Automatic assessment of voice pathologies on the GRB scale, based on multimodal deep learning architectures

Julián D. Arias-Londoño, *Senior Member, IEEE*, Jorge A. Gómez-García

***Abstract*—This paper addresses the automatic assessment of voice quality according to the GRB scale, based on the use of various deep learning architectures for prediction purposes. The proposed architectures are multimodal, because they employ multiples sources of information, and also multi-output because they simultaneously predict all the traits of the GRB scale. A feature engineering approach is followed, based on the use of deep neural networks and a set of well-established features such as MFCC, perturbation and complexity characteristics. Likewise, a representation learning is considered, using convolutional neural networks feed on modulation spectra extracted from voices. Finally, a variety of loss functions are also investigated, including two surrogate ordinal classification, a conventional weighed categorical cross-entropy, and a mean square error function. Experiments are carried out in a dataset containing registers of the sustained phonation of three vowels. The best deep learning architecture provides a relative performance improvement of 6.25% for G, 14.1% for R and 18.1% for B, in comparison with recently published results using the same dataset.**

***Index Terms*—Automatic voice quality analysis, Perceptual voice assessment, GRB scale, Deep Neural Networks**

I. Introduction

THE perceptual evaluation of voice consists on a subjective assessment, which is intended to describe qualitatively the vocal quality and degree of hoarseness that is present in the voice. The literature reports a variety of scales that are employed for perceptual evaluation purposes, which differ on the number of traits, levels within the trait, and the evaluation procedures that are followed. They also differ in the speech tasks that are considered, ranging from the phonation of sustained vowels, reading of predefined phonetically balanced sentences, to free monologues. Some scales are clinician-based while others are patient-based. In the first case, the specialist (e.g. ENT physician, phoniatrician or speech therapist) evaluates the patients voice, and reports the results according to the traits that are evaluated, whereas in the second case, the patient itself documents his/her perception about the presence, severity and impact of voice disorders on his/her own life [1]. The most used scale for perceptual assessment of dysphonic voices is the GRBAS scale [2]. This scale is composed of five categorical traits (or descriptors) ranging from 0 to 3, where 0 refers to normophonia and 3 to grave dysphonia. The categorical traits of the GRBAS scale are *Grade* (G), *Roughness* (R), *Breathiness* (B), *Asthenia* (A) and *Strain* (S), although a simplified scale limited to G, R and B, named GRB, is frequently found in the literature due to the unreliability of the A and S traits [3]. Typically, perceptual evaluations using then GRBAS scale employ sustained phonation of vowels /a/ and/or /i/ along with connected speech samples.

Despite perceptual evaluations are still widely used in the clinical management of voice disorders to quantify the extent of dysphonia [4], [5], they have been widely criticized because of the subjective process on which they stand and the lack of reliability they offer [6]. Indeed, perceptual assessments can be confounded by factors such as the listeners perceptual bias, experience, the type of rating scale that is used, fatigue, the perceptual sensitivity of the evaluator to a particular voice feature or to the voice sample being evaluated [7], [5]. Under these circumstances, voice assessments based on acoustic features and automatic systems are gaining attention due to the advantages they offer in regards to make the evaluation process objective. Certainly, automatic systems might provide accurate and reproducible graded measures of a patients voice quality, representing an objective help for the patients treatment and rehabilitation [8]. With this aim, some works have addressed the automatic assessment of voice quality. For instance in [8], an approach to automatically assess voice quality based on a seven-labels ranking scale was presented. A detector based on *Artificial Neural Networks* (ANN) was investigated in conjunction to a combination of short-term and long-term time-domain and frequency-domain parameters extracted from *Electroglotographic* (EGG) signals. The experiments were carried out in a corpus composed of 77 abnormal speech signals, using only one training/validating procedure. The best result was obtained with 21 features, yielding an average accuracy of 92%. Despite of the results, it should be noted that only the intra-speaker variability was considered during the cross validation step. When the inter-speaker variability was taken into account, the average accuracy decreased to a modest but remarkable 40% [9]. Another approach presented in [10], used three voice quality measures extracted from the spectral envelope to classify speech signals into a three-level rating scale, considering only the G trait of the GRBAS scale. The dataset that was employed contained recordings of 10 Parkinsons disease patients and 4 normophonic speakers. The authors concluded that the Itakura-Saito distortion provides good correlation with the perceptual evaluation and hence, it might be used for its prediction. It is worth noting though,

J.D. Arias-Londoño is with the Department of Systems Engineering, Universidad de Antioquia. Calle 67 No. 53 - 108, 050010, Medellín, Colombia. e-mail: julian.ariasl@udea.edu.co

J.A. Gómez-García is with the Bioengineering and optoelectronics lab (ByO). Universidad Politécnica de Madrid. Ctra. Valencia, km. 7, 28031. Madrid, Spain. e-mail: jorge.gomez.garcia@upm.es

