An emotional modulation model as signature for the identification of children developmental disorders

Arianna Mencattini, Francesco Mosciano, Maria Colomba Comes, Tania Di Gregorio, Grazia Raguso, Elena Daprati, Fabien Ringeval, Björn Schuller, Corrado Di Natale, Eugenio Martinelli, et al.

SCIENTIFIC REPORTS

OPEN

An emotional modulation model as signature for the identification of children developmental disorders

Received: 21 February 2018
Accepted: 6 August 2018
Published online: 27 September 2018

Arianna Mencattini[1], Francesco Mosciano[1], Maria Colomba Comes[1], Tania Di Gregorio[2], Grazia Raguso[2], Elena Daprati[3], Fabien Ringeval[4], Bjorn Schuller[5,6], Corrado Di Natale[1] & Eugenio Martinelli[1]

In recent years, applications like Apple's Siri or Microsoft's Cortana have created the illusion that one can actually "chat" with a machine. However, a perfectly natural human-machine interaction is far from real as none of these tools can empathize. This issue has raised an increasing interest in speech emotion recognition systems, as the possibility to detect the emotional state of the speaker. This possibility seems relevant to a broad number of domains, ranging from man-machine interfaces to those of diagnostics. With this in mind, in the present work, we explored the possibility of applying a precision approach to the development of a statistical learning algorithm aimed at classifying samples of speech produced by children with developmental disorders(DD) and typically developing(TD) children. Under the assumption that acoustic features of vocal production could not be efficiently used as a direct marker of DD, we propose to apply the Emotional Modulation function(EMF) concept, rather than running analyses on acoustic features per se to identify the different classes. The novel paradigm was applied to the French Child Pathological & Emotional Speech Database obtaining a final accuracy of 0.79, with maximum performance reached in recognizing language impairment (0.92) and autism disorder (0.82).

Star Trek fans will remember EMH, the Emergency Medical Holographic program that had the appearance of a reliable, middle aged family doctor. Even if we are miles away from developing an artificial healthcare practitioner, in recent years, significant advancements have been made in computer-aided diagnosis, digital technology and artificial-intelligence support to clinical practice[1–8].

Promising opportunities come from the domain of machine learning and, specifically, from supervised[9] and unsupervised[10] learning machines. Such approaches typically require extraction of a set of features that characterize the items at study (e.g., colour, frequency, wavelength …) and involve a classification method capable of distinguishing and assigning items to separate classes. Through machine learning, algorithms can automatically extract features from the available data and implement classification by using a reference data set (*training set*) for which labelling is known. Once trained, the classifier works on its own, allowing for huge amounts of data to be rapidly labelled, an approach that has proved successful in a number of domains (e.g., diagnostic imaging, remote sensing imaging …).

In recent years, the interest for Big Data analysis has extended to the area of psychiatry research[5], providing novel ways to classify brain disorders from abnormalities in neuroimaging and/or genomic data[11–14] and introducing new methods to predict the outcome of therapeutic approaches[15]. The obvious advantage of integrating clinical practice with information drawn from statistical learning rests on the opportunity to speed up the entire diagnostic procedure, which – in turn – can reduce frustration, adverse outcomes and prolonged disability in patients. These benefits are especially relevant to the paediatric population since early diagnoses of

[1]Department of Electronic Engineering, University of Rome Tor Vergata, via del Politecnico 1, 00133, Roma, Italy. [2]Faculty of Science MM.FF.NN., University of Bari, Aldo Moro, University Campus Ernesto Quagliariello, Via Edoardo Orabona 4, 70126, Bari, Italy. [3]Department of Systems Medicine, CBMS, University of Rome Tor Vergata, via Montpellier 1, 00133, Roma, Italy. [4]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, 38401, St Martin d'Hères, France. [5]GLAM – Group on Language, Audio & Music, Imperial College London, SW7 2AZ, London, UK. [6]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159, Augsburg, Germany. Correspondence and requests for materials should be addressed to E.M. (email: martinelli@ing.uniroma2.it)

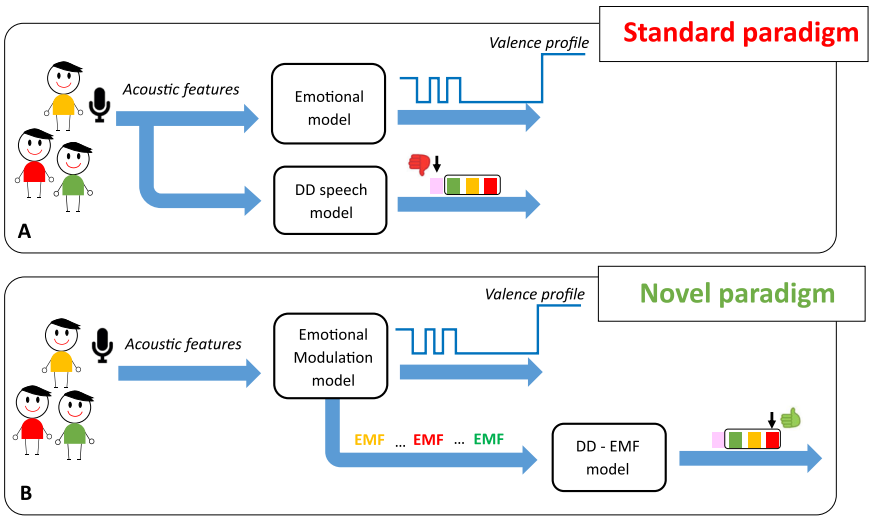

Figure 1. Comparison of standard and novel paradigm. Panel A. Acoustic features extracted from the recorded speech are used to recognize the expressed emotion (Emotional model) (e.g., valence profile) and the pathology (DD speech model). Panel B. Acoustic features are used to construct the emotional modulation model. The Emotional Modulation Function (EMF) of different subjects is then used to train an DD – EMF model.

neurodevelopmental disorders, such as Autism Spectrum Disorders (ASD) or Attention Deficit Hyper-Activity Disorder (ADHD), can promote timely intervention, positively influencing the children's future lives.

When data from the clinical domain is concerned, a problem faced by most machine learning methods is the heterogeneity of presentation of most diseases. For example, in the case of ASD, highly heterogeneous patterns are described for genetic profiles[16], gender-specific effects[17] language phenotypes[18] and more, to the point that it has often been suggested that autism should not be considered as a single disorder but rather as '*the autisms*'[19] – hence the term *spectrum* in ASD. In terms of machine learning applications, heterogeneity can hinder efficiency of classifiers, making predictions less reliable. Whereas some recent approaches exploit generative adversarial networks to augment artificially the data space[20], other methods can be drawn from the fast-developing area of '*personalized medicine*', i.e. the growing knowledge that diagnostic and therapeutic strategies should take variability into account, thence applying highly individualized approaches to patients. This message – that has been largely received in the domain of oncology[21] – is becoming increasingly more relevant also to studies in other areas, including neuropsychiatry (see for instance[22–24]).

Stemming from this idea, we explored the possibility of applying an emotion-driven approach to the development of a personalized statistical learning algorithm aimed at classifying samples of speeches produced by typically developing children (TD) and by children with autism disorder (AD), specific language impairment (SLI), Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). The latter three conditions are characterized - to different degrees - by severe deficits in social interactions and communication skills, as well as by stereotyped behaviors[25]. In addition, when speech is concerned, all three conditions are known to induce a flat, monotone intonation and anomalies in the use of volume, pitch and stress[26–32]. Accordingly, for the purpose of this investigation, children with AD, SLI, and PDD-NOS will be considered as a single group, broadly labeled here as Developmental Disorders (DD)[32]. We used acoustic features to automatically differentiate children with DD from children with TD diagnosis using speech recordings[2,33]. A recent meta-analysis[34] has pointed out that acoustic features of vocal production can indeed be used as a marker of ASD, even if the over 30 papers reviewed failed to identify a single characterizing feature. Recently, a study has described a machine learning strategy that recognizes spontaneous emotional expressions in the voice and discriminates DD individuals from TD children based on speech features[35]. However, the proposed method did not account for the variability in emotional expression within individuals although this aspect can impact on characterization of the pathology. To counteract this limitation, we introduce a novel signal processing paradigm that exploits the individual emotional modulation occurring during speech in order to model atypical behaviors that are symptomatic of DD. The present paradigm thus departs from traditional approaches that directly learn acoustic models from the speech signal of DD children, while treating the corresponding valence profile in parallel. For a better outline of the novel paradigm, Fig. 1 compares the standard and the novel paradigm.

