Interspeech 2018
2-6 September 2018, Hyderabad

Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks

Shahin Amiriparian[1], *Alice Baird*[1], *Sahib Julka*[1], *Alyssa Alcorn*[2], *Sandra Ottl*[1], *Sunčica Petrović*[3], *Eloise Ainger*[2], *Nicholas Cummins*[1], *Björn Schuller*[1,4]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Centre for Research in Autism and Education, UCL Institute of Education, U. K.
[3]Serbian Society of Autism, Belgrade, Serbia.
[4]GLAM – Group on Language, Audio & Music, Imperial College London, U. K.

shahin.amiriparian@tum.de

Abstract

Autism spectrum conditions (ASC) are a set of neuro-developmental conditions partly characterised by difficulties with communication. Individuals with ASC can show a variety of atypical speech behaviours, including echolalia or the 'echoing' of another's speech. We herein introduce a new dataset of 15 Serbian ASC children in a human-robot interaction scenario, annotated for the presence of echolalia amongst other ASC vocal behaviours. From this, we propose a four-class classification problem and investigate the suitability of applying a 2D convolutional neural network augmented with a recurrent neural network with bidirectional long short-term memory cells to solve the proposed task of echolalia recognition. In this approach, log Mel-spectrograms are first generated from the audio recordings and then fed as input into the convolutional layers to extract high-level spectral features. The subsequent recurrent layers are applied to learn the long-term temporal context from the obtained features. Finally, we use a feed forward neural network with softmax activation to classify the dataset. To evaluate the performance of our deep learning approach, we use leave-one-subject-out cross-validation. Key results presented indicate the suitability of our approach by achieving a classification accuracy of 83.5 % unweighted average recall.

Index Terms: autism spectrum conditions, vocal abnormalities, echolalia, convolutional recurrent neural network

1. Introduction

Patterns in speech and atypical vocal behaviours are extremely informative for the further understanding of an array of health-related disorders including depression [1], Parkinson's disease [2], and Autism Spectrum Conditions (ASC) [3]. ASC are a grouping of neuro-developmental conditions which are defined in the literature by limitations in two primary domains: (i) social and communicative behaviours, and (ii) restricted and repetitive behaviours [4]. Often, ASC becomes noticeable in early childhood, as a divergence from typical developmental milestones, with many children with an ASC having limited verbal ability. Diagnosis methods for an ASC are based on direct behavioural observation or reports, e. g., [5], with a number of assessed behaviours relating to the ability for verbal language production.

Common to ASC (amongst other conditions), *echolalia* is an occurrence in which, with similar intonation, words or phrases are repeated, from what a conversational partner has said [6]. In this study, we explore the efficacy of the automatic recognition of echolalic occurrences. To the best of the authors' knowledge, there have been no previous attempts to use advanced computational methods for this task. Obtaining a quantifiable account of occurrences across samples of a child's behaviour would take a substantial time effort from multiple expert annotators.

The literature describes six key functions for echolalic vocalisations used by those with an ASC. These include, turn taking, assertion, affirmative answers, and self-regulation [7]. In this regard, Gleitman et al. [8] explored context discrimination in relation to ASC echolalic vocalisations. It has also been discussed that echolalic speech of individuals with an ASC can contribute to the general picture of their language development [9], and additionally having a quantifiable measure of occurrences could potentially be a marker for the presence of anxiety or stress [10].

Speech, prosodic cues in particular, have previously been utilised to explore typically developing children against those an ASC [11]. Such a task was also presented as an Interspeech Computational Paralinguistics ChallengE (ComParE) task in 2013 [12]. Additionally, studies have also investigated the efficacy of state-of-the-art machine learning approaches to classify ASC severity through vocalisation [13]. With work in this specific area seemingly sparse, the proposed non-invasive audio-based system for automatic detection of echolalic occurrences could be a step forward to assist ASC experts in assessing more subtle ASC behavioural changes, as well as contributing to language-use and development profiling over time.

