Deep Acoustic Embeddings for Identifying Parkinsonian Speech

Zafi Sherhan Syed[1], Sajjad Ali Memon[2], Abdul Latif Memon[3]
Mehran University
Pakistan

***Abstract*—Parkinson's disease is a serious neurological impairment which adversely affects the quality of life in individuals. While there currently does not exist any cure for this disease, it is well known that early diagnosis can be used to improve the quality of life of affected individuals through various types of therapy. Speech based screening of Parkinson's disease is an active area of research intending to offer a non-invasive and passive tool for clinicians to monitor changes in voice that arise due to Parkinson's disease. Whereas traditional methods for speech based identification rely on domain-knowledge based hand-crafted features, in this paper, we investigate the efficacy of and propose the deep acoustic embeddings for identification of Parkinsonian speech. To this end, we conduct several experiments to benchmark deep acoustic embeddings against handcrafted features for differentiating between speech from individuals with Parkinson's disease and those who are healthy. We report that deep acoustic embeddings consistently perform better than domain-knowledge features. We also report on the usefulness of decision-level fusion for improving the classification performance of a model trained on these embeddings.**

***Keywords*—Affective computing; deep acoustic embeddings; Parkinson's disease; social signal processing**

I. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder caused by decay of neurons in the area of the brain which controls body movements [1]. It manifests as muscle rigidity, slowness of body movement, compromised gait, and involuntary shaking amongst other symptoms. Individuals with Parkinson's disease also suffer from vocal impairments such as impoverished speech prosody, hoarse voice quality, and imprecise articulation [2].

According to a handbook by the World Health Organization on public health challenges caused by neurological disorders [3], Parkinson's disease contributes approximately 2% of the total global burden of diseases. In terms of disability-adjusted life year (DALY) score, a commonly used metric which quantifies the number of years lost due to ill-health, the burden of PD is on a rise, with the number of DALYs increasing from 1,617,000 in 2005 to 1,762,000 in 2015, and is expected to increase up to 2,015,000 by the year 2030. Parkinson's disease does not currently have a cure and improving the quality-of-life of patients is of prime importance. According to Yousefi et al. [4], early detection can help improve the patients' quality of life through physiotherapy, mental health counseling, and in some cases surgery. There also do not exist specific tests for diagnosis of Parkinson's disease, therefore, patients are diagnosed by trained clinicians based on common signs and symptoms for the disease through physical and neurological examination as well as medical history [5]. It is common to recommend brain imaging tests to patients for differential diagnosis in order to rule diseases than Parkinson's [6], [7]. While brain imaging has indeed been a successful tool, it is invasive. Such tests also require patients to visit special facilities, which may not be convenient for the elderly.

Recently, there has been a growing interest in developing voice-based screening tools that can identify patients with Parkinson's disease based on the characteristics of their voice alone. For example, Tsanas et al. [8] showed that speech-based tools can be used to recognize the progression of Parkinson's in a telemonitoring setup. Traditional methods for speech based identification of Parkinson's disease mostly rely on domain-knowledge based hand-crafted features [9], [10], [11], [12]. However, advances in the field of natural language processing have shown that embeddings from pre-trained deep neural networks often perform better than hand-crafted features. To this end, we investigate the efficacy of deep acoustic embeddings generated from pre-trained deep neural works for the task of automated identification of Parkinsonian speech. Whilst using domain-knowledge based features to create a baseline classification performance, we show that these embeddings can achieve a better classification performance than those hand-crafted features. Moreover, we also show that upstream training tasks for these embeddings are not a limiting factor for its downstream task of speech paralinguistics. Finally, we report that decision-level fusion is an effective method to improve the classification performance of machine learning models trained to identify Parkinsonian speech.

The rest of the paper is structured as follows: In Section II, we introduce the concept of deep acoustic embeddings for the task at hand and briefly describe the four deep neural models used in our work. In Section IV, we discuss the methodology followed for data-driven analysis. In Section V, we report the results of experiments and provide a discussion for each aspect of the experimentation. A conclusion of our work is provided in Section VI and supplementary data to support our work is provided as appendices.

II. Deep Acoustic Embeddings

A major limitation of domain knowledge based hand-crafted features is that they are narrow in scope and often require subject expertise in order to be used in the correct context. For example, while Mel Frequency Cepstral Coefficients (MFCCs) is a popular acoustic feature for representing cepstral characteristics of audio signals, it is still represented by a relatively small number of coefficients as compared to say a Mel spectrogram, which offers a rich time-frequency representation of an audio signal. Deep neural network models

for applications based on audio modality are trained to learn useful features from Mel spectrogram representations of audio signals, similar to how deep learning based image classification models learn to recognize useful features from an image. In the case of audio, spectrograms serve as images. The word deep acoustic embeddings refer to features learned from deep neural networks that are trained for audio applications. These embeddings are typically extracted from the penultimate layer of the model meaning that these features are representative of the characteristics of audio signals that were used to train the models.

In our work, we experiment with embeddings generated from four deep neural networks which were optimized for applications based on audio modality. These are VGGish, YAMNet, openl3: Music, and openl3: Environment sounds.

A. VGGish Embeddings

The VGGish is a deep convolutional neural network proposed by Hershey et al. [13] for large scale audio classification of Youtube videos based on their audio content. As the name suggests, VGGish is based on the famous VGGNet [14] which was once the state-of-art model for image classification and remains amongst performing models in computer vision. VGGish's network architecture consists of four blocks, each with convolutional kernels and maxpooling layers, which serve as a feature extractor. These are followed by two fully connected layers that serve as the classifier.

