HAL
open science

Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations

Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, Björn Schuller

To cite this version:

Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, et al.. Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations. 7th International Digital Health Conference, Jul 2017, Londres, United Kingdom. pp.53-57, 10.1145/3079452.3079492 . hal-02080880

HAL Id: hal-02080880

https://hal.science/hal-02080880v1

Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations

Jun Deng
Chair of Complex & Intelligent Systems,
University of Passau
jun.deng@uni-passau.de

Nicholas Cummins
Chair of Complex & Intelligent Systems,
University of Passau
nicholas.cummins@uni-passau.de

Maximilian Schmitt
Chair of Complex & Intelligent Systems,
University of Passau
maximilian.schmitt@uni-passau.de

Kun Qian
Institute for Human-Machine Communication,
TUM
andykun.qian@tum.de

Fabien Ringeval
Laboratoire d'Informatique de Grenoble,
Université Grenoble Alpes
fabien.ringeval@imag.fr

Björn Schuller
DoC, Imperial College London
& CIS, University of Passau
schuller@IEEE.org

ABSTRACT

Machine learning paradigms based on child vocalisations show great promise as an objective marker of developmental disorders such as Autism. In conventional detection systems, hand-crafted acoustic features are usually fed into a discriminative classifier (e. g., Support Vector Machines); however it is well known that the accuracy and robustness of such a system is limited by the size of the associated training data. This paper explores, for the first time, the use of feature representations learnt using a deep Generative Adversarial Network (GAN) for classifying children's speech affected by developmental disorders. A comparative evaluation of our proposed system with different acoustic feature sets is performed on the Child Pathological and Emotional Speech database. Key experimental results presented demonstrate that GAN based methods exhibit competitive performance with the conventional paradigms in terms of the unweighted average recall metric.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; **Health informatics**; • **Computing methodologies** → *Neural networks*;

KEYWORDS

Autism Spectrum Condition; automatic diagnosis; generative adversarial networks; representation learning

ACM Reference format:
Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller. 2017. Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations. In *Proceedings of DH '17, London, United Kingdom, July 2–5, 2017,* 5 pages.
DOI: http://dx.doi.org/10.1145/3079452.3079492

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DH '17, London, United Kingdom
© 2017 ACM. 978-1-4503-5249-9/17/07…$15.00.
DOI: http://dx.doi.org/10.1145/3079452.3079492

1 INTRODUCTION

In recent years, there has been a notable increase in research focused on identifying biological and behavioural markers to aid the early detection of *Autism Spectrum Conditions* (ASC). ASC is a group of conditions characterised by social, language and communications impairments as well as, repetitive stereotyped behaviours [1]. Early diagnosis is important for increased positive outcomes from therapy, as well as for reducing parental stress [6, 15].

Autism is known to manifest in different ways in the speech of children and adults [4, 15, 19]. Commonly reported linguistic peculiarities include echolalia, out of context phrasing, as well as pronoun and role reversal [5, 15, 21]. However, language skills in autism show several varying subtypes within the spectrum [13, 14]. Thus, linguistic based markers may not be reliable for the automatic diagnosis of ASC. Since abnormal prosody has also been reported as a core marker of ASC [12], paralinguistic cues appear, on the other hand, better suited for the automatic detection. Suprasegmental acoustic features relating to articulation, loudness, pitch, and rhythm have indeed shown promising results for children's speech [4, 19, 21, 25]. These acoustic features have also been successfully used in speech-based interaction systems for improving social skills of children suffering from ASC [18, 20].

Investigations have been undertaken with machine learning paradigms relying on acoustic and prosodic feature sets to automatically detect autism [24, 28, 32]. Whereas results show that high levels of accuracy can be achieved for a task like discriminating typically developing children from children with ASC, performance obtained by such systems have been evaluated on rather small datasets, and may lead to potential confounds [3]. The small size of currently available ASC related datasets represents a major obstacle in the development of robust models which are sufficiently reliable for clinical practice [31].

