Late Fusion of the Available Lexicon and Raw Waveform-based Acoustic Modeling for Depression and Dementia Recognition

Esaú Villatoro-Tello[1,2], *S. Pavankumar Dubagunta*[2,3], *Julian Fritsch*[2,3], *Gabriela Ramírez-de-la-Rosa*[1], *Petr Motlicek*[2], *Mathew Magimai.-Doss*[2]

[1] Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico City, Mexico
[2] Idiap Research Institute, Martigny, Switzerland
[3] École polytechnique fédérale de Lausanne (EPFL), Switzerland

evillatoro@cua.uam.mx, pavankumar.dubagunta@idiap.ch, julian.fritsch@idiap.ch, gramirez@cua.uam.mx, petr.motlicek@idiap.ch, mathew.magimaidoss@idiap.ch

Abstract

Mental disorders, e.g. depression and dementia, are categorized as priority conditions according to the World Health Organization (WHO). When diagnosing, psychologists employ structured questionnaires/interviews, and different cognitive tests. Although accurate, there is an increasing necessity of developing digital mental health support technologies to alleviate the burden faced by professionals. In this paper, we propose a multi-modal approach for modeling the communication process employed by patients being part of a clinical interview or a cognitive test. The language-based modality, inspired by the Lexical Availability (LA) theory from psycho-linguistics, identifies the most *accessible* vocabulary of the interviewed subject and use it as features in a classification process. The acoustic-based modality is processed by a Convolutional Neural Network (CNN) trained on signals of speech that predominantly contained voice source characteristics. In the end, a late fusion technique, based on majority voting, assigns the final classification. Results show the complementarity of both modalities, reaching an overall Macro-F1 of 84% and 90% for Depression and Alzheimer's dementia respectively.

Index Terms: Depression Detection, Alzheimer's Disease, Mental Lexicon, Raw Speech, Multi-modal Approach.

1. Introduction

Mental disorders represent a major public health concern, with considerable associated socio-economic costs, and are recognized as a major cause of disability affecting a great number of people. According to the World Health Organization (WHO), depression and dementia are among the main types of mental disorders and are categorized as priority conditions [1, 2]. Although the severity of suffering a mental illness is well known by psychologists, there is an acknowledged necessity for digital solutions for addressing the burden of mental health diagnosis and treatment. It is recognized that won't be possible to treat people by professionals alone, and even if possible, some people might require to use alternative modalities to receive mental health support [3]. Such situation has become more evident with the current COVID-19 pandemic. Interested readers are referred to [4, 5, 6] to know efforts towards this direction.

Accordingly, the research community has been interested in making first steps towards computer-supported detection of mental disorders during face-to-face interviews/tests [7, 8, 9]. The underlying hypothesis of most of previous work relies on the notion of the language as a powerful indicator about our personality, social, or emotional status, and mental health [10, 11]. In dementia, for instance, previous research indicates that assessing the language production represents a useful strategy in detecting early markers of dementia [12]. Thus, designed tests for evaluating the language production in elderly patients such as word association tasks, description of objects in pictures, elicitation exercises, etc., aim at measuring the expository speech, oral expression, as well as comprehension. Similarly, for depression, previous research suggest that using excessive self-focused language, and negative emotions represent important markers for screening depressed users [10, 13, 14] and, recent studies have documented how depressed users suffer some kind of impairment in their speech motor control [15], such as prosodic abnormalities, articulatory and phonetic errors.

Although multi-modal approaches have been explored before [7, 16, 17], the key novelty of our work is to leverage the psycholinguistics theory for approximating the *mental lexicon*[1] of analyzed subjects for processing the language-based modality. The acoustic-based modality aims at modeling the patients' speech in an end-to-end fashion from raw waveform-based CNNs. In conjunction, both modalities allow modeling the language production process employed by subjects with a mental disease during a clinical interview/test. We performed experiments in two well-known clinical datasets, using individual modalities, and in a multi-modal fashion, where a voter makes the final decision through a majority voting mechanism.

