Speech signal analysis as an aid to clinical diagnosis and assessment of mental health disorders

Ester Bruno[1,2,]*, Émilie Martz[3,4], Luisa Weiner[3,4], Alberto Greco[1], Nicola Vanello[1,2]

[1]Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
[2]Research Center "E. Piaggio" - University of Pisa, Pisa, Italy
[3]Department of Psychiatry, University Hospital of Strasbourg, Strasbourg, France
[4]Laboratoire De Psychologie Des Cognitions, University of Strasbourg, Strasbourg, France

*ester.bruno@phd.unipi.it

Abstract

Objective: In this study, we estimate speech features from different Verbal Fluency Tests (VFT) conditions to distinguish comorbid Bipolar Disorder (BD) in adults suffering from Attention Deficit Hyperactivity Disorder (ADHD) and to identify ADHD subtypes such as the inattentive (ADHD-I) from the combined one (ADHD-C).
Methods: Prosodic and spectral features in five conditions of VFTs were extracted and selected for the classification performed with machine learning methods. Specifically, a Support Vector Machine exploiting Recursive Features Elimination (SVM-RFE) has been trained with clinical scores and exploiting the within subject variability of speech features across VFT conditions. The final classification was optimized by combining the marginal classification outcomes obtained from the different VFTs using a voting scheme.
Results: Our results show that we successfully classify the ADHD+BD comorbidity and the ADHD subtypes according to clinician diagnosis. The results are discussed in the light of possible benefits of developing such approach within clinical research.
Conclusion: Significant information is carried out by speech audio features acquired with VFTs, allowing to classify ADHD subtypes and comorbid patterns. This work clearly shows that the audio analysis of speech, along with properly designed speech tasks, is a candidate for the development of clinical decision support systems in psychiatry.
Significance: This work represents a major contribution to the applications of speech analysis in ADHD subjects and could support clinicians by identifying objective biomarkers.

Keywords: acoustic analysis, Attention Deficit Hyperactivity Disorder, Bipolar Disorder, Verbal fluency tests, speech features, SVM-RFE.

1. Introduction

The development of approaches to speech and voice analysis for the evaluation of the emotional and mood state of the speaker [1], [2] raised great interest about its application for the development of clinical decision support systems. Since then, several new applications in mental health research have been proposed [2]–[4].

Interviews and questionnaires used by clinicians to estimate mood states in psychiatry and neurodevelopmental disorders often lead to misdiagnosis when they are used alone, given that they are prone to different methodological and subject response biases [5]. Clinical assessment is often complicated by the possible intermittent nature of symptoms as well as the presence of comorbidities that decrease inter-observer and test-retest reliabilities[6]. For these reasons, automatic approaches for speech analysis could be a powerful tool in psychiatry to identify objective markers to support the diagnosis and assess its severity, to detect prodromal signs and finally to formulate prognosis and modulate interventions [7]. Moreover, currently available technologies, such as smartphone devices, offer the possibility of monitoring subject status outside the clinical setting, e.g. between two visits, possibly detecting meaningful events or changes or to monitor treatment results [8].

Different scenarios were identified where automatic analysis of speech could contribute to a clinical decision support system, the more frequently proposed being bipolar depression, major depressive disorders and post-traumatic stress disorder [6]. Automatic speech analysis methods have been proposed for neurological disorders such as for Parkinson's, Alzheimer and MCI [9]–[11].

One of the main contributions of automatic speech analysis methods is the possibility of determining biomarkers for differential diagnosis. In fact, the diagnosis based on symptoms that can be determined by multiple factors, as well as the presence of comorbidities may lead to a lack of diagnostic specificity [7], [12]. For instance, in bipolar disorders, the use of automatic methods could give the clinician further information to discriminate mixed symptoms that occur when both manic and depressive symptoms are present, and require specific interventions [2], [8], [13].

Model based automatic approaches could be a powerful and easy-to-apply complement to help clinicians in the diagnosis of neurodevelopmental and psychiatric disorders, to explore and investigate different clinical presentations among diseases and to identify necessary interventions. This study deals with the design of an automatic tool to aid clinicians in the diagnosis and monitoring of Attention Deficit Hyperactivity Disorder (ADHD). Our approach is based on machine learning and on the analysis of speech acquired in controlled conditions, using specific speech tasks.

ADHD is a neurodevelopmental disorder affecting up to 4% of adults [14]–[16] co-occurring with Bipolar Disorder (BD) in 5 to 32% of cases [17]–[20] ADHD is characterized by pervasive and persisting inattention, excessive activity and impulsivity. Recent studies identified additional ADHD characteristics such as emotion dysregulation and a "ceaseless unfocused thinking", akin to racing thoughts [21]. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) identifies three presentations of ADHD: predominantly inattentive, predominantly hyperactive-impulsive and combined presentation. Inattentive subjects have trouble focusing their attention and concentrating. They may scarcely adhere to directions and miss important details, they seem absent-minded and lose track of their things. Hyperactive subjects are fidgety, restless and easily bored. They are constantly in motion, have difficulty performing quiet activities and they often interrupt conversations or others' activities. Individuals with the combined ADHD presentation display a mixture of all the symptoms outlined above. These deficits in social interactions present a central problem causing social, occupational and emotional disadvantages [22]. Symptoms of ADHD are treated with the combination of pharmacological and behavioral interventions. The greatest difficulty for clinicians is to identify ADHD and its subtypes, given that ADHD is often comorbid with other disorders, such as BD. The combined form in particular shares several symptoms with bipolar disorder, e.g., emotion dysregulation. This makes the differential diagnosis in clinical settings particularly difficult [6]. The identification of ADHD subtypes has important consequences on care due to the different treatment options that can be used according to the clinical presentation of patients.