Whilst the collection of more data is the straightforward solution for this issue of data scarcity, the high costs associated with obtaining clinical data, from a population that further includes children, limits the practicality of this approach. Another option is the augmentation of the training data by artificially generated samples [31]. The potential of this approach has already been shown for speech-based emotion detection systems [17, 30].

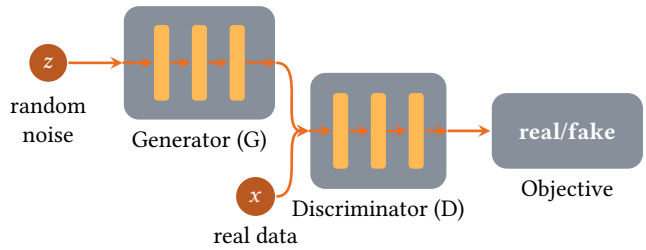

Figure 1: Diagram of a Generative Adversarial Networks (GANs). A GAN basically involves a generator and a discriminator competing against each other in a zero-sum game framework. The generator takes random noise as input and tries to generate data in the hope of fooling the discriminator. Simultaneously, the discriminator tries to classify samples as either coming from the training data or the generated samples.

Inspired by the recent success in representation learning associated with advances in deep learning [7, 8, 16], we propose to learn feature representations by leveraging deep *Generative Adversarial Networks* (GANs) for automatic diagnosis of ASC in children's voices. The deep GANs, a recently proposed unsupervised learning algorithm [11, 23, 26], are used as a means of learning intrinsic representations from unlabelled complex speech data. The resulting GAN representations are input to a traditional classifier in an attempt to facilitate the learning process.

Whilst predominantly used in image processing, the use of GAN's to learn invariant feature representations has started to be explored in speech processing tasks such as speech enhancement [22] and, automatic speech recognition (ASR) in noisy conditions [33]. To the best of the author's knowledge this is the first time it has been explored in computational paralinguistics.

The rest of this paper is laid out as follows; Section 2 introduces the GAN based classification framework; Section 3 sets out the key experimental setting and present the results; a succinct conclusion and future work plans are given in Section 4.

2 METHODOLOGY

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs) have recently attracted considerable attention in the field of deep learning [11, 23, 26]. A GAN, as illustrated in Figure 1, consists of two competing networks in a zero-sum game framework. A generator network performs a data generating process, which takes random noise sampled from a pre-defined distribution (e. g., a uniform distribution or a unit Gaussian distribution) and maps them to a given true training data distribution. A second discriminator network receives samples from the generator and the training data, and then is forced to predict samples as either coming from the training data or the generated samples. The two networks pay a MinMax or zero-sum game, where the discriminator is learning to differentiate between the two sources as accurately as possible. The generator is simultaneously learning to fool the discriminator by producing realistic samples. By the end of the 'game', the generator is able to perfectly synthesise the training data, and the discriminator is unable to find a difference between 'fake' samples synthesised by the generator and real samples from the dataset.

Mathematically, in order to learn the generator's distribution p_g over data $\boldsymbol{x}$, we note $p_{\boldsymbol{z}}(\boldsymbol{z})$ as a prior on input noise variables, $G(\boldsymbol{z};\theta_g)$ a mapping to the data space, and G a differentiable function represented by a deep neural network with parameters θ_g. Similarly, we note $D(\boldsymbol{x};\theta_d)$ as a second deep neural network with parameters θ_d. $D(\boldsymbol{x})$ indicates the probability of $\boldsymbol{x}$ belonging to the data rather than p_g. Therefore, a loss function $V(G,D)$ of the two-player MinMax game [11] is defined as:

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] \\ &+ \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))],\end{aligned} \tag{1}$$

where $\mathbb{E}$ denotes the expected value. As shown in [11], for a fixed generator G, the optimal discriminator D is:

$$D^*_G(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}. \tag{2}$$

It is noteworthy that this *minimax* rule has a global optimum for $p_g = p_{\text{data}}$, i. e., the generative model perfectly replicating the data distribution [11].

