Speech Intelligibility Classifiers from 550k Disordered Speech Samples

Subhashini Venugopalan, Jimmy Tobin, Samuel J. Yang, Katie Seaver, Richard J.N. Cave, Pan-Pan Jiang, Neil Zeghidour, Rus Heywood, Jordan Green, Michael P. Brenner

ICASSP 2023

Google Research

Why study speech intelligibility?

how well speech is understood by a human listener.

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.

Data

Project Euphonia

focused on helping people with atypical speech be better understood

g.co/euphonia, g.co/projectrelate

Euphonia-SpICE dataset: >750K utterances, 650+ speakers

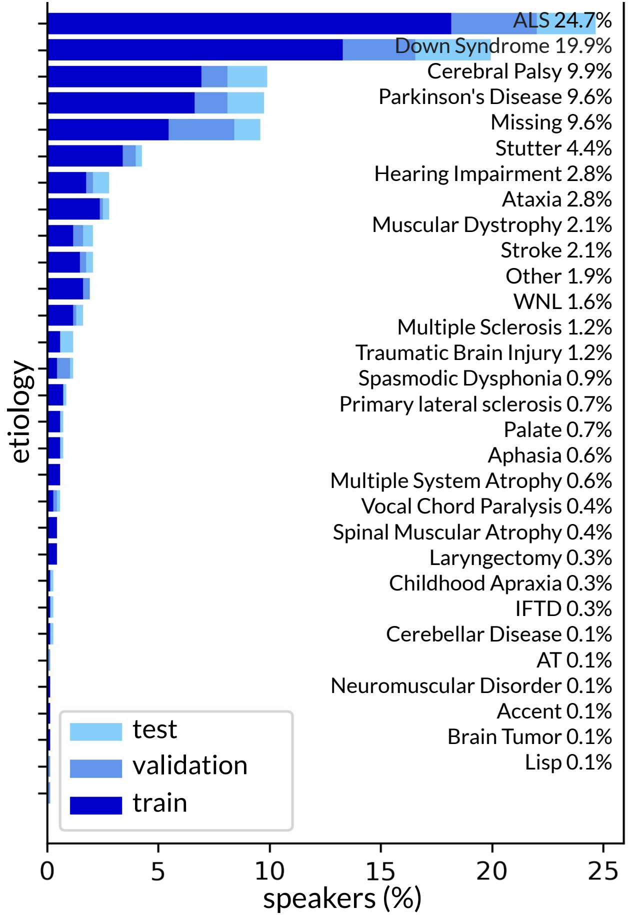
Table 1: *Count of speakers and utterances in Euphonia-SpICE.*

Intelligibility	# speakers			# utterances		
	Train	Val.	Test	Train	Val.	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

All roughly similar distribution

Google Research

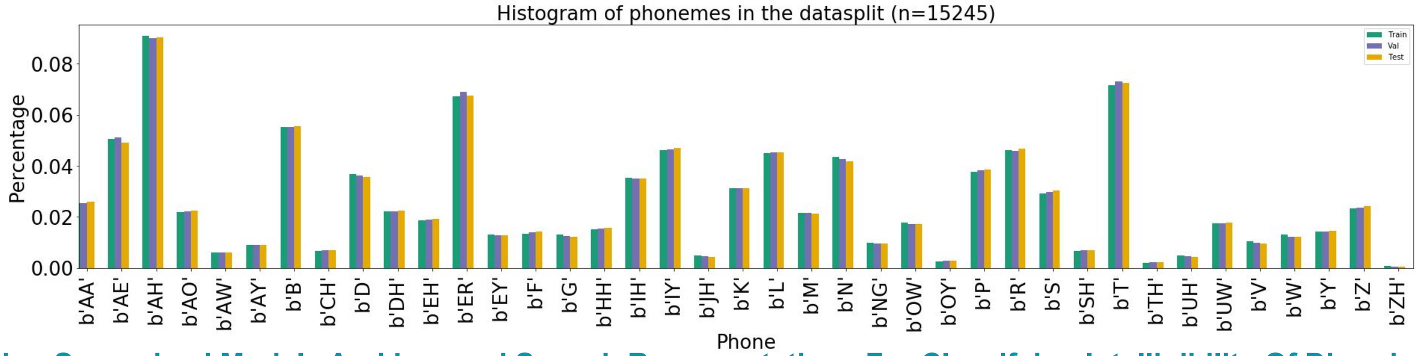
The Euphonia-SpICE dataset: Diverse etiologies