the reduced number of recordings, and the lack of reported classification results. The work in [11], employed *Higher-Order Statistics* (HOS) -estimated from the *Linear Prediction Coding* (LPC) residual- and a detector based on decision trees for the classification of the G trait. The dataset was composed of 83 speech recordings distributed as follows: 20 normophonic voices, 17 recordings graded as G=1, 26 graded G=2, and 20 graded G=3. Experiments were carried out following a 5-folds cross-validation scheme with 70% of the recordings used for training and 30% for testing, yielding a 92.9% accuracy. The authors also compared their methods to those presented in [10] obtaining a decreased accuracy of 75.7%. In [12], a preliminary study for the automatic evaluation of the five traits of the GRBAS scale was presented. Characterization was carried out using short-time analysis of speech, computing the energy of each frame along with 15 *Mel-Frequency Cepstral Coefficients* (MFCC) and their first and second derivatives. Experiments were performed in private dataset composed of 433 normophonic and 215 pathological recordings, where 70% of them were used for training a detector based on *Learning Vector Quantization* (LVQ) and the remaining 30% for testing. The overall accuracy was 65%, but the results were evaluated without cross-validation. More recently, in [5] the automatic assessment of the G and R scales was addressed using MFCC and a group of *Modulation Spectra* (MS) morphological parameters. The experiments were performed using recordings of the sustained vowel /a/ extracted from the well-known MEEI database [13]. The authors reported an impressive accuracy of 81.6% for G and 84.7% for R, but when a careful selection of recordings (based on inter agreement criteria among several raters) was carried out. When such a selection was not taken into account, the performance dropped about 20%. In [14] the automatic assessment of the G trait was addressed using registers of Mandarin speakers. The proposed system employed cepstral coefficients, perturbation, energy and complexity measures. By using an extreme learning machine an accuracy of 80% was obtained. A similar set of features was included in [15] for the evaluation of G trait in Mandarin voices. The set of features included MFCC, MS, *Smoothed Cepstral Peak Prominence* (CPPS) and long-term average spectrum. Unlike the previous work, their main purpose was to evaluate the usefulness of a deep belief network in comparison to a more classical approach based on Gaussian mixture models, such as the one used in [5]. A relevant element in this case, is the fact that some of the features were extracted from the sustained vowel /a/ and others from running speech samples, although slightly differences in performance were reported in comparison to [14]. In [16] the automatic assessment of the scales G,R and B, was addressed using perturbation, spectral/cepstral, MS and complexity features for the characterization of the sustained phonations of the vowel /a/ in 3 datasets. The goal of the paper was to emulate the perceptual capabilities of a single evaluator that performed assessments on different corpora. In this case, the authors considered that the voice quality assessment according to the GRB scale, is indeed, an ordinal regression problem and treated it as such. Experiments were carried out using regression techniques and performance measures more suitable for the evaluation of this problem, evaluating the proposed approach in three cross-dataset scenarios and in a clinical setting. One of experiment, is related to the *Saarbrücken Voice Database* (SVD) that is freely available online [17]. The best results reported in the paper range between 0.5 to 0.7 according to an ordinal *Mean Absolute Error* (MAE) measure, indicating that in average, predicted labels deviate half an unit from the perceptual evaluation provided by the speech therapist.

In view of the state of the art, we can conclude that the attempts to objectively evaluate the quality of the speech, or to try to emulate perceptual capabilities of evaluators are quite heterogeneous, and their comparison is far from trivial. The lack of a common consensus regarding the use of a certain assessment scale, and the absence of labeled corpora available for the reproduction of results are clearly two of the main difficulties present in the field. Another issue often encountered is that, in most of the cases, the corpora used for the automatic voice quality assessment is not uniformly distributed in relation to their labels [5]. This might certainly bias the detection systems. In fact, in the GRB scale, the label 0 is commonly the most abundant and therefore the best predicted of the labels [18]. However, in most of the analyzed works the performance measures did not take into account this imbalance problem, or even worse, no information was even given about the distribution of the labels. Likewise, the most recent approaches agree on the use of multiple voice features to characterize the diverse phenomena involved in the voice production process, and to provide as much information as possible to the detection systems, but all the consulted learning models followed a classical approach based on a feature vector representation of the samples. This prevents the configuration of a multimodal approach based on concurrent and heterogeneous sources of information, e.g., spectra-based information extracted from voice or speech signals, which takes the form of a matrix, and feature vectors extracted from the spectra. In a similar fashion, there is a reported correlation between traits that is often ignored [19]. Indeed, in the analysis using the GRBAS scale, G is often considered a superclass that embodies R and B [3], but in most of literature this relationship is simply ignored and each one of the traits is studied separately. In light of these antecedents, this paper exploits the most recent advances in *Deep Learning* (DL), and proposes the use of a multimodal, multioutput neural network architecture for the automatic assessment of the GRB scale. DL models are basically ANN architectures with multiple and diverse layers, capable of learning arbitrary representations of the data. They have been successfully employed in different contexts, such as image processing and speech recognition [20], but also to automatic pathological voice detection [21]. The proposed DL architectures is multimodal since it combines dense layers to process, on one hand, vector shaped features with convolutional layers and to process, on the other, spectra-based representations of the voices signals. The vector features include MFCC, perturbation, spectral/cepstral and complexity features, while the spectra-based representations are MS multidimensional matrices. It is also multimodal as it allows to consider acoustic material coming from different

sources such as the phonation of different vowels. Indeed, one common approach that is followed in the automatic assessment of voice quality consists on using the single vowel /a/ for analysis. However, there are evidences indicating the usefulness of including vowels /i/ and /u/ for the assessment of certain traits such as R [22] For this reason, the present paper includes three different vowels (/a/, /i/ and /u/) for the purposes of performance improvement. The proposed architecture is also multi-output, since the prediction of the G, R and B descriptors is carried out simultaneously (with one output layer per feature of the GRB scale). This procedure permits to exploit the correlation that exists between all the traits in the decision making process. This paper, also addresses the prediction problem differently by considering a classification, regression or ordinal regression scenario, and evaluating different configurations for the output layers. Classification and regression are addressed using the well known categorical cross-entropy and *Mean Square Error* (MSE) loss functions respectively. For the ordinal regression case, two surrogate ordinal loss functions are evaluated. The approaches based on classification and ordinal regression incorporate strategies to compensate the imbalance problem in the database.

The paper is organized as follows: section 2 describes the methods followed to process the speech signals and the proposed DL architecture; section 3 presents the database used, the experimental setup and the obtained results; finally, the discussion and conclusions of the work are presented in section 4.

II. Methods

The proposed approach for voice quality assessment is composed of two main stages: characterization and decision making.

A. Characterization

Before the actual characterization, the speech signal is framed and windowed following a short-time analysis approach. The reliability of the short-time analysis for the automatic detection of pathological voices has been widely demonstrated before with successful results [23], [24]. After this procedure, two different approaches are followed to extract features and perform the actual characterization, one based on feature engineering and another on representation learning.