2. Methodology

The proposed language-based modality aims at modeling the vocabulary production of subjects suffering from a mental disorder through the Lexical Availability (LA) theory [19]. The LA test is associated with the category fluency tests and the free word association tasks, which taps directly into the semantic information of the *mental lexicon* [20]. Hence, our main hypothesis establishes that it could be possible to approximate the *available lexicon* for a group of people suffering from a mental disease. Contrary to the traditional LA elicitation test, we aim to demonstrate that it is possible to approximate the available lexicon by analyzing subjects' responses in a semi-structured communication process (e.g, a clinical interview/test). To the best of our knowledge, this is the first time the LA theory is adapted to: *i)* obtain the available lexicon from utterances produced during a clinical test and use the extracted features in a traditional

[1] The *mental lexicon* of a community reveals the type, size, and richness of their vocabulary as well as provides evidence of the community member's understanding of a particular culture, or the structure of their context and the existing regularities present [18].

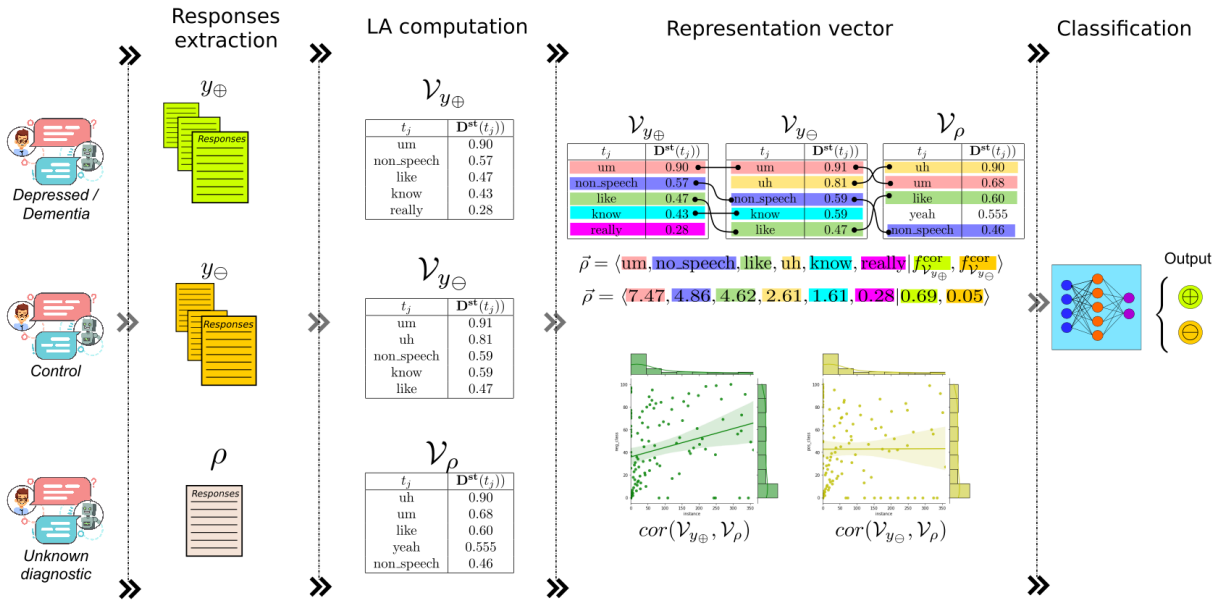

Figure 1: *General overview of the proposed language-based modality.*

classification pipeline; *ii)* fuse its predictions in a multi-modal fashion with a raw waveform-based acoustic approach.

Figure 1 shows the main components of our LA method. First, we identify the available lexicon from each population (i.e., subjects with a mental disorder and control subjects) and then use it to generate a non-sparse text representation to train a classification model to distinguish between *mentally ill* (*D*) and *control* (*C*) subjects. More formally, let $\mathcal{D} = \{(d_1, y_1), \ldots, (d_h, y_h)\}$ be a training set of h-pairs of documents[2] d_i and class labels $y_i \in \mathcal{Y} = \{y_\oplus, y_\ominus\}$. The first step consists of obtaining the available lexicon ($\mathcal{V}$) for each category, i.e., $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$ for the documents belonging to *D* and *C* categories respectively. The resultant available lexicon for each category y_i is a list of n-pairs of the form $\mathcal{V}_{y_i} = \{(t_1, \mathbf{D^{st}}(t_1)), \ldots, (t_n, \mathbf{D^{st}}(t_n))\}$, where each term t_j is accompanied by its lexical availability score $\mathbf{D^{st}}(t_j)$. Details on how to compute the availability score are depicted in §2.1.

