INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Comparing Supervised Models And Learned Speech Representations For Classifying Intelligibility Of Disordered Speech On Selected Phrases

Subhashini Venugopalan[1], *Joel Shor*[2], *Manoj Plakal*[1], *Jimmy Tobin*[1], *Katrin Tomanek*[1], *Jordan R. Green*[3,4], *Michael P. Brenner*[1,4]

[1]Google Research, USA
[2]Google Research, Japan
[3]MGH Institute of Health Professions, USA
[4]Harvard University, USA
{vsubhashini,mbrenner}@google.com

Abstract

Automatic classification of disordered speech can provide an objective tool for identifying the presence and severity of a speech impairment. Classification approaches can also help identify hard-to-recognize speech samples to teach ASR systems about the variable manifestations of impaired speech. Here, we develop and compare different deep learning techniques to classify the intelligibility of disordered speech on selected phrases. We collected samples from a diverse set of 661 speakers with a variety of self-reported disorders speaking 29 words or phrases, which were rated by speech-language pathologists for their overall intelligibility using a five-point Likert scale. We then evaluated classifiers developed using 3 approaches: (1) a convolutional neural network (CNN) trained for the task, (2) classifiers trained on non-semantic speech representations from CNNs that used an unsupervised objective [1], and (3) classifiers trained on the acoustic (encoder) embeddings from an ASR system trained on typical speech [2]. We found that the ASR encoder's embeddings considerably outperform the other two on detecting and classifying disordered speech. Further analysis shows that the ASR embeddings cluster speech by the spoken phrase, while the non-semantic embeddings cluster speech by speaker. Also, longer phrases are more indicative of intelligibility deficits than single words.

Index Terms: atypical speech, classification, intelligibility, speech disorders

1. Introduction

Speech production requires the activation of dozens of muscles responsible for coordinating the respiratory, phonatory, resonatory, and articulatory speech subsystems. While speech disorders such as dysarthria are caused by neurological diseases such as amyotrophic lateral sclerosis (ALS) and Parkinson's disease (PD), impaired speech can also manifest from a wide variety of conditions such as cleft lip and palate, hearing loss, or stuttering. Regardless of the underlying etiology, speech problems are currently assessed using subjective ratings scales and observations. More objective approaches are, however, needed to address pressing clinical needs such as improving early detection, diagnostic accuracy, and clinical monitoring [3, 4]. Speech analytics that leverage machine classification are particularly well suited to address these diagnostic challenges. Recent reports have demonstrated, for example, the added value of speech analytics for early detection of speech decline in ALS [5], stratifying patients into fast or slow progressing groups [6], and documenting the response to drug interventions designed to slow the rate of disease progression [7].

Automatic classification of impaired speech could also play a critical role in reducing the performance gap in automatic speech recognition (ASR) between typical and disordered speech [8, 9]. Although ASR technologies could have a major impact on the quality of life in persons with severe speech impairments, these are the speakers for which ASR systems are most likely to fail [10]. Although, some techniques can improve performance on these speakers [11], none so far have bridged the gap. Poor performance has largely been attributed to lack of sufficient diversity in the training data [12]. Hence, automatic detection of disordered speech can enable large scale collection of hard-to-recognize speech samples that can help teach ASR systems about the variable manifestations of severely impaired speech.

Long-Short-Term-Memory(LSTM) [13] based models have been extensively used in the context of speech recognition, while classification of non-semantic audio, disordered speech, and in particular dysarthric speech, is often based on convolutional neural network (CNN) models [14] or machine learning models trained on handcrafted acoustic features [15, 16]. CNNs are often preferred for general audio and dysarthria classification [17] since their training process is (i) more suited to overcome the issue of limited training data available for such tasks, and (ii) better capture acoustic features associated with the underlying disorder as opposed to focusing on long-range content. More specifically, when training CNN models on audio, signals are often split into short segments (in our case we use 960ms) that may or may not overlap, and the label for the entire sample is propagated to each of the segments, thus boosting training data. Further, such short segments often lack sufficient context to capture semantic content, but are better suited to capture acoustic features that may exhibit characteristics of the disorder. However, many of the short segments may contain silence or noise or not exhibit characteristics of the impairment and may still require longer context. With clever techniques to increase training samples, there have also been LSTM-based dysarthria classifiers [18, 19, 20] that have performed well.