1) Feature engineering: The feature engineering methodology is based on a careful selection of curated characteristics that are employed for the training of the decision making machines. In this paper, the feature engineering approach is based on the extraction of a set of well know features often used in voice pathology detection or assessment tasks. This features are employed to train a *Deep Neural Network* (DNN) that carries out the automatic decisions. The features that are employed have been grouped into different sets, as in [24], [16], according to the signal processing techniques that are used or to the voice properties that these features are intended to measure: perturbation, spectral/cepstral and complexity. These features are computed following a short-time analysis methodology, having to set the length and type of the window in accordance to the type of characteristic that are extracted. Following guidelines of other works in literature, Hamming windows of 40 ms are employed for the perturbation and spectral/cepstral features to ensure that each frame contains at least three pitch periods, whereas windows of 55 ms length are used with the complexity features as suggested in [25]. The different sets of features -all descriptors of vocal condition- that are considered in this work are presented next.

a) Perturbation features: measure the presence of additive noise resulting from an incomplete glottal closure of the vocal folds, and the presence of modulation noise which is the result of irregularities in the movements of the vocal folds. These include *Normalised Noise Entropy* (NNE) [26], *Cepstral Harmonics-to-Noise Ratio* (CHNR) [27] and *Glottal-to-Noise Excitation Ratio* (G E) [28]. NNE and CHNR rely on the calculation of the energy of the noise, which is compared to the total energy of voice; whereas GNE is based on quantifying the loss of correlation between Hilbert envelopes of different frequency bands.

b) Spectral and cepstral features: measure the harmonic components of the voice. These include MFCC (with no derivatives), CPPS and *Low-to-High Frequency Spectral Energy Ratio* (LHr). MFCC are very well known in the field and can be considered as the gold standard characterization approach in speech technologies. CPPS is a normalized measure of the cepstral peak amplitude, which compares the level of harmonic organization of the speech to the cepstral background noise resulting from aspiration [29]. It is also considered one of the strongest correlates of breathiness [30]. Finally, LHr -a feature that often accompanies CPP- is the ratio between the average spectral energy below 4 kHz to the average energy above 4 kHz.

c) Complexity features: characterize the dynamics of the system and its structure. Several sets of complexity features are extracted. These include classical dynamic invariants such as the *Correlation Dimension* (D2), the *Largest Lyapunov Exponent* (LLE), and the *Recurrence Period Density Entropy* (RPDE) [31]; features which measure long-range correlations, such as *Hurst Exponent* (He) and *Detrended Fluctuation Analysis* (DFA) [31]; regularity estimators such as *Approximate Entropy* (ApEn) [32], *Sample Entropy* (SampEn) [33], *Modified Sample Entropy* (mSampEn) [34], *Gaussian Kernel Sample Entropy* (GSampEn) [10] and *Fuzzy Entropy* (FuzzyEn) [35]; and other entropy/complexity estimators such as the *Permutation Entropy* (PE) [36], the *Lempel-Ziv Complexity* (LZC) and the Shannon (s) and Rényi (r) estimators of the *Markov Chain Entropy* (H_{MC}), *Conditional Hidden Markov Process Entropy* (H_{HMP}) and *Recurrence State Entropy* (H_{RSE}) [25], [37]. Similarly to the ApEn and mSampEn estimators, which use the correlation sum for two different embedding dimensions, some modifications of the measures H_{MC}, H_{HMP} and H_{RSE}, which consist of averaging the entropy estimations over two different embedding dimensions are also considered. These measures are called *Averaged Markov Chain Entropy* (A_{MC}), *Averaged Conditional Hidden Markov Process Entropy* (A_{HMP}) and *Averaged Recurrence State Entropy* (A_{RSE}).

2) Representation learning: In a representation learning approach (or feature learning) a multilayer system is feed with the raw signal or its transformation, in the hope of finding representations that are suitable for decision making purposes. The idea is that in this system the higher layers of representation amplify aspects of the input that are relevant for discrimination while suppressing irrelevant variations [38]. This process is automatic, in the sense that the system itself is in charge of finding the most pertinent characteristics for classification.

For the purposes of this paper, a representation learning approach based on MS is employed to characterize modulation and acoustic frequencies of input voices [39], following a short-time basis using frames of 180 ms as proposed in [5], [40]. The MS have been successfully used in different works related with the characterization of pathological voices, but because of the large amount of data they contain, it is always necessary to extract some hand tuned statistics [5], [40] or to use feature selection techniques [41]. In the representation learning approach considered in this paper, *Convolutional Neural Network* (CNN) are used to automatically extract information from MS in the context of voice quality assessment. This is itself a challenging task for the CNN as the energy patterns in MS extracted from a sustained phonation are not as clear as the patterns of words found in spectrograms for other speech processing applications.

NO ENTIENDO PATTERN OF WORDS

B. Decision making

Given a certain input signal, the goal of the decision making stage is to convey a final decision about the label to which the signal belongs. In other words, the decision making stage is in charge of learning a mapping from the input signal (characterized either through a feature engineering or a representation learning approach) to an output label. This decision can be addressed in, at least, three different ways according to hypothesis followed in relation to the labels. In this manner, if the labels are categorical or nominal, the learning task is known as *classification*; if these are continuous the learning task is known as *regression*; if on contrary these are ordinal and discrete the learning task receives the name of *ordinal classification*. These three decision making tasks can be addressed using DL architectures that employ different types of loss functions. The loss functions that are used to deal with each one of the aforementioned decision making approaches are presented in the following sections.

In a similar fashion, three different DL architectures for the automatic assessment of voices are presented. These include a DNN architecture based on feature engineering, a CNN architecture based on representation learning and a CNN-DNN architecture that combines both of them. These are also described in the next sections.

1) DNN architecture: It is designed to process the set of features listed in the section II-A1 in the feature engineering approach. This architecture is multimodal, as it receives the individual information of the three vowels simultaneously, and multi-output because it predicts the levels of G, R and B simultaneously.