Then, for generating the representation of subject ρ, we define two sets of features: the *availability degree* (f^{avail}), and the *correlation degree* attributes (f^{cor}). Thus, we first compute the available lexicon of subject ρ, referred as $\mathcal{V}_\rho$, and we calculate its *availability* features (f^{avail}) by means of a fusion strategy among the top k terms from $\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus}$, and $\mathcal{V}_\rho$ (see §2.2). For obtaining the *correlation* features (see §2.3) we compare the data distributions between ρ and the two classes ($y_\oplus$ and $y_\ominus$), resulting in a representation vector with the following form:

$$\vec{\boldsymbol{\rho}} = \langle f^{\text{avail}}_{t_1}, \ldots, f^{\text{avail}}_{t_j}, \ldots, f^{\text{avail}}_{t_k} | f^{\text{cor}}_{\mathcal{V}_{y_\oplus}}, f^{\text{cor}}_{\mathcal{V}_{y_\ominus}} \rangle \quad (1)$$

Once we have this representation, we can follow the traditional machine learning pipeline for training a classifier.

2.1. Lexical availability computation

Traditionally, the LA test produces a single word list, i.e., the available lexicon (with its corresponding availability scores), for each analyzed community. To compute the availability scores of this available lexicon, we have to analyze the responses of each individual in that population (see Fig. 1, columns 1-3); to that end, we use the formulation proposed by [21], defined as follows:

$$\mathbf{D^{st}}_{w,k,m}(t_j) = \sum_{i=1}^{n} w^{\left(\frac{i-1}{k-1}\right)^m} \times \frac{f_{ji}}{I} \quad (2)$$

where t_j represents the lexical term for which we want to know its availability score; i is the position indicator where t_j is mentioned in the considered individual responses; n is the maximum position reached by term t_j in all the considered responses; I serves as a normalization factor and is defined as $I = max\_freq$, which depicts the highest frequency found in the vocabulary of the population being analyzed; f_{ji} is the number of participants who produced term t_j at position i in their respective responses; k indicates the position value where the score will be equal to w; w is the desired weight (normally a value close to 0) for position k, and m is a parameter that modulates the weight decay across terms in the final mental lexicon.

Eq. 2 represents a standardized LA metric that allows direct comparisons among studies independently from the size of the produced vocabulary lists of different communities [21]. Accordingly, the $\mathbf{D^{st}}$ equation will assign higher scores (close to 1) to the most available words produced by the analyzed subjects. Conversely, it assigns progressively lower scores to less accessible words until reaching value w in position k, at a weight decay intensity defined by the parameter m. Intuitively, the smaller the value of m, the faster the weight decay across words in consecutive positions. For all our experiments, we defined $w = 0.0001$ and $m = 0.8$.

2.2. Availability features

We defined the *availability* features (f^{avail}) as the single (most representative) LA score for each term $t_j \in (\mathcal{V}_{y_\oplus} \cup \mathcal{V}_{y_\ominus})$. Thus, to obtain the $f^{\text{avail}}_{t_j}$ score of term t_j we apply the CombMNZ [22] data-fusion strategy. Data-fusion strategies aim at integrating many possible answers (scores) for an object into a single

[2] We'll refer as documents to the transcribed text obtained from the subjects' utterances.

best representative score. Therefore, to compute the representative score of t_j we first obtain the available lexicon $\mathcal{V}_\rho$ of the instance ρ applying Eq. 2. Then, for obtaining the $f_{t_j}^{\text{avail}}$ we fuse the scores of word t_j from the list $\mathcal{V}_\rho$ with the available lexicons $\mathcal{V}_{y_\oplus}$ and $\mathcal{V}_{y_\ominus}$. For this process, we do as follows:

$$f_{t_j}^{\text{avail}} = \text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\}) \quad (3)$$

where t_j is the word for which we want a fused score, k indicates the maximum position where t_j will be searched in the input lists, and the $\mathcal{V}$'s are the set of lists to be considered for the fusion process. Notice that k has the same interpretation of that in Eq. 2; intuitively, it indicates the number of words (features) to be considered for building the representation vector.