2.2 Generative Adversarial Networks based Automatic Diagnosis System

Drawing inspiration from the extensive success of deep representation learning for classification tasks with small data [2, 7, 8, 16], our proposed automatic diagnosis system, uses features learnt from unlabelled data generated by a GAN, instead of directly using handcrafted acoustic features. The resulting features are used as an input to a supervised classifier (in this work, Support Vector Machines (SVM)), that performs classification modelling on the generated data as well as making predictions on the (real) test utterances.

As explained in the previous section (see Section 2.1), there are two different types of neural networks in a typical GAN. To learn meaningful representations, only GAN discriminators, which take *acoustic features* as input, can act as a non-linear feature extractor in our system: the output of an intermediate layer can be treated as a representation of the original input data. As a result, the input data are mapped to a feature space through a non-linear feature mapping, which is learnt by exploring non-linear structures of the

Table 1: ComParE acoustic feature set: 65 low-level descriptors (LLD).

4 energy related LLD	**Group**
RMS energy, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
55 spectral LLD	**Group**
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 voicing related LLD	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice qual.
log. HNR, jitter (local and δ), shimmer (local)	Voice qual.

data. In sum, our proposed system consists of three main modules: acoustic features extraction, a GAN, and a linear SVM.

2.2.1 Acoustic Features. The aim of acoustic feature extraction is to provide compact and discriminant representations of the speech signal on the basis of expert knowledge. For transparency and reproducibility, we exploited the openSMILE feature extraction toolkit [10] to extract two widely used audio feature sets in the field of computational paralinguistic; the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) and the large-scale *Interspeech 2013 Computational Paralinguistics Challenge* feature set ComParE [10]. Both sets have been successfully utilised in the field of affective computing [8], and recently investigated for the automatic diagnosis of ASC in children's voices [25, 28].

eGeMAPS is a *knowledge driven* data set that exploits the first two statistical moments (*mean* and *coefficient of variation*) to capture the distribution of *low-level descriptors* (LLDs) describing spectral, cepstral, prosodic and voice quality information, creating an 88 dimensional acoustical representation of an utterance. It was specifically designed by a small group of experts to be a basic standard acoustic parameter set for voice analysis tasks including paralinguistic and clinical speech analysis. For full details the reader is referred to [9]. ComParE, on the other hand, is a large-scale brute forced acoustic feature set which contains 6 373 features representing prosodic, spectral, cepstal and voice quality LLDs. A detailed list of all LLDs for ComParE is given in Table 1. For full details on ComParE the reader is referred to [10].

2.2.2 Generative Adversarial Network. In our GAN model, the generator consists of a deep feed-forward neural network with one input layer of 100 neural units, three hidden layers of 256 hidden units, and one output layer. As already mentioned in Section 2.1, a uniform distribution is selected to provide random noise samples as input to the generator. Note that the number of neural units in the output layer is dependent on the selected acoustic feature sets. Similarly, the discriminator is a deep feed-forward neural network with four hidden layers of 256 hidden units which lead into a sigmoid activation function which outputs the probabilities of whether the input utterance is real or artificial. The binary cross entropy is chosen as the objective function of the discriminator. The *tanh* activation function is adopted for the output layer of the generator. For the remaining layers of these models, *LeakyReLU* activations [34] and batch normalisation are employed to stabilise the training. Finally, we train an SVM with a linear kernel using the representations from a hidden layer of the discriminator.