In this paper, we utilise a convolutional recurrent neural network (CRNN) deep learning approach for recognition of echolalia. CRNNs were first proposed for document classification [14] and can be regarded as state-of-the-art in many audio tasks [15, 16]. Convolutional neural networks (CNNs) themselves are well known for their ability to learn a robust, task specific feature representation and have been successfully applied in similar tasks such as speech recognition [17], audio analysis [18, 19], and phoneme sequence recognition [20]. At the same time, recurrent neural networks (RNNs) are well known for their strengths in modelling temporal sequences. Long-short-term memory (LSTM) RNNs, in particular bidirectional LSTM (BLSTM) RNNs, have also been used in related tasks, e. g., social signal classification [21] and speech recognition [22]. In this work, the proposed system to recognise the vocalisations of interest is a combination of a CNN and an RNN, that allows for the global temporal context to be taken into account, while efficiently extracting features [15, 23], and thus reducing the network complexity.

The rest of this paper is structured as follows: the characteristics of echolalia vocalisations are detailed in Section 2, the corpus used is outlined in Section 2.1. We describe our pro-

10.21437/Interspeech.2018-1772

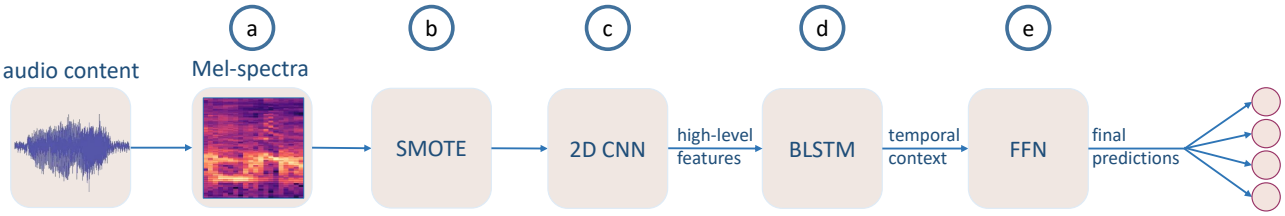

Figure 1: *Illustration of the proposed deep learning approach composed of convolutional and recurrent neural networks for feature extraction and a feed forward network for the classification. SMOTE is only used for the training data. A detailed account of the procedure is given in Section 3.*

posed CRNN system in Section 3. The experimental settings are then presented in Section 4, and the results and discussion in Section 5, followed by concluding remarks in Section 6.

2. Echolalic Vocalisations

In relation to ASC, echolalia has been defined in the literature as an occurrence with similar intonation, of repeated (from what others have said) words or phrases [6]. There are two key subcategories for echolalia, immediate and delayed.

Put simply, immediate echolalia is a vocalisation which is immediately repeated from a current interaction with another person in the session. A common example found across our dataset is the repetition of a question, e. g., researcher: "*shall we play with Zeno?*" child: "*play with Zeno*". On the contrary, delayed echolalia refers to heard vocalisations from a previous interaction, e. g., catch-phrases heard on television, or from public announcements.

2.1. The Echolalic Vocalisation Data

The data collected for this study has been provided through the DE-ENIGMA, Horizon 2020 initiative. This data has been collected in Belgrade, Serbia, from children with varying degrees of autism severity. For this study, we have used instances collected during the Serbian studies[1]. A dataset of 36 sessions from 15 children (12 h: 21 m), containing the presence and absence of echolalic speech, have been chosen. Ages range from 7–12 years, and ± 1.91 within this corpus gender is split 4:11 (female:male), an (unavoidable) gender bias; however, this is somewhat representative of current diagnosis rates in autism, which have been reported to be 3.5 times higher in males [24]. It has also been shown that gender has a minimal effect on the voice during childhood [25], and so we do not foresee that this will bias our results.

The audio dataset was collected over 1–5 short daily sessions (each child participated in an average of 2.7 sessions). The children took part in a *human-led* or *robot-led* emotion-recognition training programme, which is adapted from the "Teaching Children with Autism to Mind–Read" programme [26]. In robot-led sessions, the child interacted with Zeno–R2, a humanoid-robot (controlled using a *Wizard-of-Oz* interface). Six children are in human-led sessions, and through a qualitative analysis of the corpus, we see no considerable difference in the quantity of echolalic vocalisations between *human-led* or *robot-led* children.

[1]Full ethical approval for the Serbian data collection was approved by the Ethics Committee of the Institute for Mental Health, approval No. 30/66 DE-ENIGMA Multi–Modal HRI for Expanding Social Imagination in Autistic Children.