The VGGish was trained on an initial version of AudioSet corpus [15] consisting of more than 2 million video clips from Youtube which were manually annotated into 527 categories. Examples of these categories include male/female adult voice, infant babbles, animal sounds, and sounds produced by various types of machines. VGGish was envisaged as a model that can learn a meaningful representation of audio signals for these classification categories but in our recent work, we showed it to be useful for speech paralinguistic tasks such as identification of Alzheimer's dementia [16].

As with most deep neural networks that generate deep acoustic embeddings, the VGGish accepts spectrogram-based representation of audio clips. To begin, each audio clip is segmented into chunks of 1 second in duration and a Mel spectrogram is computed over short-time frame duration of 25ms and frame hop-duration of 10ms whilst using Mel frequency shaping filter with 96 bins. Our objective is to only compute deep acoustic embeddings from the VGGish model (and not to classify our dataset into 527 classes of the AudioSet corpus), we take the output from the model before the final classifier. With spectrograms as input, the VGGish model is therefore used as a feature extractor that produces 128-dimensional embeddings as its output. An illustration of this workflow is provided in Figure 1. We posit that these semantically relevant embeddings are useful for our downstream task of Parkinsonian speech classification. A pre-trained model for VGGish has been made available by Google for academic research [1] and we make use of this model in our work.

[1]https://github.com/tensorflow/models/tree/master/research/audioset/vggish

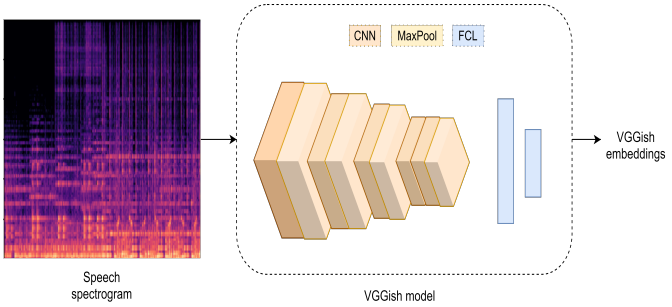

Fig. 1. Illustration of Feature Extractor for VGGish Embeddings

B. YAMNet Embeddings

A major drawback of the VGGish model is that with more than 72 million parameters, it has high computational complexity. This inhibits the use of VGGish in most applications that are based on mobile embedded systems. The YAMNet model was developed by Ellis and Chowdhry [2]. as a computationally efficient model for classification of audio events for the AudioSet corpus. It is based on the MobileNet architecture proposed earlier by Howard et al. [17] that used depth-wise separable convolutional kernels to create lightweight models that can be used for mobile and embedded vision applications. As a result, the YAMNet model has 4.7 million parameters versus the 72 million required for VGGish.

The network architecture of YAMNet model consists of 14 blocks of convolutional layers where all except the first layer are based on depth-wise convolutional kernels. While there are no differences between VGGish and YAMNet models in terms of input spectrograms, there are some differences in terms of training data: the YAMNet is trained with the larger AudioSet corpus but it has a slightly smaller number of classes (521 versus 527 for VGGish) since some classes were removed from AudioSet corpus due to ethical considerations. Therefore, it is not possible to compare, in a fair manner, the classification performance of VGGish and YAMNet models on the AudioSet corpus. Nevertheless, we shall compare their classification for recognition of Parkinsonian speech through data-driven analysis and report results in Section V.

C. openL3 Embeddings for Music and Environmental Sounds

In addition to deep acoustic embeddings generated from VGGish and YAMNet models, which are trained to recognize human voice amongst various other types of audio events, we also make use of embeddings generated from openl3 models [18] [3], which are optimized to identify types of music and environmental sounds. Our motivation to use these embeddings is to investigate whether the upstream training task matters for downstream classification performance for deep acoustic embeddings. More specifically, we seek to answer whether embeddings from models trained to recognize music and environment sounds can be used to recognize characteristics of speech paralinguistics which are present in Parkinsonian speech.

[2]https://github.com/tensorflow/models/tree/master/research/audioset/YAMNet
[3]https://github.com/marl/openl3

The openl3 embeddings are based on the Look, Listen, and Learn (L3) concept which was proposed by Arandjelovic et al. [19] for training neural networks to learn meaningful audio representations in a self-supervised manner through audio-visual correspondence tasks. Their approach seeks to alleviate stringent requirements for manual annotation of training data. Cramer et al. [18] developed openl3 models as an extension to the work from [19] and investigated various network architecture choices, such as the choice between short-time-Fourier transform based spectrograms or Mel frequency scaled spectrograms and a different size for deep acoustic embeddings. It is important to mention here that the video recordings used to train openl3 models were also curated from within the AudioSet corpus, however, the scope of these models is much smaller than VGGish and YAMNet which seek to classify between all available classes in the AudioSet corpus. The network architecture for openl3 models consists of four blocks of convolutional layers which are used as feature extractors from spectrograms that are fed to the model as input. MaxPooling operation is performed to the output of feature extractor with an option to yield either an embedding of size 512 or an embedding of size 6144.