In the standard paradigm (Panel A), the acoustic features extracted from the recorded speech are used to construct the emotional model that estimates a valence profile of an individual[34,35]. The same features are then used to train the DD speech model. Pink, green, yellow and red boxes represent the different subcategories of children (TD, PDD-NOS, SLI, and AD respectively). The limits of this approach will be outlined in section 2.2. In the novel paradigm (Panel B), the acoustic features are used to construct the emotional model as before, if needed. However, a different concept, the Emotional Modulation function (EMF) represented by the regression coefficient of the emotional model, is used to train a model of the pathology. The EMF represents the way each individual modulates his/her emotional response to a given known stimulus. More specifically, EMF is the leading concept in our rationale that fosters new scenario in understanding atypical behaviors that are symptomatic of DD.

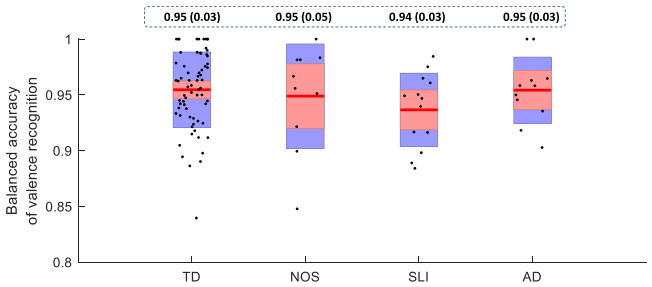

Figure 2. Average balanced accuracy of valence level recognition. Boxplots of the ACC representing the valence level recognition performance on three classes (chance score is 0.33). From left to right: TD, PDD-NOS, SLI, and AD subjects. Dots represent the different participants, central pink area represents the interquartile range [25th–75th] percentiles and the blue dashed area represent the [10th–90th] percentiles range. Values in the dashed board represent average (standard deviation) values.

Numerically, EMF quantifies the *weights* that each individual spontaneously gives to his/her voice alterations in order to encode non-verbal information in speech. The weights are the core of the new way of thinking since we believe their dynamic is relevant to the different DD manifestation and can be then used to learn an appropriate representation. Most importantly, this novel approach ensures the possibility to construct a personalized model of emotion first, and successively to use the associated EMF to predict the class to which the child belongs (e.g. AD, PDD-NOS, SLI, or TD).

In the present work, we will apply the novel paradigm to the French Child Pathological & Emotional Speech Database (CPESD)[35]. For compilation of the database, children aged 6–18 were involved in an unconstrained task: a story telling of a pictured book[36]. It was assumed that the children's production of prosodic cues during the telling of the story was correlated to the level of emotional valence elicited by each picture of the book, which was assessed in three categories by a psychologist (negative, neutral and positive). The dataset includes 102 individuals reported respectively as DD (N = 34) and TD (N = 68), with a ratio of two TD for one DD child of same age and sex. Acoustic features contained in each utterance produced by children during the story telling were then extracted for further model development.

Results

In order to evaluate the performance obtained by the proposed methodology, we ran three different tests. The first presents the personalized model of emotion constructed for each participant, independently from the corresponding diagnosis. The second provides evidence for the general failure of the standard paradigm (Panel A Fig. 1), as shown by the computation of the balanced accuracy (ACC) on the children's groups when acoustic descriptors are solely used to construct the recognition system; ACC is an evaluation metric that compensates uneven class distribution by computing the average recall per class. Finally, as a third test, we present the performance of classification when the novel paradigm (DD-EMF model) is applied in recognizing TD from DD subjects. In the dataset, children with a diagnosis of a disorder belonged to one of the three following categories: AD (N = 11), PDD-NOS (N = 13) or SLI (N = 10), which, for the sake of simplicity, will henceforth be all labelled comprehensively as DD. For the third test, the low number of participants in each subcategory does not justify development of a four-classes recognition problem. Therefore, we investigated the two-class problem of recognizing typical vs atypical children from their voice (i.e., TD vs DD).

Test 1. Performance of the valence recognition model. To demonstrate the appropriateness of the acoustic features in describing the emotional valence in an individual's speech sample, we constructed a three-class classifier based on Linear Discriminant Analysis (LDA). The three classes, labelled as −1, 0, and 1, represent negative, neutral, and positive valence, as codified in the dataset. Features are preliminarily selected using thresholding of the individual Pearson's correlation coefficient (ρ_c) with respect to the valence level assessed for a given subject. Only features with an absolute ρ_c value larger than 0.7 in the training dataset will be kept and used for further analysis. The accuracy of the three classes, computed using a leave-one-utterance-out cross-validation procedure, is estimated separately for each disorder and for the TD subjects and collected in the boxplots shown in Fig. 2. It can be observed that accuracy is around 0.95 in all categories of children indicating a very strong effectiveness of the selected features in modelling the emotional valence, independently from the presence of DD disorder. On the other hand, this is an indirect demonstration of the fact that acoustic features equally behave with respect to the presence of disorder.

Low performances of the standard paradigm in DD recognition. To highlight the importance of the novel paradigm for DD recognition, we first provide results obtained by using the standard approach (Fig. 1 Panel A). In particular, we consider two different model settings (henceforth described as FS1 and FS2). In the first (FS1), the binary classification model for discrimination of DD from TD individuals was trained directly on the acoustic features extracted from each utterance. In this scenario, each utterance is a row and each acoustic feature is a column of the data matrix. Features were selected by implementing a two-sided Student t-test ranking criterion with respect to the diagnosis collected in the training partition and the first 100 ranked features were retained to construct the classification model. In the second test (FS2), the binary classification model was trained

