COMPARATOR LOSS: AN ORDINAL CONTRASTIVE LOSS TO DERIVE A SEVERITY SCORE FOR SPEECH-BASED HEALTH MONITORING

Jacob J Webber¹, Oliver Watts¹, Lovisa Wihlborg¹, David Wheatley¹, Johnny Tam^{2,4}, Christine Weaver^{2,4}, Suvankar Pal^{2,4,5}, Siddharthan Chandran^{2,4,5}, Cassia Valentini-Botinhao¹

¹SpeakUnique Ltd., UK, ²Anne Rowling Regenerative Neurology Clinic,
University of Edinburgh (UoE), UK, ³Institute of Adaptive and Neural Computation, UoE, UK,

⁴Euan MacDonald Centre for MND Research, UoE, UK, ⁵UK Dementia Research Institute, UK

ABSTRACT

Monitoring the progression of neurodegenerative disease has important applications in the planning of treatment and the evaluation of future medications. Whereas much of the state-of-the-art in health monitoring from speech has been focused on classifying patients versus healthy controls, or predicting real-world health metrics, we propose here a novel measure of disease progression: the severity score. This score is derived from a model trained to minimize what we call the comparator loss. The comparator loss ensures scores follow an ordering relation, which can be based on diagnosis, clinically annotated scores, or simply the chronological order of the recordings. In addition to giving a more detailed picture than a simple discrete classification, the proposed comparator loss-based system has the potential to incorporate information from disparate health metrics, which is critical for making full use of small health-related datasets. We evaluated our proposed models based on their ability to affirmatively track the progression of patients with motor neuron disease (MND), the correlation of their output with clinical annotations such as ALSFRS-R, as well as their ability to distinguish between subjects with MND and healthy controls.

Index Terms— Health monitoring, severity prediction, contrastive loss, representation learning

1. INTRODUCTION

Neurodegenerative disorders are becoming more prevalent globally, posing a significant health challenge. They include progressive and fatal diagnoses affecting cognition such as Alzheimer's disease, and neuromuscular disorders such as motor neuron disease (MND), the most common subtype of which is amyotrophic lateral sclerosis (ALS). There is a growing need for diagnostic and monitoring methods that are both accurate and accessible [1]. Recent advances in speech technology and machine learning have highlighted the potential of using speech for detecting diseases and predicting their severity [2, 3]. Speech offers great potential as an objective digital biomarker for these conditions, being non-invasive and easily collected remotely without extensive expertise. State-of-the-art speech-based health monitoring methods focus on either categorising conditions (typically as healthy versus diagnosed) or predicting scores equivalent to those from existing medical tests. Classification tends to be the focus of most work in the area, epitomised by Challenge tasks such as [4]. For an overview of recent progress in predicting health status between healthy controls and patients with MND and Alzheimer's, see [2] and [3]. Speech can also be used to monitor progression of a disease, in which case the task would entail regressing to a clinical score. These scores include, for ALS, ALSFRS-R [5] and ECAS [6]; and for dementia, ACE-III [7].

Although classifying speech as healthy or non-healthy might be simpler, this result alone offers limited insight and becomes particularly problematic if inaccurate. Conversely, an automatic tool that replicates a clinician's score is only as effective as the score itself. Clinically annotated scores have been found to be unreliable and subjective. Authors in [8] found poor correlation between ALSFRS-R and another clinical score as the disease progresses. The work in [9] shows how different trial centers use the ALSFRS-R scale. In [10], the authors discuss the problems with the ALSFRS-R being aggregated from several items (like speech, stairs and food) with different metric properties. Rather than learning annotated scores directly, the authors in [11] propose a model that predicts longitudinal shifts in depression from two speech samples. Even though the model is not learning to predict clinical scores, the shift directions used for training were obtained from clinically annotated data, the eight-item patient health questionnaire [12].

Our proposed method follows a similar concept. Instead of predicting severity scores through regression to clinical scores, we aim to learn scores indirectly based on the order we believe governs the data. To do that we introduce a new contrastive loss for neural network training, the comparator loss, aimed at operating on a continuous-valued score that can be used to indicate disease severity, progression, or simply diagnosis status. The comparator loss relies on the existence of an ordering system. An ordering system can be derived from clinical test scores, more simply from the time since diagnosis, or in its most simple form from the presence of a diagnosis. Our loss is designed to impose an order upon these scores based on the relationship between speech samples. Specifically, it enforces that samples that are ranked higher, based on a particular ordering system, are also scored higher and vice versa.