The analysis of voice and speech is a good candidate to measure the emotion dysregulation characteristic of ADHD. In fact, the change of speech features with subject emotional and mood status has been deeply described [1], [23]–[25]. In particular, the relevance of prosodic features in emotional

expression was stressed [26]. In the field of psychiatric disorders, a special focus was put on the depressive condition [27], [28], showing a complex behaviour of speech features often, but not always as in the case of agitated depression, resulting in a low rhythm and flat intonation [3].

More recently, several studies have been performed using voice features to characterize subjects suffering from depression and bipolar disorders [1], [23]. To our knowledge, very few studies have investigated the utility of voice or speech features for the classification of ADHD and ADHD subtypes [29], [30]. Specifically, these works highlighted the potential relevance of speech analysis to differentiate ADHD from both normal [26], [27] and dyslexic [27] subjects. Moreover, in [27] the possibility of studying speech features changes with symptoms severity and with comorbidity was preliminarily evaluated. Overall, both studies emphasized how speech features may carry relevant information in both the diagnosis and classification of ADHD subjects. In this work, we propose the use of speech features related to prosody, as those describing F0 temporal dynamics and intonation, and of spectral features describing both vocal tract characteristics and speech quality, to classify ADHD subtypes and ADHD subject with comorbid bipolar disorder.

Noticeably, we used five different verbal fluency tests (VFTs) that have been designed to elicit different strategies and cognitive approaches. Indeed, according to the VFT condition considered, different dimensions of the executive functioning are measured. More specifically, the free, letter and semantic conditions of the task involve the "cold" executive functioning (i.e., initiation and flexibility of a search strategy and inhibition of irrelevant words [31], [32]. Whereas the emotions and action verbs VFT present the particularity of tackling the "hot" component of executive functioning, involved in emotions processing [33]. A recent study shows how VFTs could be an economic, fast and standardized means for acquiring speech samples overcoming disadvantages of studies using natural free speech such as the acquisition of a large amount of data or the ethical concerns lead to the recording of personal conversations. Speech features obtained from VFTs seem to be valid and informative potential biomarkers regarding mood dysregulations of neurodevelopment disorders [34].

The proposed approach will exploit a speech feature normalization that will highlight the within subject feature variability across verbal fluency conditions. The hypothesis is that the different groups of patients will differ in the response to different verbal fluency tests. Moreover, this approach will allow to remove the bias of subject-specific speech features, such as those related to gender and individual characteristics.

The present study aims at exploring whether speech features can be used to classify ADHD subtypes and the ADHD+BD comorbidity starting from speech signals recorded during VFTs, speech features were extracted and used to train classifiers through the application of robust and well-validated machine learning methods. The model was trained using the information of clinical labels provided by the physicians. This approach could support physicians in formulating a diagnosis and monitor patient status, also when the subject is not hospitalized. The major limitation of our classifiers is the small number of subjects (training samples) compared to the large number of features. We faced this issue by applying a feature selection algorithm with a backward elimination procedure which allowed us to identify the most informative subsets of features for the distinction of ADHD subtypes and ADHD+BD comorbidity.

2. Methods

2.1 Participants

67 patients with ADHD (age: 33±11 mean±dev.st.) and 30 patients presenting with the ADHD + BD comorbidity (age: 32±11 mean±dev.st.) were recruited from the outpatient psychiatry clinics at the University Hospital of Strasbourg.

ADHD diagnoses were established by senior psychiatrists according to the DSM-5 criteria for ADHD (APA,2013). The diagnosis was retained if patients present at least 5 inattentive and/or 5 hyperactive symptoms. Among the ADHD group, 11 patients were identified as inattentive and 41 as combined. Most ADHD patients (82%) and ADHD+BD patients (60%) were not taking any psychostimulant drugs. Those who were under psychostimulant drugs were asked to interrupt their treatment 24 h prior to the neuropsychological assessment. 4 patients, 3 ADHD and one ADHD+BD, with current history of other neurological or psychiatric disorders, such as substance use disorders (SUD), autism spectrum disorder (ASD), current depression or borderline personality disorder were discarded. Other 12 ADHD and 3 ADHD+BD patients were discarded for noisy audio quality, for having less than 5 VFTs and for no consent.

All subjects provided written informed consent prior to inclusion in the study in accordance with the Declaration of Helsinki. This study was approved by the regional ethics committee of Eastern France (CPP EST IV).

2.2 Verbal fluency tasks

Following the clinical assessment, voice signals were recorded during verbal fluency tasks (VFT) in a quiet and low reverberation room using Audacity© software (fs=44100 Hz, 24-bit resolution PCM). A Tascam DR-05 microphone was connected to a laptop and kept approximately 60 cm away from the subject. In VFT subjects were asked to complete different tasks belonging to five conditions following a fixed order: (1) the free fluency condition, (2) the letter fluency condition, (3) the semantic fluency condition, (4) the emotion fluency condition and finally (5) the verb fluency condition.

- *Free fluency condition*: in the free fluency condition, subjects were asked to produce as many words as possible during 150 seconds with their eyes closed [34];
- *Letter fluency condition:* in the letter fluency condition, subjects were asked to produce as many words as possible starting with the letter 'P', except for proper nouns, during 120 seconds [35];
- *Semantic fluency condition:* participants were asked to produce as many words as possible belonging to a specific category, 'animals', during 120 seconds [35];
- *Emotion fluency condition:* in this condition the subjects were asked to produce as many words as possible belonging to the emotion category, during 60 seconds;
- *Action verbs fluency condition:* in the action verbs fluency condition, subjects were asked to produce as many words as possible belonging to the verb category, during 60 seconds.

2.3 Extraction of vocal features

Recordings from 78 different subjects resulted in 390 audio signals. Prosodic features and spectral features were extracted and investigated. Prosody describes the rhythm, the stress and the intonation of the speech. Spectral features are related to the voice quality.