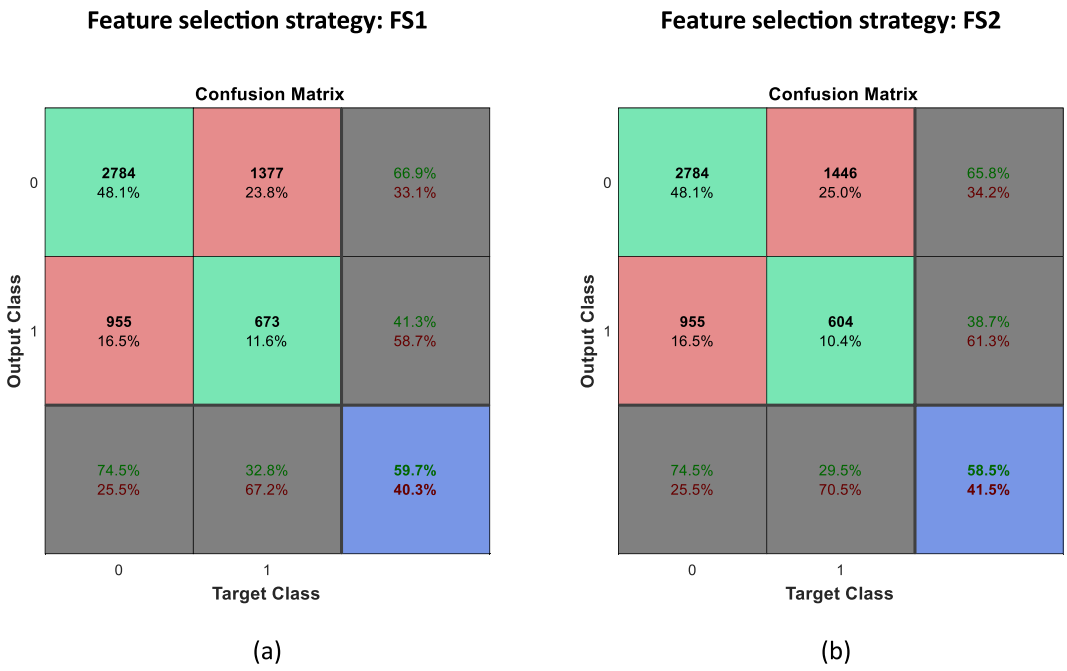

Figure 3. Confusion matrix in the standard paradigm. Confusion matrix obtained (**a**) using an emotional-driven feature selection strategy (FS1) and Support Vector Machine (SVM) classification, and (**b**) implementing a diagnosis-based feature selection approach (FS2) in the training set and Support Vector Machine (SVM) classification. A blue box represents the overall accuracy of the results (green boxes = percentage of success and red boxes = percentage of failure).

on the features selected according to the maximum correlation with the valence annotation of the corresponding subject in the training partition. The rationale behind these tests is to present a way to improve the standard paradigm regarding the direct recognition of DD through the child's vocal expression. A Support Vector Machine (SVM) with a linear kernel and default parameters setting (box-constraint parameter value set to one and feature weighted standardization) has been implemented in both tests to assign the final label of TD or DD to each utterance. Figure 3 shows the confusion matrices obtained in the two cases.

Results clearly show that in both situations, the average recognition rate is unacceptably low (59.7% and 58.5% respectively), and hence discredits the assumption of a direct relationship between acoustic features and the child's diagnosis.

Performance of the novel DD recognition paradigm. The novel approach stems from the assumption that the way the child modulates his/her emotional speech towards a given emotional stimulus could be characteristic for the subjects suffering from DD. One risk with this approach is the high heterogeneity in the children's capacity to verbally express their emotional attitude. To overcome this problem, we begin by computing average descriptors of all the utterances by the same subject. To do this, after having collected features that mostly correlate with the valence level (see Section 3.3), we combine the values of such descriptors over all the utterances assessed with the same valence level for one subject (all utterances assigned a negative valence, all with assigned neutral valence and all assessed with positive valence levels, respectively). This step is motivated by the underlying assumption that not all of a child's utterances could be considered as correlated with a DD diagnosis. Moreover, this strategy allows us to combine verbal productions from different individuals, regardless of the speech length. Biasing in valence subjectivity of each participant is reduced by the expert-based assessment of the valence level performed by psychologist. By combining the information extracted from all the utterances for the same individual, this novel method further compensates for the unavoidable inaccuracy of this task. In addition, this combination is performed by using distribution descriptors such as the mean value, skewness and kurtosis computed over the utterances of the same individual labelled with the same valence category (e.g., all the utterances of an individual labelled with positive valence). The last two parameters (skewness and kurtosis) allow us to add to the average value of the feature over different utterances, the dispersion of the feature over the utterances of the same valence level. Recalling again that only 19 acoustic feature were extracted at the first step, then 57 high-level descriptors resulted (19 average values, 19 skewness values, and 19 kurtosis values). These descriptors are fed into a Multilinear Regression Block (MLR) block (see Section 3.3) with expected output valence category −1, 0 and 1. The corresponding 57 MLR coefficients are estimated and used. Such coefficients represent the EMF through which each child reacts to the administered emotional stimulus, and that, in our rationale, is expected to carry on the disorder symptoms (if any). The procedure is performed over each individual. Acoustic feature were selected using valence annotations and no other information regarding diagnosis. The DD model is indeed constructed using a supervised learning approach ran over different individuals, thus requiring a cross-validation procedure

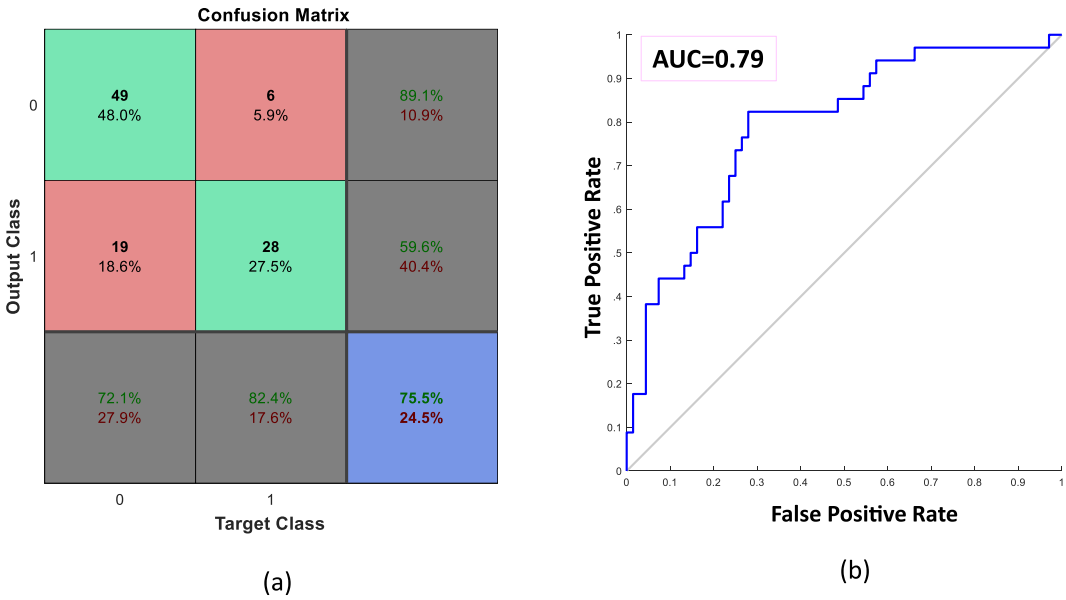

Figure 4. Results of the proposed EMF paradigm. (**a**) Confusion matrix of the recognition performance obtained by the proposed approach. Grey boxes at the bottom indicate the average recognition rate of TD subjects (72.1% Specificity) and of subjects with DD (82.4% Sensitivity or Recall), leading to an unbalanced accuracy equal to 75.5% (blue box). Balanced accuracy ACC is equal to 77%. Positive Predictive Value (PPV) is equal to 59.6% and false omission rate (FOR) is equal to 89.1%. (**b**) The Area under the Receiving operating curve (AUC) value equals to 0.79.

for training and validation. In particular, by implementing a leave-one-patient-out cross validation procedure, we ran the dynamic feature selection (DFS) approach (see Section 3.3) to dynamically select the features according to each test data[37,38]. A binary classification model based on an SVM, with linear kernel and default parameters setting, is then trained on the EMF matrix of all the subjects except for one that, in turn, is left out for test. The ACC is computed along the area under the Receiving Operating Characteristic (ROC) curve (i.e., AUC). Figure 4(a) shows the confusion matrix obtained in our test while Fig. 4(b) reports the ROC curve, with the corresponding AUC indicated.

EFM coefficient and disorder's diagnosis. Figure 5 describes the percentage of selection for each of the 57 features obtained using the DFS approach (DD vs TD recognition) and displayed according to diagnosis. Recall that the 57 features have been extracted by computing standard high-level statistical descriptors from 19 acoustic features. The 19 selected features belong to the well-defined groups described in Table 1[39]. All retained features are related to spectral and cepstral acoustic descriptors.