The proposed DNN is composed of three dense layers connected to three inputs, one for each of the vowels. The output of those layers is concatenated and processed through two additional dense layers. Finally three output layers provide the automatic voice quality assessment of the subject. In order to reduce the computational complexity of the model, the input layer receives, for a certain subject, a single feature vector that corresponds to the average and standard deviation of all the feature vectors calculated in the short-time analysis of speech. Fig. 1 illustrates the structure of this network.

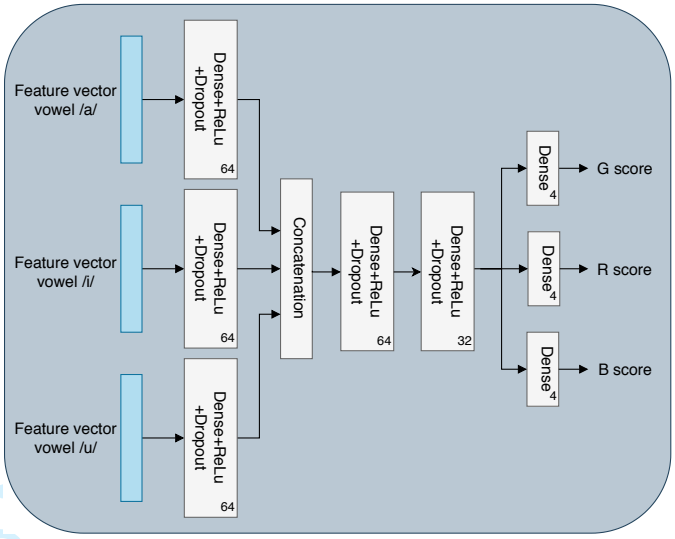


Fig. 1: DNN architecture based on the feature engineering approach

2) CNN architecture: It is designed to process the MS extracted from the three vowels, in a representation learning approach. The proposed architecture uses two parallel pipelines of convolutional layers, emulating the idea followed recently in different speech processing tasks [42], on which 1-dimensional convolutions are employed to process spectrograms. In this manner, the first pipeline performs convolutions in the acoustic axes whilst the other one performs convolutions in the modulation axis. Fig. 2 depicts a diagram of the convolutional module used in this CNN approach.

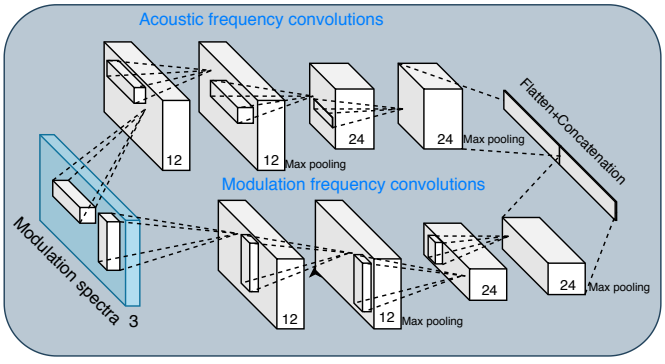


Fig. 2: Convolutional module used by the CNN architecture

For this particular architecture, the input layer is a 3 channel modulation spectra which include the maximum, mean and

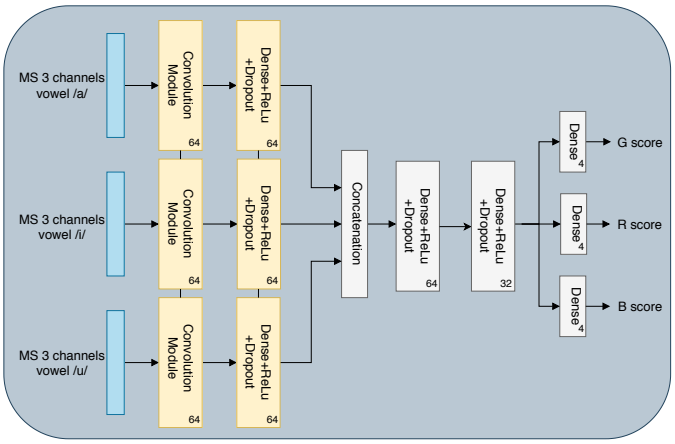


Fig. 3: CNN architecture based on the representation learning approach

standard deviation of the spectra obtained in a frame by frame basis. Every convolutional block is composed of the convolution layer itself, a batch normalization component and a ReLu activation function. The filter sizes are of [1,8] and [8,1] for acoustic and modulation frequency convolutions respectively. The complete CNN architecture is depicted in Fig. 3.

It is important to highlight that due to the size of the database, it is not possible to train individual convolutional modules for every vowel, and that the convolutional modules and the first dense layer of each input share their weights. That, in combination to a data augmentation strategy using translation and scaling operations, provide a more stable training of the network and reduce the chances of over-fitting.

3) CNN-DNN architecture: It is a combination of the architectures depicted in Figures 1 and 3 working in conjunction. This architecture is multimodal as it combines information from different vowels and also processes data from heterogeneous sources. In the CNN-DNN architecture, the concatenation layer combines all the information sources simultaneously, in a similar fashion than the concatenation layers of the previous architectures. After this information fusion stage, the CNN-DNN architecture follows a structure similar to the former cases: two dense layers and three output layers are added to the model.

4) Loss functions: The output layers of the proposed architectures vary whether the problem is assumed to be a classification, an ordinal regression or a regression; and therefore the corresponding loss functions must be changed accordingly. In a purely *classification* task, the activation function of the output layers corresponds to a softmax function. Therefore the most suitable loss function should be a standard categorical cross-entropy. However, due to the dataset imbalance (see section III-B), the loss function is replaced by a *Weighted Categorical Cross-Entropy* (WCC) as given by:

$$\mathcal{L}_{WCC} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\mathbf{1}_{y_i \in C_j}\omega_j \log p_{model}[\hat{y}_i \in C_j] \quad (1)$$

The term $\mathbf{1}_{y_i \in C_j}$ is the indicator function of the i-th observation belonging to the j-th category. The $p_{model}[\hat{y}_i \in C_j]$ is the predicted probability for the i-th observation to belong to the j-th class. When there are more than two classes, the neural network outputs a vector of C probabilities, where each value in the vector refers to the probability that the network input is classified as belonging to the respective category. ω_j is the weight associated to the error when the actual class is j. In this work the weights ω_j, $j \in [0, 3]$ are adjusted to balance the importance of all the classes during training.