In this work we compare aspects of these different and commonly used deep learning techniques and study them in the context of automatically detecting intelligibility deficits and classifying intelligibility of disordered speech. Specifically, we compare the performance of 3 types of models:

- A CNN model similar to those used for audio classification [21], but directly trained to predict intelligibility ratings on our dataset of disordered speech samples.

http://dx.doi.org/10.21437/Interspeech.2021-1913

- Classifiers trained on non-semantic speech representations [1] learned by CNNs using an unsupervised objective. These have performed consistently well on several non-semantic speech tasks [1].
- Classifiers trained on representations extracted from the LSTM encoder of a production quality ASR model [2]. In particular, we consider the embeddings extracted from the encoder portion alone since it captures the acoustic features.

We compare these models on a labeled dataset of 15,246 disordered speech samples from 661 participants speaking 29 selected phrases. We analyze performance on different classification subtasks and across speaker cohorts of different intelligibility classes. Further, we study representations from the different models to compare their ability to discriminate between semantic content and intelligibility. We also examined if specific phrases were more indicative of intelligibility deficits.

2. Automatic classification of intelligibility

Early works on classification of disordered speech focused on handcrafted acoustic features [15, 16]. CNNs [17] and LSTMs [20] and other machine learning based models [22] have also been used to classify speech disorders. However, previous works have been based on smaller sets of speakers or etiologies than that studied here. Such works have also focused largely on discriminating dysarthric speech from typical speech e.g., from speakers with Parkinson's or ALS [17, 23, 24, 25] or dysarthria from apraxia [22].

In this work we focus particularly on detecting and classifying *intelligibility* of speech samples from speakers with a variety of self-reported speech disorders. Intelligibility measures how well speech is understood by a human listener [26]. In our dataset, each individual speaker is scored for intelligibility on a five-point Likert scale by speech-language pathologists (SLPs). We consider 3 different approaches to training classifiers to discriminate intelligibility. Our first model uses a CNN trained specifically to predict intelligibility. Our second approach also uses CNNs but pretrained with an unsupervised objective; we then build classifiers on embeddings extracted from the CNNs. Our third approach is LSTM based: we use representations learned by the LSTM encoder of an ASR system and build classifiers on those representations.

2.1. Convolutional neural network (CNN-ResNetish)

Our first approach is a convolutional network trained specifically for the classification task. Our CNN model is based on the ResNetish variant of the standard ResNet-50 architecture used for large scale audio classification [21]. The model is initialized with random weights. As is typical for CNNs for audio tasks, the model takes the spectrogram of the audio waveform as input. The speech sample (one recording) is split into non-overlapping 960ms segments that are then decomposed with a short-time Fourier transform (STFT) applied to 25ms windows every 10ms. The resulting spectrogram is integrated into 64 mel-spaced frequency bins, and the magnitude of the bins are log-transformed to obtain 96x64 log-mel spectrogram windows of the input. We train a CNN model on these spectrograms with a logistic loss for each score class. During training, all segments of an input are given the same label as the underlying speech sample. The model outputs a probability distribution for each score class and for each segment. We use the ADAM optimizer with a learning rate of 3e-5, mini-batch size of 64 segments (selected randomly across training speech samples) and train for 100 epochs. We choose the best performing model on the validation set. Since the label distribution across the score classes are different,we weight the loss for each segment inversely to the frequency of the class label at the segment level. During inference, we aggregate the segment level scores by taking a mean across all spectrogram segments to get a probability distribution of the scores over the entire speech sample. The argmax of these scores determines the final class prediction.

2.2. Non-semantic speech representations (TRILL)

For our second approach, we use representations learned by a convolutional network that is trained with an unsupervised objective. Specifically, we consider the TRILL and TRILL-distilled models [1]. The backbone architecture of TRILL is also based on the ResNetish [21] CNN model, whereas the TRILL-distilled model is based on the more efficient MobileNet [27]. These models also take input spectrogram context-windows from 960ms segments of audio/speech.

Unlike the fully-supervised training objective, the TRILL models are trained using an unsupervised triplet loss. They consider 3 segments, where 1 segment is the anchor, a second segment which comes from the same audio clip forms the positive example, and a third segment from a different audio clip constitutes the negative example. The model is trained with a hinge loss such that the distance (L_2) between the representations of the anchor and negative example is larger than the distance between the representations of the anchor and positive example by at least some margin. The TRILL models were trained on a subset of AudioSet [28] training set clips possessing the "speech" label. We consider the pre-ReLU output of the first 512-depth convolutional layer in ResNetish (referred to as TRILL layer19), and the penultimate average pool layer in MobileNet (referred to as TRILL-distilled in [1]). TRILL layer19 yields a 12288-d embedding, and TRILL-distilled yields a 2048-d embedding.

