INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automated Screening for Alzheimer's Dementia through Spontaneous Speech

Muhammad Shehram Shah Syed[1], *Zafi Sherhan Syed*[2], *Margaret Lech*[1], *Elena Pirogova*[1]

[1]RMIT University, Australia
[2]Mehran University, Pakistan

muhammad.shehram.shah.syed@rmit.edu.au

Abstract

Dementia is a neurodegenerative disease that leads to cognitive and (eventually) physical impairments. Individuals who are affected by dementia experience deterioration in their capacity to perform day-to-day tasks thereby significantly affecting their quality of life. This paper addresses the Interspeech 2020 Alzheimer's' Dementia Recognition through Spontaneous Speech (ADReSS) challenge where the objective is to propose methods for two tasks. The first task is to identify speech recordings from individuals with dementia amongst a set of recordings which also include those from healthy individuals. The second task requires participants to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech alone. To this end, we investigated characteristics of speech paralinguistics such as prosody, voice quality, and spectra as well as VGGish based deep acoustic embedding for automated screening for dementia based on the audio modality. In addition to this, we also computed deep text embeddings for transcripts of speech. For the classification task, our method achieves an accuracy of 85.42% compared to the baseline of 62.50% on the test partition, meanwhile, for the regression task, our method achieves an RMSE = 4.30 compared to the baseline of 6.14. These results show the promise of our proposed methods for the task of automated screening for dementia based on speech alone.

Index Terms: Social signal processing, Computational paralinguistics, Alzheimer's disease

1. Introduction

Dementia is an umbrella term for diseases which causes significant and continual cognitive and physical impairments. Individuals who are affected by dementia experience decline in language, thinking ability, and memory along with deterioration in their ability to perform day-to-day tasks in order to take care of themselves at a level which is beyond what is expected for ageing. According to the World Health Organization (WHO), there are around 50 million people worldwide who suffer from dementia and this number is increasing, with 10 million new cases every year [1]. Although there are various causes of dementia, Alzheimer's disease is the most prominent one, accounting for $60 - 70\%$ of total cases [1]. Alzheimer's disease is also known to adversely affect the mental health of care givers [2] such that they may require psychiatric interventions themselves.

It is known that cognitive impairments such as those caused by dementia affect the speech production system [3]. In [4], Yu et al. reported the use of vocal biomarkers for prediction of cognitive decline in the elderly population. They investigated the efficacy of a variety of acoustic features such as pitch variance, syllable rate, phoneme-based measures, and formant-based articulatory coordination features for automated cognitive impairment diagnosis. Ivanov et al. [5] developed phoneme-conditioned statistical models for cognitive impairment diagnosis and found them to be useful for the task at hand. Fraser et al. [6] consider a large number of features (370 in total) such as part-of-speech information, grammatical constituents, and vocabulary richness to capture linguistic phenomena which can identify subjects with dementia amongst a corpus which also includes healthy subjects. Luz et al. [7] used turn-taking patterns, speech rate, and other speech parameters which are essentially "content-free" for Alzheimer's disease recognition and report that their method achieves better accuracy than lexical, syntactic and semantic features.

In [8], Mirheidari et al. explored the use of word vector representations based on word2vec and GloVe embeddings for dementia recognition based on speech-transcripts and reported high accuracy. The authors hypothesized that since these embeddings can capture the semantics and syntax of words in a text, they will be useful for detecting diminished articulation from subjects with dementia. Haider et al. [9] investigate the efficacy of various types of speech paralinguistic features for voiced based screening from spontaneous speech. We find that the ADReSS challenge baseline closely follows the methodology proposed in [9].

In this paper, we propose methods for speech based screening of Alzheimer's dementia. To this end, we first train machine learning models which seek to model differences in characteristics of speech paralinguistics between subjects with dementia and those from the control group. Next, we conduct an exploratory analysis to generate numerical representations for speech transcripts based on recently developed deep language models. Our proposed models perform significantly better than the ADReSS challenge baselines for classification and regression tasks.