The proposed method offers several advantages over state-ofthe-art classification techniques. The comparator loss enables the development of scoring models that can learn from different sources of potentially noisy information, extracting deeper insights from them. It provides not only a more detailed picture than a discrete classification loss, but also facilitates fine-tuning across different tasks, which is crucial when dealing with small health-related datasets. Typical classifier-based models output a set of logits, meaning the number of classes is hard-coded into the model and its weights. Conversely, the comparator loss requires only the existence of an order, regardless of the number of classes. A model trained with this loss does not require 'structural' changes to handle data with a different number of classes.

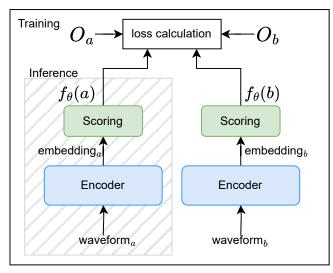


Fig. 1: Model structure.

2. BACKGROUND

2.1. Category learning

Categorical losses, such as cross-entropy, are well-established in the field of machine learning and commonly used in medical tasks. These make no assumption on the ordering of classes, meaning there is no inductive bias encoded in the model that categories are sequential. However, cross-entropy has also been applied to ordinal classes, such as the bits in binary encoding [13].

Authors in [14] propose the NRRank method, an extension of cross-entropy loss that incorporates ordering knowledge. The original cross-entropy loss requires the model to assign a probability to each class, with the condition that these sum to one. The loss is calculated as the difference between the predicted distribution and an ideal distribution, where the ideal distribution is taken to be a one-hot vector encoding the correct category. Instead of this one-hot encoding, the NRRank method involves constructing a target vector where all categories less than or equal to the ground-truth category, according to an ordering scheme, are assigned a one, whereas all subsequent categories in the vector are assigned a zero. The loss can be calculated by comparing this target vector with a vector predicted by the model using either a mean square error or the relative entropy. NRRank has recently been applied to a speech task, specifically predicting Mean Opinion Score (MOS) for automatic speech synthesis evaluation [15].

2.2. Representation learning and contrastive losses

Representation learning is a suite of techniques used to learn representations, typically a vector assigned to each sample, for later downstream tasks. The relationship between different data samples can be used to guide representation learning. For instance, in contrastive representation learning [16], similar samples are encouraged to be close in the embedding space, while dissimilar samples are encouraged to stay apart. An early example is the work described in [17] where authors introduce a contrastive loss for dimensionality reduction. Their proposed model is trained to learn a low-dimensional manifold from images. Pairs of images that are of similar categories are mapped to nearby points in that manifold. To prevent the representation space collapsing, a contrastive term is introduced, forcing image pairs from different categories apart. Authors in [18] extend

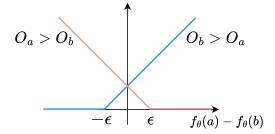


Fig. 2: The comparator loss function. O_a , O_b is the ground-truth ordinal value associated with sample a and b. f_{θ} is the model as applied to the sample. ϵ is a hyperparameter. The curves show the loss in each ordinal case for a pair of samples. The loss penalises situations where the orders of the ordinal values and the orders of the model outputs are not in agreement.

this work by proposing a triplet-based contrastive loss, whereby at each training step a training sample is compared to both a positive and negative sample (i.e. a sample within and outwith the category of the sample under consideration respectively). This was further extended into the so-called lifted structured loss [19], which compares all possible pairs in a batch for the purposes of calculating the contrastive loss. Unsupervised learning methods use contrastive learning by relying on data augmentation for creating positive and negative sample pairs [16].

3. METHOD

3.1. Model

Fig. 1 illustrates the general structure of the proposed comparator model. The model is composed of an encoder that outputs an embedding given a speech waveform and a scoring network that takes this embedding and converts it into a scalar value $f_{\theta}(.)$. We refer to this value as the 'score' as it can be used as a proxy for a severity score. During training the model learns to generate a score for an input waveform by minimising a 'comparator' loss. This loss is updated based on pairs of predicted scores and the 'orders' O associated with the inputs that produced them. While training requires pairs of stimuli, during inference the model produces scores from a single input waveform.

