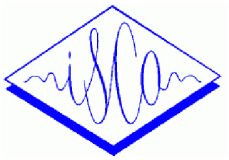
ISCA Archive
http://www.isca-speech.org/archive



Third Workshop on Child, Computer and Interaction (WOCCI 2012)
Portland, OR, USA
September 14, 2012

Emotion in the Speech of Children with Autism Spectrum Conditions: Prosody and Everything Else

Erik Marchi[1], *Björn Schuller*[1], *Anton Batliner*[1,2], *Shimrit Fridenzon*[3], *Shahar Tal*[3], *Ofer Golan*[3]

[1]Institute for Human-Machine Communication, Technische Universität München, Munich, Germany
[2]Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany
[3]Department of Psychology, Bar-Ilan University, Ramat Gan, Israel

(erik.marchi|schuller)@tum.de, anton.batliner@lrz.uni-muenchen.de,
shimfri@gmail.com, shahar0190@gmail.com, ofer.golan@biu.ac.il

Abstract

Children with Autism Spectrum Conditions (ASC) may experience significant difficulties to recognise and express emotions. The ASC-Inclusion project is setting up an internet-based digital gaming experience that will assist children with ASC to improve their socio-emotional communication skills, combining voice, face, and body gesture analysis, and giving corrective feedback regarding the appropriateness of the child's expressions. The present contribution focuses on the recognition of emotion in speech and on feature analysis. For this purpose, a database of prompted phrases was collected in Hebrew, inducing nine emotions embedded in short-stories. It contains speech of children with ASC and typically developing children under the same conditions. We evaluate the emotion task over the nine categories including the binary valence/arousal discrimination. We further investigate the discrimination of each emotion against neutral. The results show performances for arousal and valance of up to 86.5% and for nine emotions including neutral of up to 42% unweighted average recall. Moreover we compare and analyse manually selected prosodic features with automatic selected features with respect to their relevance for discriminating each of the eight emotion classes.

Index Terms: Autism Spectrum Conditions, emotion recognition, prosody, feature analysis

1. Introduction

Three decades of research have shown that children and adults with Autism Spectrum Conditions (ASC) may experience significant difficulties in recognising and expressing emotions from facial expressions, speech, gestures, and body language. Attempts to teach emotion and mental state recognition, either on an individual basis or as a part of social skills group training, have shown mixed results. A solution for the shortage of trained therapists for individuals with ASC may be found in Information and Communication Technology (ICT), which enables users everywhere to enjoy state-of-the-art professional support on-line. The computerised environment is especially appealing for individuals with ASC, due to its predictable, controllable and structured nature, which facilitates them to use their strong systemizing skills. Existing systems, such as the Rachel Embodied Conversational Agent (ECA) [1] and the Mind-Reading software [2], aim to elicit the targeted emotion through an interactive agent in order to study the interaction patterns of children with ASC and to teach people in the spectrum to recognise complex emotions using interactive multimedia. The ASC-Inclusion project aims to create an internet-based platform that will assist children with ASC to improve their socio-emotional communication skills. Unlike the past ICT solutions, the project will address the recognition and the expression of socio-emotional cues, by providing an interactive-game that gives scores on the prototypicality and on the naturalness of child's expressions. It will combine several state-of-the-art technologies in one comprehensive virtual world environment, combining voice, face and body gesture analysis, giving corrective feedback as for the appropriateness of the child's expressions. The present study focuses on the recognition of emotional vocal expressions and on features analysis, in order to investigate the behaviour of prosodic features against large sets of features that include a vast number of acoustic, spectral and cepstral features. The importance of prosody with respect to several aspects of voice and language impairment in Autism Spectrum Conditions is addressed in [3], [4], [5], [6].

We are interested in classification as well as in analysing to what extent prosodic features are relevant when the child is expressing his or her emotional state. Furthermore, given that prosodic features such as energy, pitch, and duration are easier to show and to convey as feedback than spectral and cepstral features, the child can interact and intuitively manipulate these parameter during the game. Prosodic features can be used both for automatic modelling and for demonstrating to the children how to employ them, and they will be used as consistent parameters for the corrective feedback that will be given to the children for improving the appropriateness of their emotional expressions. This study further focuses on other aspects, such as discrimination of typicality between typically developing children and children with ASC, and diagnosis discrimination within the focus group. For that, a database of prompted phrases was collected, inducing nine emotions embedded in short-stories. The utterances were produced by children with Autism Spectrum Conditions as well as by typically developing children. The article is structured as follows: first, a detailed description of the database is given (Section 2); then we define the experimental tasks, features and set-up (Section 3). We next comment on the evaluation results (Section 4) before concluding the paper in Section 5.