As can be seen from Fig. 5, on average, a very homogeneous group of features were selected for TD subjects (Fig. 5a). Similarly, selected features were almost the same across the three different groups of DD (Fig. 5b–d), thus reinforcing the assumption that the EFM coefficients can represent the different facets of the disorder. Moreover, due to the DFS mechanism, features were selected according to test data. Hence, features were differently chosen in presence of different diagnosis. For example, note that feature 30 has been selected only in presence of AD or TD, while feature 57 is totally absent in SLI and TD. Most relevantly, selection of features 20–28 (skewness of features RAST-PLP computed over utterances with equal valence) in AD subjects manifests a strong deviation from that of TD subjects, indicating a high specific behaviour with respect to the diagnosis.

As an example, in Fig. 6, we further show the features selected for a TD subject compared with those selected for an SLI subject. Pink bars denote features selected for a TD subject (subject n. 2), yellow bars identify features selected for an SLI subject (subject n. 90), whereas cyan bars locate features selected for both. The limited number of cyan bars underlines the evident difference in features selected for the two subjects thus confirming the usefulness of the DFS approach.

Finally, in support to the efficiency of the novel paradigm, Table 2 lists the percentage of individuals from the different pathological subgroups (PDD-NOS, SLI, and AD) that were correctly included in the DD class.

Discussion

Experienced psychologists and psychiatrists can easily remark whether a subtle change in a patient's voice betrays a change of mood or suggest a specific disorder. Present machine learning algorithms are not so good. Indeed, the experimental results described in Section 2.2 pose a dilemma to the domain of speech analysis: can acoustic features be used directly for classification of psychiatric disorders? The standard paradigm presents strong criticisms that have been previously reported in[35] and are further demonstrated by the present test (Section 2.2), suggesting that the answer to the question is more likely 'no'. The novel paradigm described here opens a promising

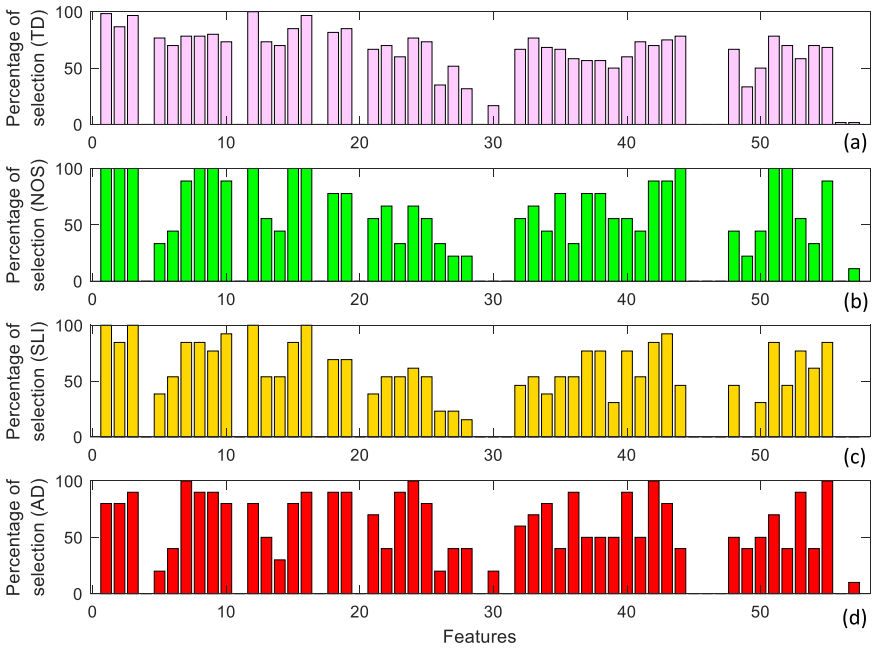

Figure 5. Percentage of feature selected. Percentage of feature selection for each of the 57 features obtained using the DFS approach (DD vs TD recognition) and displayed according to diagnosis. (**a**) TD children, (**b**) NOS children, (**c**) SLI children, (**d**) AD children. It is interesting to note that a very homogeneous group of features were selected for TD subjects (**a**) and similarly, selected features were almost the same across the three different groups of DD (**b**–**d**). This evidence reinforces the assumption that the value of the EFM coefficients can represent the different facets of the disorder.

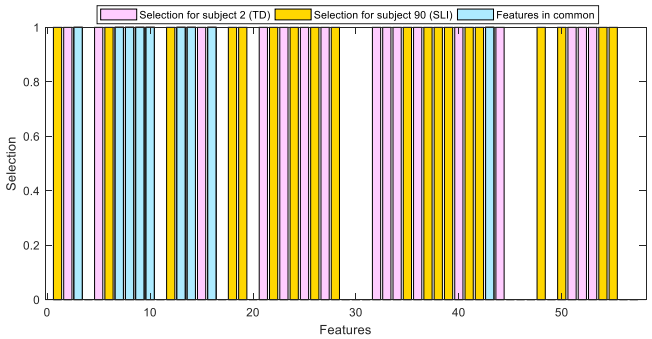

Figure 6. Two examples of feature selection. Pink bars denote features selected for a TD subject, yellow bars identify features selected for a SLI subject, whereas cyan bars locates features selected for both.

Feature N.	Features Group
1–16	RelAtive Spectral TrAnsform Perceptual Linear Prediction (RASTA-PLP) that considers temporal properties of the human hearing and speech production systems.
17	Functionals computed over spectral moments.
18–19	Functionals computed over Mel-Frequency Cepstral Coeffi cient (MFCC).

Table 1. Features group according to the categorization defined in[39].

	PDD-NOS	SLI	AD
Rate of classification	7/10 (70%)	12/13 (92%)	9/11 (82%)

Table 2. Rate of correctly classified individuals in each subcategory of DD: PDD-NOS, SLI, and AD. TOT quantifies the total number of cases for each subcategory.

alternative, indicating that by applying the Emotional Modulation function (EMF), rather than running analyses on acoustic features *per se*, the answer to the question could be turned to 'yes'. The EMF is a quantitative way to model the emotional speech reaction of a single individual to a given stimulus. The rationale behind this method is that differently from acoustic features, the EMF 'keeps' the disorder's traits. This specificity arises from the fact that EMF is a personalized concept, different for each individual. It encompasses the natural heterogeneity that individuals manifest during storytelling, verbal responding, and speech production in general.

As shown in Fig. 4, this novel approach allows to clearly separate traits that belong to TD individuals from those that characterize DD individuals. Moreover, although here, we only explored the binary classification problem of recognizing TD vs DD individuals, Table 1 clearly shows an excellent performance in the number of true positive subjects (TPs) included in each subcategory of DD. It is interesting to note that based on the approach used here, the PDD-NOS category is the class most likely to be confused with TD children. PDD-NOS corresponds to Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS)[35], which is characterised by social, communicative and/or stereotypic impairments that are less severe than in AD and appear later in life. Hence, it describes individuals likely to display more common features with TD participants. On the contrary, SLI impairment appears as the most evident disorder category, since Selective Language Impairment directly impacts on speech, and hence on the EMF. AD diagnosis reaches a recognition rate equal to 82%. Taken together, these results encourage research in this direction, with the acquisition of a larger dataset of children, thus, allowing the development of a 4-class recognition model.

By modelling the emotional speech reaction of each participant individually, the present method overcomes a frequent limitation of statistical approaches to clinical samples, i.e. their generalizability. ASD is a profoundly heterogeneous condition, meaning that sampling methods are likely to impact on the models that are developed. For example, a meta-analysis on studies on emotion recognition in ASD recently pointed out that females and individuals outside the typical IQ range are poorly represented in the common study populations[40]. Similarly, among the examined studies, half had been conducted in the US, the others in the UK, Australia and Ireland, suggesting that only limited socio-demographic characteristics are likely to be represented in the sampled population. When this is added to the variability intrinsic to the disorder (symptoms severity, co-morbidity, etc.) it becomes clear that statistical models averaging across data samples are liable to show underlying biases and be poorly generalizable. Conversely, an algorithm that models each individual's emotional speech reaction has more probability of succeeding in providing a method that is be less encumbered by the disadvantages linked to heterogeneous samples.