III. Dataset

We make use of the GITA corpus of Parkinsonian speech in our experiments which was published by Orozco-Arroyave et al. [20] as part of work carried out at Applied Telecommunications Group (GITA) at Universidad de Antioquia, Colombia. The GITA corpus is one of the most prominent publicly available datasets on Parkinsonian speech. It consists of speech recordings from 100 native speakers of Spanish language amongst whom 50 subjects were diagnosed with Parkinson's disease as per the Unified Parkinson's Disease Rating Scale (UPDRS) scale [21]. These subjects were matched in terms of age and gender with their respective healthy counterparts. The GITA corpus consists of three main speaking tasks which vary in duration and phonetic content. The first speaking task is based on diadochokinetic non-words, short-duration utterances based on the pronunciation of words and phrases, and long duration utterances which require subjects to read out a section of text or monologue. A summary of statistics for time durations for each task is provided in Table I. For a more thorough description of the dataset such as the age and gender distribution of subjects, we refer the reader to [20].

TABLE I. Summary of time-duration statistics for each speaking task in the GITA corpus

Speaking Task		*Time Duration Statistics*		
Category	**Task**	**Min**	**Avg**	**Range**
DDK	ka	0.81	2.76	8.02
	pa	0.76	2.94	6.77
	pakata	1.39	4.16	7.63
	pataka	1.39	4.38	9.13
	petaka	1.10	4.14	7.70
	ta	0.77	2.92	8.36
Short-duration	juan	1.81	3.21	4.48
	laura	1.27	2.18	2.65
	loslibros	1.84	3.43	5.64
	luisa	2.21	4.01	7.42
	micasa	1.20	1.96	2.35
	omar	1.56	2.65	3.38
	preocupado	2.30	4.32	6.16
	rosita	2.61	4.36	5.58
	triste	1.74	3.26	4.34
	viste	4.81	7.88	16.90
Long-duration	readtext	10.35	18.12	34.91
	monologue	14.10	47.11	149.99

IV. Methodology

The process flow diagram for automated identification of Parkinsonian speech is illustrated in Fig. 2. Here, one starts with raw audio recordings from subjects that are preprocessed into a standard format used in audio signal processing (16 KHz sampling rate, mono-channel, and amplitude normalized between +/-1). The next step is to compute domain-knowledge features such as ComParE and eGeMAPS.

In order to compute deep acoustic embeddings such as VGGish, YAMNet, openl3:Music, and openl3:Environment Sounds, a Mel spectrogram based representation is generated for the recording and passed down to feature extractors based on these models (details of these models were provided in Section II). Since deep acoustic embeddings are computed over chunks of audio recordings, these embeddings need to be summarized in order to generate a global representation for the audio recording. In this work, we use three functionals of descriptive statistics, namely, average, maximum, and range to pool a global feature vector for deep acoustic embeddings.

Finally, given the relatively small number of examples per speaking task (50 each for subjects with Parkinson's disease and those who are healthy), we conduct experimentation through leave-one-subject-out (LOSO) cross-validation. In each iteration, 99 examples are used to form the training set and the one remaining example is used for testing. Therefore, in total, the classification performance with each acoustic feature (domain-knowledge based as well as deep acoustic embeddings) is computed over 100 examples. We train a logistic regression classifier to differentiate between features from healthy and Parkinson's disease groups. Logistic regression has a hyper-parameter called *complexity* which needs to be optimized in order to tune its classification performance. To this end, we integrate hyper-parameter optimization within the LOSO cross-validation (we found this strategy to be successful in [16]) and optimize complexity over a logarithmically spaced grid between 10^{-7},10^{-6},...,10^{3}.

In subsequent paragraphs, we provide details of domain-knowledge and deep acoustic embedding based features which are used in our experiments.

A. Domain-knowledge based Handcrafted Features

Evidence suggests that muscular dystrophy, where muscles shrink and weaken, due to Parkinson's disease causes aphasia, which leads to changes in paralinguistic characteristics of speech in terms of prosody, voice quality, and voice spectra [22], [23], [24], [25], [26]. Prosody is defined as the intonation or melodic contour speech and abnormal prosody is a recognized marker of individuals with Parkinson's disease [25]. Voice quality describes the degree of hoarseness, breathiness, or tenseness of voice [22]. It is known that muscular tightening causes glottis to function improperly which leads to poorer quality of voice for those with Parkinson's disease as compared to individuals who are healthy [11], therefore, acoustic features which quantify voice quality can be useful for the task of