3 EXPERIMENTS

3.1 ASC Corpus

We exploited the *Child Pathological & Emotional Speech Database* (CPESD) [24, 25] to conduct the empirical evaluations of our diagnosis system. This dataset includes spontaneous speech recordings inducing three emotion categories of valence (positive, neutral, and negative) from 34 monolingual children. All participants were recruited in two university departments of child and adolescent psychiatry located in Paris, France. All children were equipped with communicative verbal skills, and diagnosed with one of the following conditions: *autism disorders* (AD; 11 children), *pervasive developmental disorders not otherwise specified* (PDD-NOS; 10 children), or *specific language impairment* (SLI; 13 children), according to DSM–IV criteria [1]. All patients were matched for age, sex, academic grades, and lexical abilities. For the control group, 68 *typically developing* (TD) children from elementary schools were recruited. Participants were also matched for age and sex (two TD for one patient). Their teacher was asked to fill in a questionnaire to exclude children with learning disorders. In total, almost 12 hours of audio were recorded: 7 h 38 min for TD children, 1 h 35 min for children with AD, 1 h 12 min for children with PDD-NOS, and 1 h 56 min for children with SLI. Those recordings were then segmented into utterances, providing in total 6 380 segmented utterances from 102 children. The corpus was further divided into partitions for training (3 692 utterances, approximately 60 % of data), validation (1 281 utterances, approximately 20 % of data) and test (1 407 utterances, approximately 20 % of data). To ensure speaker identity is not a confounding factor, this partitioning was done in an completely child independent fashion i. e., all utterance from any given child are contained completely within one partition. All parameters of the models were optimised on the validation set, whereas the test partition is solely used for the purpose of performance evaluation on *unseen* children.

3.2 Key Experimental Settings

For the GAN models, the training data was scaled to the range of the *tanh* activation function [-1, 1]. The models were trained with *stochastic gradient descent* (SGD) with a mini-batch size of 128. In order to gain insights into the optimal representation, we investigate the outputs from each hidden layer of the discriminator network.

Note that the present study focusses on the recognition of diagnosis condition – as provided by clinicians – from speech recordings of AD, PDD-NOS, SLI, and TD children, which leads to a 4-way imbalanced classification task. Hence, performance is evaluated by *unweighted average recall* (UAR), which is suitable for imbalanced

Table 2: Results in terms of UAR (%) for the 4-way speech-based diagnosis ASC task on CPESD with the eGeMAPS acoustic feature set. l_o indicates the index of the selected hidden layer used to compute the GAN representation. Maximum test UAR is highlighted in bold. Significant results (p-value < 0.05, one-sided z-test) are marked with an asterisk.

Method	Validation	Test
Linear SVM	41.06	39.91
SVM (RBF)	40.83	39.09
MLP	40.77	40.97
GAN (l_o = 1)+SVM	39.33	43.13*
GAN (l_o = 2)+SVM	40.92	42.76
GAN (l_o = 3)+SVM	40.59	**44.06***
GAN (l_o = 4)+SVM	39.35	43.29*

Table 3: Confusion matrix of the best system with eGeMAPS on the CPESD test set. Abbreviations: *TD* typically developing; *NOS* pervasive developmental disorders not-otherwise specified; *SLI* specific language impairment; *AD* autism disorders.

		Predicted Labels			
		TD	NOS	SLI	AD
True Labels	TD	849	28	25	26
	NOS	27	19	12	31
	SLI	70	37	114	20
	AD	36	9	80	24

classes. In addition, significance tests are conducted by computing a *one-sided z-test* in order to compare two different diagnosis systems.

For comparison purposes, three representative methods including a linear SVM, an SVM with the Radial Basis Function (RBF) kernel, and a Multi-Layer Perceptron (MLP) with four hidden layers, which are fed with the eGeMAPS and ComParE feature set, respectively, serve as baseline systems. In these approaches, all features are standardised w. r. t. the mean and the standard deviation of each feature derived from the training set. Note that the complexity parameter of the SVM was optimised w. r. t. the highest UAR on the validation set.

3.3 Results

For eGeMAPS, we first observe that all systems achieved promising performances far above the chance level UAR of 25.00 % (cf. Table 2). We also observed that the outputs from each of the hidden layers of our proposed system achieved notable increases in performance on the test data. The outputs of the third hidden layer achieved the best test UAR of 44.06 %, which is a relative increase of 10.40 % over the Linear SVM baseline; the confusion matrix for this system is presented in Table 3. Further, except for the second layer, our proposed system with the GAN representations and SVM significantly outperforms the Linear SVM baseline system.