For each model, we extract embeddings for each 960ms segment of the speech sample, and consider the average-pooled (over time) embedding as the representation of the speech sample. We then train a Logistic Regression model on the embeddings to predict intelligibility, using multinomial loss (identical to [1]) to get a distribution over classes. Additionally, we also evaluate the performance of a balanced Random Forest with the embeddings as features, implemented by scikit-learn [29]

2.3. ASR system encoder representations (ASR-enc)

Our third set of models are based on the representations from an LSTM encoder that models acoustic inputs in a production quality ASR system. The ASR system we consider is a recurrent neural network transducer (RNN-T) [30] model. Specifically, the architecture is based on He et. al. [31] and has been trained with modifications on more diverse acoustic data to improve long-form speech recognition [2].

The RNN-T model consists of an *encoder* and a *prediction* network that use LSTMs, and a *joint network* that uses feed-forward layers. In this work, we only consider the *encoder* which models the acoustic inputs. The *encoder* consists of 8 unidirectional LSTM layers. The acoustic frontend is modeled using 128-dimensional log-mel filterbank energies computed on 32ms windows with a 10ms hop. Features from 4 contiguous segments are stacked, and then sub-sampled by a factor of 3. This is input to the encoder network LSTMs which have 2048 units each. The outputs are then projected down to 640 units after each layer. The encoder network also uses a time-reduction

layer after the 2^{nd} LSTM layer, which stacks output features from 2 contiguous timesteps and subsamples them by a factor of 2; the final encoder outputs are 640-d and at a 60ms frame rate.

Similar to our treatment of the TRILL models, we consider the average-pooled (over time) embeddings as the final representation of the speech sample, and train logistic regression and random forest models on the embedding features to predict class scores.

3. Experiments

We now describe the dataset, the classification tasks and objectives, and the evaluation metrics used in our experiments.

3.1. Dataset of speakers and selected phrases

Our dataset is a human-rated small subset of [32]. We use data from 661 speakers with a diverse set of self-reported speech disorders depicted in Fig. 1, including Down syndrome (26.8%), ALS (24.6%), Cerebral palsy (13.3%), Parkinson's disease (7.6%), hearing impairments (4.2%) and others. The speakers each record 29 distinct words/phrases (Fig. 2) that have a mixed distribution of phonemes (Fig. 3). Speech Language Pathologists (SLPs) listened to the recordings for each speaker and assessed the overall intelligibility of the speaker on a five-point Likert scale. The scale was mapped to 5 classes - *typical, mild, moderate, severe,* and *profound*. We then split the speakers randomly into training, validation (val.) and test sets, with a distribution of 70:15:15. All our models were trained on the same splits. The number of speakers and utterances in each split for each label, along with the overall count is shown in Tab. 1.

Table 1: *Count of speakers and utterances in the data splits.*

Intelligibility	# speakers			# utterances		
	Train	Val.	Test	Train	Val.	Test
TYPICAL	160	30	23	3,875	734	544
MILD	153	35	36	3,343	817	788
MODERATE	87	25	18	1,969	567	471
SEVERE	54	12	14	1,113	316	388
PROFOUND	10	1	3	224	9	87
OVERALL	464	103	94	10,524	2,443	2,278

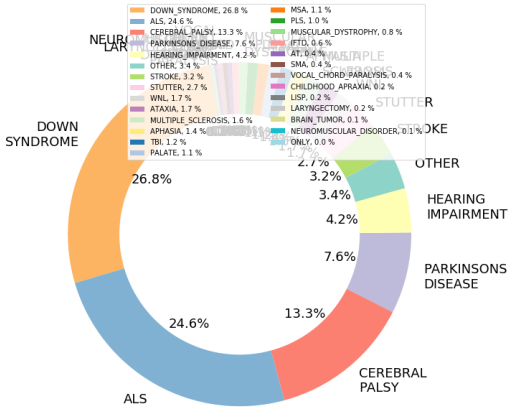

Figure 1: *Distribution of etiologies in the dataset.*

3.2. Classification objectives

We train all our models on 4 different classification tasks based on the intelligibility ratings. Since the ratings are for each speaker, all utterances from a speaker have the same rating.