2. ASC-Inclusion children's emotional speech database

As an evaluation database for the recognition of emotions and for the analysis of speech features that are modulated by emotion, a database of prototypical emotional utterances containing

Table 1: *Number of utterances per emotion category (# Emotion), binary arousal/valence, diagnosis and overall number of utterances (# All) for the two groups. Emotion classes: happy (Ha), sad (Sa), angry (An), surprised (Su), afraid (Af), proud (Pr), ashamed (As), calm (Ca), neutral (Ne). Diagnosis categories: Asperger Syndrome (AS), High-Functioning autism spectrum disorders (HF).*

# utterances	# Emotion									# Arousal		# Valence		# Diagnosis		# All
	Ha	**Sa**	**An**	**Su**	**Af**	**Pr**	**As**	**Ca**	**Ne**	**-**	**+**	**-**	**+**	**AS**	**HF**	
Focus group	30	21	20	21	18	21	17	14	16	67	111	76	102	88	90	178
Control group	49	38	38	38	38	46	37	27	40	142	209	151	200	-	-	351
Total	79	59	58	59	56	67	54	41	56	209	320	227	302	-	-	529

sentences spoken in Hebrew by children with ASC and typically developing children has been created. The focus group consists of nine children (8 male and 1 female) at the age of 6 to 12, all diagnosed with an autism spectrum condition by trained clinicians. 11 typically developing children (5 female and 6 male) at the age of 6 to 9 were selected to form the control group. In order to limit the effort of the children, the experimental task was designed to focus on the six "basic" emotions except *disgust*: *happy, sad, angry, surprised, afraid* plus other three mental states: *ashamed, calm, proud,* and *neutral*. During a 2 hour meeting with the child and his/her parents, a semi-structured observation was conducted which included free-play in a virtual environment, followed by a directed play in pre-selected games, and by an interview with the child. Only then, the recording session was held, since it requires a good rapport with the child. The recordings took place at the children's home according to the following set-up: the child and the examiner sat at a table in front of a laptop. The microphone stood next to the laptop, about 20 cm in front of the child. As recording device, a Zoom H1 Handy Recorder was used. Recordings were taken in wav format at a sampling rate of 96 kHz and a quantization of 16 bits and stored directly on the microphone's internal SD memory card. The examiner read to the child a sequence of short stories from a power point presentation. The stories were simple and short. The child was asked to imagine that he/she was the main character in the story. The stories contained, every few sentences, a quotation of an utterance by the story's main character. Each of these quotations related to a specific emotion, which was explicitly stated. For example: [Danny said happily: *"It was the best birthday I ever had!"*] or [Jain was very surprised. She looked at the box and said: *"What is that thing?"*]. When the examiner read the stories, he read the sentence on a flat, unnatural tone. Then he asked the child to say the sentence as the child in the story would have said it. Each slide that contained an emotional utterance to be said by the child also showed a photograph of a person expressing the same emotion through his facial expressions. The photos were taken from the Mind-Reading database [2]. The text material used for the task consists of nine stories. Each story aims to elicit some of the target emotions as described above and contains from 3 to 7 different emotional utterances. In total, the nine stories contain 37 utterances.

An example for one of the nine stories is:

Happy - Today it's a special day for Danny: it's his birthday! Danny was very happy - a birthday is an especially enjoyable and fun day. Danny went into his sister's room and said **happily**: *"Today's my birthday!"*.

Sad - Afterwards he entered the kitchen. He noticed his mother was preparing a simple breakfast for him and a not a birthday's one. Danny was very sad. He was convinced his family had forgotten his birthday. In school no one had congratulated him either, not even his teacher! Tears flooded his eyes, and so he looked for his sister on break time. When he found her, he told her **sadly**: *"No one had remembered"*.

