Speech Intelligibility Classifiers from 550k Disordered Speech Samples

Subhashini Venugopalan[1], Jimmy Tobin[1], Samuel J. Yang[1], Katie Seaver[1,2], Richard J.N. Cave[1], Pan-Pan Jiang[1], Neil Zeghidour[1], Rus Heywood[1], Jordan Green[1,2], Michael P. Brenner[1,3]; [1] *Google Research*, [2] *MGH Institute of Health Professions*, [3] *Harvard University*
{vsubhashini, jtobin}@google.com

Introduction

Intelligibility classifier uses

- Atypical speech can manifest from a variety of conditions, including neurological diseases such as ALS, Parkinson's Disease, and Cerebral Palsy.
- They can also be used to detect such speech in YouTube, to allow better transcriptions from specialized Automatic Speech Recognition (ASR) systems, or used by researchers as an objective measure to monitor decline in speech, e.g., in ALS.
- Automatic assessments of speech intelligibility can help predict how well voice-based assistive technologies might aid a person with speech disorders.
- Such classifiers can also help identify variable manifestations of impaired speech, to enable automatic collection of such data at scale to teach and improve ASR systems.

Will ASR on device work for you?
Or do you need a custom model?

Can users monitor deterioration?
Across different speaking disorders.

Improve video transcriptions.
Collect disordered speech at scale.

Dataset and Method

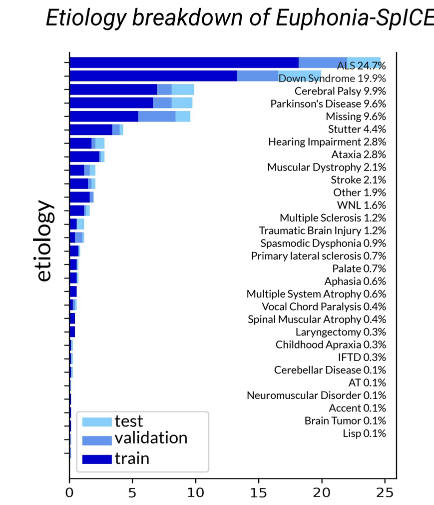
The Euphonia-SpICE Dataset

The Euphonia-SpICE dataset is a subset of the Euphonia dataset. It contains data from 677 speakers (756,147 utterances) who were rated by speech-language pathologists (SLPs) on a five-point Likert scale of intelligibility. The scale was mapped to five classes: typical, mild, moderate, severe, and profound. All utterances from a speaker are labeled with the same rating.

Table 1: *Count of speakers and utterances in Euphonia-SpICE.*

Intelligibility	# speakers Train	# speakers Val.	# speakers Test	# utterances Train	# utterances Val.	# utterances Test
TYPICAL	161	41	25	149,941	24,142	10,664
MILD	161	29	37	208,843	22,532	39,007
MODERATE	83	23	19	124,984	48,814	21,214
SEVERE	54	12	15	60,692	13,868	22,397
PROFOUND	9	4	4	6,716	1,691	642
OVERALL	468	109	100	551,176	111,047	93,924

Etiology breakdown of Euphonia-SpICE



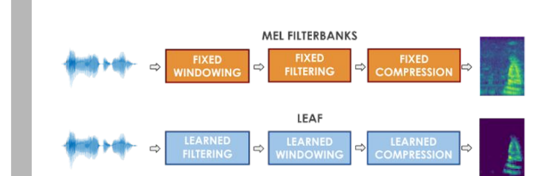
Datasets for Generalization

- UASpeech: Speech produced by speakers with CP
- TORGO: Speech produced by speakers with either CP or ALS
- ALS-TDI PMP dataset: Speech produced by speakers with ALS.
- SpICE-V: A dataset of unprompted speech from speakers with different disorders, curated from a collection of web videos.