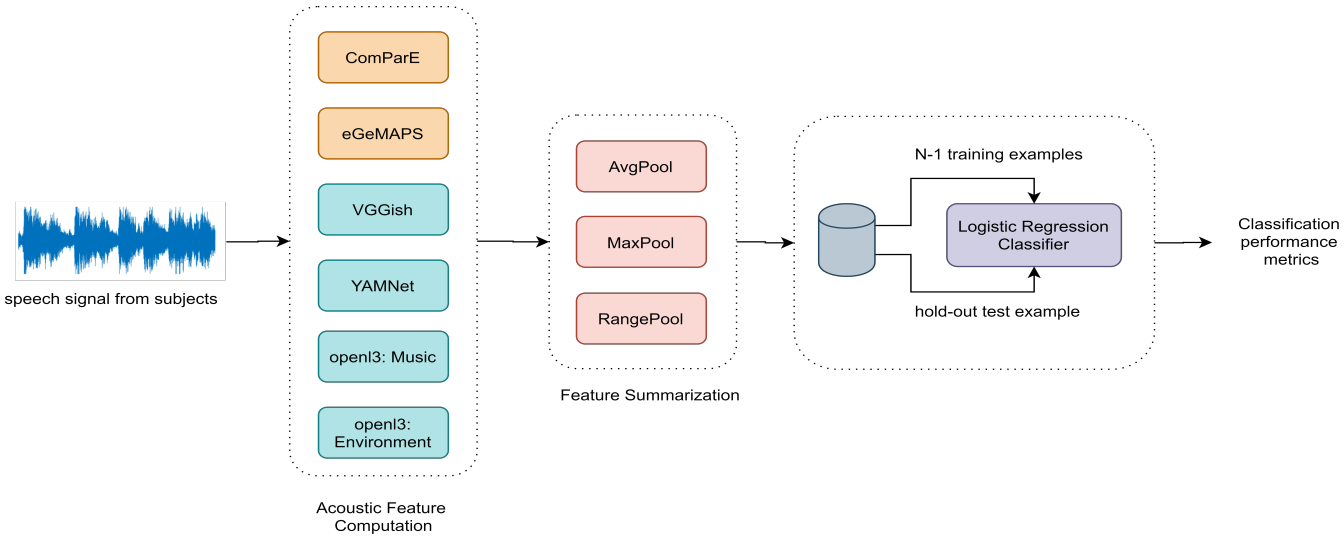

Fig. 2. Process Flow Diagram for Automated Identification of Parkinsonian Speech

identifying Parkinsonian speech. The final characteristic of voice that is popular in speech paralinguistics is voice spectra. It is reminded that spectral characteristics of voice are shaped by the movement of the vocal tract. One can surmise that this vocal tract becomes rigid due to muscular dystrophy therefore the speech of individuals with the disease will lack in spectral richness [26]. This suggests that spectral analysis of speech can be useful for the identification of Parkinsonian speech.

In order to represent these three paralinguistic characteristics of speech quantitatively, we shall compute acoustic features from two feature sets, (a) Computational Paralinguistics Challenge (ComParE) and (b) Extended Geneva Acoustic Minimalistic Feature Set (eGeMAPS) using the openSmile toolkit [27]. These feature sets are de facto standard in the field of social signal processing for quantification of characteristics of speech paralinguistics [28], [29].

The ComParE feature set consists of 65 acoustic low-level-descriptor (LLDs) amongst which 4 LLDs characterize voice energy, 6 features characterize voice quality, and 55 features represent voice spectra. It is used to develop baselines classification performance for the popular Interspeech Computational Paralinguistic Challenges and as a testament to its effectiveness, ComParE features have achieved better performance than deep learning methods as well [30]. The eGeMAPS feature set is considered as a lower-dimensional alternate to the ComParE feature set (88 vs 6373-dimensional feature vectors). eGeMAPS consists of 23 acoustic LLDs amongst which 13 features describe voice quality, 9 features describe voice spectra, and one acoustic feature is dedicated to voice energy. Further details of these features have been provided in Tables VII and VIII.

B. Deep Acoustic Embeddings based Features

As mentioned earlier, we compute three types of pooling methods in order to summarize deep acoustic embedding features which are computed for short-duration chunks of audio recordings. Therefore, the pooling method is one of the hyper-parameters which needs to be optimized for deep acoustic embeddings based features. Furthermore, there exist two further hyper-parameters for embeddings based on openl3 models. The first hyper-parameter amongst these determines whether a linear spectrogram or Mel spectrogram should be used as a representation of audio signal. The second hyper-parameter determines the dimensionality of openl3 embeddings with options of either a 512-dimensional embedding or a 6144-dimensional embeddings. Given the data-driven nature of machine learning, these hyper-parameters also need to be optimized using the cross-validation process.

V. Experimentation, Results and Discussion

In this section, we summarize and report on results from experiments conducted to determine the efficacy of deep acoustic embeddings for a variety of speaking tasks, along with analysis into fusion and performance comparison of deep acoustic embeddings.

A. Diadochokinesis Tasks

In Table II, we summarize results for experiments performed to identify individuals with Parkinson's disease based on six Diadochokinesis (DDK) tasks using two domain-knowledge based features (which serve as a baseline) and the four deep acoustic embeddings. Amongst the six tasks, one can note that `music_linear_512 embedding` achieves the best classification performance thrice, i.e. for *ka*, *pakata*, and *pekata* utterances. This includes the overall best classification accuracy of 85.0% amongst the six DDK tasks and the highest average classification accuracy of 79.3% for the six tasks. One can also note that `environment_mel256_512` embedding offers a competitive performance, with best classification accuracy for *ka* utterance as well as an average classification accuracy of 78.0% compared to 79.3% of `music_linear_512` embedding. Meanwhile, VGGish and YAMNet embeddings, which are trained on audio recordings that also contain human voice perform worst for the DDK tasks amongst all deep acoustic embeddings.

TABLE II. CLASSIFICATION ACCURACY FOR VARIOUS ACOUSTIC FEATURES FOR SIX DIADOCHOKINESIS TASKS, ALONG WITH THE AVERAGE ACCURACY OVER ALL DIADOCHOKINESIS TASKS

Feat	*Diadochokinesis*						*Average*
	pa	**ka**	**ta**	**pakata**	**pekata**	**pataka**	
ComParE	79.0	72.0	**79.0**	74.0	76.0	74.0	75.7
eGeMAPS	78.0	79.0	67.0	72.0	74.0	71.0	73.5
VGGish	65.7	66.3	67.0	71.0	65.0	70.0	67.5
YAMNet	66.0	73.0	76.0	69.0	68.0	71.0	70.5
environment_linear_512	77.0	77.0	74.0	79.0	77.0	75.0	76.5
environment_linear_6144	81.0	81.0	70.0	72.0	72.0	76.0	75.3
environment_mel256_512	**84.0**	77.0	74.0	79.0	79.0	75.0	78.0
environment_mel256_6144	79.0	75.0	74.0	77.0	75.0	**81.0**	76.8
music_linear_512	78.0	**82.0**	77.0	**85.0**	**80.0**	74.0	***79.3***
music_linear_6144	74.0	81.0	68.0	77.0	75.0	73.0	74.7
music_mel256_512	73.0	81.0	74.0	74.0	79.0	70.0	75.2
music_mel256_6144	75.0	81.0	75.0	78.0	73.0	73.0	75.8

Another interesting observation from Table II is that `environment_mel256_6144` embedding performs best for *pakata* utterance with a classification accuracy of 81.0%. In fact, the second placed acoustic feature is also based on environmental sounds, that is, `environment_linear_6144` embedding which achieves an accuracy of 76.0%. This suggests that some characteristics of environmental sounds are also useful for identifying Parkinsonian speech based on *pakata* utterance.