2. Dataset

The dataset for the Interspeech 2020 ADReSS challenge consists of speech recordings elicited for the Cookie Theft picture description task from the Boston Diagnostic Aphasia Exam [10]. This data was explicitly balanced by the organizers in terms of age, gender, and the distribution of labels between the training and test partitions in order to minimize the risk of bias in the prediction tasks. The dataset has labels for machine learning tasks of binary classification and regression. As the name suggests, labels for the binary classification include Alzheimer's dementia and healthy control, whereas the labels for the regression task are Mini-Mental State Examination (MMSE) scores [11] which provide a means for dementia diagnosis based on linguistic tests. For further details regarding the dataset, we refer the reader to the ADReSS challenge baseline paper [12].

http://dx.doi.org/10.21437/Interspeech.2020-3158

3. Methodology

As part of our investigation into automated recognition of dementia with spontaneous speech as the input, we follow a two-pronged approach which includes voice-based screening and speech transcripts based screening as illustrated in Figure 1. For voice-based screening, we investigate the efficacy of acoustic features which are known to represent paralinguistic characteristics of prosody, voice quality, and spectra. Such categorization has previously proved to be useful for automated recognition of depression [13, 14] and bipolar disorder [15]. Meanwhile, our work on speech-transcripts based screening is largely exploratory such that we investigate the efficacy of deep language embeddings such as Bidirectional Encoder Representations from Transformers (BERT) [16] and its derivatives for generating a numerical representation of speech-transcripts.

3.1. Voice based screening

Here, we hypothesize that subjects with dementia have unique characteristics to their voice, given that the disease causes cognitive impairments, which can be quantified using acoustic descriptors of speech-paralinguistics. Following the approach of Horwitz et al. [13] for depression recognition, we propose to investigate the efficacy of acoustic features which characterize prosody, voice quality, and voice spectra. Prosody defines patterns of stress and intonation and is likely to be affected due to cognitive impairments. Voice quality analysis seeks to quantify changes at the vocal source level (glottis). It has been shown that the perceptual quality of voice changes on a scale between breathy and tense depending on the available cognitive resources [17]. Finally, acoustic descriptors of voice spectra have the potential to provide vital insights into muscular changes due to dementia at the vocal-tract level.

To this end, we compute prosody, voice quality, and spectral features using the openSmile [18] and COVAREP [19] toolkits. These toolkits have become the standard tools for computation of acoustic features for tasks related to social signal processing. These are not only open source but also freely available for academic research. In addition to the mentioned features, we use the (a) ComParE-2016 feature-set, (b) IS10-Paralinguistics feature-set, and (c) VGGish acoustic embeddings as part of our investigation of acoustic descriptors. The Computational Paralinguistics Challenge 2013 feature set (ComParE) is a brute-force feature set which has proved to be useful for a variety of speech paralinguistic tasks and is regularly used to set a baseline for Interspeech ComParE challenges [20, 21, 22]. The most recent version of the ComParE feature set was released as part of the 2016 edition of the ComParE challenge. The IS10-Paralinguistics feature set was introduced as part of the 2010 edition of Interspeech ComParE challenge and can be considered as a low-dimensional alternate to the ComParE feature set (6373 features vs 1582 features). Recently, we have found this feature set to be useful for tasks related to the recognition of bipolar disorder from speech [15] and emotion recognition [23]. Finally, we use VGGish embeddings [24] since they provide an alternative to domain-knowledge features such as those computed using openSmile and COVAREP toolkits.

The six types of acoustic features are computed as low-level-descriptors which means that they only represent the acoustic characteristics of a small chunk of the audio file. There is a need for these features to be aggregated using an appropriate method in order to generate a global acoustic representation for the speech recording. For this purpose, we use three types of feature aggregation methods: (a) functionals of descriptive statistics, (b) Bag-of-Audio-Words (BoAW), and (c) Fisher Vector encoding. These feature aggregation approaches are relatively well known in the research community and (mainly due to a requirement of brevity here) we refer the reader to [25, 26, 27, 28, 29] for details.