Features extraction was performed with BioVoice, a Matlab® toolbox developed at the Biomedical Engineering Lab, Firenze University [36]. BioVoice performs time and frequency analysis of audio signals concerning the human voice, from the newborn to the elder, estimating more than 20 acoustical parameters. BioVoice first identifies voiced/unvoiced (V/UV) audio segments (Fig.1 a) and then extracts from each voiced segment a set of features of interests, including among others the number and length of voiced segments, the number and length of pause segments, percentage of voiced segments in the time domain, speech fundamental frequency (F0), noise level (Normalized

Noise Energy), jitter and formant frequencies (F1, F2, F3) in the frequency domain (formants are the resonance frequencies generated by the laryngeal cavities) (Fig. 1 b). Along with these features, we used BioVoice to estimate statistical descriptor of F0 and of the formants such as mean, median, standard deviation, maximum and minimum values.

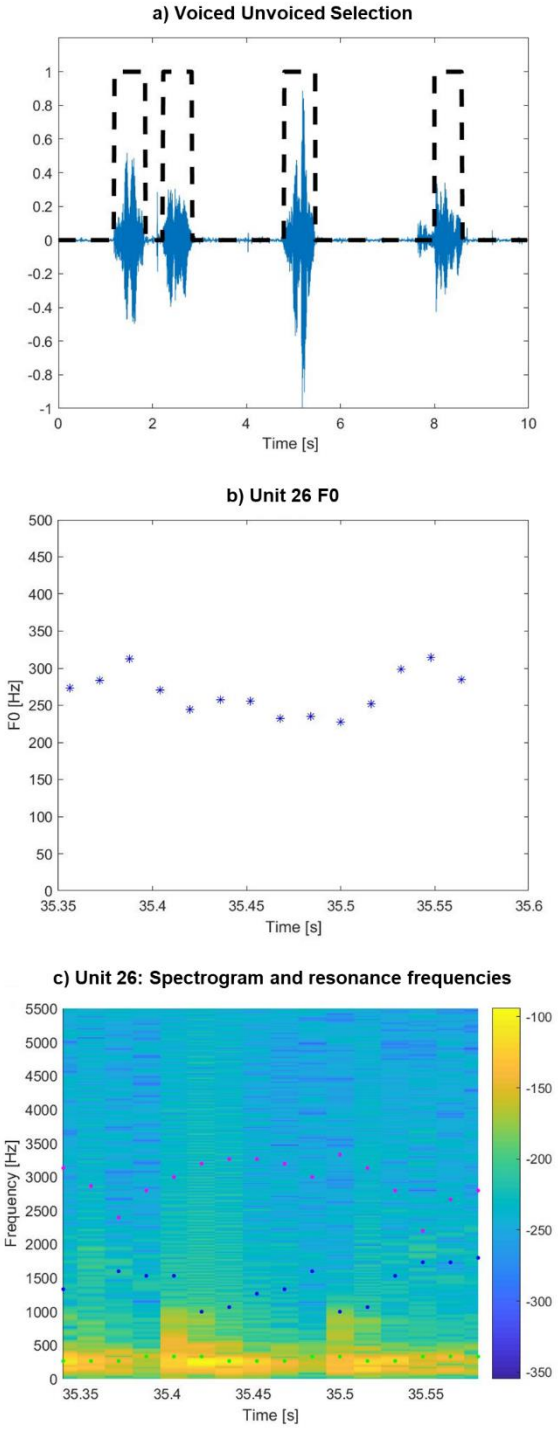

Fig. 1. a) Voiced/UnVoiced (V/UV) selection obtained with BioVoice. Four voiced units are found. They are indicated with a black dotted line, are found. b) ADHD subject combined presentation: F0 plot of the 26th voiced unit computed with BioVoice. F0= 196.363 ± 359.32 Hz. c) ADHD subject combined presentation: Spectrogram with formants F1-F3. F1= 377.109 ± 80.061 Hz; F2= 1636.77 ± 310.202 Hz; F3= 3068.713 ± 591.847 Hz

A second set of features describing the prosodic behavior in each word were estimated. Specifically, this set of features describes the F0 contour using an approach borrowed from Taylor's tilt intonational model [2]. The F0 contour was estimated using the Camacho's SWIPE algorithm [37], which compares the spectral content of the audio signal with a spectral template of a sawtooth waveform, mimicking the glottal source signal characteristics. Spectral features related to voice quality such as Long Term Average Spectrum (LTAS) were also estimated [2].

Finally, all the 49 features extracted were normalized with a z-score normalization, i.e., subtracting the mean and dividing by the standard deviation estimated across each of the five tasks.

2.4 Statistical analysis

Once all the features have been extracted, a classification analysis was performed to distinguish ADHD from ADHD+BD comorbidity and inattentive ADHD from combined ADHD. Supervised classifiers were trained using the above-mentioned features and the clinical label as *a-priori* information. Specifically, our algorithm is based on a Support Vector Machine model (SVM). The SVM finds the decision boundary that maximizes the margin separating the two classes of training data points. The choice of SVM was based on the poor sample-to-feature ratio of our dataset and on the possibility of implementing an effective embedded feature selection strategy. Particularly, we chose an embedded Recursive Features Elimination (SVM-RFE) strategy [38] to maximize the accuracy and to mitigate the risk of overfitting. SVM-RFE is an algorithm that combines SVMs with a backward variable selection of features and has a unique characteristic of incorporating an embedded correlation bias reduction. The final output of this algorithm is a ranked list with features ordered according to their relevance in separating the two classes through the SVM classifier [38].

The classifier was tested using the leave-one-subject-out (LOSO) cross validation to obtain a nearly unbiased estimation of the out-of-sample error. The LOSO technique involves using one observation as the validation set and the remaining observations (N-1) as the training set, where N is the total number of subjects. The process is repeated for the number of subjects present in the dataset, each time leaving out a different one to use as the single test case, until a predicted label is obtained for each subject.