Thus, assuming $N = \text{len}(\{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\})$, D^c as the score of t_j in list c, and $|D^c > 0|$ as the number of non-zero scores given to t_j by any list c, the final score for each unique term t_j is computed as follows:

$$\text{CombMNZ}(t_j, k, \{\mathcal{V}_\rho, \mathcal{V}_{y_\oplus}, \mathcal{V}_{y_\ominus}\}) = \sum_c^N D^c \times |D^c > 0| \quad (4)$$

Broadly speaking, the $f_{t_j}^{\text{avail}}$ of term t_j represent a weight value indicating to what category it adjust the best.

2.3. Correlation degree features

The *correlation degree* features aim at measuring the relationship between the two sets of paired words, particularly we compute $cor(\mathcal{V}_{y_\ominus}, \mathcal{V}_\rho)$, and $cor(\mathcal{V}_{y_\oplus}, \mathcal{V}_\rho)$. The correlation ($cor$) value will be an indicator of the association between the available lexicon form subject ρ and the corresponding $\mathcal{V}_{y_\ominus}$ and $\mathcal{V}_{y_\oplus}$ categories. For the experiments performed in this paper, every $f_{t_j}^{\text{cor}}$ feature is formed by two values, the Spearman's correlation coefficient and its corresponding p-value.

2.4. Acoustic based method

The acoustic based method directly models raw waveforms to predict the class-conditional probabilities using a CNN-based architecture. As described in [23], the architecture consists of four 1-D convolutional layers, followed a hidden layer and an output-layer. In order to guide the learning procedure, depending on the task, different approaches were previously proposed: We distinguish between sub-segmental and segmental filtering (see [23] Table 1); raw waveforms can be filtered to extract voice-source related characteristics to guide the learning procedure. Specifically, for the depression detection task, the primary method (denoted as 1stAcoustic) uses zero frequency filtering to get a signal that characterizes the glottal excitation. The secondary method (denoted as 2ndAcoustic) consists of modeling speech at a frame level using linear prediction and subtracting it from the original speech to get the linear prediction residual, which contains voice source related characteristics, while both use an input length of 250ms. However, for Alzheimer's detection, both systems use 4 second length inputs of zero frequency filtered signals, where the primary method (denoted as 1stAcoustic) applies a sub-segmental filtering stage, the secondary method (denoted as 2ndAcoustic) a segmental filtering stage.

2.5. Late fusion

Once both the language-based and acoustic-based modalities are trained independently, the late fusion approach consists of a voter that takes as inputs the predictions made by the language-based and acoustic-based approaches. The final decision is made by means of a majority voting mechanism, where if tied, the output will be always labeled as *C* (i.e., control).

Mod.	Approach	DAIC-WOZ Class.	DAIC-WOZ F1-score *O*	*D*	*C*	ADReSS Class.	ADReSS F1-score *O*	*D*	*C*
Textual	BoW	MLP	0.65	0.48	0.83	SVC	0.84	0.83	0.86
	LIWC	MLP	0.53	0.34	0.72	LR	0.70	0.70	0.70
	BERT	SVC	0.70	0.53	0.86	MLP	0.73	0.74	0.72
	LA-A$_{100}$	PER	0.58	0.40	0.77	SVC	0.77	0.74	0.79
	LA-A$_{500}$	MLP	**0.71**	**0.58**	0.84	LR	0.84	0.83	0.86
	LA-A$_{1000}$	MLP	0.71	0.56	**0.87**	LR	0.84	0.83	0.86
	LA-AC$_{100}$	PER	0.57	0.41	0.73	MLP	0.77	0.75	0.80
	LA-AC$_{500}$	MLP	0.68	0.53	0.83	LR	0.86	0.85	0.87
	LA-AC$_{1000}$	MLP	0.66	0.45	0.86	LR	**0.87**	**0.86**	**0.88**
Acoustic	1stAcoustic	-	**0.58**	**0.41**	**0.76**	-	**0.76**	**0.69**	**0.90**
	2ndAcoustic	-	0.52	0.32	0.71	-	0.76	0.71	0.88