TABLE I: Output codification for the first surrogate ordinal regression loss function OC_1

label	Network's output
0	1,0,0,0
1	1,1,0,0
2	1,1,1,0
3	1,1,1,1

Bearing in mind that a voice quality assessment based on the GRB scale is an *ordinal classification* problem, two surrogates ordinal loss functions are also investigated. In this respect, the first ordinal loss function, denoted *Ordinal Classification One* (OC_1), codifies the network's target cumulatively as in Table I. The activation functions for the output layer are sigmoid functions and the loss function is a weighted binary cross-entropy (WBC), given by:

$$\mathcal{L}_{OC_1} = -\frac{1}{N}\sum_{i=1}^{N}\omega_{C_i}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij}) \quad (2)$$

where y_{ij} is the j-th ouput of the network for the sample i and ω_{C_i} is the weight of the class to which the sample i belongs.

The second surrogate ordinal regression function uses a regular softmax activation function in conjunction with a double weighted categorical cross-entropy loss function. This function is denoted as *Ordinal Classification two* (OC_2) and is as follows:

$$\mathcal{L}_{OC_2} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\mathbf{1}_{y_i \in C_j}\upsilon_{ij}\omega_j \log p_{model}[\hat{y}_i \in C_j] \quad (3)$$

where $\upsilon_{ij} = 1+|C_i - j|$ and C_i is the actual class of the sample i. The first weight is the same ω_j incorporated in Eq. (1) to compensate for the imbalance problem. The second weight penalizes the errors of the model in accordance to how far the predicted class is from the ground truth.

Lastly, when the prediction problem is assumed as a pure *regression*, the loss function is a MSE and the final labels are obtained by rounding the output to its nearest integer. In this case, the network has only one neuron in the output layers instead of four as depicted in Fig. 1, 2 and 3. In fact, the implementation could use a single output layer with 3 neurons instead of three layers of 1 output.

III. Experiments and Results

A. Setup

Experiments are carried out using a 5 folds cross-validation strategy. To evaluate the performance of the proposed

approach, two metrics are employed: *Balanced Accuracy* (BACC) and *Average Mean Absolute Error* (AMAE).

BACC is a classification measure normalized with respect to the number of samples per class, which is defined as:

$$BACC = \frac{1}{C}\sum_{j}\frac{1}{N_j}\sum_{\forall i \in C_j}[\![\hat{y}_i = y_i]\!] \quad (4)$$

where $[\![\cdot]\!]$ is and indicator function giving 1 if the condition is satisfied and 0 otherwise. N_j is the number of samples in the j-th class.

AMAE is a balanced measure computing the average deviation between the predicted and the true class, and which is defined as follows [43]:

$$\text{AMAE} = \frac{1}{C}\sum_{j}\frac{1}{N_j}\sum_{\forall i \in C_j}|\mathcal{O}[y_i] - \mathcal{O}[\hat{y}_i]| \quad (5)$$

where y_i are the actual and $\hat{y}_i$ the predicted labels; and $\mathcal{O}[\cdot]$ is an operator indicating the position of the label in the ordinal rank (i.e., if a certain label y_i can take up values $0, 1, 2$ and the label is 2, its position is 3).

B. Database

The Saarbrücken Voice Database [17] is used for experimentation. It contains registers of more than 2000 German speakers phonating different vowels and uttering a short sentence. Registers were recorded at a sampling frequency of 50 kHz and 16 bits of resolution. For this paper purposes, the same subset of 568 normophonic and 970 pathological subjects used in [44] is employed. To include similar material to that used by speech therapists during the evaluation of a patient according to the GRB scale [22], the sustained phonations of the vowels /a/, /i/ and /u/ were used in the experiments. Figure 4 shows the distribution of the samples for each one of the traits and their respective labels, including the recordings of the vowels /a/, /i/ and /u/.

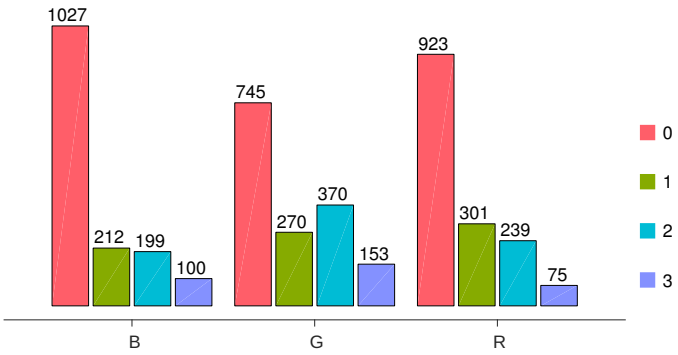


Fig. 4: Histogram summarizing the distribution of the labels associated to G, R and B