Table 4: Results in terms of UAR (%) for the 4-way speech-based diagnosis ASC task on CPESD with ComParE. l_o indicates the index of the selected hidden layer used to compute the GAN representation. Maximum test UAR is highlighted in bold. Significant results (p-value < 0.05, one-sided z-test) are marked with an asterisk.

Method	Validation	Test
Linear SVM	64.92	42.83
SVM (RBF)	50.07	40.40
MLP	41.67	37.61
GAN (l_o = 1)+SVM	51.09	**46.93***
GAN (l_o = 2)+SVM	50.64	44.27
GAN (l_o = 3)+SVM	51.38	45.29
GAN (l_o = 4)+SVM	47.06	43.83

Table 5: Confusion matrix of the best system with the ComParE feature on the CPESD test set.Abbreviations: *TD* typically developing; *NOS* pervasive developmental disorders not-otherwise specified; *SLI* specific language impairment; *AD* autism disorders.

		Predicted Labels			
		TD	NOS	SLI	AD
True Labels	TD	830	54	25	19
	NOS	46	24	5	14
	SLI	54	20	146	21
	AD	47	9	77	16

Similarly when using ComParE features (cf. Table 4), we observed a test set performance increase for the GAN representations when compared to the baseline systems. Furthermore all GAN ComParE representations achieve stronger performances than their corresponding eGeMAPS representations. Although, it is worth noting that the outputs computed from the bottom hidden layer give the best ComParE test UAR of 46.93 %, which is different from the eGeMAPS case where the top layers were observed to be more suitable. The confusion matrix obtained by the strongest ComParE system is presented in Table 5.

4 CONCLUSIONS AND OUTLOOK

Whilst research into using biological markers such as speech to aid the diagnosis of autism spectrum conditions is steadily increasing, the relative small size of the associated corpora is an important limiting factor in the creation of robust models with clinical utility. In this regards, this paper explored whether a deep *Generative Adversarial Network* (GAN) could serve to generate additional data representations that are suitable for recognising children with ASC

from TD. Results reported on the CPESD dataset indicate that a system trained with our generated features could achieve comparable accuracies with a similar system trained using state-of-art-feature representations. This result is, to our knowledge, the first of its kind, not just in speech-based autism detection, but in Computational Paralinguistics in general, and shows great promise for many tasks where data scarcity is an issue in general.

Future work will include testing our GAN based system in range of other health-based tasks such as depression detection. We will also test the GAN system capability for generating other commonly used behavioural signals such as video descriptors or physiological features such as electrocardiogram and electrodermal activity representations. Furthermore, we will compare and combine GAN-based feature learning with other unsupervised representation learning techniques, such as *bag-of-words* [27, 29].

5 COMPETING INTERESTS

The authors have declared that no competing interests exist.

6 ACKNOWLEDGEMENTS

This work was supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

REFERENCES

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. Washington, D.C., 4th edition, 2000.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[3] D. Bone, T. Chaspari, K. Audkhasi, J. Gibson, A. Tsiartas, M. V. Segbroeck, M. Li, S. Lee, and S. Narayanan. Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In ISCA, editor, *Proceedings of INTERSPEECH*, pages 182–186, Lyon, France, 2013.

[4] D. Bone, C.-C. Lee, M. Black, M. Williams, S. Lee, P. Levitt, and S. Narayanan. The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research*, 57(4):1162–1177, 2014.

[5] M. Carpenter, M. Tomasello, and T. Striano. Role reversal imitation and language in typically developing infants and children with autism. *Infancy*, 8(3):253–278, 2005.

[6] N. Davis and A. Carter. Parenting stress in mothers and fathers of toddlers with autism spectrum disorders: Associations with child characteristics. *Journal of Autism and Developmental Disorders*, 38(7):1278–1291, 2008.