Table 1: *Performance under a 10-CFV strategy on train sets.*

3. Experimental Setup

For the experiments, we use the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset [24] and the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) dataset [9]. The DAIC-WOZ dataset contain semi-structured clinical interviews, performed by an (human controlled) animated virtual interviewer, designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic disorder. This dataset was used during the AVEC 2016 challenge [7], and contains audio-visual interviews of 189 participants: 107 for training, 35 for development, and 47 for test. The ADReSS data, introduced for the Interspeech 2020 ADReSS challenge [9], consists of speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [25]. It contains speech and transcripts information from 156 participants: 108 for training, and 48 for test. In the DAIC-WOZ dataset approximately $\approx$30% of the subjects are labeled as depressed (*D*), while the ADReSS data is perfectly balanced. It is worth mentioning that the labeling of each datset was done by expert mental healthcare providers, interested reader is referred to [24, 9].

We evaluate the performance of three well-known text-based methods. First, a traditional Bag-of-Words (BoW) using the top 1000 most frequent words under a Term Frequency Inverse Document Frequency *tf-idf* weighting scheme. Secondly, we use the Linguistic Inquiry and Word Count (LIWC) [26] categories for representing the documents. LIWC psychological categories capture the semantic content of the language produced [27], e.g., allow to detect positive vs. negative emotions, words referencing family/friends/society, pronouns which can capture inclusive language vs. exclusive language, and words referencing how the person is feeling.

As third baseline, we evaluate the impact of recent transformer-based models [28] as a language representation strategy. For our experiments we test an English pre-trained BERT model. As known, the [CLS] token acts an "aggregate representation" of the input tokens, and is considered as a sentence representation for many classification tasks [29]. Accordingly, for generating the representation of each document, we split the document into smaller chunks (max length of 512 tokens), obtain the [CLS] encoding of each chunk, and we apply

Mod.	Approach	DAIC-WOZ Class.	DAIC-WOZ F1-score *O*	DAIC-WOZ F1-score *D*	DAIC-WOZ F1-score *C*	ADReSS Class.	ADReSS F1-score *O*	ADReSS F1-score *D*	ADReSS F1-score *C*
Textual	BoW	MLP	0.53	0.32	0.75	LR	0.85	0.84	0.86
	LIWC	MLP	0.49	0.29	0.69	SVC	0.62	0.57	0.67
	BERT	MLP	0.51	0.30	0.72	SVC	0.81	0.80	0.82
	LA-A_{100}	DT	0.63	0.54	0.73	SVC	0.73	0.70	0.76
	LA-A_{500}	MLP	0.54	0.36	0.71	MLP	0.85	0.86	0.85
	LA-A_{1000}	DT	0.58	0.40	0.76	LR	0.85	0.85	0.86
	LA-AC_{100}	SVC	**0.70**	**0.64**	0.76	MLP	0.75	0.71	0.79
	LA-AC_{500}	MLP	0.51	0.25	**0.79**	LR	**0.87**	**0.88**	**0.86**
	LA-AC_{1000}	PER	0.60	0.48	0.71	LR	0.81	0.82	0.80
Acoustic	1stAcoustic	-	**0.69**	**0.65**	**0.73**	-	**0.79**	**0.82**	**0.75**
	2ndAcoustic	-	0.55	0.53	0.57	-	0.68	0.72	0.65

Table 2: *Obtained performance over the dev and test partitions for DAIC-WOZ and ADReSS datasets respectively.*

Dataset	Fused approaches	F1-score *O*	F1-score *D*	F1-score *C*
DAIC-WOZ	[LA-AC_{100}, LA-A_{100}, 1stAcoustic, 2ndAcoustic]†	**0.84**	**0.80**	**0.89**
	[LA-AC_{100}, LA-A_{100}, BoW, 1stAcoustic, 2ndAcoustic]†	0.82	0.77	0.86
	[LA-AC_{100}, LA-A_{100}, BERT, 1stAcoustic, 2ndAcoustic]	0.79	0.74	0.84
	Al Hanai,T., *et al.* (2018) [16]	*0.77*	-	-
ADReSS	[LA-AC_{500}, LA-A_{500}, 1stAcoustic, 2ndAcoustic]	**0.90**	**0.90**	**0.89**
	[LA-AC_{500}, LA-A_{500}, BoW, 1stAcoustic, 2ndAcoustic]	0.85	0.87	0.84
	[LA-AC_{100}, LA-A_{100}, BERT, 1stAcoustic, 2ndAcoustic]	**0.90**	**0.90**	**0.89**
	Mahajan, P. & Baths, V., (2021) [17]	-	*0.70*	*0.75*

Table 3: *Obtained performance of the late fusion approach. The reported performance in [7] for depression was F1=0.58, while for ADReSS, in [9] the best reached score was F1=0.75.*

a *mean* pooling to obtain the final representation.