3.2. Screening based on Speech-Transcripts

The availability of speech transcripts provides a second modality which can be used alongside voice for the development of a multimodal framework for automated screening for dementia. This has been our objective, as illustrated in Figure 1. To this end, we conduct an exploratory analysis in order to determine the efficacy of pre-trained embeddings from deep language models for the task at hand. It must be mentioned here that these embeddings have already been shown to be useful for a large variety of tasks in the field of natural language processing [30, 31]. More specifically, we compute embeddings from eight models i.e. BERT base cased, BERT large cased, BERT large uncased, distilbert cased, distilbert uncased, distilroberta base, roberta base, and the biomed roberta base using the Huggingface Transformers library [32]. These embeddings are computed for each word of every transcript. In order to generate a transcript-level representation for transcripts we use four types of pooling functions which are average pooling (AvgPool), maximum value pooling (MaxPool), outlier-robust percentile-based range pooling (RangePool), and the coefficient of deviation (StdDevNormPool). The resultant feature vector is passed down to the machine learning pipeline as shown in Figure 1.

4. Experiments and Results

In this section, we present results for our experiments on speech based screening for Alzheimer's dementia. We used two types of algorithms each in order to predict labels for the classification and regression tasks. For the classification task, we used support vector machine classifier with a linear kernel (SVC) and logistic regression classifier. A grid search was carried out to optimize the model using leave-one-subject-out (LOSO) cross-validation whilst using the training partition. The optimization parameter *complexity* was tuned for both of these methods between a logarithmically-spaced range of 10^{-7} and 10^{3}. For the regression task, we used support vector machines based regression (SVR) (again with a linear kernel) whose hyperparameters were tuned using the same method as the classifier. In addition to SVR, we used a partial least squares regressor (PLSR) which has been shown to be useful for tasks related to speech paralinguistics [33]. A grid search was carried out to optimise the number of components for PLSR between 1 and 20. The results summarized in this section report the best performing models.

4.1. Voice based screening

A summary of classification results for voiced based screening has been provided in Table 1, where one finds that the IS10-Paraling.-BoAW model achieves the highest classification accuracy of the training partition with 76.85%, which is significantly better than the challenge baseline of 56.50%. This result is closely followed by the VGGish-BoAW model which achieves the second-best performance with an accuracy of 75.00%. Furthermore, the best performing models for Prosody, Voice Quality, and Spectra achieve a classification accuracy of 67.59%, 72.22%, and 71.30% respectively. This suggests that demen-

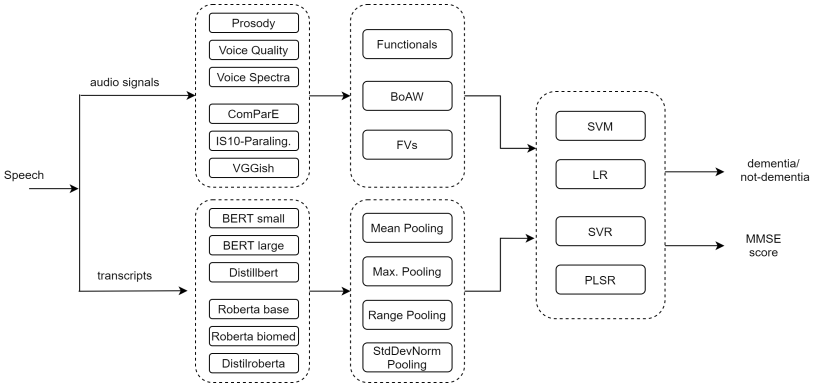

Figure 1: *Multimodal framework for automated screening of Alzheimer's' dementia*

tia may cause changes at voice source and vocal tract level, although a detailed investigation across datasets is required to support this observation. The best performing model based on ComParE features achieves an accuracy of 69.44%. It is important to note that all of these models achieve a better performance than the challenge baseline. The most interesting result from this table is that VGGish features provide better accuracy than most models trained on domain-knowledge based acoustic features such as Prosody features, Voice Quality features, Spectral features, and the ComParE features.