5-class This task is the 5-way classification task of labeling each utterance with one of the 5 ratings of *0-typical, 1-mild, 2-moderate, 3-severe,* or *4-profound*.

3-class This task is a 3-way classification task where *typical* and *mild* are each a separate class of their own, but ratings of *moderate, severe* and *profound* are grouped into a single class. Hence, the label for each utterance corresponds to one of 3 ratings *0-typical, 1-mild, 2-(moderate, severe,* or *profound)*.

```
'Buy Bobby a puppy.'
'I owe you a yo-yo today.'
'The police helped a driver.'
'The boy ran down the path.'
'The fruit came in a box.'
'The shop closes for lunch.'
'Strawberry jam is sweet.'
'Flowers grow in a garden.'
'He really scared his sister.'
'The tub faucet was leaking.'
'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.'
```

```
'Sadder.'    'Banter.',
'Chatter.'   'Shatter.'
'Batter.'    'Tatter.',
'Meaner.'    'Patter.',
'Eater.'     'Ladder.'
'Manner.'    'Bladder.'
'Platter.'   'Banner.'
'Heater.'

'She looked in her mirror.'
'A match fell on the floor.'
```

Figure 2: *List of 29 words and phrases recorded by speakers.*

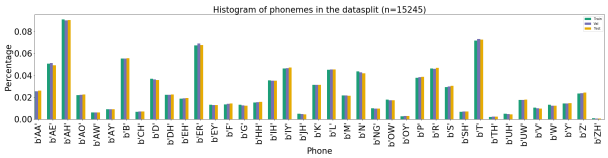

Figure 3: *Distribution of phonemes in the datasplits. x-axis represents phonemes and y-axis the percentage. (Best viewed in high resolution)*

2-class MODR+ This task is a 2-way classification task where *typical* and *mild* are grouped into a single class and *moderate, severe, profound* are grouped into another.

2-class MILD+ This is a 2-way classification task where *typical* is a class of its own and *mild, moderate, severe, profound* are grouped into the other class.

3.2.1. Evaluation metrics.

We report performance of our models on 3 evaluation metrics. For all models we re-sample recordings to 16 kHz mono audio.

1-vs-rest AUC (AUC) This computes the Area Under the Receiver Operating Characteristic Curve for each class vs the rest.

F1 score (F1) This is the balanced F-score which is the weighted average of the precision and recall of the model.

Accuracy (Acc.) Accuracy simply measures the number of correctly predicted samples over the total number of samples.

4. Results

The performance of all our models across the different classification tasks is reported in Tab. 2. For the TRILL and ASR-enc-embeddings based models, we only show performance of the logistic regression model, which was consistently better than random forest for all representations and all classification tasks.

ASR-enc representations perform best. Surprisingly, the simpler classification models trained on the representations from the ASR *encoder* achieve the best performance across all tasks. While we might expect the supervised CNN (CNN-ResNetish) model trained explicitly for the task to perform well, its performance is a distant second, particularly in terms of mean AUC. The models based on ASR representations also maintain a high mean AUC on the 3-class and 5-class tasks.

Table 2: *We report the mean 1-vs-rest AUC values, F1 score, and accuracy (Acc.). Higher is better. bold indicates highest value.*

Models	2-class MILD+			2-class MODR+			3-class			5-class		
	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
CNN-ResNetish	0.761	0.761	0.757	0.706	0.671	0.673	0.679	0.499	0.498	0.677	0.341	0.405
TRILL (layer 19)	0.723	0.708	0.709	0.631	0.643	0.650	0.638	0.445	0.448	0.603	0.373	0.387
TRILL-distilled	0.717	0.729	0.732	0.654	0.645	0.654	0.627	0.443	0.444	0.582	0.367	0.381
ASR-enc	**0.820**	**0.776**	**0.776**	**0.812**	**0.754**	**0.763**	**0.749**	**0.544**	**0.544**	**0.771**	**0.448**	**0.459**