C. Results

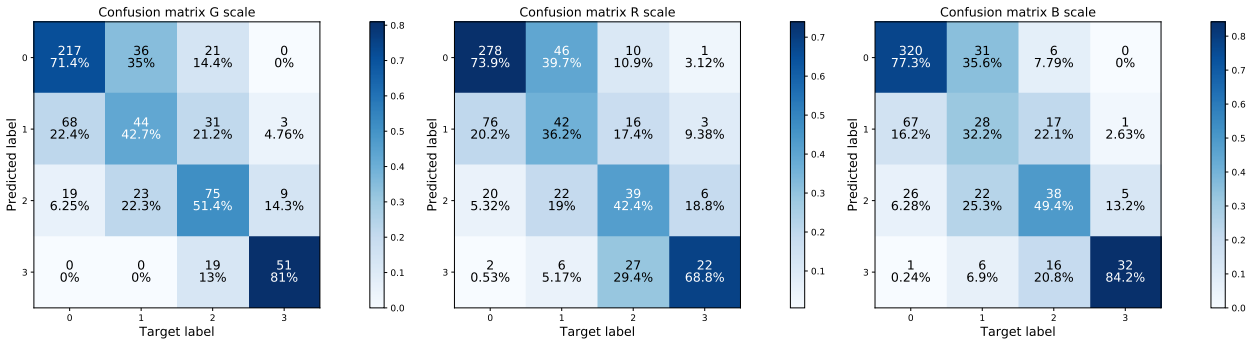
Table II presents the performance metrics of all the architectures and the loss functions described in section II-B. As observed from the table, the DNN_{OC_2} model presents the best outcomes, slightly outperforming those of the DNN_{WCC} architecture. Comparing the results to the most recent approach in the state of art, which also employs the same subset of the Saarbrücken dataset [16], it can be observed that the DNN architecture improves the performance of all the analyzed traits. In absolute terms, AMAE is reduced in 0.03 points for G, 0.09 points for R and 0.09 points for B, which means a relative improvement of 6.25%, 14.1% and 18.1% respectively. This performance improvement could be related to the fact that the proposed DL architectures carried out a simultaneous prediction of all the traits of the perceptual scale, hence, exploiting the well-known correlation that exists among them. It is worth noting that considering the simultaneous information provided by all the traits by means of DL architectures, constitutes a major novelty of this paper. Most of works in literature disregard the possible correlations among traits, and certainly none of them exploits them to improve performance.

TABLE II: Results for all the proposed architectures. WCC is the loss function when following a classification approach, OC_1 and OC_1 an ordinal regression, and MSE a regression.

Architecture	Loss function	$BACC$			$AMAE$		
		G	R	B	G	R	B
DNN	WCC	0.61	0.55	0.60	0.46	0.53	0.46
	OC_1	0.56	0.39	0.48	0.53	0.77	0.63
	OC_2	**0.62**	**0.56**	**0.60**	**0.45**	**0.53**	**0.45**
	MSE	0.57	0.40	0.44	0.49	0.73	0.61
CNN	WCC	0.59	0.54	0.56	0.49	0.57	0.54
	OC_1	0.54	0.43	0.52	0.55	0.68	0.61
	OC_2	0.57	0.52	0.56	0.50	0.59	0.54
	MSE	0.51	0.43	0.44	0.54	0.70	0.63
CNN-DNN	WCC	0.57	0.52	0.54	0.52	0.62	0.58
	OC_1	0.44	0.41	0.41	0.66	0.77	0.71
	OC_2	0.56	0.54	0.56	0.52	0.59	0.55
	MSE	0.44	0.43	0.46	0.60	0.75	0.65

Figure 5 depicts the confusion matrices of the best performing architecture, DNN_{OC_2}, for all the traits of the GRB scale. It is possible to observe that most of the errors committed by the system occur in the labels adjacent to the diagonal, affecting mainly the labels 1 and 2. The errors committed in non-adjacent labels correspond to less than 1% of all the cases and for all traits, except for label 3 of the R trait that reached 3.1%. This can be confirmed by the use of the *weighted-AMAE* as in [16]. This measure permits to measure the influence of the errors around the diagonal of the confusion matrices, by varying a parameter in the range 0 -where an error around the diagonal is regarded as accurate- to 1 -the usual case-. A graphic depicting the weighted-AMAE for the confusion matrices of Fig. 5 is presented in Fig. 6. Results indicate that by allowing errors in the adjacent class to be treated as accurate, the AMAE decreases to a mere 0.18 for R, 0.12 for B and 0.13 for G. This is a considerable performance improvement that indicates the good behaviour of the system as most of the errors are located in the vicinities of the actual class.

Despite the imbalance of the corpus, it is worth to note that the predictions of the label 3 are better than those of the label 0 for the G, and B traits, and just slightly worse for



(a) Confusion matrix of for feature G. (b) Confusion matrix of for feature R. (c) Confusion matrix of for feature B.

Fig. 5: Confusion matrices of the best model for every parameter of the GRB scale. Each cell contains the number of incidents on top, and the label percentage at the bottom.

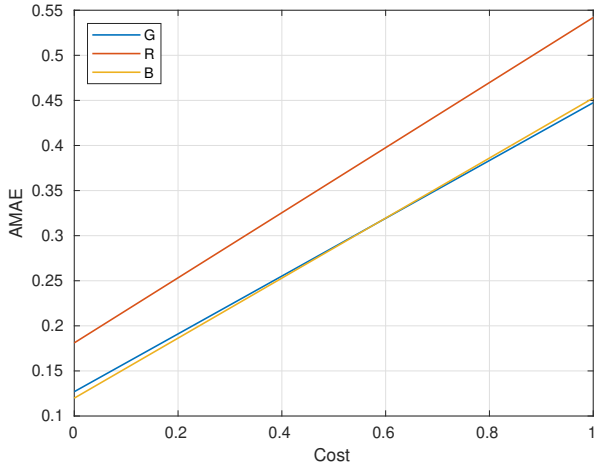


Fig. 6: Weighted AMAE for G, R and B

the trait R. These results reflect the success of using weights-based loss functions to compensate for the imbalance in the dataset, an scenario that is particularly harmful for neural network based predictors. When comparing the performance of the systems in relation to the trait that is analyzed, it can be observed from Table II, that in general all the architectures performed better when using the trait G, rather than B or R. Despite that, the DNN_{OC_2} model achieved similar prediction performance for G and B features. The better performance of the G trait in comparison to B or R is a common result reported in literature. One possible explanation of that might be related to the assessment process itself, as it could be easier to characterize vocal quality globally using the G parameter, instead of disregarding individual components of vocal quality such as B and R.