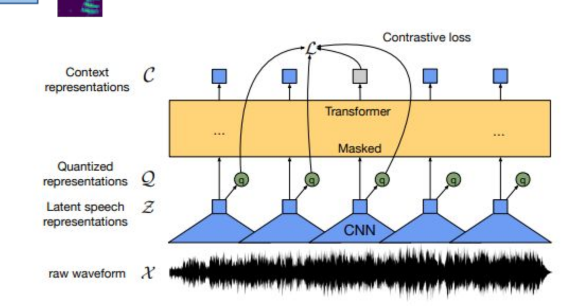
Different representation backbones: CNN, CNN+Transformers, RNN-T

LEAF + CNN: This model trains a fully learnable convolutional classifier with a LEAF frontend which jointly learns filtering, pooling, compression and normalization from data.
wav2vec 2.0: This model uses self-supervised representations from the final layer of the wav2vec 2.0 model, which is publicly available on HuggingFace.
ASR-enc: This model uses an LSTM encoder that models acoustic inputs in an ASR system based on an RNN transducer (RNN-T) model.

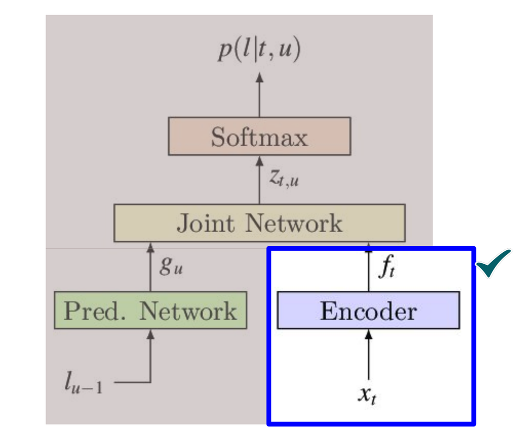
LEAF + CNN
Learnable frontend [4]



wav2vec2
Transformer+CNN [5] and is open-source and includes model weights.



ASR encoder representations
RNN-T model trained on typical speech [3]



[4] LEAF: A Learnable Frontend for Audio Classification ICLR '21
[5] wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations NeurIPS '20
[3] Narayanan et. al. Recognizing longform speech in end-to-end models ASRU '19

Results

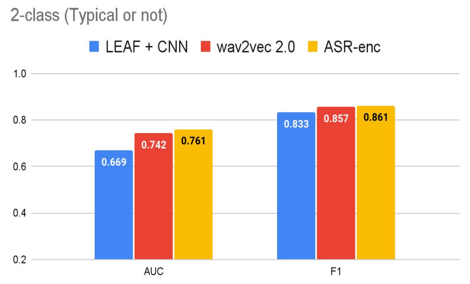
Euphonia SpICE performance

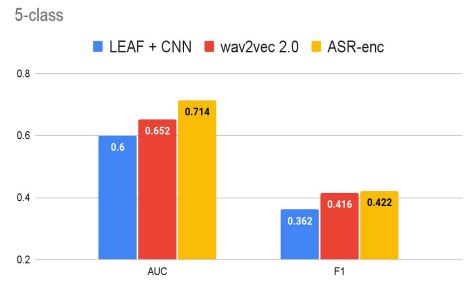
Performance on two classification tasks:

Task 1: 2-class
0: {TYPICAL} or 1: {MILD, MODERATE, SEVERE, PROFOUND}
Task 2: 5-class
0: {TYPICAL} or 1: {MILD} or 2: {MODERATE} or ...

Evaluation metrics: AUC, F1 and Accuracy

The ASR-enc model had the best performance on both tasks, followed by the wav2vec 2.0 model. LEAF + CNN model performed comparably worse.





Generalization

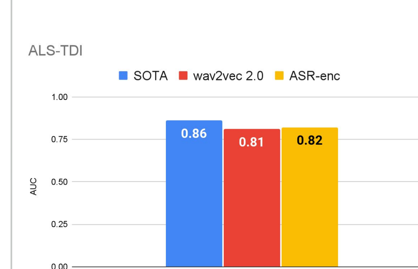
TORGO

14 speakers
7 controls, 7 - CP/ ALS

Speaker	wav2vec 2.0	ASR-enc
FC01	typ. (96.2)	typ. (96.2)
FC02	typ. (95.9)	typ. (100)
FC03	typ. (83.2)	typ. (78.4)
MC01	typ. (96.6)	typ. (92.4)
MC02	typ. (94.3)	typ. (92.6)
MC03	typ. (98.3)	typ. (98.3)
MC04	typ. (98.3)	typ. (99.2)
F03	mild (87.0)	mild (88.0)
F04	typ. (91.8)	typ. (74.2)
M03	typ. (98.9)	typ. (100)
F01	mod. (100)	mod. (100)
M02	mild (100)	mild (100)
M04	sev. (100)	mod. (100)
M05	sev. (100)	mod. (100)