The performance of the classification was estimated through the confusion matrix and complementary measures of classification such as accuracy, F1 score and Matthew's correlation coefficient (MCC) have been extracted and investigated. Accuracy is an index of the closeness between the value found and the real value and is defined as the sum of true positive and true negative results divided by the total positive and negative results. F1 score considers both precision p and recall r of the test to compute the score: precision is the number of correct positive results divide by the total number of the positive results returned by the classifier, and the recall measures the proportion of actual positives that are correctly identified as positives. MCC is a measure of the correlation between the observed and the ideal results, and an MCC higher than 0.5 reflects a strong correlation. MCC is not influenced both by the number of the observations and the target choice and, with F1 score, they can be used in case of unbalanced number of subjects between the two groups respect to the accuracy which could be biased by the group with more subjects.

The final classification was optimized by combining the marginal classification outcomes obtained from all the five different tasks by using a voting scheme. Specifically, the final classification was obtained after selecting the label that was predicted more frequently than a selected number of times N. For instance, in the ADHD versus ADHD+BD, the predicted label for a give subject was selected to be ADHD when at least N marginal classifiers were predicting such label. We highlight that this

means that ADHD+BD was selected as the predicted labels when at least 5-N+1 classifiers were predicting ADHD+BD condition. The threshold N was independently selected for each of the two classification analysis, by choosing the one that optimized F1 score.

3. Results

3.1 ADHD versus comorbidity ADHD+BD classification results

The highest accuracy, of about 0.88, was achieved with the emotions and action verbs conditions (task 4 e task 5) in which the algorithm selected a subset of 10 and 12 features respectively. For the verb actions condition, Fig. 2 shows the accuracy trend of the SVM-RFE learning algorithm as a function of selected features, which increase in number at each step (according to the RFE algorithm ranking). The maximum was reached with 10 features. Table 1 shows the accuracy and F1 score values for each subset of features until the maximum selected by the algorithm.

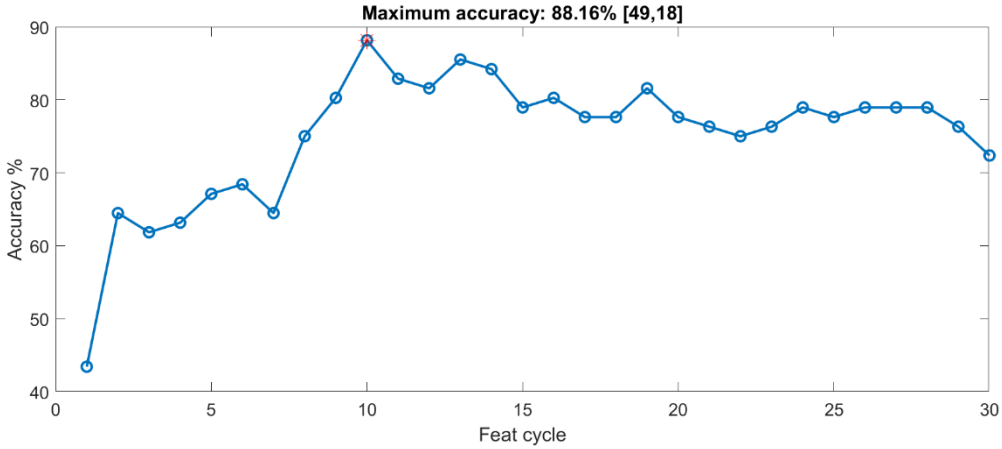

Fig. 2. Classification accuracy trend of ADHD versus ADHD+BD model as a function of the selected speech features. Verbs condition (task 5). Maximum accuracy of 88.16% combining the first 10 ranked features.

Features										Acc	F1
F_1 median										43,42	42,67
F_1 median	voiced duration									64,47	70,97
F_1 median	voiced duration	%voiced								61,84	68,82
F_1 median	voiced duration	%voiced	F_3 median							63,16	72,00
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean						67,11	75,73
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean	LTAS Ratio_Max					68,42	76,47
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean	LTAS Ratio_Max	Pause duration_std				64,47	73,27
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean	LTAS Ratio_Max	Pause duration_std	F_2 max			75,00	81,90
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean	LTAS Ratio_Max	Pause duration_std	F_2 max	Median F_0		80,26	85,44
F_1 median	voiced duration	%voiced	F_3 median	F_3 mean	LTAS Ratio_Max	Pause duration_std	F_2 max	Median F_0	F_1 max	88,16	91,59

Table 1 Accuracy and F1 score values for each subset of features until the maximum selected by the algorithm (task 5).

Among all the tasks, accuracy ranged from 0.79 to 0.88 across the five verbal fluency conditions, and F1 score from 0.85 to 0.92 (see Table 2). After applying the voting scheme procedure, accuracy and F1 score increased to 0.96 and 0.97, respectively. Among the ADHD subjects, a correct classification was achieved in 50 out of 52 patients and in 25 out of 26 subjects with ADHD+BD (see Table 3).

VFTs	Max ACC	F1 score	MCC
Task 1 - Free	0,79	0,85	0,52
Task 2 - Letter	0,81	0,86	0,54
Task 3 - Animals	0,81	0,87	0,54
Task 4 - Emotions	0,88	0,91	0,74
Task 5 - Verbs	0,88	0,92	0,73

Table 2 Statistical parameters obtained in each VFT condition.