If the present findings may upset the traditional way of reasoning, on the other hand, it opens up new clinical and diagnostic scenarios through a change of perspective. The EMF can play a crucial role in diverse diagnostic tools, beyond DD disorder, as a way to extrapolate the hidden traits of a given disorder – bypassing, but embedding, speech. Moreover, the potentiality of EMF spreads over any kind of communicative act (multimodal emotional cues – audio, video, physiological, etc.) and hence, its efficacy can be proved in very diversified contexts.

Conclusion

In the present work, we explored the possibility of applying a precision approach to the development of a statistical learning algorithm aimed at classifying samples of speech produced by children with developmental disorders (DD) and typically developing (TD) children. Under the assumption that acoustic features of vocal production could not be efficiently used as a direct marker of DD, we propose here a radical change of paradigm. The novel way of reasoning described here opens a promising alternative, by suggesting to apply the Emotional Modulation function (EMF) concept, rather than running analyses on acoustic features *per se*. The EMF is a quantitative way to model the emotional speech reaction of a single individual to a given stimulus. It 'keeps' the disorder's traits while encompassing the natural heterogeneity that individuals manifest during speech production in general. Recognition performance along with comparative results with standard approaches demonstrates the efficacy of the proposed methodology opening up new clinical and diagnostic scenarios through a change of perspective.

Material and Methods

Database. In this study, we used the French Child Pathological & Emotional Speech Database (CPESD)[35] who received the approval by the Ethical Committee of the Pitié-Salpétrière Hospital to conduct recruitment and speech recording of children (as already illustrated in details in[35]). All the thirty-four monolingual participants with communicative verbal skills were recruited in two University departments of child and adolescent psychiatry located in Paris, France. They were diagnosed as AD (Autism Disorders), PDD-NOS (Pervasive Developmental Disorder-Not Otherwise Specified), or SLI (specific language impairment), according to DSM IV criteria[41]. An additional group of 68 TD (Typically Developing) children was recruited in elementary schools. The 102 participants included 21 girls (mean age 11.09; std 4.15) and 81 boys (mean age 9.24; std 2.94). Mean age equally distributed over the three distinct DD diagnoses and the control subjects. Average Verbal Intelligence Score (VIQ) and of Performance Intelligence Quotient (PIQ) resulted 50 ($\pm$8.3), 85($\pm$14.4), and 71(11.7) and 77 ($\pm$15.3), 76.8($\pm$10.5), 95.4($\pm$14.5) for for AD, PDD-NOS, and SLI subjects, respectively. Further details can be found in[42].

We will denote as Developmental Disorders (DD) children belonging to categories (AD, PDD-NOS, and SLI). A questionnaire was used to exclude children with learning disorders, a history of speech, language, hearing, or general learning problems. The task administered to the 102 participants was based on a story-telling of a pictured book "Frog where are you?"[36], wherein a little boy tries to find his frog, escaped during the night. The underlying assumption was that the child is supposed to produce prosodic cues during the story-telling that are correlated to the levels of the emotional valence, which was categorized in three categories by a psychologist: Negative/Neutral/Positive. In total, the pictured book included 15 emotionally negative, six emotionally neutral

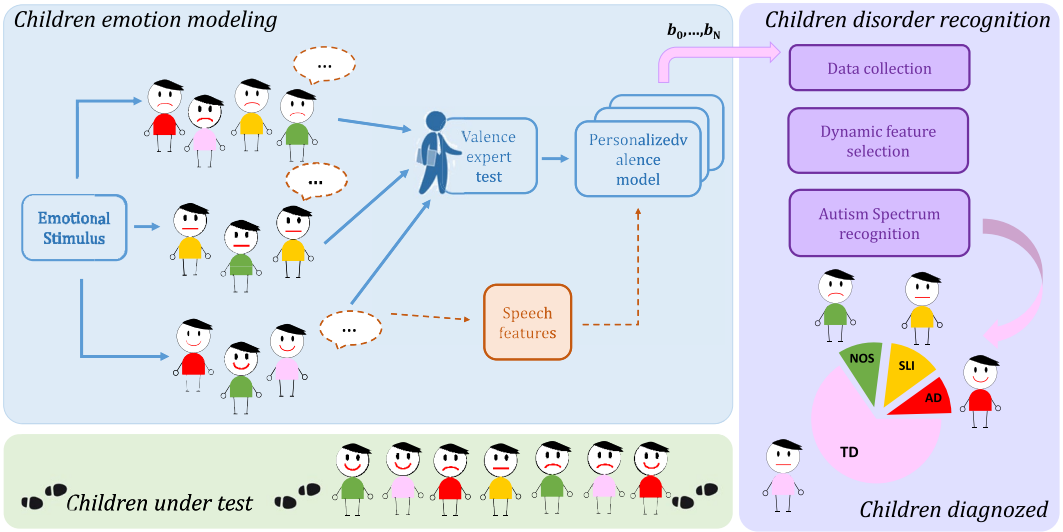

Figure 7. Schematic representation of the whole approach. Scheme of the proposed approach for DD recognition in children using an emotional-valence based speech modelling.

and five emotionally positive pictures. In the database, nearly 10 hours of recording were collected: 7 h 38 min for TD children, 1 h 35 min for children with AD, 1 h 12 min for children with NOS, and 1 h 56 min for children with SLI. Recordings were then segmented automatically into groups of breaths, using the energy contour. To eliminate sources of perturbation appearing during the recordings (e. g., false-starts, repetitions or environmental noise), the speech segments were further manually processed; only utterances with a complete prosodic contour, i.e., whatever the pronounced words, were kept. For each utterance valence was assessed in three categories by a psychologist: *negative* (labelled as −1), *neutral* (labelled as 0), and *positive* (labelled as +1). Further statistics (number, relative proportion, and mean duration) on those utterances, provided for each valence category, can be found in[35].

Speech Analysis. Acoustic features were automatically extracted from the speech waveform on the utterance level using the 2.2 release of the open-source openSMILE feature extractor[43]. Five different feature sets were investigated: large brute-forced feature sets (IS09, IS11, and ComParE), which have all been used for paralinguistic information retrieval, and a smaller, expert knowledge based feature set (eGeMAPS). Those feature sets cover spectral-, source- and duration-related feature space with different levels of detail, cf. Table 1. The first four sets, i.e., IS09, IS11, and ComParE (IS13 in the following), show a clear tendency in enlarging the feature space over the years, by including further low-level acoustic descriptors and associated functionals. Recently, this "brute-forcing" approach has been revisited, with investigations on a small, expert knowledge based feature set, eGeMAPS[44]. A detailed description and implementation of these feature sets is given in[39]. Aggregation of all the available descriptors leads to a feature set of 11227 different descriptors (384 from IS09, 4368 from IS11, 6373 from IS13, and 102 from eGeMAPS). Features redundancy has been solved using mutual correlation analysis. Feature values were then averaged over each utterance hence providing a feature vector for each sentence, leading to a variable number of data for each participant and therein for each affective condition.

Methods. The procedure is summarized by the pictorial scenario shown in Fig. 7. From left to right, we have a group of participants (children), each with a reported diagnosis (red for AD, yellow for SLI, and green for PDD-NOS) and a number of TD subjects (pink dressed) as assessed by experienced psychiatrists. All children were presented with emotional stimuli (here the story-telling, but it can be formulated on different kind of tasks) and their valence attitude was assessed by an expert evaluator during the test administering. The facial expressions in the drawings identify different valence attitude. As observed, there is no a-priori apparent correlation between valence attitude and disorder. Utterances pronounced during the emotional stimulation are recorded (dashed brown arrows) and sent to an automatic speech analysis tool that extracts the acoustic descriptors described in Section 2.1. Descriptors were used to train a personalized valence recognition model. Model coefficients estimated for each participant are collected in a data matrix (data collection block) as the individual signature of emotional attitude in response to a known unique stimulus. The emotional signature of participants with known DD diagnosis are used to train a model, for the automatic discrimination of DD subjects vs control (TD) subjects (pie chart at the bottom-left). Let us consider now in detail each session of the whole methodology.