Amongst domain-knowledge based features, ComParE offers competitive performance overall as evident from the average classification accuracy for all tasks. In fact, ComParE features even achieved best performance for the *ta* task. Overall, however, it is clear that deep acoustic embeddings are a better alternative to domain-knowledge based features if the objective is to maximize classification performance based on diadochokinesis speaking tasks.

B. Short-Duration Utterance Tasks

The results of Parkinsonian speech classification based on short-duration utterances are summarized in Table III. On the basis of average classification computed for the ten utterances, one can note that `environment_mel256_512` embedding achieves the best performance overall (77.3%), closely followed by `music_linear_512` (76.5%). Here, `environment_mel256_512` achieved top performances for most tasks (five out of ten), including *laura*, *luisa*, *micasa*, *preocupado*, *rosita*, and *viste*) as well as the highest classification accuracy of 85.0%, that was achieved with *triste* utterance. The second placed acoustic features, `music_linear_512`, achieved top performances for *juan*, *loslibros*, and *omar* utterances.

Amongst the two domain-knowledge features, ComParE again performs much better than eGeMAPS – 74.6% average classification over the ten utterance tasks for ComParE versus 67.5% for eGeMAPS. Moreover, VGGish and YAMNet embeddings also perform poorly with an average classification accuracy of 70.7% and 69.1%, respectively.

C. Long-Duration Utterance Tasks

The classification results for identification of Parkinsonian speech for *readtext* and *monologue* tasks are summarized in Table IV. A special characteristic of these tasks is that they are of a relatively long duration as compared to *DDK* and *short utterance* tasks, therefore, more speech data is available per subject.

To begin, one can note that `music_linear_512` embedding achieves the best classification performance overall, with top classification accuracy of 84.0% for the *readtext* task and a competitive 79.0% for the *monologue* task. The best performance for *monologue* task is achieved with `environment_linear_512` feature but it lags behind other features for the reading task. As an example, consider `environment_mel256_512`, which achieves 79.0% and 78.0% for reading and monologue utterances to achieve a higher average accuracy than `environment_linear_512`. Furthermore, it is pertinent to mention here that `environment_mel256_512` was also the best performing model for short-duration utterance tasks as well which shows that the embedding is consistent in its efficacy for the task at hand.

D. Decision-Level Fusion

From Tables II–IV, one can note that embeddings from different upstream tasks achieve varying degrees of success at classification of Parkinsonian speech. For example, `environment_mel256_512` and `music_linear_512` are embeddings with different upstream tasks but yield top classification performance for a variety of speaking tasks. This suggests that machine learning models trained on these embeddings carry complimentary information which can be fused together in order to achieve improved classification accuracy. This is a well known premise of decision-level fusion [31] and we have had success at improving the quality of machine learning models using fusion in our previous works [32], [16].

To this end, we use two types of decision-level fusion, i.e. confidence based fusion and majority-vote based fusion for top three acoustic features for each speaking task. In confidence based fusion, the confidence scores of classifiers for predicting each class are averaged and judgement about class labels is made using the averaged confidence scores. Meanwhile, in majority-vote approach, each classifier makes its own judgement about class labels before a majority-vote is carried out to make final judgement about the class labels using information from all models. It must be mentioned here that due to data-driven nature of fusion process, one needs

TABLE III. CLASSIFICATION ACCURACY FOR VARIOUS ACOUSTIC FEATURES FOR TEN SHORT DURATION SPEECH UTTERANCE TASKS, ALONG WITH THE AVERAGE ACCURACY OVER ALL SHORT UTTERANCE TASKS

Feat	*Short Utterances*										*Average*
	juan	**laura**	**loslibros**	**luisa**	**micasa**	**omar**	**preocupado**	**rosita**	**triste**	**viste**	
ComParE	73.0	77.0	74.0	76.0	74.0	71.0	81.0	74.0	76.0	70.0	74.6
eGeMAPS	69.0	59.0	69.0	67.0	70.0	69.0	68.0	65.0	70.0	69.0	67.5
VGGish	72.0	70.0	67.0	67.0	69.0	73.0	76.0	69.0	76.0	68.0	70.7
YAMNet	69.0	66.0	70.0	71.0	66.0	74.0	70.0	67.0	69.0	69.0	69.1
environment_linear_512	68.0	71.0	72.0	76.0	68.0	68.0	78.0	70.0	**85.0**	78.0	73.4
environment_linear_6144	67.0	70.0	**76.0**	69.0	68.0	67.0	66.0	72.0	73.0	73.0	70.1
environment_mel256_512	72.0	**74.0**	73.0	**78.0**	**76.0**	75.0	**84.0**	**83.0**	79.0	**79.0**	**77.3**
environment_mel256_6144	68.0	71.0	73.0	74.0	72.0	73.0	76.0	74.0	79.0	77.0	73.7
music_linear_512	**78.0**	71.0	**76.0**	76.0	75.0	**77.0**	79.0	77.0	82.0	74.0	76.5
music_linear_6144	68.0	71.0	75.0	72.0	71.0	74.0	70.0	71.0	75.0	70.0	71.7
music_mel256_512	68.0	63.0	74.0	75.0	73.0	67.0	71.0	68.0	79.0	72.0	71.0
music_mel256_6144	69.0	68.0	73.0	70.0	71.0	71.0	75.0	70.0	79.0	77.0	72.3