		PREDICTED		
		ADHD	**ADHD + BD**	**Total**
ACTUAL	**ADHD**	50	2	52
	ADHD + BD	1	25	26
	Total	51	27	78

Table 3 Confusion matrix after majority voting

3.2 ADHD-inattentive versus ADHD-combined classification results

The highest accuracy, of about 0.96, was achieved with the free task (task 1) in which the algorithm selected a subset of 9 features. Fig. 3 shows an example of the accuracy trend of the SVM-RFE learning algorithm in the free task. The algorithm is a function of selected features, which increase in number at each step (according to the RFE algorithm ranking). An accuracy of 0.91 is achieved with the first 4 features of the subset selected by the algorithm. The maximum was reached with 9 features. Table 4 shows the accuracy and F1 score values for each subset of features until the maximum selected by the algorithm.

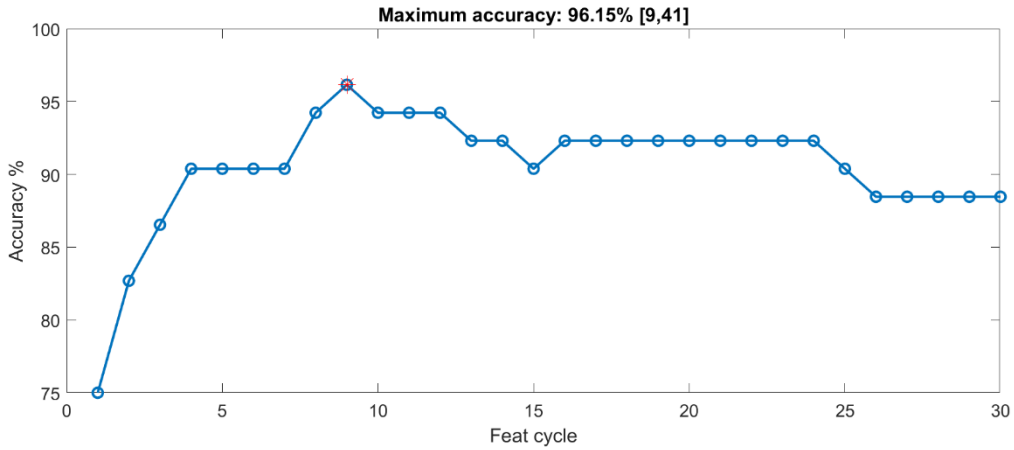

Fig. 3. Classification accuracy trend of the Inattentive versus Combined model as a function of the selected speech features. Free condition (task 1). Maximum accuracy of 96.15% combining the first 9 ranked features.

Features									Acc	F1
Tilt duration									75,00	55,17
Tilt duration	Number pauses								82,69	47,06
Tilt duration	Number pauses	F_3 mean							86,54	66,67
Tilt duration	Number pauses	F_3 mean	F_1 mean						90,38	73,68
Tilt duration	Number pauses	F_3 mean	F_1 mean	AbsNegSlope					90,38	73,68
Tilt duration	Number pauses	F_3 mean	F_1 mean	AbsNegSlope	F_3 median				90,38	73,68
Tilt duration	Number pauses	F_3 mean	F_1 mean	AbsNegSlope	F_3 median	Signal duration			90,38	73,68
Tilt duration	Number pauses	F_3 mean	F_1 mean	AbsNegSlope	F_3 median	Signal duration	Mad F_0		94,23	84,21
Tilt duration	Number pauses	F_3 mean	F_1 mean	AbsNegSlope	F_3 median	Signal duration	Mad F_0	F_3 max	96,15	90,00

Table 4 Accuracy and F1 score values for each subset of features until the maximum is selected by the algorithm (task1).

In general, accuracy ranged from 0.90 to 0.92 across the five verbal fluency conditions, and F1 score from 0.71 to 0.90 (see Table 5). After applying the majority voting a correct classification was achieved for all patients, 11 for the inattentive patients and 41 for the combined presentation of ADHD, increasing both accuracy and F1 score to 1 (see Table 6).

VFTs	Max ACC	F1 score	MCC
Task 1 - Free	0,96	0,90	0,88
Task 2 - Letter	0,92	0,78	0,76
Task 3 - Animals	0,94	0,84	0,82
Task 4 - Emotions	0,90	0,71	0,70
Task 5 - Verbs	0,94	0,84	0,82

Table 5 Statistical parameters obtained in each VFT condition.

		PREDICTED		
		INATTENTIVE	**COMBINED**	**Total**
ACTUAL	**INATTENTIVE**	11	0	11
	COMBINED	0	41	41
	Total	11	41	52

Table 6 Confusion matrix after majority voting

Table 7 shows the comparison between the different feature ranks for the two-classification models, ADHD versus ADHD+BD and inattentive versus combined ADHD, in its best performing configuration. Specifically, for each subset of features selected by the algorithm the table shows the features (both prosodic and spectral features) ranked for each task (from T1 to T5) in each classifier. Features are labeled by using their position in each subset in terms of their relevance.

			ADHD – ADHD+BD					Inattentive ADHD – combined ADHD				
	Features		T 1	T 2	T 3	T 4	T 5	T 1	T 2	T 3	T 4	T5
Prosodic features	F_0	Mean	8									
		Median										
		Std										
		Max										
		Min			5							3
		T_0min							6	7		
		T_0max	9		17						5	
	F_1	Mean						4	5		9	4
		Median					1			1	10	
		Std										
		Max	12		8	2	10					
		Min			11					5		
	F_2	Mean	7		18					2		
		Median			15				7		2	8
		Std										
		Max		2	9	3	8					
		Min										
	F_3	Mean			14		5	3				
		Median					4	6				
		Std	1									
		Max			10			9		6		
		Min							10			2
	Pauses	Number	11	8				2	9	3		
		Duration_mean				6						
		Duration_std					7					
	Signal duration		2		2	7		7				
	Voiced parts	Duration_mean		1								6
		Duration_std		9								
	Vibrato	Number Units	3	3								
		%Voiced					3					5
		Voiced_duration			7		2				7	
	Jitter					1						
	Median_F_0			10			9					
	Mad_F_0		4	4				8	2		1	9
	Tilt										8	
	Tilt_amplitude			7	12	5	4		3			7
	Tilt_duration							1				
	PosSlope				1							
	AbsNegSlope		10	6		4		5			4	
	SumDer											
	GlobalSlope		13		6					4	3	
	SumAbsDer		5		13							
Spectral features	LTAS_F_median			5	4							
	LTAS_A_median		6		16				4			
	LTAS_A_Max								8			1
	LTAS_Max_A_F								1			
	LTAS_slope											
	LTAS_Ratio_Max			11	3		6				6	
	LTAS_Ratio_Median											