Angry - On his way home the sad feeling had faded away, and anger burned inside of him. He was so angry of his mom and classmates, and said **angrily** to his sister: *"I won't remember their birthday either!"*.

Surprised - When he got back home, there was a complete silence. He went into the dark kitchen, lit up the light and suddenly heard: "surprise"! He saw there his parents and classmates holding balloons! He was very **surprised –** and said: *"What's going on?"*.

Happy - Danny was happy, they haven't forgotten him, they planned him a surprise birthday party. After a party, he went to his sister and said **happily**, *"It was the best birthday I ever had!"*.

The 37 utterances were not collected for each subject since the task was new for the children and it required both a strong sense of comfort and a high level of cooperation. In particular, in the focus group, two children were not recorded because they found the task not comfortable and other three of them were partially recorded since they wanted to stop their participation. In the control group, one child found the task not comfortable and recordings were not held. Furthermore, some samples belonging to the control group were left out because of the high level of background noise. Hence, the actual focus group consists of seven children (6 male and 1 female) at the age of 6 to 10 (M=8.1, SD=1.6). Three of them were diagnosed with an Asperger Syndrome (AS) and the other four were diagnosed with High-Functioning (HF) autism spectrum disorder. The actual control group is composed by 10 typically developed children (5 male and 5 female) at the age of 5 to 9 (M=7.2, SD=1.8).

Since the recordings were held at the children's home, they are partly affected by background noise. Compared to the standards of present day databases used for automatic speech processing, this is a small database; however, taking into account the difficulties to recruit children from the envisaged population, to successfully conduct all the experimental tasks, and in comparison to other studies within the fields of ASC and emotion modelling for specific and less-studied populations, it can be taken as fairly representative, especially for a pilot study aiming at setting the field and defining the roadmap for collecting a larger database. It comprises 529 utterances with a total duration of 16 min 24 sec and an average utterance length of 1.8 sec. 178 utterances contain emotional speech of children with ASC with a total recording time of 7 min 1 sec and an average utterance duration of 2.37 sec. Within this group, 90 and 88 utterances are performed, respectively, by children with Asperger syndrome and high-functioning diagnosis. The remaining 351

Table 2: *Arousal and valence mapping.*

AROUSAL		VALENCE	
Low	**High**	**Negative**	**Positive**
sad	happy	sad	happy
ashamed	angry	angry	surprised
calm	surprised	afraid	proud
neutral	afraid	ashamed	calm
	proud		neutral

utterances are produced by the control group with a total duration of 9 min 23 sec and an average utterance recording time of 1.61 sec. Since different class problems have been performed on this database, Table 1 shows the number of utterances for each classification task.

3. Experiments

In this part we describe the classification tasks in Section 3.1, the feature sets in Section 3.2, the experimental set-up (Section 3.3) and our evaluation and analysis criteria (Section 3.4).

3.1. Tasks

Six tasks were evaluated: typicality and diagnosis, emotion, valence, arousal, and every emotion-against-neutral.

The **typicality** task concerns the classification of typically developing children and children with ASC. The **diagnosis** task aims to distinguish between Asperger syndrome and high-functioning diagnosis. The **emotion** task covers the recognition of the nine target classes (eight emotions plus "neutral"). We further evaluated the discrimination between high and low **arousal** as well as between positive and negative **valence**. Additionally, we evaluate the **emotion-against-neutral** task in order to analyse the differences and discriminate across each of the eight emotions against the neutral state.

The typicality task was performed on the full database, and the diagnosis task on the focus group. All the emotion related tasks (emotion, valence, arousal and emotion-against-neutral) were performed on the focus and control group subsets separately. The mapping of the emotion categories onto the binary arousal/valence labels is shown in Table 2; a detailed description of the number of instances belonging to the classes of each task per subset is given in Table 1.

3.2. Features

For a better readability we grouped all the features into three categories: **Spectral** such as functionals of auditory spectrum at different frequency bands with or without RASTA filtering, magnitude spectrum and Mel Frequency Cepstral Coefficients (MFCCs), **Voice Quality** that comprises functionals of jitter, shimmer and Harmonic to Noise Ratio (HNR), and **Prosodic** such as functionals of energy, loudness, duration, fundamental frequency contour, voice probability and zero-crossing rate. In the following sections we will refer to the features by using this taxonomy.

