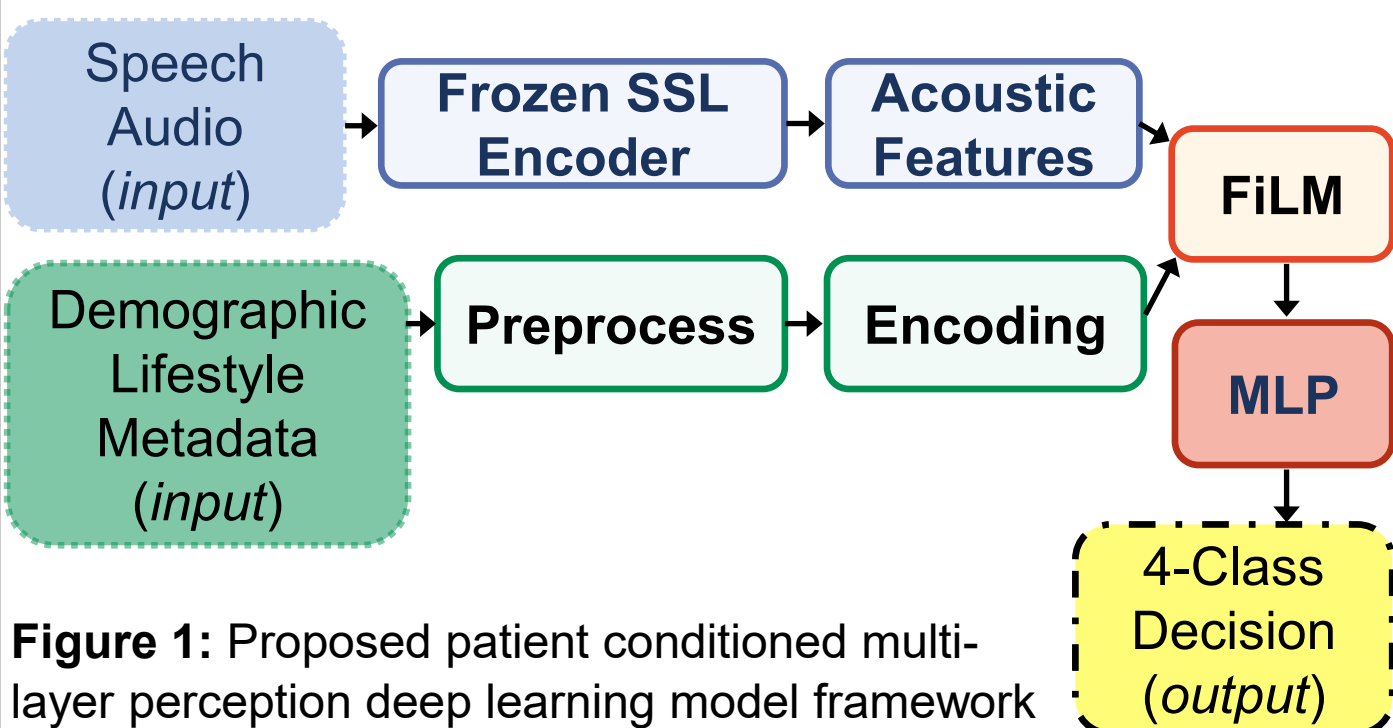


Figure 1: Proposed patient conditioned multi-layer perception deep learning model framework

Results

Table 1: Automatic 4-class voice disorder classification across six self-supervised speech (SSL) model types and four different experimental fusion settings. Per dataset, balanced accuracy (%) is shown for each SSL model feature type.