Table 7 Comparison of the two classification models (ADHD versus ADHD+BD and inattentive versus combined ADHD) performance in terms of different features selected (both prosodic and spectral features). For each task (from T1 to T5)

features are labeled with a number corresponding to their position in each subset of features selected by the algorithm in terms of their relevance. Blank cells indicate that that feature has not been selected for the corresponding classification.

4. Discussion

The results here obtained indicate that speech signals acquired from VFTs contain relevant information about ADHD type both when a differential classification of inattentive versus combined subtype is needed and when bipolar comorbidity should be highlighted.

Our results are obtained performing an automatic selection of the combination of speech features that provides the best classification results. Noticeably, the final classification result was obtained by combining the classifiers using the normalized features of each task. This allowed to improve the classification, using a voting scheme across the classifiers that was also designed to optimize the results provided by the model.

Given the low number of samples and the high number of features, risk of overfitting was faced in the model. For this reason, SVM with a proper recursive feature elimination scheme was adopted. Specifically, RFE reduces the problem dimensionality by selecting the features which maximize the accuracy, thus mitigating the risk of overfitting [38]. The analysis of classification accuracy, as a function of number of features indicates that a lower number of features could be selected to further reduce the overfitting risk.

One of the main findings of this study is the relevance of speech feature normalization adopted, that was designed to exploit the within-subject different behavioural and cognitive dynamics across different VFT conditions. In fact, by applying a z-score like transformation of each subject feature, the classification is based on the feature changes that are observed in each subject due to a specific task execution with respect to the average feature value across all the conditions. This normalization allows to remove the bias due to subject specific speech characteristics, as those related to gender and identity. Furthermore, it will not be feasible to evaluate possible absolute speech feature differences among different groups of patients.

It is interesting to note that the two VFT that better distinguished ADHD and ADHD+BD comorbidity are those that measure emotion dysregulation and embodied cognition specifically [40]. This suggests that the emotion eliciting nature of these tasks might be involved in specific speech patterns that contribute to the distinction of the two disorders. Regarding ADHD subtypes, the free task allowed to better distinguish inattentive and combined presentations of ADHD. This is in line with previous research by our group which found, using a semantic analysis approach, that people with the combined presentation switched more often from one concept to another, akin to racing thoughts [39]. Our study therefore suggests that both emotion processing and emotion dysregulation, on the one hand, and racing thoughts on the other hand, manifest also in speech. Moreover, they are relevant for the distinction between ADHD subtypes and comorbid conditions. An identification of the most informative VFTs could lead to optimize the classification results using a voting procedure based on VFTs weighted for their relevance. Further studies could try to acquire a larger number of subjects in order to include ADHD hyperactive impulsive subtype, thus exploring the possibility of classifying the three subtypes in ADHD.

Nonetheless, looking at the features that were selected by the RFE scheme, as the most important for a correct classification of each subject, it is possible to gain insight into the speech dynamics that are more relevant for distinguishing the different subgroups. Among the ADHD group, prosodic features were more frequently selected by the algorithm to classify inattentive and combined ADHD. Specifically, formants and number of pauses seem to add the most significant information for the distinction of the two groups. Furthermore, spectral features play a key role in the classification

of ADHD and ADHD+BD comorbidity reflecting the relevance of these features to characterize Bipolar Disorder [2].

The unbalanced number of participants across the different classes might pose some issues in the classifier design, as a potential bias towards the majority class. We have to point out that the use of performance metrics, i.e., F1 score ad MCC, that are robust with respect to imbalanced data, partially mitigate this problem.

Although promising, our results have been found on a low number of subjects, so we have to stress that caution must be taken in generalizing our findings. Although the use of leave-one- subject-out cross validation and recursive feature elimination allow to reduce some issues related to selection bias and overfitting, we have to point out that a larger dataset would have allowed a more robust generalization including a test set. to adopt a train-validation and testing strategy, by splitting the available data in sufficiently large independent samples. This would allow to explore other validating scheme as a n-fold cross validation, thus optimizing the trade-off between model variance and bias [40]. We have to point out that the limited availability of studies exploiting a large dataset of high quality recordings, as those that can be obtained in a clinical setting, testifies the difficulty of acquiring such kind of data. In fact, data quality and reliability is a relevant factor that has to be taken into account to design validated decision support systems based on speech analysis [41], [42]. However, several researchers are exploring the possibility of using large speech corpora acquired in ecological scenarios, with promising results in specific applications, such as [43], [44]. In this study, we are exploiting a structured speech task, i.e., the VFT, that we delivered in controlled conditions, also allowing to obtain a high quality speech dataset. We believe that the obtained results are leveraging upon the cognitive aspects of such task execution. Further experiments should be designed to verify whether such task could be designed to be performed at home with widely available recording devices, such as smartphones. Moreover, it would also be interesting to include recordings from different laboratories/clinics and verify the robustness of our findings with respect to inter-observer and inter-laboratory variability. Such a research activity could lead to a more robust approach for the classification of ADHD could lead to an improvement of information for the clinicians and help them to identify the correct diagnosis.