Table 1 also provides a summary of results for the regression task. Here one finds that the best performing model i.e. VGGish-BoAW achieves an RMSE = 5.95 which is better than the challenge baseline of 7.28. Furthermore, while MAE metric was not provided as part of the ADReSS challenge baseline, we find that the VGGish-FV BoAW model also achieves the smallest MAE of 4.49. These results are particularly interesting since they show that deep-learning based acoustic embedding can achieve a better performance than domain-knowledge based features and compliments our observation from the classification task. The performance of VGGish-BoAW is closely followed by BoAW and FV models based on IS10-Paralinguistic features. These models achieve an RMSE = 6.02 and RMSE = 6.04 respectively. The ComParE-FV model also achieved an RMSE = 6.04. Amongst the models which explicitly focus on characteristics of speech paralinguistics, we found that the Voice Quality-BoAW model achieved the smallest RMSE of 6.22, the Spectra-BoAW model achieved an RMSE = 6.12, and the Prosody-functionals model achieved an RMSE = 7.17 – all of these models achieve a smaller RMSE than the challenge baseline. This shows that modelling speech paralinguistics for recognition of dementia speech has promise, although, if the aim is to minimize the error between MMSE scores then the VGGish features with BoAW feature aggregation should be chosen.

4.2. Screening based on speech-transcripts

In Table 2 we provide a summary of classification results for the top-10 performing models based on text modality. Here, one can observe a notable improvement in the classification accuracy as compared to the challenge baseline accuracy of 62.5%, although it needs to be reminded that the challenge baseline was computed using audio modality [1]. The best performing model

Table 1: *Summary of results for classification and regression tasks using acoustic features for the training partition with LOSO cross-validation*

Feature Class	Feat. Agg.	Acc. (%)	RMSE	MAE
Prosody	Functionals	67.59	7.18	6.20
Voice Quality	Functionals	63.89	7.08	6.10
	BoAW	69.44	6.22	5.17
	FVs	72.22	6.52	5.49
Voice Spectra	Functionals	60.19	7.74	6.70
	BoAW	71.30	6.12	5.24
	FVs	71.30	6.12	4.89
IS10-Paraling.	Functionals	70.37	6.66	5.74
	BoAW	**76.85**	6.02	5.04
	FVs	66.67	6.04	5.21
ComParE	Functionals	68.52	7.16	5.69
	BoAW	65.74	6.90	6.13
	FVs	69.44	6.04	5.21
VGGish	BoAW	75.00	**5.92**	4.69
	FVs	62.96	6.75	5.53
Challenge baseline		56.50	7.29	—

i.e. *biomed roberta base* embedding with RangePool achieves an accuracy of 89.81%, which is followed by *roberta base* with RangePool which achieves an accuracy of 87.96%. Interestingly, we do not observe a difference in performance due to case and uncased versions of deep language models. For example, both *distilbert uncased* and *distilbert cased* models achieve the same accuracy, and the cased and uncased versions of the *BERT large* models achieve the same accuracy.

Table 3 summarizes the results for the top-10 performing models for MMSE scores prediction from the text modality. Here, one finds that the *BERT base uncased* embedding with MaxPool provides the best results in terms of the RMSE, achieving an RMSE = 4.32 which is better than the challenge baseline for regression of RMSE = 7.28. This is followed by the same BERT model but with RangePool which achieved an RMSE = 4.39. One can also note that all of the top-10 models based on text modality achieve a significantly better performance than the challenge baseline. It must be mentioned here for the sake of clarity that the baseline RMSE was computed us-

[1] A text modality baseline was added in the final version of the baseline paper with a classification UAR for train/test = $77.00\%/75.00\%$ and regression RMSE for train/test = $4.38/5.20$. As the reader shall note, our proposed methods still beat the updated challenge baseline.

ing audio features only (the organizer did not provide an RMSE computed using text features). Nevertheless, a comparison of results from Tables 1 and 3 makes it clear that the text modality is better for the task at hand.