With regards to the use of a feature engineering or a representation learning approach, it was found that the best results were obtained with the feature engineering scheme and the DNN model. Notwithstanding, the performance provided by the best CNN architecture (CNN_{WCC}) in the representation learning approach, between 0.04 and 0.09 AMAE points worse than the results of the best DNN model. If this performance is contrasted with the results presented in the literature in [16], a 11% of improvement is achieved for the trait R, and a 10% for trait B. This is an interesting result considering that the CNN architecture was compelled to automatically extract relevant features from the MS spectra. It is also impressive if we consider that the results published in [16], also included a set of features extracted from MS which were carefully selected for the assessment task. The combination of the feature engineering and representation learning, using the CNN-DNN model, did not yield to a performance improvement in comparison to using the architectures individually. Indeed, the CNN-DNN architecture is considerably much bigger than the CNN and the DNN architectures alone, and the increment in the number of parameters is of several orders of magnitude. As it is well known, an increased number of parameters provides the model of more flexibility but also makes it more susceptible to overfitting, especially in scenarios where the training data is scarce, just as in this particular case.

With respect to the decision making scenarios, it was found the performance was quite similar when the problem is treated as a classification (using WCC) or an ordinal classification (with OC_2). The worse results were obtained when the decision making task was a pure regression, utilizing a loss function based on MSE. In this particular case, the use of the round function that was employed to produce a final label can be considered a naive approach. Literature has already reported results related with ordinal regression problems where poor performance is achieved with such rounding function. There exist alternative methodologies in order to learn the thresholds that define the limits of every label [45], however these cannot be straightforwardly extrapolated to neural networks.

IV. Discussions and Conclusion

This paper has been devoted to the automatic assessment of voice pathologies based on DL. Experiments were carried out in a subset of the Saarbrücken voice dataset, testing out three DL architectures suitable for the exploration of three scenarios based on feature engineering, representation learning and their combination. In a similar way, three decision making approaches were tested out: regression, classification and

ordinal regression. Each approach was defined by considering different loss functions.

Previous studies have shown that the automatic assessment of voice quality requires multidimensional approaches capable of characterizing the different phenomena involved in the voice production process. This paper not only follows that line of thought by proposing multidimensional architectures for the automatic voice quality assessment, but it also takes a step forward and incorporates multimodalities for the sake of performance improvement. Current advancements in DL allow the definition of complex systems that can be trained as a whole. This work explored the performance of different neural network architectures, which fused the information extracted from sustained phonations of the vowels /a/, /i/, and /u/, and provided a prediction for every trait of the GRB scale simultaneously. One of the proposed DNN architectures achieved, in terms of AMAE, a relative improvement of 6.25% for G, 14.1% for R and 18.1% for B, in comparison with recently published results using the same database. This results demonstrate the usefulness of treating the GRB assessment as a multi-output problem, taking advantages of the correlation among the traits, establishing a new line of work that could be explored further in the future. This approach can also be easily extended to other traits such as A and S of the GRBAS scale, or directly to other perceptual scales.

This paper also evaluated the capabilities of a CNN to extract relevant information in the context of the perceptual assessment of voice, following the strategy known as representation learning. In this case, the network was trained using a three channel MS, which provides information about perturbations in amplitude and frequency modulation of the voice signal, and that has demonstrated to be valuable for the analysis and characterization of pathological voices. The outcomes provide evidence about the capability of the CNN to learn and extract automatically patterns in the MS spectra, that were useful for the automatic assessment of the voice signals. Even though the CNN in the representation learning approach did not perform better than the DNN models in the feature engineering approach, the results of the CNN models were better than others reported in literature, specially when R and B where considered. These outcomes indicate the usefulness of incorporating representation learning techniques in the development of automatic pathological speech classification and assessment systems. This requires more investigations with larger and more heterogeneous datasets. It is important to highlight, though, that representation learning using speech spectrograms is currently the state of the art in speech processing task, such as keyword spotting or emotion recognition based on speech. However, the analysis of pathological speech constitutes a more challenging approach as much of the information used for pathological voice detection and assessment is extracted from sustained phonations of vowels, or diadochokinetic exercises. These contain spectral information that is poorer in terms of the multiplicity of energy components in comparison to word utterances or sentences, and therefore the patterns that must be discovered by the network are less evident. Even though the weight sharing strategy among the convolutional modules in the CNN architecture helps to stabilize the network's training, at the end it might have affected its performance. This is because, the CNN should look for different patterns for each one of the vowels, but the sharing strategy imposes strong restrictions to the network during this data discovery process.

When feature engineering and representation learning were used in conjunction through a CNN-DNN, there was not a performance improvement in comparison to treating CNN and DNN separately. We hypothesize that this is the result of an increment in the complexity and number of parameters of the resulting model, with a consequent increase in susceptibility of the models to overfitting. To deal with this scenario, a larger databases is needed, along with more aggressive data augmentation strategies. As an alternative, Bayesian approaches using variational inference layers could be evaluated in order to compensate for the scarcity of data. The lack of available datasets for the study of pathological speech has been an open and known problem since early times of the field. In particular for the automatic assessment of pathologies, it is also required that the voice signals are properly labeled, maintaining consistency as it has been demonstrated to be a crucial element for the proper emulation of the perceptual capabilities of human evaluators [16].

Regarding the decision making strategies that were followed, the results showed similar performance between the loss function $\mathcal{L}_{OC_2}$ (based on a double WCC) and a more conventional $\mathcal{L}_{WCC}$, for the ordinal regression and classification approaches respectively. The use of a MSE loss function following a regression approach yields the worse results. Similar behaviors have been reported in the literature and new ideas regarding the learning of thresholds that define every label have also been already proposed. However, these cannot be straightforwardly extrapolated to loss functions of DL models, which could be matter of further research.

Acknowledgment

This work was supported by the Universidad de Antioquia, Medellín, Colombia, and the Ministry of Economy and Competitiveness of Spain under grant DPI2017-83405-R1.

References

[1] M. Karnell, S. Melton, J. Childes, T. Coleman, S. Dailey, and H. Hoffman, "Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders," *Journal of Voice*, vol. 21, no. 5, pp. 576–590, 2007.

[2] M. Hirano, *Psycho-Acoustic Evaluation of Voice*. New York, USA: Springer-Verlag, 1981.