ALS-TDI

Test set: 90 speakers,
~1330 recordings
"I owe you a yoyo" x 5



UASpeech

28 speakers
13 - controls, 15 - CP
765 words per speaker

SpICE-V

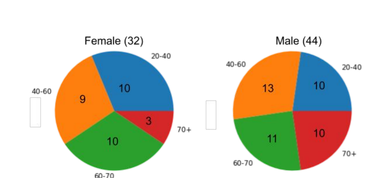
106 Dysarthric speakers + 76 Controls sourced from AudioSet

Sourcing dysarthric speech from the web

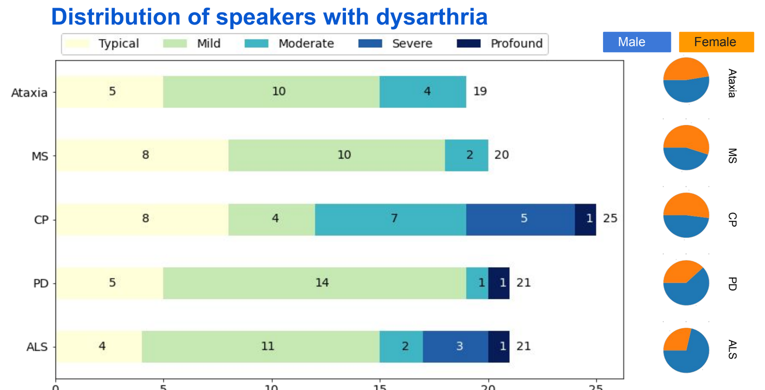
SLPs label

- ROI - time segments when dysarthric speaker is speaking
- severity and intelligibility - 5-point Likert
- inferred gender (to help balance)

Distribution of SpICE-V control videos/speakers



Distribution of speakers with dysarthria



Results

Comparing accuracy of identifying atypical speech

Group	w. Typ. non-ctrl	# Utts.	Total (Atyp.) # Spkr	wav2vec 2.0 Acc. (%) spkr	wav2vec 2.0 Acc. (%) utt.	ASR-enc Acc. (%) spkr	ASR-enc Acc. (%) utt.
Controls	×	76	76 (0)	76.32	76.32	**96.42**	**96.42**
Dysarthric (-Typ.)	×	1489	76 (76)	**93.42**	**94.83**	63.16	66.92
Dysarthric (all)	✓	2221	106 (76)	**77.36**	**75.64**	68.65	67.92
All (-Typ.& Dys.)	×	1565	152 (76)	**84.87**	**93.93**	78.29	68.21
All	✓	2297	182 (76)	76.92	**75.66**	**78.57**	69.47

Sliced by Etiology

Etiology	# Utt.	# Spkr Total (Typ.)	wav2vec 2.0 Acc. (%) spkr	wav2vec 2.0 Acc. (%) utt.	ASR-enc Acc. (%) spkr	ASR-enc Acc. (%) utt.
ALS	443	21 (4)	90.5	87.6	76.2	76.0
PD	498	21 (5)	85.7	84.9	61.9	73.0
CP	620	25 (8)	72.0	69.8	72.0	74.5
MS	352	20 (8)	55.0	57.5	60.0	48.6
Ataxia	308	19 (5)	84.2	75.6	68.4	62.1

Conclusion

Takeaways

- We developed & compared different approaches to classifying intelligibility of speech
- Our models were trained on utterances from over 650 speakers.
- The models generalized well to different datasets - TORGO, ALS-TDI and UASpeech.
- Collected SpICE-V dataset of realistic speech from videos.
- Dysarthric speakers with typical speech are harder to classify.
- Models do well on ALS, PD, CP and Ataxia.

Links

- Link to paper: https://arxiv.org/abs/2303.07533
- Github: https://github.com/google-research/google-research/tree/master/euphonia_spice