Table 2: *Summary of results for top-10 performing models based on text modality for the classification task*

Feature Class	Pooling meth.	Accuracy (%)
biomed roberta base	RangePool	**89.81**
roberta base	RangePool	87.96
distilbert uncased	MaxPool	86.11
distilbert cased	MaxPool	86.11
BERT base uncased	MaxPool	86.11
BERT large uncased	AvgPool	86.11
BERT large cased	AvgPool	86.11
biomed roberta base	MaxPool	85.19
BERT base uncased	RangePool	85.19
BERT large cased	MaxPool	85.19

4.3. Predictions for the test partition

The ADReSS challenge baseline for the test partition is 62.50% and each participant has five attempts at predicting the labels of the test partition. A summary of the baseline and our results for the classification task is provided in Table 4. For our first attempt, we use predictions from the *biomed roberta base RangePool* model which was the best performing model for the training partition by achieving an accuracy of 89.81%. On the test partition, this model achieved an accuracy of 77.08% only which suggests that the model may have overfitted the training partition.

The second attempt used label fusion from the top-5 performing models from the text modality for the training partition (see Table 2). The resultant predictions for the test partition achieved an accuracy of 85.45%. This is not only our best result but also a large improvement from the challenge baseline of 62.50%. Our third attempt used label fusion from the top-5 performing models from the audio modality for the training partition (see Table 1). The resultant predictions for the test partition achieved an accuracy of 64.58% which is slightly better than the challenge baseline, although it does show that the audio modality offers weaker classification performance than the text modality. The fourth attempt used label fusion from the top-5 performing models from audio and text modalities (top-5 from each modality). The resultant predictions for the test partition achieved an accuracy of 79.17% which is an improvement over the results from the first and third attempt. For the final attempt, we used label fusion from the top-10 performing models overall (see Tables 1 and 2). Incidentally, all ten models are based on text modality. The resultant predictions for the test partition achieved an accuracy of 85.45% which is the same as the accuracy achieved by a fusion of top-5 models for text modality.

Table 3: *Summary of results for top-10 performing models based on text modality for the regression task*

Feature class	Pool meth.	RMSE	MAE
BERT base uncased	MaxPool	**4.32**	3.57
BERT base uncased	RangePool	4.39	3.62
distilbert uncased	RangePool	4.49	3.62
roberta base	AvgPool	4.49	3.48
BERT large cased	MaxPool	4.49	3.64
BERT large uncased	MaxPool	4.49	3.64
distilbert uncased	MaxPool	4.51	3.70
allenai biomed roberta base	AvgPool	4.51	3.68
allenai biomed roberta base	MaxPool	4.55	3.69
distilbert cased	AvgPool	4.57	3.51

Table 4: *Summary of results on the test partition for our proposed methods*

	Accuracy (%)	RMSE
Attempt 1	77.08	4.83
Attempt 2	**85.42**	6.91
Attempt 3	64.58	5.18
Attempt 4	79.17	4.91
Attempt 5	**85.42**	**4.30**
Challenge baseline	62.50	6.15

Similar to the classification task, each participant of the regression task has five attempts at predicting the MMSE scores. The challenge baseline for the regression task is an RMSE = 6.14. Our first attempt used predictions from the *BERT base uncased MaxPool* model, which was the best model on the training partition with an RMSE = 4.32. We find that this model achieved an RMSE = 4.83 on the test partition. The second attempt used test partition predictions from the *VGGish-BoAW* model which achieved an RMSE = 5.92 on the training partition but ends up achieving an RMSE = 6.91 on the test partition. This result is poorer than the challenge baseline.