2.2. Annotation and Classes

In collaboration with the Serbian Autism Society and the University College London (UCL) Institute of Education, an annotation protocol was developed, with the intention of providing data-informed insights into this currently under-represented population. Annotation of the audio data has been made in a tier-like manner by 3 native Serbian-speaking ASC experts. Speaker diarisation of each speaker was made manually and set as Tier 0. Annotators were then targeting instances of child vocalisation and labelling in the following manner.

- **Tier 1:** Speech Type, e. g., Speech, Non Speech, Speech-Like.
- **Tier 2:** ASC Specific, e. g., Echolalia, Another ASC Behaviour, Unsure ASC.

A majority vote *gold-standard* was produced from the 3 annotators. For this task, we are using the classes within Tier 2 of the dataset, and our system is designed to recognise 4 particular classes, which have have been defined as follows:

- **Immediate echolalia (ec-im)** – a vocalisation, which is immediately repeated from current interaction with another person in the session.
- **Delayed echolalia (ec-de)** – a vocalisation, which is repeated from a previous interaction or heard sound.
- **Other ASC vocal behaviours (other)** – another stereotyped behaviour. Not echolalia, but a behaviour which could be unique to ASC – for example, pronoun inversion or stereotyped speech.
- **No specific ASC event (no-ev)** – is a grouping of all other child speech events, labelled in Tier 2, including the labels such as 'not-specific to ASC' and 'unsure'.

3. System Architecture

A high-level overview of our deep learning approach to classify the proposed dataset is given in Figure 1. Motivated by the systems presented in both [27, 28], we implemented a convolutional and recurrent neural network composed of five main components: first, log Mel-spectrograms are extracted from the audio recordings (cf. Figure 1a). We then over- and under-sample all four classes in order to balance the class ratio (cf. Figure 1b). These samples are then fed as input into the convolutional layers to extract high-level, shift-invariant spectral features (cf. Figure 1c). Afterwards, recurrent layers are used to learn the long-term temporal context from the obtained features (cf. Figure 1d). Finally, we use a feed forward neural network with softmax activation to classify the dataset (cf. Figure 1e).

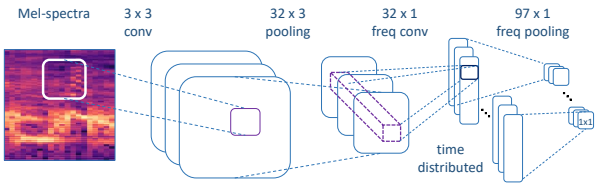

Figure 2: *The structure of the 2D convolutional neural network applied for extracting high-level features from the input Mel-spectrograms. Figure adopted from [23].*

3.1. Spectrogram Extraction

The Mel-spectrograms are computed with a window size of 46 ms and an overlap of 23 ms from the log-magnitude spectrum by dimensionality reduction using a Mel-filter. We apply 128 filter banks equally spaced on the Mel-scale. The Mel-scale is based on the frequency response of the human ear that has better resolution at lower frequencies. The log Mel-spectrograms are then divided into chunks of a desired time step τ (for our experiments $\tau = 41$), and after augmentation (cf. Section 3.2) fed into the CNN.

3.2. Data Augmentation

The (training) data augmentation method used in this study is a combination of under-sampling and the synthetic minority over-sampling technique (SMOTE) [29]. This combination has been shown to perform stronger than simple over-sampling [30, 29]. SMOTE augments the minority classes by creating synthetic examples, rather than by over-sampling with replacement. The algorithm selects user defined k-nearest neighbours and uses this information to augment each minority class along the line segments joining any or all of the k minority class samples. For our experiments, we used the default value for $k = 5$. In order to approximately balance the number of samples, we under-sample the majority class (class 'no-ev') by a factor of 0.1 and over-sample the minority classes by factors of 2.0, 2.0, and 1.2 for classes 'ec-im', 'ec-de', and 'other', respectively.