N subjects have been registered collecting a set of speech sequences N_{si} i = 1, …, N, whose number can be different for each subject. From each sequence, a set of acoustic descriptors are extracted, namely $x_1(k)$, …., $x_{Nf}(k)$, where $k = 1, \ldots, N_{si}$ indicates the sequence and N_f indicates the number of features originally measured. For each subject i and for all the relative sequences, we have a feature matrix and an emotion sequence, indicated respectively with $\mathbb{X}_{N_{si}}$ and $\mathcal{E}_i$, defined as

$$\mathbb{X}_{N_{si}} = \begin{bmatrix} x_1(1) & \cdots & x_{N_f}(1) \\ \vdots & \ddots & \vdots \\ x_1(N_{si}) & \cdots & x_{N_f}(N_{si}) \end{bmatrix} \mathcal{E}_i = \begin{bmatrix} e(1) \\ \vdots \\ e(N_{si}) \end{bmatrix} e(k) \in \{-1, 0, 1\}, \tag{1}$$

where label "−1" corresponds to negative valence, label "0" corresponds to neutral valence, and label "+1" corresponds to positive valence. Such labels have been assessed by an expert evaluator.

Emotion-related feature selection. The first step aims to select features (i.e., column vector in matrix $\mathbb{X}_{N_{si}}$) that mostly correlate with the emotion vector $\mathcal{E}_i$ for each subject *i*.

To do this, we computed the Pearson correlation coefficient $\rho_{j,i}$ between columns of matrix $\mathbb{X}_{N_{si}}$, X_j, $j = 1, \ldots, N_f$, and $\mathcal{E}_i$ as follows

$$\rho_{j,i} = \frac{E[(X_j - \mu_{X_j})(\mathcal{E}_i - \mu_{\mathcal{E}_i})]}{\sigma_{X_j} \cdot \sigma_{\mathcal{E}_i}}, \tag{2}$$

where μ_{Xj} and σ_{X_j} are the average and the standard deviation of values in column X_j, while $\mu_{\mathcal{E}_i}$ and $\sigma_{\mathcal{E}_i}$ are the average and the standard deviation of values in vector $\mathcal{E}_i$, and $E[]$ indicates the expected value. The absolute value of $\rho_{j,i}$ is indicative of the degree of correlation each feature vector X_j has with the corresponding sequence of emotion $\mathcal{E}_i$.

Then, for each subject *i*, we finally select those features having an absolute value of $\rho_{j,i}$ larger than 0.7, experimentally set, and achieved the following subset of selected features

$$S_i = \{x_j | |\rho_{j,i}| > 0.7\}, \ j = 1, \ldots, N_f, \tag{3}$$

Hence, by the union of the features selected for each subject, we finally obtained a set of selected features for the entire dataset of subjects

$$S = \bigcup_i S_i. \tag{4}$$

Let us indicate in the following with N_{opt} the number of selected features, where in general $N_{opt} < N_f$, and with $\mathcal{F}_i = [f_{i1}, \ldots, f_{iN_{opt}}]$ the features for all the subjects. Actually, the set of features are always the same for all the subjects in order to derive an equal number of descriptors for each individual.

Personalized Emotional model. In order to construct a personalized model of emotion for each subject *i*, we preliminary divided the matrix $\mathcal{F}_i$ into the three submatrices, $\mathcal{F}_i|_{\mathcal{E}_i=-1} \triangleq \mathcal{F}_i^{-1}$, $\mathcal{F}_i|_{\mathcal{E}_i=0} \triangleq \mathcal{F}_i^{0}$, and $\mathcal{F}_i|_{\mathcal{E}_i=-1} \triangleq \mathcal{F}_i^{-1}$ that represent the selected features extracted from the sequences of negative, neutral and positive valence, respectively. Correspondingly, let us denote with N_{si}^{-1}, N_{si}^{0} and N_{si}^{+1} the number of sequences for emotions labelled as "−1", "0" and "+1" respectively, for subject *i* and with f_{ij}^{-1}, f_{ij}^{0} and f_{ij}^{+1} the feature values of each submatrix, $j = 1, \ldots, N_{\text{opt}}$.

In order to provide a synthetic representation of the emotions picture for a subject, we described the distribution of feature values for each emotion by computing the first, the third and the fourth statistical moments, i.e., the *mean*, the *skewness* and the *kurtosis*, respectively. The skewness parameter is usually used to evidence deviation from Gaussian nature, since it provides a degree of asymmetry of a given distribution of values. The kurtosis instead, also named tailedness, is related to the amount of tails the distribution has with respect to the Gaussian. Both the moments are descriptors of the shape of a distribution more than being descriptors of their localization in the feature space, as conversely the first moment is. Mean μ, skewness *sk* and kurtosis *ku* are defined as follows:

$$\mu_{ij}^{-1} = \frac{\sum_{k=1}^{N_{si}^{-1}} f_{ij}^{-1}(k)}{N_{si}^{-1}}, \ \mu_{ij}^{0} = \frac{\sum_{k=1}^{N_{si}^{0}} f_{ij}^{0}(k)}{N_{si}^{0}}, \ \mu_{ij}^{+1} = \frac{\sum_{k=1}^{N_{si}^{+1}} f_{ij}^{+1}(k)}{N_{si}^{+1}}, \tag{5}$$

$$sk_{ij}^{-1} = \frac{\sum_{k=1}^{N_{si}^{-1}} \left(f_{ij}^{-1}(k) - \mu_{ij}^{-1}\right)^3}{\left(\sigma_{ij}^{-1}\right)^3}, \ sk_{ij}^{0} = \frac{\sum_{k=1}^{N_{si}^{0}} \left(f_{ij}^{0}(k) - \mu_{ij}^{0}\right)^3}{\left(\sigma_{ij}^{0}\right)^3}, \ sk_{ij}^{1} = \frac{\sum_{k=1}^{N_{si}^{+1}} (f_{ij}^{+1}(k) - \mu_{ij}^{1})^3}{\left(\sigma_{ij}^{1}\right)^3}, \tag{6}$$

$$ku_{ij}^{-1} = \frac{\sum_{k=1}^{N_{si}^{-1}} \left(f_{ij}^{-1}(k) - \mu_{ij}^{-1}\right)^4}{\left(\sigma_{ij}^{-1}\right)^4}, \ ku_{ij}^{0} = \frac{\sum_{k=1}^{N_{si}^{0}} (f_{ij}^{0}(k) - \mu_{ij}^{0})^4}{\left(\sigma_{ij}^{0}\right)^4}, \ ku_{ij}^{1} = \frac{\sum_{k=1}^{N_{si}^{+1}} (f_{ij}^{+1}(k) - \mu_{ij}^{1})^4}{\left(\sigma_{ij}^{1}\right)^4}, \tag{7}$$

where σ_{ij}^{-1}, σ_{ij}^{0}, σ_{ij}^{1} are given by

$$\sigma_{ij}^{-1} = \frac{\sum_{k=1}^{N_{si}^{-1}} (f_{ij}^{-1}(k) - \mu_{ij}^{-1})^2}{N_{si}^{-1}}, \ \sigma_{ij}^{0} = \frac{\sum_{k=1}^{N_{si}^{0}} (f_{ij}^{0}(k) - \mu_{ij}^{0})^2}{N_{si}^{0}}, \ \mu_{ij}^{-1} = \frac{\sum_{k=1}^{N_{si}^{+1}} (f_{ij}^{+1}(k) - \mu_{ij}^{+1})^2}{N_{si}^{+1}}$$