Our third attempt used prediction for the test partition from the *BERT base uncased RangePool* model. This model was the second-to-best performing model for the training partition by achieving an RMSE = 4.39 and ends up achieving an accuracy of 5.18 on test partition which is still better than the challenge baseline. For the fourth attempt, we submitted the average value of predictions from our first and third attempt, the resultant predictions achieved an RMSE = 4.91 on the test partition. It is important to note that this score is slightly larger than the RMSE achieved from the first attempt. Finally, for our last attempt, we submitted an average of MMSE score predictions for the test partition from the top-10 performing models for the training partition. Interestingly, this setup produced our best RMSE score for the test partition with 4.30. This score easily beats the challenge baseline of 6.14.

5. Conclusions

In this paper, we investigated the efficacy of speech based automated screening of Alzheimer's dementia, a disease which significantly deteriorates the quality of life of affected individuals. From voiced based analysis we report that voice quality and voice spectral features perform better than features which characterise speech prosody. However, the best performing model from voice modality was based on VGGish deep acoustic embeddings. Overall, we report that the text modality which is available in the form of speech-transcripts perform the best by achieving an accuracy of 89.91% for the training partition. On training and test partitions, our methods outperformed the challenge baselines for both classification and regression tasks.

6. References

[1] World Health Organisation, "Dementia: Key Facts," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] A. S. Alfakhri, A. W. Alshudukhi, A. A. Alqahtani, A. M. Alhumaid, O. A. Alhathlol, A. I. Almojali, M. A. Alotaibi, and M. K. Alaqeel, "Depression among caregivers of patients with dementia," *Inquiry (United States)*, vol. 55, no. 1, pp. 1–6, 2018.

[3] G. W. Ross, J. Cummings, and D. F. Benson, "Speech and language alterations in dementia syndromes: Characteristics and treatment," *Aphasiology*, vol. 4, no. 4, pp. 339–352, 1990.

[4] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *INTERSPEECH*, 2015, pp. 3734–3738.

[5] A. V. Ivanov, S. Jalalvand, R. Gretter, and D. Falavigna, "Phonetic and anthropometric conditioning of MSA-KST cognitive impairment characterization system," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 228–233.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[7] S. Luz, S. de la Fuente, and P. Albert, "A Method for Analysis of Patient Speech in Dialogue for Dementia Detection," in *International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 35–42.

[8] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting signs of dementia using word vector representations," in *INTERSPEECH*, 2018, pp. 1–5.

[9] F. Haider, S. de la Fuente, and S. Luz, "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[10] H. Goodglass, E. Kaplan, and B. Barresi, *BDAE-3: Boston Diagnostic Aphasia Examination – Third Edition*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[11] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state". A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[13] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in *IEEE International Conference on Body Sensor Networks (BSN)*, 2013, pp. 1–6.

[14] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[15] 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.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint:1810.04805v2*, vol. 1, no. 1, pp. 1–16, 2018.

[17] A. S. Cohen, J. E. McGovern, T. J. Dinzeo, and M. A. Covington, "Speech Deficits in Serious mental Illness: A Cognitive Resource Issue?" *Schizophrenia research*, vol. 160, no. 0, pp. 173–179, 2014.

[18] 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.

[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP — A collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[20] 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.

[21] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Noth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *INTERSPEECH*, 2019, pp. 1–5.

[22] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *INTERSPEECH (to appear)*, 2020, pp. 1–5.

[23] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the Urdu-Sindhi Speech Emotion Corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1–6, 2020.

[24] 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.

[25] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Lecture Notes in Computer Science*, vol. 6314, 2010, pp. 143–156.

[27] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native language," in *INTERSPEECH*, 2016, pp. 2001–2005.

[28] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[29] Z. S. Syed, J. Schroeter, K. Sidorov, and D. Marshall, "Computational Paralinguistics: Automatic Assessment of Emotions, Mood, and Behavioural State from Acoustics of Speech," in *INTERSPEECH*, 2018, pp. 511–515.

[30] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," *IEEE Access*, vol. 1, no. 1, pp. 100 943 – 100 953, 2019.

[31] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challengesand New Directions in Sentiment Analysis Research," *arXiv:2005.00357*, vol. 1, no. 1, pp. 1–26, 2020.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv:1910.03771*, pp. 1–11.

[33] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 37–43.