3.3. Convolutional Neural Network

The extracted Mel-spectrograms are – after training data augmentation – fed into a convolutional layer with 2D filters. As depicted in Figure 2, the frequency time convolution is followed by non-overlapping pooling to ensure no shrinking in time. Subsequently, a 1D convolution along the spectral domain is applied, which is then followed by max pooling along the frequency domain. Rectified linear unit (ReLU) activation is used in the convolutional layers, and batch normalisation (BN) [31] is applied between them. Furthermore, a dropout of 30 % is used for all layers to add regularisation and also to minimise the potential overfitting problems caused by non-overlapping max pooling [32]. However, it has been shown that convolutional layers work best with small filter sizes, i. e., the captured temporal context is small (often in the range < 200 ms) [33]. Thus, giving way to the need for recurrent networks with memory cells for longer temporal modelling.

3.4. Recurrent Neural Network

Long short-term memory (LSTM) recurrent neural networks (RNNs) were introduced by Hochreiter and Schmidhuber in order to solve error back-flow problems on gradient descent optimisation [34]. LSTM RNNs have been successfully applied in a range of audio-based tasks, including acoustic modelling [35], speech recognition [36], and audio-based multimedia event detection [37]. During our initial experiments, we observed that for BLSTM nets, which propagate errors in both time directions, the recognition rate is better than for LSTM nets. This might be due to the repeated echolalic imitation which occurs throughout the time domain [6]. As shown in Figure 3, we use two hidden layers in backward and forward directions. Hyperbolic tangent is used as the activation function on each layer, and a dropout of 30 % is applied on the activations passed from the BLSTM nets. In the final step, 128-dimensional output vectors are returned for each frame, which are passed into a fully connected feed forward layer for the classification.

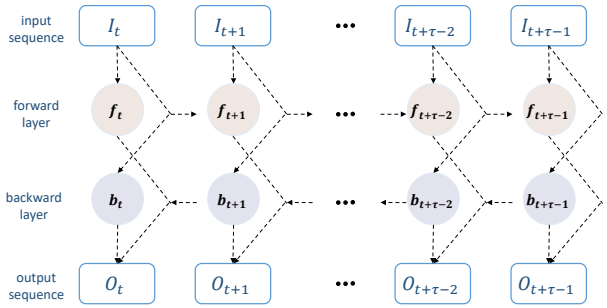

Figure 3: *The BLSTM RNN structure applied in our CRNN approach. Two hidden layers, one in the forward direction (f) and another one in the backward direction (b) are used. For all input CNN features ($I_{t+\tau}$) during the timestep (τ), the returned outputs from each layer are then concatenated ($O_{t+\tau}$).*

3.5. Fully Connected Layer

The concatenated vectors obtained from the RNN are passed through a fully connected feed forward network with 128 hidden units. We normalise the outputs using batch normalisation, so that the mean is close to 0 and the standard deviation is close to 1. Then, we apply ReLU activations, which adds non-linearity to the outputs. The updated output features are then used for classification by passing through a layer consisting of four output cells and softmax as the activation function.

4. Key Experimental Settings

4.1. Evaluation Metric

We use unweighted average recall (UAR) to evaluate our leave-one-speaker-out cross-validation (LOSO-CV) experiments, as this metric gives equal weight to all four classes and is accordingly more suitable than a weighted metric (e. g., accuracy) for datasets which have imbalanced class ratio. The classification chance level for a four-class classification problem is 25 % UAR.

4.2. Deep Learning Hyperparameters

We tested three CRNNs with various hyperparameters. The loss function applied for training is the categorical cross entropy, which minimises the loss obtained on the categorical outputs. We use the Adam algorithm [38] to optimise the deep learning models. The networks differ in terms of their hyperparameters, including the number of hidden units, learning rate, learning rate decay, dropout, and number of epochs (cf. Table 1).

Table 1: *Hyperparameters selection for our CRNNs together with their approach identifiers. lr : learning rate; lrd : learning rate decay. All CRNNs have six layers, two for each neural network. All other hyperparameters have been applied equally for all networks in each CRNN.*

Parameters	CRNN1	CRNN2	CRNN3
layers	6	6	6
hidden units	64	96	128
dropout	40 %	35 %	30 %
lr	0.01	0.0001	0.001
lrd	0.01	0.01	0.01
epochs	25	35	60

Table 2: *Classification results obtained from three CRNNs with various hyperparameters (cf. Table 1) using LOSO-CV. ec-im: immediate echolalia; ec-de: delayed echolalia; other: other ASC vocal behaviour; no-ev: no specific ASC event; overall: the overall classification result for all classes. The best result is highlighted with light grey shading.*