3.2. Proposed loss function

Fig. 2 illustrates the comparator loss curves given the relative difference between predicted scores $f_{\theta}(a) - f_{\theta}(b)$. The comparator loss is designed to ensure that predicted scores reflect the ordering of the elements, so that an element that is ranked higher has a greater score. The proposed loss function is similar to a traditional contrastive loss [17], with two key differences. The first is that instead of predicting a vector and using cosine differences, it uses a model that outputs a single scalar value, assigning a single score to samples. The second is that rather than assigning similarity or difference based on whether a pair of samples are from similar or different categories, it instead depends on an assigned *order* to the categories, penalizing results where for a pair of samples, a higher score is given to the sample of a lower order of category.

For each pair of samples (a, b), there are three possible scenarios: $O_a > O_b$, $O_a = O_b$ or $O_b > O_a$, where the in/equalities are determined solely by the order of the samples' classes. By swapping the ordering of the samples in the pair we can ensure that $O_b \ge O_a$,

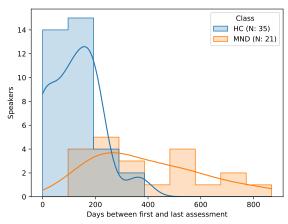


Fig. 3: Date range of assessments for test-set speakers.

reducing the number of cases to two. For pairs of samples in the same class, we define the loss as zero. Otherwise, for when $O_b > O_a$, the loss is defined as:

$$J = \max((f_{\theta}(a) - f_{\theta}(b)) + \epsilon, 0) \tag{1}$$

where ϵ is a hyper-parameter designed to stop rewarding the model once the scores are correctly ordered beyond a certain threshold. We can see in Fig. 2 that when O(b) > O(a) (blue curve) the loss is zero when $f_{\theta}(b) - f_{\theta}(a) > \epsilon$.

4. EVALUATION

4.1. Data

We used the Anne Rowling Neurological Speech Corpus [20] for training and evaluation of the proposed system. The corpus contains speech recordings of 780 individuals, grouped according to their documented diagnosis. We used data from the healthy control group (HC) of 170 individuals and the motor neuron disease (MND) group of 103. Determining whether the severity score could accurately track the progression of MND (i.e. determining if the severity score increases with time) was of particular interest. To maximise the longitudinal data available for analysis of results, the dataset was split by keeping the top decile of speakers in terms of number of assessments (where an assessment is a set of recordings gathered from a subject on the same day) as a held-out test set. This results in a test set (shown in Fig. 3) with 56 speakers and 1,112 utterances, a validation set with 27 speakers and 219 utterances, and a training set of 197 speakers and 1,094 utterances. The decision to hold out speakers with the largest number of assessments is made due to the particular difficulty in evaluating systems according to their ability to monitor progression of disease, which is of particular importance to clinicians. Longitudinal health data is particularly challenging to gather, so this choice of split enables us to measure the progression performance while training on a relatively limited amount of longitudinal data, modelling a challenging, but likely real-world scenario. Each assessment includes a number of speech tasks, including reading fixed passages, and freer picture description tasks. Other work conducted as part of this project has indicated that fixed passage reading is particularly useful for MND detection, but a detailed analysis of these subtasks is outside the scope of this paper. We therefore used all subtasks from each assessment as separate recordings. The corpus contains annotation for disease severity (ALSFRS-R) that is given at the assessment level; we evaluate the correlation of our learned severity scores with these clinical ratings.

4.2. Proposed model

The encoder we used in our model is based on TitaNet [21]. TitaNet is a model trained to produce speaker embeddings. The model is trained on the additive angular margin loss [22] that optimises the cosine distance between embeddings by 'enforcing higher similarity for intra-class samples and diversity for inter-class samples'. We modified TitaNet by replacing the final layer (that converts the 192-size embedding into a logit output that encodes speaker-id) with a linear layer that converts embeddings into a single scalar value, the score. TitaNet was adopted as an exemplar of a successful speech embedding model capable of representing speaker information in a compact way. We note however that this choice is not crucial – the comparator loss we present can be used in conjunction with any neural architecture that can convert speech into embeddings of fixed dimensionality.

We generated scores for every example in a batch and calculated the comparator loss for every possible pair within the batch. The sum of these pairwise loss terms was then used to update the model. We set the margin value ϵ to that used in [17]. We trained all models for 100,000 training batches of size 96 using Adam as the optimizer. Early stopping was performed using a separate validation set, where checkpoints were only saved when the validation accuracy exceeded all previously achieved values. This accuracy was determined by training a threshold classifier on the output severity scores as described in section 4.4.