The experiments were conducted using four feature sets: IS12, IS12-CFS, IS12-IG and PROS. The **IS12** features set, from the INTERSPEECH 2012 Speaker Trait Challenge [7], contains 6128 features (84.6% spectral, 9.4% prosodic and 6% voice quality) and is taken as reference in our experiments. Next, we applied feature selection to IS12 using two methods: by considering the individual predictive ability of each feature using correlation-based selection (**IS12-CFS**) and by measuring the information gain (**IS12-IG**). While the former selected a variable number of features for each task (up to 140 features), for the latter we selected the best 15 features in order to have a set of features of equal size to compare with our manually selected prosodic feature set comprising 15 features. The prosodic set (**PROS**) consists of statistical functionals of: **Energy** such as the sum of auditory spectrum at different frequency bands (from 20Hz to 8kHz) and root-mean-square signal frame energy; **Pitch**: fundamental frequency contour; and **Duration** by modelling temporal aspects of F0 values, such as the F0 onset segment length. We applied mean, standard deviation, 1st percentile and 99th percentile to Energy and Pitch, and only mean and standard deviation to Duration.

As mentioned before, we choose these three prosodic low level descriptors (Energy, Pitch and Duration) with their basic functionals (mean, standard deviation, maximum and minimum values) as simplest prosodic parameters that can be easily conveyed to the children. They enable the child to manipulate them intuitively throughout the game, for instance, by modulating pitch in order to accomplish a simple task such as moving a graphical object to a target, or by increasing/decreasing energy in order to jump over an obstacle. Such intuitive and easy interaction would be hardly provided by spectral features and cepstral features such as MFCCs. It can be expected that automatically selected features yield a better performance than pure prosodic features; however, these might be correlated up to some extent with the automatically selected ones, and thus still be good candidates for our envisaged game. All features were extracted with openSMILE [8].

3.3. Setup

Since all data sets are unbalanced (i.e. one class is underrepresented in the data), the unweighted average recall (UAR) of the classes is used as scoring metric. Adopting the Weka toolkit [9], Support Vector Machines (SVMs) with linear kernel were trained with the Sequential Minimal Optimization (SMO) algorithm. SVMs have been chosen as classifier since they are a well known standard method for emotion recognition due to their capability to handle high and low dimensional data. The SVM training has been made at different complexity constant values $C \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 1.5, 2\}$. To ensure speaker independent evaluations, Leave-One-Speaker-Out (LOSO) cross-validation has been performed. In order to balance the class distribution, we applied the Synthetic Minority Over-sampling Technique (SMOTE) for all the evaluation experiments.

Furthermore, we adopt the speaker z-normalisation (SN) method since it is known to improve the performance of speech-related recognition tasks, as described in [10]. With such a method, the feature values are normalised to a mean of zero and a standard deviation of one for each speaker. For typicality and diagnosis tasks, we do not apply speaker z-normalisation since centring and scaling the feature space in such tasks is not effective because the phenomena considerably vary in the range across subjects. By applying this technique the relevant features able to characterise the subject are flattened, making the classification performances not acceptable and below the chance level.

3.4. Evaluation

For each task, we first perform classification experiments using the four different feature sets, in order to evaluate the performances over decreasing dimensional feature spaces. Then we

analyse the selected feature sets with a detailed description of the differences/similarities across the IS12-IG and PROS sets. For that, we compute the correlation between the features belonging to the two sets and adopt the average mean correlation coefficient $\bar{r}$ to identify the level of correlation across the two sets with a unique parameter. Note that we first compute the absolute value of the correlation coefficients $r_{i,j}$ and then we calculate the mean, since we are interested in both decreasing and increasing linear relationships between the features. This analysis has the goal to bring to light if and which prosodic features are relevant for each task and what further prosodic functionals we should include in our manually selected features set.

4. Results

This section shows evaluation and feature analysis for the targeted tasks: typicality and diagnosis (Section 4.1), emotion, arousal, valence, and emotion-against-neutral (Section 4.2).