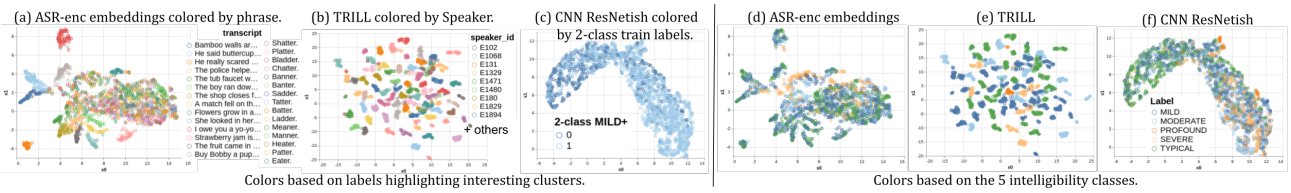

Figure 4: *UMAP clusters of model representations. Samples are colored based on specific labels to highlight interesting clusters: (a) ASR embeddings cluster by phrase, (b) TRILL clusters by speakers, and (c) CNN-ResNetish clusters on training class labels. (d), (e) and (f) depict coloring based on the 5 intelligibility class labels for the 3 representations. (Best viewed in color and high-res.)*

However, in all tasks and for all models, there appears to be much room for improvement, since the best scores are around 0.8 for the 2-way classification tasks.

2-way tasks are simpler for all models. All models perform substantially better on the 2-class MILD+ and 2-class MODR+ classification tasks than the more fine-grained 3-class and 5-class classification tasks. Also, between the MILD+ and MODR+ tasks, models consistently perform better on the MILD+ task. One might have expected that the task of discriminating *moderate* or above rating (MODR+) is easier than the (MILD+) task, since distinguishing mildly impaired speech from typical speech is more nuanced than identifying speech that is considerably impaired. The result indicates that the models can detect mild disordered speech, but this might be an indicator that the range of mild intelligibility scored by the speech-language pathologists is quite large.

5. Analysis

Performance by intelligibility group: Table 3 compares the performance of the 2-class MILD+ classifiers based on ASR encoder embeddings, TRILL (layer 19) and the CNN-ResNetish model grouped by intelligibility class labels. Utterances from speakers labeled moderate, profound or severe are easier to discriminate, and those labeled typical and mild are the hardest to discriminate.

Table 3: *Performance of 2-class MILD+ models grouped by intelligibility class. Scores increase with severity. Typical and mild are hardest to discriminate.*

2-class MILD+	ASR-enc-embs.		TRILL (l.19)		CNN-ResNetish	
	F1	Acc.	F1	Acc.	F1	Acc.
TYPICAL	0.766	0.621	0.659	0.491	0.706	0.546
MILD	0.860	0.755	0.880	0.786	0.882	0.788
MODERATE	0.981	0.963	0.944	0.894	0.899	0.817
SEVERE	0.959	0.921	0.904	0.825	0.934	0.876
PROFOUND	1.000	1.000	0.977	0.954	0.976	0.954

Longer phrases are more indicative of intelligibility: Next, we examine the performance at the phrase level for the best model i.e., ASR-enc-embeddings logistic regression model on the 2-class MILD+ task. The top two phrases with highest AUC for the model are: (1) "He said buttercup, buttercup, buttercup, buttercup all day." This had (AUC: 0.901, F1: 0.849, accuracy=0.851), and (2) "Bamboo walls are getting to be very popular because they are strong, easy to use, and good-looking." (AUC:0.88, F1: 0.793, accuracy: 0.8). In contrast, the phrases with the lowest scores consist of single words, with the lowest scoring words (1) "Tatter." (AUC: 0.624, F1: 0.661, accuracy: 0.671), and (2) "Chatter." (AUC: 0.764, F1: 0.744, accuracy: 0.746). Irrespective of the actual content longer phrases had better performance than single words, indicating that more signal/context could help.

Comparing representations on speaker, intelligibility and content Finally, Figure 4 shows clusters using UMAP [33] on representations from TRILL (layer 19), ASR-enc embeddings, and CNN-ResNetish (from the 2048-d fully connected layer of the 2-class MILD+ model). In the ASR-enc based embeddings (Fig. 4a), longer sentences have clearer clusters, whereas single words are more spread out. In Fig. 4d, we can observe that utterances by speakers labeled severe and profound (orange, yellow) tend to cluster closer towards center, likely because of their poor intelligibility. In contrast, the TRILL embeddings (Fig. 4b) cluster based on the identity of the speaker, and not so much by content (not depicted) or intelligibility class (Fig. 4e). The CNN-Resnetish model's embedding (Fig. 4c) appears to cluster more based on the training labels (here on the 2-class MILD+ labels) and doesn't seem to have any observable pattern on the full scale of 5 classes (Fig. 4f).