For each subject i, a matrix $\mathbb{M}_i$, $3 \times 3\ N_{opt}$ of synthetic descriptors and a corresponding vector of emotions $\mathbb{V}_i$, 3×1, are built as follows

$$\mathbb{M}_i = \begin{bmatrix} sk_{i1}^{-1} & \cdots & sk_{iN_{opt}}^{-1}\ \mu_{i1}^{-1} \cdots \mu_{iN_{opt}}^{-1}\ ku_{i1}^{-1} \cdots ku_{iN_{opt}}^{-1} \\ sk_{i1}^{0} & \cdots & sk_{iN_{opt}}^{0}\ \mu_{i1}^{0} \cdots \mu_{iN_{opt}}^{0}\ ku_{i1}^{0} \cdots ku_{iN_{opt}}^{0} \\ sk_{i1}^{+1} & \cdots & sk_{iN_{opt}}^{+1}\ \mu_{i1}^{+1} \cdots \mu_{iN_{opt}}^{+1}\ ku_{i1}^{+1} \cdots ku_{iN_{opt}}^{+1} \end{bmatrix}, \text{ and } \mathbb{V}_i = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The coefficient vector $\mathcal{B}_i$ of a multilinear regression estimated by the orthogonal least square approach is then achieved for each subject i by

$$\mathcal{B}_i = (\mathbb{M}_i^T \mathbb{M}_i)^{-1} \mathbb{M}_i^T \mathbb{V}_i. \tag{8}$$

Coefficient vector $\mathcal{B}_i$ represents the personalized model of emotion of each subject i, his/her EMF, i.e., the way the subject reacts to specific emotional stimuli provided during the task with his/her own speech frequency alteration. The assumption is that EMF coefficients $\mathcal{B}_i$ may be used to discriminate TD subjects from DD subjects, by concealing the different emotional picture of the two groups of subjects.

An emotional-guided diagnostic tool for DD patients. By collecting as rows the coefficient vector $\mathcal{B}_i$ for all the subjects – for simplicity sorted according to the disorder (i.e., control, label 1, label 2 and label 3) – we constructed a data matrix $\mathbb{D}$, a corresponding binary disorder-labelled vector, $\mathbb{Y}$ and a 4-classes disorder-labelled vector, $\mathbb{L}$, as follows

$$\mathbb{D} = \begin{bmatrix} -\mathcal{B}_1 - \\ \vdots \\ -\mathcal{B}_i - \\ \vdots \\ -\mathcal{B}_N - \end{bmatrix}, \quad \mathbb{Y} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbb{L} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 2 \\ \vdots \\ 3 \end{bmatrix} \tag{9}$$

where values $\mathbb{Y} \equiv 1$ correspond to any $\mathbb{L} > 0$. Due to the low number of cases for each disorder, 10 (NOS), 13 (SLI) and 11 (AD) respectively, we decided to develop a binary classification model able to recognize TD subject from DD subject. Hence, we considered the labelled vector $\mathbb{Y}$ as ground truth.

Under the assumption that features selected play a crucial role in the recognition performance, especially due to the heterogeneity of the test set, we applied here a dynamic feature selection (DFS) procedure intended to optimally select model features according to each specific test data. More specifically, in line with the recently developed methodology[37,38], we design the following three-level DFS approach:

Test-independent feature elimination step: Starting from the training set, the Fisher Discriminant Score (FDS) is used to sort all the available features according to their compliance with the classification problem. In particular, a feature to be included requires that at least one class-distribution is statistically different from the others. Let us consider a two-class problem, and let us assume that each class has L_d, $d = \{1, 2\}$ training vectors $\mathcal{B}_i$ each formed by elements b_{ik}, $i = 1, .., L_d$, $k = 1, \ldots, M$, with M as the total number of features. Then, for each feature k, FDS_k is defined as the ratio between intra-class and inter-class variance and it is estimated as follows:

$$FDS_k = SB_k / SW_k, \tag{10}$$

where SB_k is the intra-class and SW_k is the inter-class variance. In particular, SB_k is defined as

$$SB_k = \sum_{d=1}^{2} \left(b_{\{Y=Y_d\}k} - \bar{b}_k \right)^2, \tag{11}$$

with $Y_d = \{0, 1\}$, $\bar{b}_k$ is the average of the feature values b_k computed over all the classes. For each feature, k, SB_k quantifies the sum of dispersions of training samples in a class (by their variance) with respect to the global average value of that feature.

On the other hand, for each feature k, we also computed

$$SW_k = \sum_{d=1}^{2} \frac{1}{L_d} \sum_{i=1}^{L_d} \left(b_{\{Y=Y_d\}_i k} - \bar{b}_{dk} \right)^2, \tag{12}$$

with $b_{\{Y=Y_d\}_i k}$ as the *i-th* element of feature *k-th* for the class labelled as Y_d and $\bar{b}_{dk}$ is the average value of feature *k-th* in the class labelled as Y_d. For each feature, SW_k quantifies the dispersion of elements in a class d, with respect to their average value, i.e., the inter-class variance.

Higher values for FDS_k indicate that the feature k is representative of at least one class and hence will be maintained. For this task, we will define a threshold value th_{FDS} and define the Condition 1 (C_1) as follows

$$C_1\text{: } b_k \text{will be kept iff } \{FDS_k > th_{FDS}\}, \tag{13}$$

Online test-dependent feature elimination step: This step utilizes two criteria for selecting a temporary subset of features. The decision is made in accordance with the sample reservoir containing the information about the class distributions and the test sample newly acquired. This step is intended to online remove the features in which either the test sample is far from all class distributions (feature outlier values) or it is surrounded by samples of different classes (high probability of misclassification). In order to achieve these targets, the algorithm selects a feature only if it fulfils the two following criteria. Let us denote with s the test sample and with s_k its k-th element, for brevity test element.

The *first* criterion is the ratio between the Mahalanobis distances of the test element s_k from the two class distributions, hereinafter denoted as MR_k (s_k). This value is calculated for each feature k as follows:

$$MR_k(s_k) = M_k(s_k,\ Y = Y_1) \cdot \frac{M_k(s,\ Y = Y_1)}{M_k(s,\ Y = Y_2)}, \tag{14}$$

where $M_k(s_k,\ Y = Y_d)$ is the Mahalanobis distance of the test element s_k from the training samples belonging to the class labelled as Y_d and it is defined as

$$M_k(s_k,\ Y = Y_d) = (s_k - \bar{b}_{dk}) * \left(Cov\left(b_{\{Y=Y_d\}k}\right)\right)^{-1} * (s_k - \bar{b}_{dk})^{\mathrm{T}}, \tag{15}$$

where $Cov(b_{\{Y=Y_d\}k})$ is the covariance of the feature matrix of samples belonging to the class labelled as Y_d. The descriptor $MR_k(s_k)$ provides a quantitative measure of the distance of the test sample from the two classes. Lower values of $MR_k(s_k)$ indicate that the test element s_k is close to a given class while being far from the other class. Hence, defined a threshold value th_{MR}, a Condition 2 (C_2) will be formulated as follows:

$$C_2\text{: } b_k \text{will be kept iff } \{MR_k < th_{MR}\}. \tag{16}$$

The *second* criterion evaluates the maximum probability for the test element s_k to belong to each class distribution. For the k-th feature, P_k is computed as follows:

$$P_k(s_k) = \max_d\left(\frac{1}{\sqrt{2\pi}\,\sigma_{dk}}\exp\left(-\frac{(s_k - \bar{b}_{dk})^2}{2\sigma_{dk}^2}\right)\right), \tag{17}$$

where σ_{dk} is the standard deviation of training sample values for feature k-th in the class labelled as Y_d. Higher values for $P_k(s_k)$ indicate that the test element s_k has a high probability to be correctly represented by a given class distribution. For this reason, defined a threshold value th_{P}, a Condition 3 (C_3) will be formulated as follows:

$$C_3\text{: } b_k \text{will be kept iff } \{P_k > th_P\} \tag{18}$$

For each test element s_k, only the features b_k in the training set that respect conditions C_1 and in cascade conditions C_2–C_3 are used to construct the predictive model. In our approach, a Support Vector Machine (SVM) with linear kernel and standard parameters setting was preferred for the scope. Features are selected at each test step; features eliminated in a step will be re-inserted in the training set and re-considered for the feature selection procedure at the next step in presence of a different test data.