TABLE IV. CLASSIFICATION ACCURACY FOR VARIOUS ACOUSTIC FEATURES FOR TWO LONG DURATION SPEECH UTTERANCE TASKS, ALONG WITH THE AVERAGE ACCURACY OVER BOTH LONG UTTERANCE TASKS

Feat	*Long Utterances*		*Average*
	readtext	**monologue**	
ComParE	69.0	72.0	70.5
eGeMAPS	72.0	**80.0**	76.0
VGGish	70.0	71.0	70.5
YAMNet	72.0	73.0	72.5
environment_linear_512	74.0	**81.0**	77.5
environment_linear_6144	79.0	76.0	77.5
environment_mel256_512	79.0	78.0	78.5
environment_mel256_6144	77.0	79.0	78.0
music_linear_512	**84.0**	79.0	**81.5**
music_linear_6144	75.0	79.0	77.0
music_mel256_512	79.0	77.0	78.0
music_mel256_6144	76.0	80.0	78.0

TABLE V. SUMMARY OF CLASSIFICTION ACCURACY FOR EACH SPEAKING TASK BEFORE AND AFTER DECISION-LEVEL FUSION

Speaking tasks	*Best result (pre-fusion)*	*Decision-level Fusion*	
		Confidence	**Majority Vote**
pa	84.0	85.0	88.0
ka	82.0	85.0	81.0
ta	79.0	77.0	80.0
pakata	85.0	84.0	84.0
pekata	80.0	76.0	80.0
pataka	81.0	81.0	79.0
juan	78.0	78.0	76.0
laura	77.0	72.0	78.0
loslibros	76.0	71.0	80.0
luisa	78.0	83.0	80.0
micasa	76.0	72.0	81.0
omar	77.0	77.0	78.0
preocupado	84.0	86.0	86.0
rosita	83.0	81.0	85.0
triste	85.0	88.0	89.0
viste	79.0	81.0	84.0
read text	84.0	84.0	82.0
monologue	81.0	84.0	83.0

to experiment with both types of decision-level fusion and determine the one which is best suited for the task hand.

In Table V we summarize the results for experiments for decision-level fusion for top-3 performing models for each speaking task. Here, one can note that decision-level fusion, in most cases, improves the classification performance. The most notable examples are speaking tasks *pa* and *triste* where classification accuracy was improved from 84.0% to 88.0% and 85.0% to 89.0%, respectively. There are also some cases where the classification accuracy after fusion actually decreases, for example, with speaking task *pakata* where pre-fusion accuracy of 85.0% decreases to 84.0%. We argue that fusion is still useful here since the slightly decreased accuracy is based on confidence of multiple models and it is more likely to be robust as compared to the accuracy achieved by a single model.

E. Performance Comparison of Deep Acoustic Embeddings

Finally, we compare the averaged classification accuracy over the eighteen speaking tasks which form the GITA corpus. A summary of results has been provided in Table VI where one can make out a ranking of deep acoustic embeddings based on their classification performance.

We note that `music_linear_512` embedding performs best with an accuracy of 78.0% and `environment_mel256_512` follows closely in second place with an accuracy of 77.7%. The third best performing embedding is `environment_linear_512` which achieves an accuracy of 74.9%. It is most interesting to note that acoustic embeddings which are trained to recognize types of music and environmental noise perform much better than `VGGish` and `YAMNet` embeddings, which are trained using data which also contains human voice. This suggests that the upstream task does not matter for downstream tasks whilst using deep acoustic embeddings. however, further testing over multiple datasets is required in order to reach a conclusion. For the sake of completeness, a summary of top-3 performing models including feature name as well as pooling method has been provided in Tables IX through XI for the three types of speaking tasks.

VI. CONCLUSION

In this work, we investigated the usefulness of deep acoustic embeddings as effective representations of speech paralinguistics for the task of identifying Parkinsonian speech, benchmarking the classification performance of these embeddings against two popular domain-knowledge based hand-crafted feature sets. Our results show that deep acoustic embeddings are indeed useful for the task at hand and perform consistently better than hand-crafted features. We also report

TABLE VI. AVERAGE CLASSIFICATION ACCURACY COMPARISON OF VARIOUS DEEP ACOUSTIC EMBEDDINGS FOR EIGHTEEN UTTERANCE TASKS FOR IDENTIFYING PARKINSONIAN SPEECH

Feat	Average Accuracy
VGGish	69.6
YAMNet	69.9
environment_linear_512	74.9
environment_linear_6144	72.7
environment_mel256_512	**77.7**
environment_mel256_6144	**75.2**
music_linear_512	**78.0**
music_linear_6144	73.3
music_mel256_512	73.2
music_mel256_6144	74.1

that the upstream training task may not be a limiting factor for the classification performance in the downstream task. For example, models trained on music and environmental sound data performed much better than embeddings which were trained on data containing human voice. Finally, we showed that decision-level fusion is an effective method to improve the stability of machine learning models for identifying Parkinsonian speech.