Dataset	SSL Model	Audio-only	Concatenation	FiLM(Ours)	Gated Fusion
FEMH	data2vec	59.80 ± 1.90	66.39 ± 2.65	66.84 ± 2.49	65.80 ± 1.77
	HuBERT	62.14 ± 1.56	68.09 ± 1.65	70.09 ± 1.23	69.73 ± 2.28
	UniSpeech-SAT	61.23 ± 1.80	68.16 ± 1.35	70.25 ± 1.20	67.91 ± 0.59
	wav2vec	58.17 ± 1.21	69.00 ± 1.18	67.40 ± 3.43	65.20 ± 2.64
	wav2vec 2.0	57.40 ± 2.20	61.02 ± 0.97	60.73 ± 1.82	62.97 ± 1.22
	WavLM Base	58.91 ± 2.13	65.23 ± 1.62	64.47 ± 1.85	66.21 ± 1.84
VOICED	data2vec	31.85 ± 2.86	33.18 ± 1.91	31.76 ± 1.47	34.88 ± 1.63
	HuBERT	22.63 ± 1.42	28.22 ± 1.88	31.92 ± 3.08	35.50 ± 1.85
	UniSpeech-SAT	29.00 ± 2.04	33.85 ± 1.96	34.19 ± 2.63	34.94 ± 1.67
	wav2vec	24.73 ± 1.48	34.00 ± 1.77	35.71 ± 1.34	35.57 ± 1.73
	wav2vec 2.0	29.73 ± 1.10	29.84 ± 1.25	33.69 ± 1.94	31.45 ± 1.01
	WavLM Base	25.94 ± 1.23	31.18 ± 1.82	30.55 ± 2.24	32.94 ± 1.81

Note: Bold values indicate the highest mean Balanced Accuracy within each dataset.