Previously - pilot study on Euphonia Quality Control data

```
'Buy Bobby a puppy.'                    'Sadder.'      'Banter.',
'I owe you a yo-yo today.'              'Chatter.'     'Shatter.'
'The police helped a driver.'           'Batter.'      'Tatter.',
'The boy ran down the path.'            'Meaner.'      'Patter.',
'The fruit came in a box.'              'Eater.'       'Ladder.'
'The shop closes for lunch.'            'Manner.'      'Bladder.'
'Strawberry jam is sweet.'              'Platter.'     'Banner.'
'Flowers grow in a garden.'             'Heater.'
'He really scared his sister.'    'She looked in her mirror.'
'The tub faucet was leaking.'     'A match fell on the floor.'
'He said buttercup, buttercup, buttercup, buttercup all day.'
'Bamboo walls are getting to be very popular because
     they are strong, easy to use, and good-looking.'
```

Euphonia- **Quality Control** dataset (29 phrases) with SLP-rated speech intelligibility.



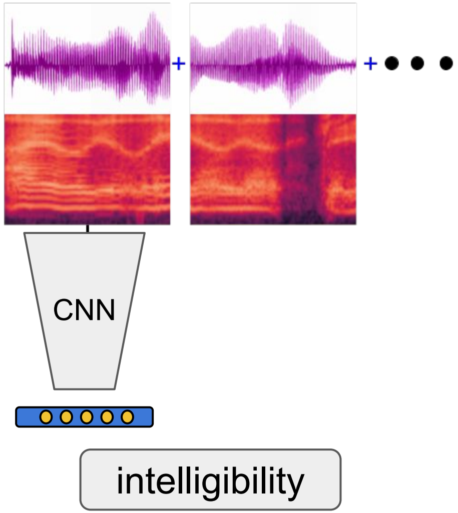
[Comparing Supervised Models And Learned Speech Representations For Classifying Intelligibility Of Disordered Speech On Selected Phrases](). S. Venugopalan, J. Shor, M. Plakal, J. Tobin, K. Tomanek, J. R. Green, M.P. Brenner. INTERSPEECH 2021

... and trained classifiers based on different approaches.

Google Research

Supervised CNN

Standard for audio classification [1]

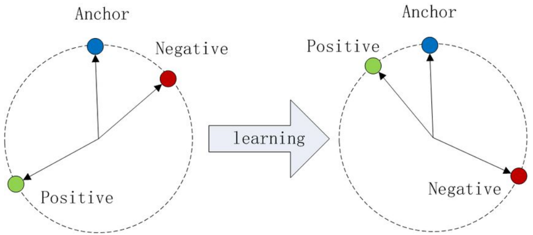


[1] Hershey et. al. CNN Architectures for Large-Scale Audio Classification ICASSP ‘17

Unsupervised representations

Classifiers on top of non-semantic speech representations (TRILL) [2]

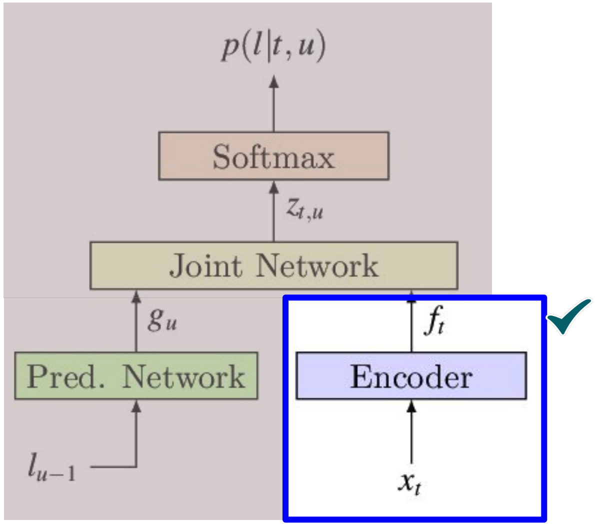
(Pre-training objective)
Triplet Loss



[2] Shor et. al. Towards Learning a Universal Non-Semantic Representation of Speech (TRILL) INTERSPEECH ‘20

ASR encoder representations

RNN-T model trained on typical speech [3]

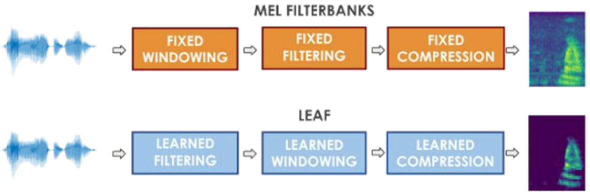


[3] Narayanan et. al. Recognizing longform speech in end-to-end models ASRU ‘19

This work - we wanted a public model competitive to ASR encoder

LEAF + CNN

Learnable frontend [4]