A key benefit of the proposed system is that it allows an arbitrary number of non-normalised training labels, each with a different range and distribution, to be used to supervise training of a single 'severity score' by calculating the mean of losses calculated for each ordinal label over each training batch. Experiments were run with two systems trained using the comparator loss. One, *Ablated*, was trained using only the category (healthy vs MND) of the sample in question, the other, *Proposed*, also made use of ALSFRS-R speech subscores.

4.3. Baselines

We constructed two baselines. For the first (*Cross-Entropy*), we adopt the cross-entropy loss to train TitaNet to classify between healthy control and diagnosed speech. The probability of MND given by this model is used as its severity score. For the second baseline (*Contrastive*), we adopt the original contrastive loss [17]. This method aims simply to predict a single-valued vector where distinct classes are maximally different, and similar classes have equal scores.

4.4. Results

We evaluate our systems based on three metrics:

- Accuracy on the task of classifying between speech from healthy
 controls and people with an MND diagnosis. We expect that a
 true severity score should be able to categorise samples into those
 classes with a high degree of accuracy. The accuracy is computed
 by determining the 'oracle' optimal threshold for the model's
 score on the test set.
- Progression of disease severity. Where longitudinal data exist for a patient, a severity score should typically increase for positive cases and remain approximately constant for healthy controls. The progression is calculated by numerically calculating the score gradient with respect to time across the entire test set.
- Correlation: the severity score should correlate with existing clinical scores for tracking disease. Correlation is measured using Spearman rank correlation.

	Classification			ALSFRS-R								ECAS	
System	Accuracy	AUC	F1	ρ_T	p-value	ρ_{sp}	p-value	ρ_{st}	p-value	ρ_f	p-value	ρ_T	p-value
Cross-entropy	0.71	0.70	0.74	0.087	0.055	-0.59	3.3e-48	0.28	2.4e-10	0.15	0.00098	-0.23	8.2e-08
Contrastive	0.54	0.52	0.64	0.059	0.19	-0.04	0.43	0.05	0.28	0.14	0.0016	0.17	7.3e-05
Ablated	0.73	0.73	0.76	-0.084	0.064	-0.53	5.3e-37	0.16	0.00038	-0.07	0.13	-0.33	1e-15
Proposed	0.74	0.74	0.77	-0.001	0.98	-0.63	1.3e-54	0.23	1.8e-07	-0.01	0.82	-0.43	6.8e-26

Table 1: Accuracy results on diagnosis detection and correlation to clinical scores. ρ_T denotes total score while ρ_{sp} , ρ_{st} and ρ_f to the speech, stairs and food items of the ALSFRS-R score. Negative correlation indicates agreement with clinical assessment. Correlations in bold are statistically significant with p > 0.05.

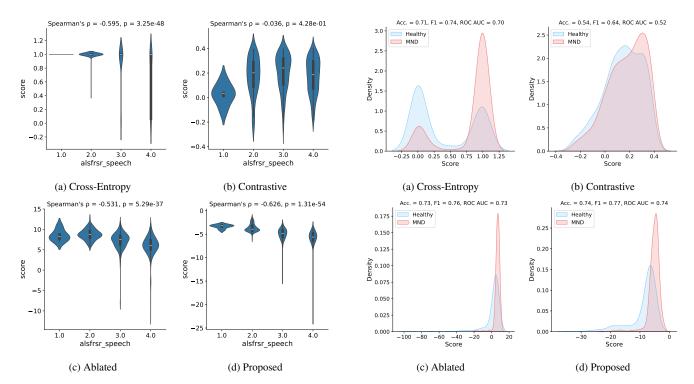


Fig. 4: Predicted score against the speech item of ALSFRS-R.

Fig. 5: Distribution of scores for healthy controls and MND.

Table 1 provides classification scores (accuracy, AUC and F1), as well as correlation values (ρ) with clinical scores. *Proposed* achieved the highest accuracy in the diagnosis classification task, outperforming even *Cross-Entropy* that is optimized specifically for classification. No significant correlation is observed in Table 1 between the predicted scores from any of the evaluated models and the ALSFRS-R total score (ρ_T) . However, for all models except *Contrastive*, the predicted scores did correlate with the ALSFRS-R speech subscale (ρ_{sp}) , and the same models' scores also correlated with ECAS ratings.