4.1. Typicality and diagnosis

For the classification of typicality and diagnosis, we perform the two tasks on the full database and on the focus group data set. Table 3 shows the best results obtained over the different complexities among the four feature sets. Applying the full set of features (IS12), we obtain up to 80.0% and 82.6% UAR for typicality and diagnosis, respectively. However, reducing the feature space led to an expected decrease of performance for both tasks; Figure 1 shows the trends over decreasing dimensional feature spaces. The correlation-based selected features set (IS12-CFS) performs quite close to the baseline (IS12).

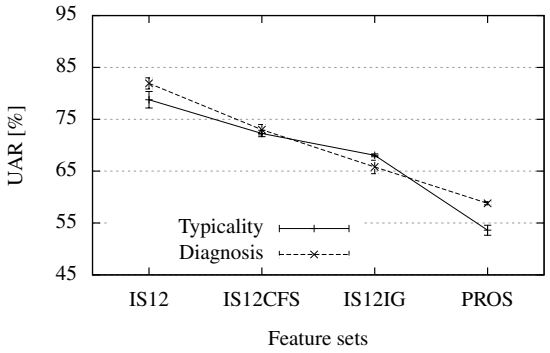
For the typicality task, the IS12-CFS set comprises mainly spectral (118) and voice quality (7) features; only 15 out of 140 features are functionals related to prosodic low level descriptors such as root-mean-square energy, fundamental frequency contour, sum of auditory spectrum, and also, zero-crossing rate and number of voiced segments, that can be considered as prosodic features to be added to our PROS set.

For the diagnosis task, the IS12-CFS contains spectral (88) and voice quality (3) features as well as 15 prosodic features (F0, RMS energy and sum of auditory spectrum) with further functionals that we did not include in the PROS set, such as quartile and range.

Then, we analysed the relationship between the IS12-IG and PROS sets. The average mean correlation coefficient, along with the standard deviation and the maximum absolute correlation value are given in Table 4. Concerning typicality the average mean correlation coefficient is very low, showing that the two feature sets are not highly correlated. In fact, the IS12-IG set comprises only spectral features. The typicality classification can obviously be better performed with spectral features than with only prosodic features, leading to 65.8% and 55.5% UAR, respectively, for IS12-IG and PROS.

Concerning diagnosis, the two feature sets are a bit more correlated since, in addition to spectral features (9), the IS12-IG comprises F0, auditory spectrum and root-mean-square energy features. In particular, the maximum absolute correlation value (1.0) holds for the 1st percentile of the sum of auditory spectrum and for the F0 standard deviation that are found in both feature sets. Diagnosis discrimination seems to rely on both prosodic and cepstral features; with the IS12-IG set and the PROS set, we achieved up to 74.0% and 59.3% UAR, respectively.

Figure 1: *Classification of typicality and diagnosis: Mean and standard deviation of UAR by average of complexity for the four different feature sets.*



4.2. Emotion related tasks

For emotion classification, we perform four different tasks: a 9-class emotion task, a 2-class arousal and valence task, and the 2-class task "e vs. Neutral", with $e \in \{$***Ha**ppy,* ***Su**rprised,* ***Pr**oud,* ***An**gry,* ***Af**raid,* ***Ca**lm,* ***Sa**d,* ***As**hamed*$\}$. All the tasks were performed both on the focus and on the control set separately. In addition to the classification, we further analyse the differences between the feature sets employed in our experiments: We adopt the same strategy as described for typicality and diagnosis discrimination, showing the best results achieved over the different complexities among the four feature sets with and without speaker z-normalisation (SN) (cf. Table 5). Since speaker normalisation led to better performances on all the tasks, we only show the speaker normalised performance trends over decreasing dimensional feature spaces in Figure 2 (emotion, arousal, and valence).

This section describes the evaluation and feature analysis for emotion (Section 4.2.1), arousal (Section 4.2.2), valence (Section 4.2.3), and emotion-against-neutral (Section 4.2.4). For all the four tasks, we first analyse the results obtained on the focus subset, and then those obtained on the control subset.