6. Conclusion

To conclude, we have demonstrated that deep learning models can classify speech samples from impaired speakers, with classifiers based on embeddings from state of the art ASR models outperforming task specific CNNs. The models are able to classify the degree of intelligibility in speakers from a wide range of etiologies. Future directions include comparing the performance of these models to SLPs, as well as assess whether such classification can be carried out on more general speech. On the modeling side, we wish to consider alternative learned frontends (e.g., LEAF [34]) and representations from wav2vec [35], as well as finetuning the speech representations from TRILL, or the ASR systems on atypical speech training data.

7. References

[1] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. A. Haviv, "Towards learning a universal non-semantic representation of speech," *ArXiv*, vol. abs/2002.12764, 2020.

[2] A. Narayanan, R. Prabhavalkar, C. Chiu, D. Rybach, T. N. Sainath, and T. Strohman, "Recognizing long-form speech using streaming end-to-end models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[3] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in als: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, pp. 494–500, 2013.

[4] Y. Yunusova, N. L. Graham, S. Shellikeri, K. Phuong, M. Kulkarni, E. Rochon, D. F. Tang-Wai, T. W. Chow, S. E. Black, L. H. Zinman *et al.*, "Profiling speech and pausing in amyotrophic lateral sclerosis (als) and frontotemporal dementia (ftd)," *PloS one*, vol. 11, no. 1, p. e0147573, 2016.

[5] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, "The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to als," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 18, 2017.

[6] P. Rong, Y. Yunusova, M. Eshghi, H. P. Rowe, and J. R. Green, "A speech measure for early stratification of fast and slow progressors of bulbar amyotrophic lateral sclerosis: Lip movement jitter," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 21, no. 1-2, pp. 34–41, 2020.

[7] J. R. Green, K. M. Allison, C. Cordella, B. D. Richburg, G. L. Pattee, J. D. Berry, E. A. Macklin, E. P. Pioro, and R. A. Smith, "Additional evidence for a therapeutic effect of dextromethorphan/quinidine on bulbar motor function in patients with amyotrophic lateral sclerosis: A quantitative speech analysis," *British journal of clinical pharmacology*, vol. 84, 2018.

[8] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6470–6474.

[9] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[10] L. De Russis and F. Corno, "On the impact of dysarthric speech on contemporary asr cloud platforms," *Journal of Reliable Intelligent Environments*, vol. 5, 09 2019.

[11] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, and Y. Matias, "Personalizing ASR for Dysarthric and Accented Speech with Limited Data," in *Proc. Interspeech 2019*.

[12] M. Moore, H. Venkateswara, and S. Panchanathan, "Whistle-blowing asrs: Evaluating the need for more inclusive speech recognition systems," *Interspeech 2018*, 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[14] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "Gmm and cnn hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial informatics*, vol. 14, 2018.

[15] J. H. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on biomedical engineering*, vol. 45, no. 3, pp. 300–313, 1998.

[16] L. Baghai-Ravary and S. W. Beet, *Automatic speech signal analysis for clinical diagnosis and assessment of speech disorders*. Springer Science & Business Media, 2012.

[17] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks," *arXiv preprint arXiv:2011.07545*, 2020.

[18] A. Mayle, Z. Mou, R. C. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks." in *INTERSPEECH*, 2019, pp. 4514–4518.

[19] M. J. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional lstm neural network." in *INTERSPEECH*, 2018, pp. 2948–2952.

[20] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *ICASSP 2019*.

[21] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[22] I. Kodrasi, M. Pernon, M. Laganaro, and H. Bourlard, "Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech," *arXiv preprint arXiv:2011.07542*, 2020.

[23] J. C. Vásquez-Correa, R. Castrillón, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth, "Speaker model to monitor the neurological state and the dysarthria level of patients with parkinson's disease," in *International Conference on Text, Speech, and Dialogue*. Springer, 2017, pp. 272–280.

[24] K. An, M. J. Kim, K. Teplansky, J. R. Green, T. F. Campbell, Y. Yunusova, D. Heitzman, and J. Wang, "Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks." in *Interspeech*, 2018.

[25] E. Vaiciukynas, A. Gelzinis, A. Verikas, and M. Bacauskiene, "Parkinson's disease detection from speech using convolutional neural networks," in *International Conference on Smart Objects and Technologies for Social Good*. Springer, 2017, pp. 206–215.

[26] K. L. Stipancic, Y. Yunusova, J. D. Berry, and J. R. Green, "Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 61, 2018.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[28] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[30] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[31] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, 2019.

[32] R. L. MacDonald *et al.*, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia," in *Proc. Interspeech 2021*.

[33] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[34] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[35] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.