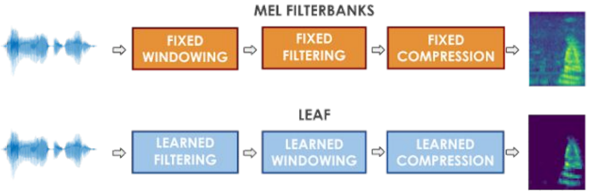
[4] LEAF: A Learnable Frontend for Audio Classification ICLR '21

Google Research

This work - we wanted a public model competitive to ASR encoder

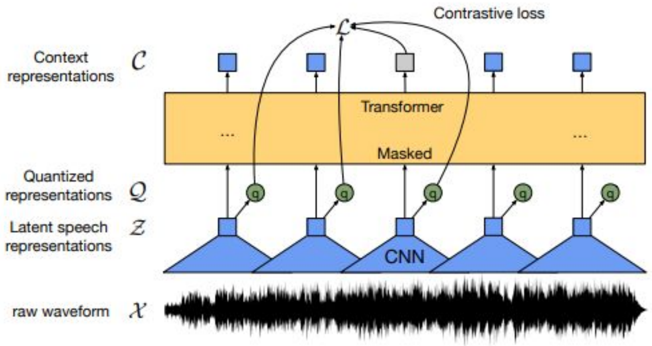
LEAF + CNN

Learnable frontend [4]



wav2vec2

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



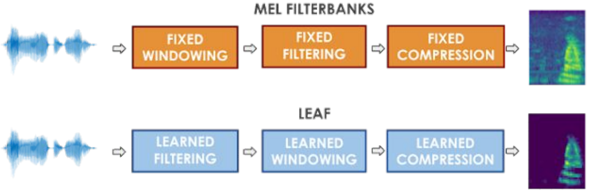
[4] LEAF: A Learnable Frontend for Audio Classification ICLR '21

[5] wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations NeurIPS '20

This work - we wanted a public model competitive to ASR encoder

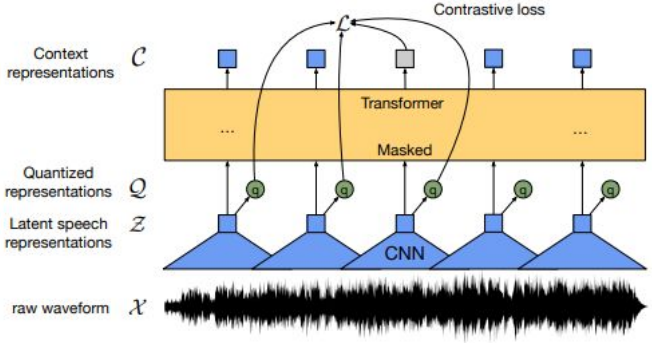
LEAF + CNN

Learnable frontend [4]



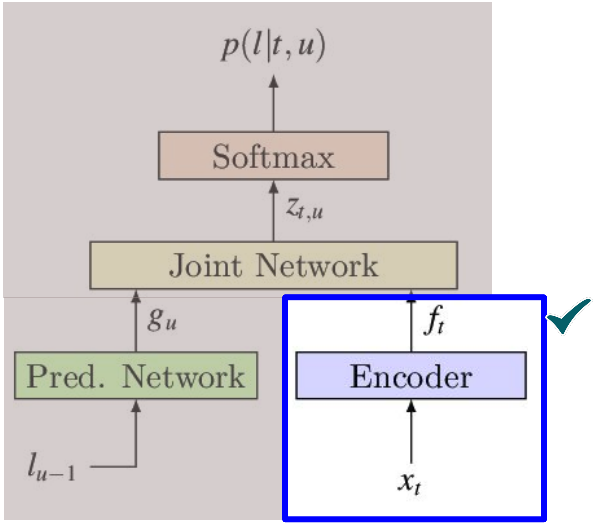
wav2vec2

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

Classification tasks and metrics

2 class MILD+: 0:{TYPICAL}, 1: {MILD, MODERATE, SEVERE, PROFOUND}

5 class classification tasks

AUC, F1 and Acc. as evaluation metrics

Google Research

Will the model generalize?

