Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks

Shahin Amiriparian\textsuperscript{1}, Alice Baird\textsuperscript{3}, Sahib Julka\textsuperscript{1}, Alyssa Alcorn\textsuperscript{2}, Sandra Ottl\textsuperscript{1}, Sunčica Petrović\textsuperscript{3}, Eloise Ainger\textsuperscript{2}, Nicholas Cummins\textsuperscript{1}, Björn Schuller\textsuperscript{1,4}

\textsuperscript{1}ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
\textsuperscript{2}Centre for Research in Autism and Education, UCL Institute of Education, U. K.
\textsuperscript{3}Serbian Society of Autism, Belgrade, Serbia.
\textsuperscript{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-
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 studies1. A dataset of 36 sessions from 15 children (12h:21m), 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.

1Full 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 Imaginaton 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 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).
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 \( \tau \) (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 oversample 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, 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 opt-

![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].](image)

![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 (\( \tau \)), the returned outputs from each layer are then concatenated (\( O_{t+\tau} \)).](image)
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.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CRNN1</th>
<th>CRNN2</th>
<th>CRNN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>layers</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>hidden units</td>
<td>64</td>
<td>96</td>
<td>128</td>
</tr>
<tr>
<td>dropout</td>
<td>40%</td>
<td>35%</td>
<td>30%</td>
</tr>
<tr>
<td>lr</td>
<td>0.01</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>lrd</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>epochs</td>
<td>25</td>
<td>35</td>
<td>60</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Network</th>
<th>Recall of each class in %</th>
<th>mean UAR %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ec-im</td>
<td>ec-de</td>
</tr>
<tr>
<td>CRNN1</td>
<td>73.2</td>
<td>88.7</td>
</tr>
<tr>
<td>CRNN2</td>
<td>73.1</td>
<td>85.7</td>
</tr>
<tr>
<td>CRNN3</td>
<td>74.4</td>
<td>96.1</td>
</tr>
</tbody>
</table>

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 ‘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 min 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