However, supervised classifiers exploit the diagnosis by the physician even if diagnosis might be also prone to classification error. A future application of unsupervised approaches could be less biased by the a-priori information by the clinician. Physicians should be deeply involved in the critical analysis of automatic or semi-automatic methods results. This could allow identifying possible specific clinical characteristics or pushing the researcher to further explore the speech features of the subjects that were classified both in agreement and disagreement with the clinicians.

5. Conclusion

In this work, a classification of ADHD subtypes and an investigation of the ADHD+BD comorbidity has been carried out exploiting speech signals acquired using VFT. Among the ADHD group, the goal was to identify inattentive and combined subtypes, since the latter need different clinical treatment. Our model allowed to obtain good classification results and showed a greater ability to classify patients according to clinician diagnosis. Regarding the comorbidity, results highlight that significant information is carried by speech features which could be a promising support for the clinical diagnosis of the ADHD+BD comorbidity.

Studies with larger samples are needed to further investigate the relationship between speech features and classification results in ADHD and to mitigate a possible risk of overfitting. Future developments will concern the critical discussion of classification performances of the approach with

the clinicians and the possible added value of supervised learning machine classification. Finally, it could be particularly relevant to determine which VFT could better provides the most relevant information, in terms of speech features, to aid the identification of the ADHD subtypes and the comorbid pattern.

Acknowledgements

Work partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence).

6. References

[1] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun*, vol. 40, no. 1–2, pp. 227–256, Apr. 2003, doi: 10.1016/S0167-6393(02)00084-5.

[2] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, and E. P. Scilingo, "Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients," *Biomed Signal Process Control*, vol. 17, pp. 29–37, Mar. 2015, doi: 10.1016/j.bspc.2014.10.011.

[3] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain Cogn*, vol. 56, no. 1, pp. 30–35, Oct. 2004, doi: 10.1016/j.bandc.2004.05.003.

[4] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *J Acoust Soc Am*, vol. 83, no. 2, pp. 716–728, Feb. 1988, doi: 10.1121/1.396114.

[5] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies.," *Journal of Applied Psychology*, vol. 88, no. 5, pp. 879–903, 2003, doi: 10.1037/0021-9010.88.5.879.

[6] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig Otolaryngol*, vol. 5, no. 1, pp. 96–116, Feb. 2020, doi: 10.1002/lio2.354.

[7] K. Dikaios, S. Rempel, S. H. Dumpala, S. Oore, M. Kiefte, and R. Uher, "Applications of Speech Analysis in Psychiatry," *Harv Rev Psychiatry*, vol. 31, no. 1, pp. 1–13, Jan. 2023, doi: 10.1097/HRP.0000000000000356.

[8] M. Faurholt-Jepsen *et al.*, "Voice analysis as an objective state marker in bipolar disorder," *Transl Psychiatry*, vol. 6, no. 7, pp. e856–e856, Jul. 2016, doi: 10.1038/tp.2016.123.

[9] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit Signal Process*, vol. 77, pp. 207–221, Jun. 2018, doi: 10.1016/j.dsp.2017.07.004.

[10] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112–124, Mar. 2015, doi: 10.1016/j.dadm.2014.11.012.

[11] B. Sonawane and P. Sharma, “Speech-based solution to Parkinson’s disease management,” *Multimed Tools Appl*, vol. 80, no. 19, pp. 29437–29451, Aug. 2021, doi: 10.1007/s11042-021-11061-1.

[12] T. R. Insel, “The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry,” *American Journal of Psychiatry*, vol. 171, no. 4, pp. 395–397, Apr. 2014, doi: 10.1176/appi.ajp.2014.14020138.

[13] L. Weiner, A. Guidi, N. Doignon-Camus, A. Giersch, G. Bertschy, and N. Vanello, “Vocal features obtained through automated methods in verbal fluency tasks can aid the identification of mixed episodes in bipolar disorder,” *Transl Psychiatry*, vol. 11, no. 1, p. 415, Dec. 2021, doi: 10.1038/s41398-021-01535-z.

[14] J. Fayyad *et al.*, “Cross-national prevalence and correlates of adult attention-deficit hyperactivity disorder,” *British Journal of Psychiatry*, vol. 190, no. 5, pp. 402–409, May 2007, doi: 10.1192/bjp.bp.106.034389.

[15] W. Retz, P. Retz-Junginger, J. Thome, and M. Rösler, “Pharmacological treatment of adult ADHD in Europe,” *The World Journal of Biological Psychiatry*, vol. 12, no. sup1, pp. 89–94, Sep. 2011, doi: 10.3109/15622975.2011.603229.

[16] V. Simon, P. Czobor, S. Bálint, Á. Mészáros, and I. Bitter, “Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis,” *British Journal of Psychiatry*, vol. 194, no. 3, pp. 204–211, Mar. 2009, doi: 10.1192/bjp.bp.107.048827.

[17] A. Halmøy, H. Halleland, M. Dramsdahl, P. Bergsholm, O. B. Fasmer, and J. Haavik, “Bipolar Symptoms in Adult Attention-Deficit/Hyperactivity Disorder,” *J Clin Psychiatry*, vol. 71, no. 01, pp. 48–57, Jan. 2010, doi: 10.4088/JCP.08m04722ora.

[18] C. Skirrow, G. M. Hosang, A. E. Farmer, and P. Asherson, “An update on the debated association between ADHD and bipolar disorder across the lifespan,” *J Affect Disord*, vol. 141, no. 2–3, pp. 143–159, Dec. 2012, doi: 10.1016/j.jad.2012.04.003.

[19] R. C. Kessler *et al.*, “The Prevalence and Correlates of Adult ADHD in the United States: Results From the National Comorbidity Survey Replication,” *American Journal of Psychiatry*, vol. 163, no. 4, pp. 716–723, Apr. 2006, doi: 10.1176/ajp.2006.163.4.716.