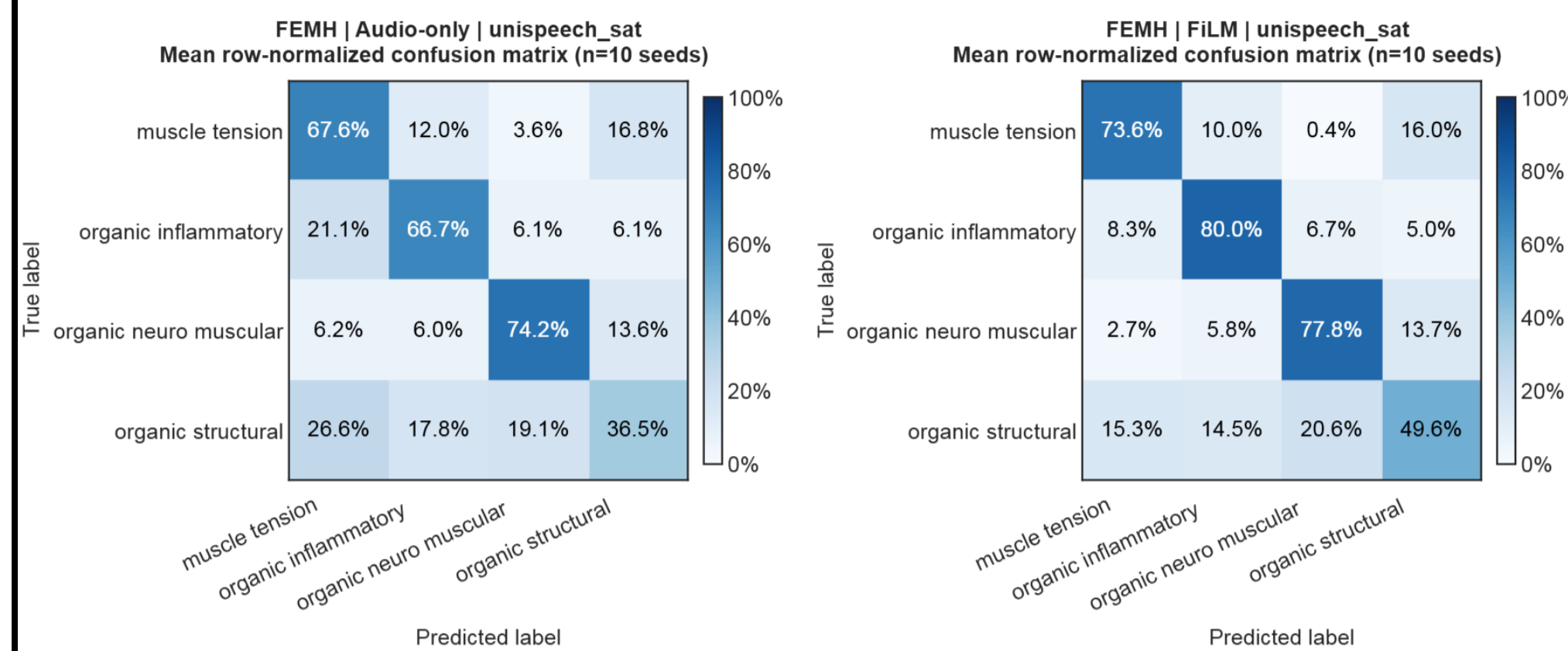


Figure 2: Comparison of FEMH dataset automatic voice disorder classification confusion matrices results for audio acoustic-only method (left) versus acoustic + FiLM demographic-lifestyle method (right).

Key Observations

- For most fusion-encoder configurations, combined acoustic + demographic feature modality improved FEMH and VOICED accuracies when compared to the acoustic-only baseline model
- The FEMH attained the highest balanced classification accuracy using the UniSpeech-SAT + FiLM model (70.25%); this was a ~9% gain over the audio-only model
- For the FEMH, the voice disorder classes with the largest recall improvements for the FiLM model were **Organic Inflammation** improved by 13.3% (from 66.7% to 80.0%); and **Organic Structural** improved by 13.1% (from 36.5% to 49.6%)
- FEMH misclassification of organic-related voice disorder types as **Organic Inflammatory** decreased from 21.1% to 8.3% and **Organic Structural** decreased from 26.6% to 15.3%
- VOICED achieved up to 35.7% balanced accuracy with the wav2vec + FiLM (35.7%); however, there are still significant performance gaps between these two datasets
- **Organic Structural** was the most frequently attributed voice disorder class – this is possibly due to comorbidity factors (i.e., MT primary diagnosis with OS disorder secondary diagnosis)

Conclusions

- ❖ The demographic and lifestyle info provides additional context for the self-supervised acoustic speaker representations, therefore improving classification compared to acoustic-only baseline models
- ❖ The FiLM model achieved the highest balance accuracy on both the FEMH and VOICED datasets; however, the optimal fusion strategy differs for different self-supervised learning encoders
- ❖ Performance gains are class-dependent; the **Organic Structural** voice disorder class remains a challenging diagnosis due to its higher degree of class confusability
- ❖ Automatic classification significantly decreased during cross-corpora experiments with the FEMH and VOICED datasets – indicating a need to address the differences in recording conditions and patient populations

Future Research

Additional validation of this automated framework on larger, more diverse clinical voice datasets, including expansion of more patient metadata (i.e., occupation, voice usage, exercise, symptom onset, health self-surveys, comorbidity), is needed, especially experimentation involving multi-class voice disorder classification and testing cross-dataset robustness.

Literature Cited

- [1] Perez, E. et al. (2018). FiLM: Visual reasoning with a general conditioning layer. AAAI 32(1).
- [2] K. Degila, et al. (2018). The UCD System for the 2018 FEMH Voice Data Challenge. 2018 IEEE Intern. Conf. on Big Data (Big Data), Seattle, WA, USA, pp. 5242-5246.
- [3] Cesari, U. et al. (2018). Voice disorder classification using the VOICED database. Computers & Electrical Engineering 69, 528–536.
- [4] Baevski, A. et al. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS.
- [5] Hsu, W.-N. et al. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM TASLP.

Contact Information

Changbao Ren (MSEng.)
E: cren0837@uni.sydney.edu.au