REFERENCES

[1] L. Marsili, G. Rizzo, and C. Colosimo, "Diagnostic criteria for Parkinson's disease: From James Parkinson to the concept of prodromal disease," *Frontiers in Neurology*, vol. 9, pp. 1–10, 2018.

[2] P. Lieberman, E. Kako, J. Friedman, G. Tajchman, L. S. Feldman, and E. B. Jiminez, "Speech production, syntax comprehension, and cognitive deficits in Parkinson's disease," *Brain and Language*, vol. 43, no. 2, pp. 169–189, 1992.

[3] World Health Organization, "Neurological Disorders: Public Health Challenges." [Online]. Available: https://www.who.int/mental{_}health/neurology/neurodiso/en

[4] B. Yousefi, V. Tadibi, A. Khoei, and A. Montazeri, "Exercise therapy, quality of life, and activities of daily living in patients with Parkinson disease: A small scale quasi-randomised trial," *Trials*, vol. 10, no. 67, pp. 1–7, 2009.

[5] J. Massano and K. P. Bhatia, "Clinical approach to Parkinson's disease: Features, diagnosis, and principles of management," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 6, pp. 1–15, 2012.

[6] U. Saeed, J. Compagnone, R. I. Aviv, A. P. Strafella, S. E. Black, A. E. Lang, and M. Masellis, "Imaging biomarkers in Parkinson's disease and Parkinsonian syndromes: Current and emerging concepts," *Translational Neurodegeneration*, vol. 6, no. 8, pp. 1–25, 2017.

[7] S. G. Ryman and K. L. Poston, "MRI biomarkers of motor and non-motor symptoms in Parkinson's disease," *Parkinsonism and Related Disorders*, vol. 73, pp. 85–93, 2019.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinsons disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.

[9] J. R. Orozco-Arroyave, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Noth, "Automatic detection of Parkinson's disease in running speech spoken in three different languages," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 481–500, 2016.

[10] D. Sztaho, M. G. Tulics, K. Vicsi, and I. Valalik, "Automatic estimation of severity of Parkinson's disease based on speech rhythm related features," in *IEEE International Conference on Cognitive Infocommunications*, vol. 2018-Janua, 2017, pp. 11–16.

[11] M. Cernak, J. R. Orozco-Arroyave, F. Rudzicz, H. Christensen, J. C. Vasquez-Correa, and E. Noth, "Characterisation of voice quality of Parkinson's disease using differential phonological posterior features," *Computer Speech and Language*, vol. 46, pp. 196–208, 2017.

[12] L. Moro-Velazquez, J. A. Gomez-Garcia, J. I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, and N. Dehak, "A forced gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing," *Biomedical Signal Processing and Control*, vol. 48, pp. 205–220, 2019.

[13] S. Hershey, S. Chaudhuri, D. P. 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. Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 131–135.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015, pp. 1–14.

[15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780.

[16] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated Screening for Alzheimer's Dementia through Spontaneous Speech," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[17] H. A. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, pp. 1–9, 2017.

[18] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 1–5.

[19] R. Arandjelovic and A. Zisserman, "Look, Listen and Learn," in *IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[20] J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, M. C. Gonzalez-Rativa, and E. Noth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *International Conference on Language Resources and Evaluation*, 2014, pp. 342–347.

[21] G. T. Stebbins and C. G. Goetz, "Factor structure of the Unified Parkinson's Disease Rating Scale: Motor Examination section," *Movement Disorders*, vol. 13, no. 4, pp. 633–636, 1998.

[22] C. Gobl and A. N. Chasaide, "Acoustic characteristics of voice quality," *Speech Communication*, vol. 11, no. 4-5, pp. 481–490, 1992.

[23] C. Dromey, L. O. Ramig, and A. B. Johnson, "Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study," *Journal of Speech and Hearing Research*, vol. 38, no. 4, pp. 751–764, 1995.

[24] B. T. Harel, M. S. Cannizzaro, H. Cohen, N. Reilly, and P. J. Snyder, "Acoustic characteristics of Parkinsonian speech: A potential biomarker of early disease progression and treatment," *Journal of Neurolinguistics*, vol. 17, no. 6, pp. 439–453, 2004.

[25] H. N. Jones, "Prosody in Parkinson's Disease," *Perspectives on Neurophysiology and Neurogenic Speech and Language*, vol. 1, no. 1, pp. 77–82, 2009.

[26] L. K. Smith and A. M. Goberman, "Long-time average spectrum in individuals with Parkinson disease," *NeuroRehabilitation*, vol. 35, no. 1, pp. 77–88, 2014.

[27] F. Eyben, M. Wollmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[28] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *ACM International Conference on Multimedia*, 2013, pp. 835–838.

[29] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge," *Computer Speech and Language*, vol. 1, no. 1, pp. 1–25, 2018.

[30] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018

Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *INTERSPEECH*, 2018, pp. 1–5.

[31] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, no. 1, pp. 115–129, 2020.