[20] A. P. Wingo and S. N. Ghaemi, “A Systematic Review of Rates and Diagnostic Validity of Comorbid Adult Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder,” *J Clin Psychiatry*, vol. 68, no. 11, pp. 1776–1784, Nov. 2007, doi: 10.4088/JCP.v68n1118.

[21] P. Asherson, “Clinical assessment and treatment of attention deficit hyperactivity disorder in adults,” *Expert Rev Neurother*, vol. 5, no. 4, pp. 525–539, Jul. 2005, doi: 10.1586/14737175.5.4.525.

[22] L. Weiner, N. Perroud, and S. Weibel, “ Attention Deficit Hyperactivity Disorder And Borderline Personality Disorder In Adults: A Review Of Their Links And Risks ,” *Neuropsychiatr Dis Treat*, vol. Volume 15, pp. 3115–3129, Nov. 2019, doi: 10.2147/NDT.S192871.

[23] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Commun*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006, doi: 10.1016/j.specom.2006.04.003.

[24] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” *Int J Speech Technol*, vol. 15, no. 2, pp. 99–117, Jun. 2012, doi: 10.1007/s10772-011-9125-1.

[25] Scherer KR., " Vocal correlates of emotional arousal and affective disturbance," in *Handbook of social psychophysiology*, H. Wagner and A. Manstead, Eds. Oxford, England, 1989, pp. 165–197.

[26] A. S. Cohen, S. Lee Hong, and A. Guevara, "Understanding emotional expression using prosodic analysis of natural speech: Refining the methodology," *J Behav Ther Exp Psychiatry*, vol. 41, no. 2, pp. 150–157, Jun. 2010, doi: 10.1016/j.jbtep.2009.11.008.

[27] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun*, vol. 71, pp. 10–49, Jul. 2015, doi: 10.1016/j.specom.2015.03.004.

[28] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech," *IEEE Trans Biomed Eng*, vol. 55, no. 1, pp. 96–107, Jan. 2008, doi: 10.1109/TBME.2007.900562.

[29] G. G. von Polier *et al.*, "Predicting adult Attention Deficit Hyperactivity Disorder (ADHD) using vocal acoustic features", doi: http://dx.doi.org/10.1101/2021.03.18.21253108.

[30] Z. Breznitz, "The Speech and Vocalization Patterns of Boys with ADHD Compared with Boys with Dyslexia and Boys Without Learning Disabilities," *J Genet Psychol*, vol. 164, no. 4, pp. 425–452, Dec. 2003, doi: 10.1080/00221320309597888.

[31] J. D. Henry and J. R. Crawford, "A Meta-Analytic Review of Verbal Fluency Performance in Patients With Traumatic Brain Injury.," *Neuropsychology*, vol. 18, no. 4, pp. 621–628, Oct. 2004, doi: 10.1037/0894-4105.18.4.621.

[32] M. Moscovitch, "Memory and Working with Memory: Evaluation of a Component Process Model and Comparisons with Other Models," in *Memory Systems 1994*, The MIT Press, 1994. doi: 10.7551/mitpress/4545.003.0010.

[33] C. A. Abeare, S. Freund, K. Kaploun, T. McAuley, and C. Dumitrescu, "The Emotion Word Fluency Test (EWFT): Initial psychometric, validation, and physiological evidence in young adults," *J Clin Exp Neuropsychol*, vol. 39, no. 8, pp. 738–752, Sep. 2017, doi: 10.1080/13803395.2016.1259396.

[34] Y. , S. B. & C. H. Joanette, "Protocole Montréal d'Évaluation de la Communication." Ortho Édition, Isbergues, France , 2004.

[35] D. Cardebat, B. Doyon, M. Puel, P. Goulet, and Y. Joanette, "[Formal and semantic lexical evocation in normal subjects. Performance and dynamics of production as a function of sex, age and educational level].," *Acta Neurol Belg*, vol. 90, no. 4, pp. 207–17, 1990.

[36] M. S. Morelli, S. Orlandi, and C. Manfredi, "BioVoice: A multipurpose tool for voice analysis," *Biomed Signal Process Control*, vol. 64, p. 102302, Feb. 2021, doi: 10.1016/j.bspc.2020.102302.

[37] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J Acoust Soc Am*, vol. 124, no. 3, pp. 1638–1652, Sep. 2008, doi: 10.1121/1.2951592.

[37] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sens Actuators B Chem*, vol. 212, pp. 353–363, Jun. 2015, doi: 10.1016/j.snb.2015.02.025.

[39] E. Martz, S. Weibel, and L. Weiner, "An overactive mind: Investigating racing thoughts in ADHD, hypomania and comorbid ADHD and bipolar disorder via verbal

fluency tasks," *J Affect Disord*, vol. 300, pp. 226–234, Mar. 2022, doi: 10.1016/j.jad.2021.12.060.

[40] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J Econom*, vol. 187, no. 1, pp. 95–112, Jul. 2015, doi: 10.1016/j.jeconom.2015.02.006.

[41] K. Huckvale, S. Venkatesh, and H. Christensen, "Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety," *NPJ Digit Med*, vol. 2, no. 1, p. 88, Sep. 2019, doi: 10.1038/s41746-019-0166-1.

[42] J. Robin, J. E. Harrison, L. D. Kaufman, F. Rudzicz, W. Simpson, and M. Yancheva, "Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations," *Digit Biomark*, vol. 4, no. 3, pp. 99–108, Oct. 2020, doi: 10.1159/000510820.

[43] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive Mob Comput*, vol. 31, pp. 50–66, Sep. 2016, doi: 10.1016/j.pmcj.2016.01.008.

[44] J. Gideon, E. M. Provost, and M. McInnis, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 2359–2363. doi: 10.1109/ICASSP.2016.7472099.