[7] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

[8] J. Deng, Z. Zhang, E. Marchi, and B. Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proceedings 5th International Conference on Affective Computing and Intelligent Interaction*, pages 511–516, Geneva, Switzerland, 2013.

[9] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[10] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings 21st ACM International Conference on Multimedia*, pages 835–838, Barcelona, Spain, 2013. ACM.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Courville, A.and Bengio. Generative adversarial nets. In *Proceedings Neural Information Processing Systems*, pages 2672–2680, Montreal, QC, Canada, 2014.

[12] L. Kanner. Autistic disturbances of affective contact. *The nervous child*, 2:217–250, 1943.

[13] M. Kjelgaard and H. Tager-Flusberg. An investigation of language impairment in autism: Implications for genetic subgroups. *Language and Cognitive Processes*, 16(2-3):287–308, 2001.

[14] M. Kjelgaard and H. Tager-Flusberg. Update on the language disorders of individuals on the autistic spectrum. *Brain & Development*, 25(3):166–172, 2003.

[15] A. Le Couteur, G. Haden, D. Hammal, and H. McConachie. Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: The ADI-R and the ADOS. *Journal of Autism and Developmental Disorders*, 38(2):362–372, 2008.

[16] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[17] R. Lotfian and C. Busso. Emotion recognition using synthetic speech as neutral reference. In *Proceedings 40th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4759–4763, Brisbane, QLD, Australia, 2015.

[18] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach. Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages. In *Proceedings of INTERSPEECH*, pages 115–119, Dresden, Germany, 2015. ISCA.

[19] E. Marchi, Y. Zhang, F. Eyben, F. Ringeval, and B. Schuller. Autism and speech, language, and emotion – a survey. In H. Patil and M. Kulshreshtha, editors, *Evaluating the role of speech technology in medical case management*. De Gruyter, Berlin, Germany, 2015.

[20] E. Mower, M. Black, E. Flores, M. Williams, and S. Narayanan. Rachel: Design of an emotionally targeted interactive agent for children with autism. In *Proceedings IEEE International Conference on Multimedia and Expo*, pages 1–6, Barcelona, Spain, 2011.

[21] D. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30):13354–13359, 2010.

[22] S. Pascual, A. Bonafonte, and J. Serrà. Segan: Speech enhancement generative adversarial network. *CoRR*, abs/1703.09452, 2017.

[23] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[24] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza. Automatic intonation recognition for prosodic assessment of language impaired children. *IEEE Transactions on Audio, Speech & Language Processing*, 19(5):1328–1342, 2011.

[25] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings of INTERSPEECH*, pages 1210–1214, San Francisco, CA, U. S., 2016. ISCA.

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Proceedings Neural Information Processing Systems*, pages 2226–2234, Barcelona, Spain, 2016.

[27] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller. A bag-of-audio-words approach for snore sounds' excitation localisation. In *Proceedings 14th ITG Conference on Speech Communication*, volume 267 of *ITG-Fachbericht*, pages 230–234, Paderborn, Germany, 2016. ITG/VDE, IEEE/VDE.

[28] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller. Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices. In *Proceedings 14th ITG Conference on Speech Communication*, volume 267 of *ITG-Fachbericht*, pages 264–268, Paderborn, Germany, 2016. ITG/VDE, IEEE/VDE.

[29] M. Schmitt, F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proceedings of INTERSPEECH*, pages 495–499, San Francisco, CA, U. S., 2016. ISCA.

[30] B. Schuller and F. Burkhardt. Learning with synthesized speech for automatic emotion recognition. In *Proceedings 35th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5150–515, Dallas, TX, U. S., 2010.

[31] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. Paralinguistics in speech and language – state-of-the-art and the challenge. *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, 27(1):4–39, 2013.

[32] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of INTERSPEECH*, pages 148–152, Lyon, France, 2013. ISCA.

[33] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio. Invariant representations for noisy speech recognition. *CoRR*, abs/1612.01928, 2016.

[34] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.