[3] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, and T. Haderlein, "Vowel- and text-based cepstral analysis of chronic hoarseness," *Journal of Voice*, vol. 26, no. 4, pp. 416–424, 2012.

[4] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[5] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: a new method to assess voice pathologies according to the GRBAS scale," *BioMed Research International*, vol. 2015, 2015.

[6] C. Sellars, A. Stanton, A. McConnachie, C. P. Dunnet, L. Chapman, C. Bucknall, and K. Mackenzie, "Reliability of perceptions of voice quality: evidence from a problem asthma clinic population," *The Journal of Laryngology & Otology*, vol. 123, no. 7, pp. 755–763, 2009.

[7] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[8] R. Ritchings, M. McGillion, and C. Moore, "Pathological voice quality assessment using artificial neaural networks," *Medical Engineering & Physics*, vol. 24, no. 8, pp. 561–564, 2002.
[9] J. I. Godino-Llorente, T. Ritchings, and C. Berry, "The effects of inter and intra speaker variability on pathological voice quality assessment," in *Third International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2003.
[10] L. Gu, J. Harris, R. Shrivastav, and C. Sapienza, "Disordered speech assessment using automatic methods based on quantitative measures," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 9, pp. 1400–1409, 2005.
[11] J. Lee and M. Hahn, "Automatic assessment of pathological voice quality using higher-order statistics in the LPC residual domain," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. ID 748207, p. 8, 2009.
[12] N. Sáenz-Lechón, J. I. Godino-Llorente, V. J. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," in *Proceedings of 28th IEEE EMBS Annual International Conference*, New York, NY, USA, Sep 2006, pp. 2478–2481.
[13] Massachusetts Eye and Ear Infirmary, "Voice disorders database, version.1.03 [cd-rom]," Lincoln Park, NJ: Kay Elemetrics Corp, 1994.
[14] Z. Wang, P. Yu, N. Yan, L. Wang, and M. L. Ng, "Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the grbas scale," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 241–251, 2016.
[15] S. Xie, N. Yan, P. Yu, M. L. Ng, L. Wang, Z. Ji *et al.*, "Deep neural networks for voice quality assessment based on the grbas scale." in *INTERSPEECH*, 2016, pp. 2656–2660.
[16] J. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, C.-D. G., and J. Godino-Llorente, "Emulating the perceptual capabilities of a human evaluator to map the grb scale for the assessment of voice disorders," *Engineering applications of artificial intelligence*, vol. 82, pp. 236 – 251, 2019.
[17] Barry, W.C. and Pützer, M., "saarbrücken voice database," Institute of Phonetics, Univ. of Saarland. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/.
[18] C. Fredouille, G. Pouchoulin, A. Ghio, J. Revis, J.-F. Bonastre, and A. Giovanni, "Back-and-forth methodology for objective voice quality assessment: from/to expert knowledge to/from automatic classification of dysphonia," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 8, 2009.
[19] M. D. Skowronski, L. M. Kopf, R. Shrivastav, and D. A. Eddins, "Acoustic models of co-varying vocal roughness and breathiness," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1832–1832, 2015.
[20] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.
[21] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice. In Press*, 2018.
[22] S. Anand, M. D. Skowronski, R. Shrivastav, and D. A. Eddins, "Perceptual and quantitative assessment of dysphonia across vowel categories," *Journal of Voice. In Press*, 2018.
[23] J. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality asessment system based on Gaussian Mixture Models and short-term cepstral parameters," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.
[24] J. Gómez-García, L. Moro-Velázquez, and J. Godino-Llorente, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.
[25] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.
[26] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *Journal of the Acoustical Society of America*, vol. 80, pp. 1329–1334, 1986.
[27] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals." *Journal of speech,language, and hearing research*, vol. 36, no. 2, pp. 254–266, 1993.
[28] D. Michaelis, T. Gramss, and H. Strube, "Glottal-to-noise excitation ratio - a new measure for describing pathological voices," *Acustica/Acta acustica*, vol. 83, pp. 700–706, 1997.
[29] J. Hillenbrand and R. A. Houde, "Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech," *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, p. 311, 1996.
[30] S. N. Awan, N. Roy, and C. Dromey, "Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model." *Clinical linguistics & phonetics*, vol. 23, no. 11, pp. 825–841, 2009.
[31] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical Engineering Online*, vol. 6, no. 23, 2007.
[32] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88, pp. 2297–2301, 1991.
[33] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal Physiol Heart Circ Physiol*, vol. 278, no. 6, pp. H2039–H2049, 2000.
[34] H.-B. Xie, W.-X. He, and H. Liu, "Measuring time series regularity using nonlinear similarity-based sample entropy," *Physics Letters A*, vol. 372, no. 48, pp. 7140–7146, 2008.
[35] W. Chen, C. Peng, X. Zhu, B. Wan, and D. Wei, "SVM-based identification of pathological voices," in *Proceedings of 29th Annual International Conference of the IEEE EMBS*, Lyon, France, 2007, pp. 3786 – 3789.
[36] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review," *Entropy*, vol. 14, no. 12, pp. 1553–1577, 2012.
[37] J. D. Arias-Londoño and J. I. Godino-Llorente, "Entropies from Markov Models as Complexity Measures of Embedded Attractors," *Entropy*, vol. 17, no. 6, pp. 3595–3620, 2015.
[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.
[39] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 7, p. 310290, 2003.
[40] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics," *Frontiers in Bioengineering and Biotechnology*, vol. 4, no. 1, 2016.
[41] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 1938–1948, 2011.
[42] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.
[43] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation Measures for Ordinal Regression," in *2009 Ninth International Conference on Intelligent Systems Design and Applications*. IEEE, 2009, pp. 283–287.
[44] J. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, and J. Godino-Llorente, "On the design of a voice pathology assessment system based on the GRB scale," in *10th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2017.
[45] R. Fathony, M. A. Bashiri, and B. Ziebart, "Adversarial surrogate losses for ordinal regression," in *Advances in Neural Information Processing Systems*, 2017, pp. 563–573.