Network	Recall of each class in %				mean
	ec-im	ec-de	other	no-ev	UAR %
CRNN1	73.2	88.7	86.6	74.5	80.1
CRNN2	73.1	85.7	82.8	73.3	74.2
CRNN3	74.4	96.1	88.6	75.1	**83.5**

5. Results and Discussion

We used LOSO-CV to reduce the potential of overfitting problems when evaluating the proposed deep learning approach. We created 15 folds of the data, each for one speaker (child). We then trained a model on 14 speakers and evaluated it on one non-data-augmented speaker. This procedure was repeated 15 times until we obtained the final classification result by averaging the UARs. We tested three CRNNs varying their hyperparameters (cf. Table 1). Our results indicate that with any of our CRNN set-ups it is possible to recognise the rare echolalic vocalisations with high UAR (cf. Table 2). CRNN3 performed the strongest on our dataset achieving a UAR of 83.5 %, followed by CRNN1 (80.1 % UAR) and CRNN2 (74.2 % UAR).

With the main target class for this system being echolalia, it is of interest that we achieve consistently lower UAR results for 'ec-im' as opposed to 'ec-de'. In particular, a substantial difference can be observed within the confusion matrix of CRNN3 Figure 4. Although there is imbalance across the classes (one possible indication for the differing results), we can also speculate that the prosodic nature of delayed echolalic utterances may vary greatly from the child's usual speech, as compared to immediate echolalia. In delayed echolalia, children may repeat language from familiar adults, media, transport announcements, or other previously-heard sources. Thus, within a behavioural session, this delayed speech may appear entirely out of context both linguistically and prosodically. To this end, it can be clearly seen from Figure 4 that there is minimal confusion across the classes for 'ec-de'.

Further, we see that there is some confusion between the classes of 'no-ev' and 'other', which may come from the inclusion of 'unsure' labels within the dataset. It is possible that such

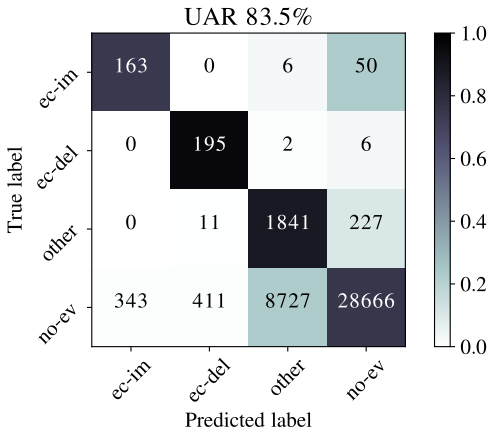

Figure 4: *Confusion matrix of the best classification result which was obtained by CRNN3.*

'unsure' labels were similar in nature, e. g., 'unsure stereotyped speech' and 'other: stereotyped speech'. Another potential cause could also be annotation error, e. g., 'not specific to ASC' could be labelled instead of finer details such as 'other; pronoun error'.

6. Conclusions and Future Work

In exploring the efficacy of automatically recognising echolalic vocalisations, we have introduced a novel dataset with a total length of 12 h: 21 m which includes recordings of 15 Serbian ASC children in a human-robot interaction scenario. Further, we have demonstrated the feasibility of using a state-of-the-art CRNN-based classification system for the task. One major advantage of our approach is the coupled modelling of spectral shifts by the convolutional layers and that of the temporal dependencies, which are inherent in these speech patterns.

In future work, we will be exploring the benefits of using extra corpora to substantially increase the amount of data for training our CRNN. In this regard, as additional data from English speakers will be made available through the DE-ENIGMA project, we also plan to evaluate this approach across such a bilingual corpus. Furthermore, we plan to add a feature quantisation step after obtaining the temporal context from the recurrent layers in order to cope with the amount of noise in the data.

Finally, the results of this study, albeit on a specific sub-set of data, indicate that integration of this system into a robot-interaction scenario or other educational technology for autism is a potential means of providing fine-grained information on a child's language use – and possibly development – over time. This type of information is a valuable resource for researchers, therapists, and teachers. However, implementing such a system in the real-world would require further study and verification.

7. Acknowledgements

This work is funded by the EU's Horizon 2020 Programme under grant agreement No. 688835 (RIA DE-ENIGMA), and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B). We thank Marina Sparić, Dragana Stojanović, and Jelena Ralevski from the Serbian Society for Autism.