[32] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

Appendix A: List of Acoustic Features in ComParE and eGeMAPS Feature Sets

TABLE VII. Acoustic low-level descriptors which form the Computational Paralinguistics Challenge (ComParE) feature set

Energy related LLD	**Group**
Sum of auditory spectrum (loudness)	Voice Spectra
Spectral LLDs	**Group**
Alpha ratio (50-1000 Hz, 1000-5000 Hz)	Voice Spectra
Energy slope (0-500 Hz, 500-1500 Hz)	Voice Spectra
Hammarberg index	Voice Spectra
MFCCs 1-4	Voice Spectra
Spectral flux	Voice Spectra
Voicing related LLDs	**Group**
Fundamental frequency (linear and semitone)	Prosodic
Formants 1-2 (frequency, bandwidth, amplitude)	Voice Quality
Harmonic differences (H1-H2, H1-A3)	Voice Quality
log. HNR, Jitter, and Shimmer	Voice Quality

TABLE VIII. Acoustic low-level descriptors which form the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set

Energy related LLD	**Group**
Sum of auditory spectrum (loudness)	Voice Spectra
Sum of RASTA-filtered auditory spectrum (loudness)	Voice Spectra
RMS energy and zero-crossing rate	Temporal
Spectral LLDs	**Group**
RASTA-filtered audio spectrum bands 1-26	Voice Spectra
MFCCs 1-14	Voice Spectra
Spectral energy 250-650 Hz, 1000-4000 Hz	Voice Spectra
Spectral roll-off at 0.25, 0.5, 0.75, and 0.9 percentage	Voice Spectra
Psychoacoustic sharpness, Harmonicity	Voice Spectra
Spectral variance, Spectral skewness, Spectral kurtosis	Voice Spectra
Voicing related LLDs	**Group**
Fundamental frequency (SHS and Viterbi smoothing)	Prosodic
Probability of voicing	Voice Quality
log. HNR, Jitter, and Shimmer	Voice Quality

Appendix B: Summary of Top-3 Performing Features (along with their Pooling Method) for Diadochokinesis, Long, and Short Duration Utterances

TABLE IX. Summary of top-3 performing features (along with their pooling method) for Diadochokinesis based speaking tasks

Speaking task	***Feature***	***Pooling***	***Accuracy***
pa	ComParE	x	79.0
	environment_linear_6144	MaxPool	81.0
	environment_mel256_512	AvgPool	84.0
	environment_mel256_6144	AvgPool	79.0
ka	environment_linear_6144	AvgPool	81.0
	music_linear_512	AvgPool	82.0
	music_linear_6144	AvgPool	81.0
	music_mel256_512	MaxPool	81.0
	music_mel256_6144	AvgPool	81.0
ta	ComParE	x	79.0
	YAMNet	MaxPool	76.0
	music_linear_512	AvgPool	77.0
pakata	environment_linear_512	MaxPool	79.0
	environment_mel256_512	AvgPool	79.0
	music_linear_512	MaxPool	85.0
pekata	environment_mel256_512	MaxPool	79.0
	music_linear_512	AvgPool	80.0
	music_mel256_512	AvgPool	79.0
pataka	environment_linear_512	AvgPool	75.0
	environment_linear_6144	AvgPool	76.0
	environment_mel256_512	AvgPool	75.0
	environment_mel256_6144	AvgPool	81.0

TABLE X. Summary of top-3 performing features (along with their pooling method) for long duration utterance tasks

Speaking task	***Feature***	***Pooling***	***Accuracy***
readtext	environment_mel256_512	AvgPool	79.0
	music_linear_512	AvgPool	84.0
	music_mel256_512	AvgPool	79.0
	environment_linear_6144	AvgPool	79.0
monologue	eGeMAPS	x	80.0
	environment_linear_512	AvgPool	81.0
	music_mel256_6144	MaxPool	80.0

TABLE XI. SUMMARY OF TOP-3 PERFORMING FEATURES (ALONG WITH THEIR POOLING METHOD) FOR SHORT DURATION UTTERANCE TASKS

Speaking task	*Feature*	*Pooling*	*Accuracy*
juan	ComParE	x	73.0
	VGGish	AvgPool	72.0
	environment_mel256_512	AvgPool	72.0
	music_linear_512	MaxPool	78.0
laura	ComParE	x	77.0
	environment_linear_512	MaxPool	71.0
	environment_mel256_512	RangePool	74.0
	environment_mel256_6144	AvgPool	71.0
	music_linear_512	MaxPool	71.0
	music_linear_6144	AvgPool	71.0
loslibros	environment_linear_6144	AvgPool	76.0
	music_linear_512	MaxPool	76.0
	music_linear_6144	AvgPool	75.0
luisa	ComParE	x	76.0
	environment_linear_512	AvgPool	76.0
	environment_mel256_512	AvgPool	78.0
	music_linear_512	AvgPool	76.0
micasa	ComParE	x	74.0
	environment_mel256_512	MaxPool	76.0
	music_linear_512	AvgPool	75.0
omar	YAMNet	RangePool	74.0
	environment_mel256_512	AvgPool	75.0
	music_linear_512	MaxPool	77.0
	music_linear_6144	AvgPool	74.0
preocupado	ComParE	x	81.0
	environment_mel256_512	AvgPool	84.0
	music_linear_512	AvgPool	79.0
rosita	ComParE	x	74.0
	environment_mel256_512	AvgPool	83.0
	environment_mel256_6144	MaxPool	74.0
	music_linear_512	MaxPool	77.0
triste	environment_linear_512	AvgPool	85.0
	environment_mel256_512	AvgPool	79.0
	environment_mel256_6144	MaxPool	79.0
	music_linear_512	AvgPool	82.0
	music_mel256_512	AvgPool	79.0
	music_mel256_6144	AvgPool	79.0
viste	environment_linear_512	MaxPool	78.0
	environment_mel256_512	AvgPool	79.0
	environment_mel256_6144	MaxPool	77.0
	music_mel256_6144	MaxPool	77.0