References

1. Delavarian, M., Towhidkhah, F., Gharibzadeh, S. & Dibajnia, P. Automatic classification of hyperactive children: comparing multiple artificial intelligence approaches. *Neurosci Lett.* **498**, 190–3 (2011).
2. Ringeval, F. *et al*. Automatic Intonation Recognition for Prosodic Assessment of Language Impaired Children. *IEEE Transactions on Audio, Speech & Language Processing* **19**, 1328–1342 (2011).
3. Petrick, N. *et al*. Evaluation of computer-aided detection and diagnosis systems. *Med Phys.* **40**, 087001 (2013).
4. Aboujaoude, E. & Salame, W. Technology at the Service of Pediatric Mental Health: Review and Assessment. *J Pediatr* **171**, 20–4 (2016).
5. Iniesta, R. *et al*. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research* **78**, 94–102 (2016).
6. Hamet, P. & Tremblay, J. Artificial Intelligence in Medicine. *Metabolism* **49**, 36–40 (2017).
7. Takahashi, R. & Kajikawa, Y. Computer-aided diagnosis: A survey with bibliometric analysis. *Int J Med Inform.* **101**, 58–67 (2017).
8. Hallgren, K. A., Bauer, A. M. & Atkins, D. C. Digital technology and clinical decision making in depression treatment: Current findings and future opportunities. *Depress Anxiety* **34**, 494–501 (2017).
9. Hastie, T., Tibshirani, R. & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag: New York (2009).
10. Ghahraman, Z. Unsupervised learning. *Advanced Lectures on Machine Learning* **3176**, 72–112 (2003).
11. Fan, Y., Shen, D., Gur, R. C., Gur, R. E. & Davatzikos, C. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging* **26**, 93–105 (2007).
12. Yang, H., Liu, J., Sui, J., Pearlson, G. & Calhoun, V. D. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. *Frontiers in Human Neuroscience* **4**, art.192 (2010).
13. Castellani, U. *et al*. Classification of schizophrenia using feature-based morphometry. *J. Neural Transm* **119**, 395–404 (2012).
14. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137–165 (2017).
15. Iniesta, R., Stahl, D. & McGuffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med.* **46**, 2455–65 (2016).
16. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol.* **14**, 1109–20 (2015).
17. Lai, M. C., Lombardo, M. V., Auyeung, B., Chakrabarti, B. & Baron-Cohen, S. Sex/gender differences and autism: setting the scene for future research. *Journal of the American Academy of Child and Adolescent Psychiatry* **54**, 11–24 (2015).

18. Lombardo, M. V. *et al*. Different functional neural substrates for good and poor language outcome in autism. *Neuron* **86**, 567–577 (2015).
19. Geschwind, D. H. & Levitt, P. Autism spectrum disorders: developmental disconnection syndromes. *Curr Opin Neurobiol.* **17**, 103–11 (2007).
20. Deng, J. *et al*. Speech-based diagnosis of Austim Spectrum Condition by Generative Adversarial Network Representations, *in Proc. of the 7th International Digital Health Conference (ACM)*, pp. 53–57 (London, UK, July 2017).
21. Abrams, J. *et al*. National Cancer Institute's Precision Medicine Initiatives for the new National Clinical Trials Network, *Am Soc Clin Oncol Educ Book*, 71–6 (2014).
22. Ozomaro, U., Wahlestedt, C. & Nemeroff, C. B. Personalized medicine in psychiatry: problems and promises. *BMC Med.* **16**(11), 132 (2013).
23. Insel, T. R. & Cuthbert, B. N. Medicine. Brain disorders? Precisely. *Science* **348**(6234), 499–500 (2015).
24. Van der Stel, J. C. Precision in psychiatry. *Acta Psychiatr Scand.* **132**, 310–1 (2015).
25. Rapin, I. & Allen, D.A. Developmental language: nosological consideration, *Neuropsychology of Language*, *Reading*, *and Speeling*, (Kvik, V. Ed.) Washington, DC: New York: Academic Press (1983).
26. Sheinkopf, S. J., Mundy, P., Oller, D. K. & Steffens, M. Vocal atypicalities of preverbal autistic children. *J Autism Dev Disord* **30**, 345–54 (2004).
27. Shriberg, L. D. *et al*. Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *J Speech Lang Hear Res.* **44**, 1097–115 (2001).
28. Hubbard, K. & Trauner, D. A. Intonation and emotion in autistic spectrum disorders. *J Psycholinguist Res.* **36**, 159–73 (2007).
29. Sharda, M. *et al*. Sounds of melody–pitch patterns of speech in autism. *Neurosci Lett.* **478**, 42–5 (2010).
30. Hubbard, D. J., Faso, D. J., Assmann, P. F. & Sasson, N. J., Production and perception of emotional prosody by adults with autism spectrum disorder, *Autism Res*., 17 Aug (2017).
31. Nakai, Y., Takashima, R., Takiguchi, T. & Takada, S. Speech intonation in children with autism spectrum disorder. *Brain Dev.* **36**, 516–22 (2014).
32. Rutter, M. J. *et al*. *Rutter's child and adolescent psychiatry*, John Wiley & Sons (2011).
33. Schuller, B. *et al*. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, *Proceedings INTERSPEECH* 2013, *14th Annual Conference of the International Speech Communication Association*, Lyon, France (2013).
34. Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M. & Gaigg, S. B. Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Res.* **10**, 384–407 (2017).
35. Ringeval, F. *et al*. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children, In 17th Annual Conference of the International Speech Communication Association (pp. 1210–1214) (INTERSPEECH 2016).
36. Mayer, M. Frog where are you?, New York: *Dial Books for young readers* (1969)
37. Magna, G., Mosciano, F., Martinelli, E. & Di Natale, C. Unsupervised On-Line Selection of Training Features for a robust classification with drifting and faulty gas sensors. *Sensors and Actuators B: Chemical* **258**, 1242–1251 (2018).
38. Mosciano, F. *et al*. An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario. *Sensors and Actuators A: Physical* **267**, 48–59 (2017).
39. Eyben, F. *Real-time speech and music classification by large audio feature space extraction*, (Springer, 2015).
40. Berggren, S. *et al*. Emotion recognition training in autism spectrum disorder: A systematic review of challenges related to generalizability. *Developmental neurorehabilitation* **21**, 141–154 (2018).
41. DSM IV, "Diagnostic and Statistical Manual of mental disorders (4th Ed.)". (Washington, DC: American Psychiatric Association, 1994).
42. Demouy, J. *et al*. Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment. *Research in Autism Spectrum Disorders* **5**, 1402–1412 (2011).
43. Eyben, F., Weninger, F., Groß, F. & Schuller,B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in Proc. of the 21st ACM International Conference on Multimedia (ACM MM), Barcelona, Spain, pp. 835–838 (2013).
44. Eyben, F. *et al*. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* **7**, 190–202 (2016).

Author Contributions

A.M., F.M., M.C.C. and T.D.G performed simulations. F.B. and B.S. provided access to the dataset. E.M. and A.M designed and performed the experiments. G.R., E.D., F.R. and B.S. revised the manuscript. E.M and A.M wrote the manuscript. All authors revised the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.