8. References

[1] N. Cummins, S. Scherer, J. Krajewski *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[2] A. Frid, A. Kantor, D. Svechin *et al.*, "Diagnosis of parkinson's disease from continuous speech using deep convolutional networks without manual selection of features," in *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Nov 2016, pp. 1–4.

[3] S. Mitchell, J. Brian, J. Brian *et al.*, "Early language and communication development of infants later diagnosed with autism spectrum disorder," *Journal of Developmental & Behavioral Pediatrics*, vol. 1, pp. 69–78, 2006.

[4] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 5th ed., Washington, D.C., 2013.

[5] C. Lord and M. Rutter, *Autism Diagnosis Observation Schedule -2 (ADOS-2)*. Western Psychological Services, 2012.

[6] H. Tager-Flusberg, R. Paul, and C. Lord, *Language and Communication in Autism*. John Wiley & Sons, Inc., 2005, pp. 335–364.

[7] B. Prizant and J. Duchan, "Language acquisition and communicative behavior in autism: Toward an understanding of the the "whole" of it," *Journal of Speech and Hearing Disorders*, no. 48, pp. 257–286, 1983.

[8] M. Gleitman, "Contextualism and echolalia," in *Modeling and Using Context*, P. Brézillon, R. Turner, and C. Penco, Eds. Cham: Springer International Publishing, 2017, pp. 267–276.

[9] S. Andressa Gouveia de Faria and G. Marcia, "Echolalia in the language development of autistic individuals: a bibliographical review," *Pr-Fono Revista de Atualizao Cientfica*, vol. 3, no. 21, pp. 255–260, 2009.

[10] B. Vicker, "Functional categories of immediate echolalia," *Indiana Resource Center Autism.*, vol. 3, 2010.

[11] F. Ringeval, J. Demouy, G. Szaszak *et al.*, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1328–1342, July 2011.

[12] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013, pp. 148–152.

[13] A. Baird, S. Amiriparian, N. Cummins *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 849–853.

[14] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[15] K. Choi, G. Fazekas, M. Sandler *et al.*, "Convolutional recurrent neural networks for music classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2392–2396.

[16] G. Parascandolo, T. Heittola, H. Huttunen *et al.*, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[17] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang *et al.*, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[18] S. Amiriparian, M. Gerczuk, S. Ottl *et al.*, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proceedings 31st International Joint Conference on Neural Networks (IJCNN)*, IEEE. Rio de Janeiro, Brazil: IEEE, July 2018, 8 pages, to appear.

[19] ——, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, ISCA. Stockholm, Sweden: ISCA, August 2017, pp. 3512–3516.

[20] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[21] R. Brueckner and B. Schulter, "Social signal classification using deep blstm recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4823–4827.

[22] M. Wöllmer, Z. Zhang, F. Weninger *et al.*, "Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6822–6826.

[23] S. Amiriparian, S. Julka, N. Cummins *et al.*, "Deep convolutional recurrent neural networks for rare sound event detection," in *Proceedings 44. Jahrestagung für Akustik, DAGA 2018*, DEGA. Munich, Germany: DEGA, March 2018, 4 pages, to appear.

[24] R. Loomes, L. Hull, and W. Mandy, "What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis." *Journal of America, Academic Child Adolescent Psychiatry*, vol. 56, no. 6, pp. 466–474, 2017.

[25] J. S. Hyde, "The gender similarities hypothesis," *The American Psychologist Association*, vol. 60, pp. 581–592, 2005.

[26] P. Howlin, S. Baron-Cohen, and J. Hadwin, *Teaching Children with Autism to Mind-Read, A Practical Guide for Teachers and Parents*. Hoboken, New Jersey: John Wiley and Sons Ltd., 1999.

[27] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," DCASE2017 Challenge, Tech. Rep., September 2017.

[28] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," DCASE2017 Challenge, Tech. Rep., September 2017.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall *et al.*, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[30] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[33] E. Cakır, G. Parascandolo, T. Heittola *et al.*, "Convolutional recurrent neural networks for polyphonic sound event detection," *arXiv preprint arXiv:1702.06286*, 2017.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[35] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[36] ——, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[37] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, March 2016, pp. 2742–2746.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.