Except for the BERT setup, we applied the following normalization steps; all the common contractions, e.g., *we'll*, *can't*, etc., are converted to its formal writing, i.e., *we will*, *can not*, etc. All disfluencies are preserved, non-speech phenomena are labeled as `<non-speech>`, punctuation marks are removed, and number occurrences are labeled as `<number>`, and, all the text is lower cased.

4. Results and discussion

As previous research [7, 9, 23, 30, 31], performance is reported in terms of the F score ($F1$) for both control (*C*) and depression/dementia (*D*) classes, and the Macro-F for the overall problem *(O)*. We acknowledge the limitations regarding the small size of the corpora, however, this is a common shortcoming of all studies that use clinical datasets. Thus, to achieve stable and robust results, we applied two validation strategies: i) the average performance over a stratified 10 cross-fold-validation using *train* partition (10-CFV), and, ii) the performance over the *dev* partition for the DAIC-WOZ[3] dataset and on the *test* partition for the ADReSS dataset.

For the proposed Lexical Availability method, we performed a series of experiments using: *i)* only the *availability degree* features (LA-A), and *ii)* the combination of availability and correlation (LA-AC) as in Eq. 1. Table 1 summarizes our results for the experiments using a 10-CFV strategy; Table 2 shows the performance of the experiments performed on the *dev* and *test* partitions, and Table 3 shows the results of the fused predictions. Given our space restrictions, we only report results from the best learning algorithm (Class. column).[4] For the experiments using the LA-A/LA-AC methods, the number in the sub-index indicates the value of the k parameter.

Clearly, from Tables 1 and 2 we conclude that our LA method outperforms all the proposed textual-based baselines, including very recent transformer-based models (i.e., BERT). Also, observe that adding the *correlation* features helps improving the classification, best performance is obtained under the LA-AC configuration for both tasks (see Table 2) with $k = 100$ for DAIC-WOZ, and $k = 500$ for ADReSS. This variation in the value of k is related to the size of the respective datasets. For instance, the DAIC-WOZ corpus, contrary to the ADReSS dataset, contains more samples of the communicative process (i.e., several utterances from interviewed subject) with a smaller variability of lexical units (i.e., small vocabulary), hence paying attention to a reduced set terms is enough for the LA method.

For the multi-modal experiments (Table 3) we took the best configurations based on the performance on the *dev/test* sets (Table 2). We compare our results against two recent multi-modal approaches. For depression, we considered the work of [16], which evaluates the performance of a multi-modal LTSM recurrent network. For dementia, [17] combines the outputs of CNN-LSTM model and a Speech-GRU cell for making the predictions. As can be observed, our late fusion strategy, outperforms very recent approaches by an important margin.[5]

5. Conclusions

We addressed the problem of detecting mental disorders from clinical tests. Inspired by the LA theory, our method approximates the *mental lexicon* through the identification of the *available lexicon* for mentally ill and control subjects, and use it in a classification process to detect depression/dementia. Additionally, based on previous studies that demonstrated the suitability of raw waveform CNNs, we designed a multi-modal approach, where a voter makes the final decision using a majority vote mechanism. A thorough evaluation in two well known clinical datasets (DAIC-WOZ and ADReSS), shows that the LA method fused with the raw waveform-based CNN is able to outperform, by a large margin, very recent deep NN techniques.

6. Acknowledgements

Esaú Villatoro-Tello, was supported partially by Idiap, SNI-CONACyT, and UAM-Cuajimalpa Mexico. S Pavankumar Dubagunta was supported by Innosuisse under the project Conversation Member Match (CMM). Julian Fritsch was funded through the EU's Horizon H2020 MSCA-ITN-ETN project TAPAS grant agreement No. 766287. Gabriela Ramírez-de-la-Rosa would like to thank UAM-Cuajimalpa for their support.