- Without any training
- On different datasets
- With different data collection processes
- Speakers with different etiologies
- Realistic speech setting

ASR-enc and SpICE wav2vec2 generalize “out-of-the-box”

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)

ASR-enc and SpICE wav2vec2 generalize "out-of-the-box"

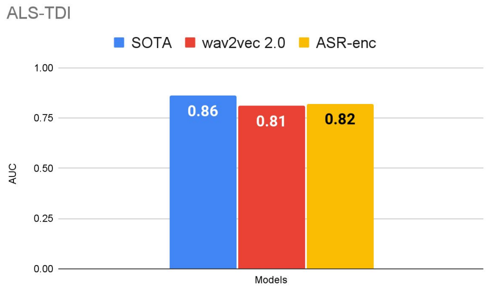
TORGO

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



ASR-enc and SpICE wav2vec2 generalize "out-of-the-box"

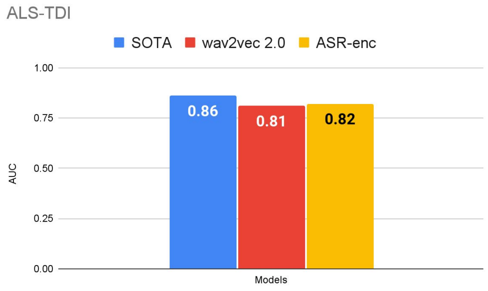
TORGO

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)

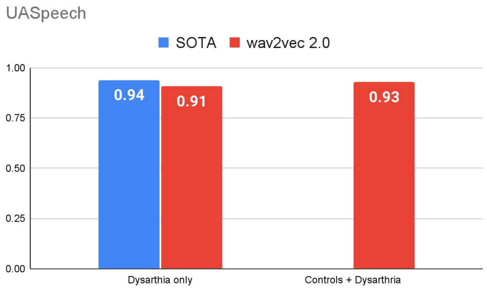
ALS-TDI

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



UASpeech

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



Google Research

Will the model generalize?

- ✅Without any training.
- ✅On different datasets
- ✅With different data collection processes
- Speakers with different etiologies
- Realistic speech setting

SpICE-V benchmark dataset

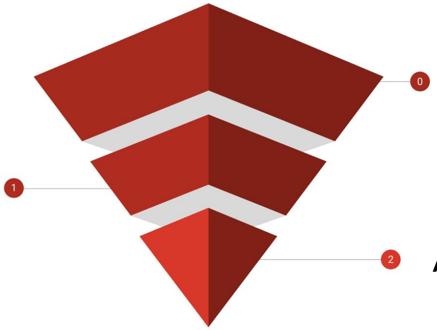
Google Research

SpICE-V data collection : 106 Dysarthric videos

1 Search to filter videos based on relevant topics.

2 Run *a different* binary classifier to tag “regions of interest” (ROIs)

ASR-enc trained additionally on Audio Set (0.5M non-speech and 0.6M typical speech utterances)



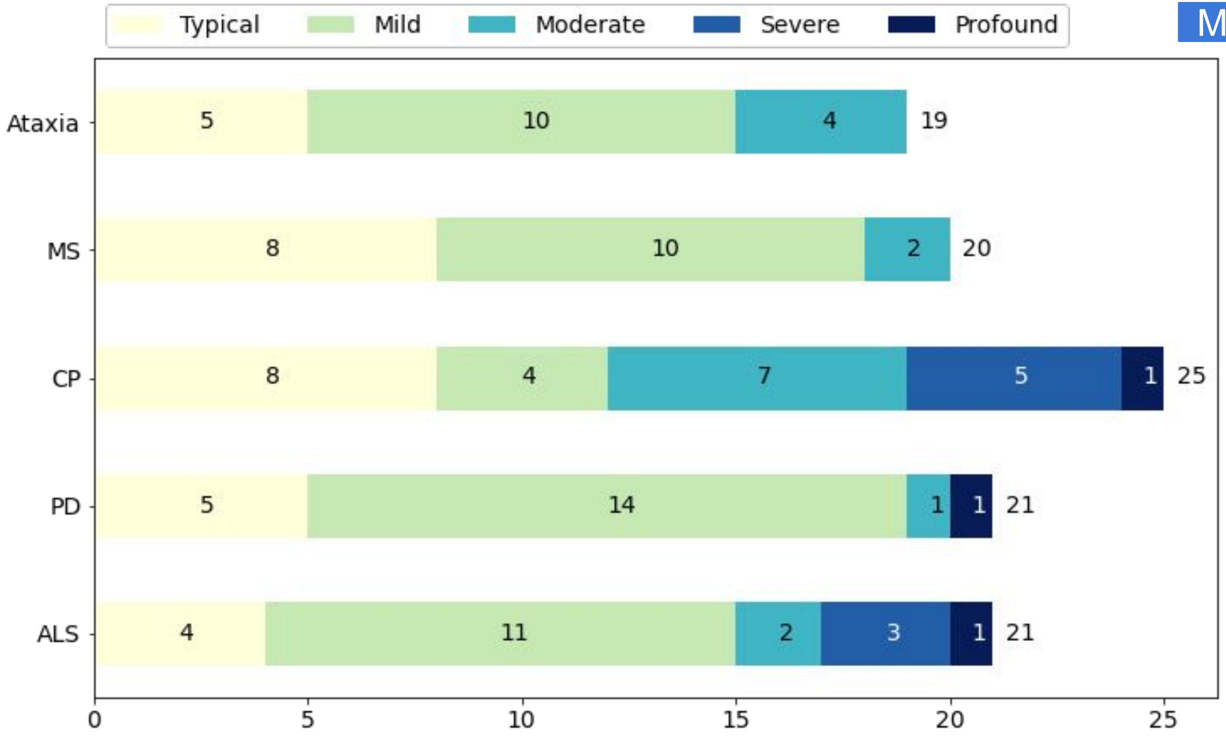
3 Further manual filtering. And SLPs tag/edit “regions of interest” (ROIs)