Fig. 4 presents the distributions of predicted severity scores against different levels of clinical assessment derived from the speech subscale of ALSFRS-R where the negative correlations can be observed. Figure 5 shows the distribution of scores for healthy controls and participants with MND. The score distribution from *Proposed* and *Contrastive* is not as bimodal as the distribution derived from *Cross-Entropy*, which suggests the proposed method's potential to provide finer grained measurements than systems based on classification.

To determine whether the predicted scores could accurately track the progression of MND we computed the gradient over time of the scores for participants with multiple assessments. We found that the computed gradients were close to zero for both healthy and MND participants regardless of the model, indicating minimal score variation over time. While expected for healthy participants, this may not apply to those with MND. This highlights the difficulty of disease progression monitoring in the typical case where longitudinal training data is not available.

5. CONCLUSIONS

A new loss has been proposed; the predictions of models trained with it correlate significantly with clinical judgments. While it is not yet possible to use the scores analysed here to confirm progression of disease using longitudinal data, the system compares favourably with strong baselines. As the design of the loss enables the exploitation of heterogenously labelled data, it is expected that further rapid improvements will be possible by increasing training dataset size and diversity.

Acknowledgement: This work is supported by NEURii, a collaborative partnership involving the University of Edinburgh, Gates Ventures, Eisai, LifeArc and Health Data Research UK (HDR UK).

6. REFERENCES

- [1] Alize J Ferrari et al., "Global incidence, prevalence, years lived with disability, disability-adjusted life-years, and healthy life expectancy for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet*, vol. 403, no. 10440, pp. 2133–2161, 2024.
- [2] Molly Bowden et al., "A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease," NPJ digital medicine, vol. 6, no. 1, pp. 228, 2023.
- [3] Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz, "Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.
- [4] Saturnino Luz et al., "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proc. Interspeech*, Shanghai, China, 2020.
- [5] Jesse M Cedarbaum et al., "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function," *Journal of the neurological sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.
- [6] Christopher Crockford et al., "ALS-specific cognitive and behavior changes associated with advancing disease stage in ALS," *Neurology*, vol. 91, no. 15, pp. e1370–e1380, 2018.
- [7] Pavagada S Mathuranath et al., "A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia," *Neurology*, vol. 55, no. 11, pp. 1613–1620, 2000.
- [8] Andrei Voustianiouk et al., "ALSFRS and appel ALS scores: discordance with disease progression," Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine, vol. 37, no. 5, pp. 668–672, 2008.
- [9] Jaap NE Bakers et al., "Using the ALSFRS-R in multicentre clinical trials for amyotrophic lateral sclerosis: potential limitations in current standard operating procedures," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 23, no. 7-8, pp. 500–507, 2022.
- [10] Franco Franchignoni, Jessica Mandrioli, Andrea Giordano, Salvatore Ferro, and ERRALS Group, "A further rasch study confirms that ALSFRS-R does not conform to fundamental measurement requirements," *Amyotrophic Lateral Sclerosis* and Frontotemporal Degeneration, vol. 16, no. 5-6, pp. 331– 337, 2015.
- [11] Paula Andrea Pérez-Toro et al., "Longitudinal modeling of depression shifts using speech and language," in *Proc. ICASSP*, 2024, pp. 12021–12025.
- [12] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1, pp. 163–173, 2009.
- [13] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *Proc.* ICML, 2018, vol. 80, pp. 2410–2419.

- [14] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri, "A neural network approach to ordinal regression," in *Proc. IJCNN*, 2008, pp. 1279–1284.
- [15] Marie Kunešová, Jindřich Matoušek, Jan Lehečka, Jan Švec, Josef Michálek, Daniel Tihelka, Martin Bulín, Zdeněk Hanzlíček, and Markéta Řezáčková, "Ensemble of deep neural network models for MOS prediction," in *Proc. ICASSP*, 2023, pp. 1–5.
- [16] Lilian Weng, "Contrastive representation learning," lilianweng.github.io, May 2021.
- [17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. CVPR*, 2006, vol. 2, pp. 1735–1742.
- [18] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, June 2015.
- [19] Hyun Oh Song et al., "Deep metric learning via lifted structured feature embedding," in *Proc. CVPR*, 2016.
- [20] Johnny Tam et al., "Anne Rowling Neurological Speech Corpus: clinically annotated longitudinal dataset for developing speech biomarkers in neurodegenerative disorders," in *Proc. Interspeech*, 2025.
- [21] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, "Ti-taNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *Proc. ICASSP*, 2022, pp. 8102–8106.
- [22] Jiankang Deng et al., "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4685– 4694.