[3] DAIC-WOZ *test* partition is not publicly available.

[4] Classifiers parameters: Logistic Regresor (**LR** - `solver=lbfgs`), Multilayer Perceptron (**MLP** - `activation=relu, alpha=1e-5, solver=lbfgs, max_iter=300`), Support Vector Machines (**SVC** - `kernel=linear`), Decision Trees (**DT** - `criterion=entropy`, and Perceptron (**PER** - `max_iter=50, tol=1e-3`). All classifiers were set with `random_state=42`.

[5] Symbol † indicates statistical significant results (based on the Wilcoxon signed-rank test with a 90% confidence) in comparison to the non-fused results.

7. References

[1] World Health Organization, *Depression and Other Common Mental Disorders Global Health Estimates*. World Health Organization, 2017. [Online]. Available: https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

[2] ——, *Towards a dementia plan: a WHO guide*. World Health Organization, 2018. [Online]. Available: https://www.who.int/mental_health/neurology/dementia/policy_guidance/en/

[3] T. Wykes, J. Lipshitz, and S. M. Schueller, "Towards the design of ethical standards related to digital mental health and all its applications," *Current Treatment Options in Psychiatry*, vol. 6, no. 3, pp. 232–242, 2019.

[4] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e19, 2017.

[5] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, p. e12106, 2018.

[6] C. Welch, A. Lahnala, V. Perez-Rosas, S. Shen, S. Seraj, L. An, K. Resnicow, J. Pennebaker, and R. Mihalcea, "Expressive interviewing: A conversational system for coping with COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, Dec. 2020.

[7] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*. ACM, 2016, pp. 3–10.

[8] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19. Association for Computing Machinery, 2019, p. 3–12.

[9] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," in *Proceedings INTERSPEECH 2020*, Shanghai, China, 2020.

[10] A. M. Tackman, D. A. Sbarra, A. L. Carey, M. B. Donnellan, A. B. Horn, N. S. Holtzman, T. S. Edwards, J. W. Pennebaker, and M. R. Mehl, "Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis." *Journal of personality and social psychology*, vol. 116, no. 5, p. 817, 2019.

[11] E. Villatoro-Tello, S. Parida, S. Kumar, P. Motlicek, and Q. Zhan, "Idiap & UAM participation at GermEval 2020: Classification and regression of cognitive and motivational style from text," in *Proceedings of the GermEval 2020 Workshop in conjunction with the 5th SwissText & 16th KONVENS Conference*, 2020, pp. 11–16.

[12] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, "Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease," *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.

[13] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[14] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes, "Detecting depression in social media using fine-grained emotions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 1481–1486.

[15] 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.

[16] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in *Interspeech*, 2018, pp. 1716–1720.

[17] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech," *Frontiers in Aging Neuroscience*, vol. 13, p. 20, 2021.

[18] N. Hernández-Muñoz, C. Izura, and C. Tomé, "Cognitive factors of lexical availability in a second language," in *Lexical availability in English and Spanish as a second language*. Springer, 2014, pp. 169–186.

[19] M. Šifrar Kalan, "Lexical availability and l2 vocabulary acquisition," *Journal of Foreign Language Teaching and Applied Linguistics*, vol. 2, no. 2, 2015.

[20] S. De Deyne, S. Verheyen, and G. Storms, "Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations," in *Towards a theoretical framework for analyzing complex linguistic networks*. Springer, 2016, pp. 47–79.

[21] F. J. Callealta Barroso and D. J. Gallego Gallego, "Medidas de disponibilidad léxica: comparabilidad y normalización (measures of lexical availability: comparability and standardization)," *Boletín de filología*, vol. 51, no. 1, pp. 39–92, 2016.

[22] E. A. Fox and J. A. Shaw, "Combination of multiple searches," *NIST special publication SP*, vol. 243, 1994.

[23] S. P. Dubagunta, B. Vlasenko, and M. M. Doss, "Learning voice source related information for depression detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6525–6529.

[24] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[25] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[26] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[27] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[30] A. Rinaldi, J. Fox Tree, and S. Chaturvedi, "Predicting depression in screening interviews from latent categorization of interview prompts," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7–18.

[31] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, "A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition," in *INTERSPEECH 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.