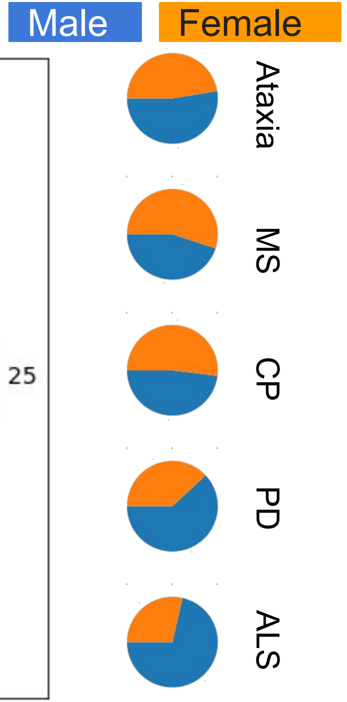
SLPs label

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

Google Research

SpICE-V distribution

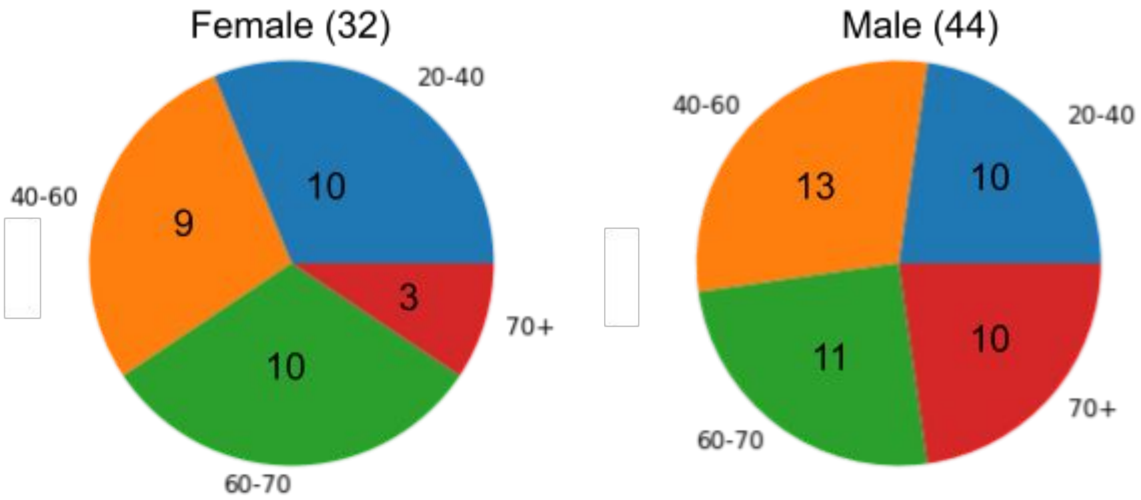




SpICE-V Controls: 76 speakers/videos

1. Select videos from AudioSet specifically the category tagged as “Speech”
2. We select from the unlabelled training set of 1M+ videos. Specifically only videos with tag
   a. Male speech, man speaking
   b. Female speech, woman speaking
   c. Optionally allowing for the tags “Narration, monologue” (and the tag speech)
   d. [detail] We looked at thumbnails of videos to determine - existence of video, confirmation of male/female speaker.
3. We watched the videos to infer age.
   a. We used the title and information tags in the video to look up speaker information as many of the speakers are somewhat public personalities e.g. sports persons, politicians featured heavily.
4. We tried to find as many videos of older people as we could.
   a. Intention to reduce bias of young adults and skew towards older age group and match gender.

SpICE-V Controls: 76 speakers/videos



Spice-V Results

Google Research

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

Google Research

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.

Model and usage

https://github.com/google-research/google-research/tree/master/euphonia_spice