4.2.1. Emotion {9-class problem}

On the focus group, we observe the influence of speaker normalisation that improves UAR by over 4%, 10% and 8% absolute, respectively, for the IS12, IS12-CFS and PROS. Applying the full set of features (IS12), we obtain up to 42.6% UAR, how-

Table 3: *Unweighted Average Recall for typicality and diagnosis tasks, respectively, on the entire dataset and on the control group subset. Typicality classes: typically developing children (C), children with ASC (F). Diagnosis classes: Asperger Syndrome (AS), High-Functioning (HF). Shown is performance obtained using SVMs with linear kernel.*

UAR[%]	**IS12**	**IS12-CFS**	**IS12-IG**	**PROS**
Full data subset				
Typicality {**F,C**}	**80.0**	77.6	65.8	55.5
Focus group subset				
Diagnosis {**AS,HF**}	**82.6**	80.0	74.0	59.3

Table 4: *Correlation of IS12-IG and PROS features for typicality and diagnosis: average mean correlation coefficient ($\bar{r}$), standard deviation (stdev) and maximum absolute correlation coefficient (max).*

	$\bar{r}$	*stdev*	*max*
Typicality {**F,C**}	0.13	0.11	0.43
Diagnosis {**AS,HF**}	0.24	0.22	1.00

ever, reducing the features space led to an expected decrease of performance (cf. Figure 2a). The IS12-CFS set performs quite close to the baseline (cf. Table 5). It consists of spectral features (23), one voice quality feature, and four prosodic features related to voicing probability and energy, such as the sum of auditory spectrum with and without RASTA filtering. We further compare IS12-IG and PROS; the average mean correlation coefficient $\bar{r}$ is 0.5 (cf. Table 6) showing that the two feature sets are correlated to some extent; the IS12-IG set, in addition to 3 spectral features, comprises 12 energy features related to the sum of auditory spectrum in the different frequency bands. In particular 1st and 99th percentiles, and the standard deviation of the sum of auditory spectrum can be found in both feature sets, therefore the maximum absolute correlation coefficient is 1.0. Thus, in the focus group, the emotion task relies on prosodic features, in particular on energy features, leading to 23.9% and 28.9% UAR, respectively, for IS12-IG and PROS. Furthermore, the prosodic feature set performs better that the automatically selected set, showing that with such a small set of features, prosody can be relevant for this task.

On the control group set, UAR is improved by speaker normalisation over 9%, 20% and 4%, respectively, for the IS12, IS12-CFS and PROS sets (cf. Table 5). Applying the full set of features (IS12), we obtain up to 55.9% UAR; Figure 2a shows the performance trends over decreasing dimensional feature spaces. IS12-CFS is close to IS12 performances. IS12-CFS contains spectral features (33); only six prosodic features such as RMS energy, sum of auditory spectrum and F0, are comprised. A more detailed comparison between PROS and IS12-IG shows that the average mean correlation coefficient is below 0.5. IS12-IG consists of 5 energy features (auditory spectrum) and 10 spectral features; in particular 99th percentile and standard deviation of the sum of auditory spectrum are used in the two sets. The two feature sets lead to similar results: 16.6% and 18.8% UAR for IS12-IG and PROS.

4.2.2. Arousal {2-class problem}

An increase in performance can be obtained by speaker normalisation among the four feature sets and on the focus and control group. On the focus data set, UAR is improved up to 86% with IS12-CFS (cf Table 5). As in the previous tasks, the reduction of the feature space led to a decrease of performance as shown in Figure 2b. The IS12-CFS and the full feature set (IS12) perform similarly; the former one consists of 94 features, comprising a significant number of spectral features (77) and only few prosodic features, such as F0 standard deviation, 1st delta coefficient of the sum of auditory spectrum, and further functionals of root-mean-square energy. The two smaller feature sets (IS12-IG and PROS) yield quite similar performance: 81.4% and 78.8% UAR; this is corroborated by a medium average mean correlation coefficient (cf. Table 6). The IS12-IG set consists of energy features and 7 spectral features. In particular the 99th percentile and the standard deviation of the sum of auditory spectrum are also found in PROS set.

On the focus group subset, we obtain up to 90.0% UAR, with the IS12-CFS set (cf. Table 5). The correlation-based selected feature set (IS12-CFS) and the full feature set (IS12) perform close to each other; the IS12-CFS comprises 135 features, including a vast number of spectral features (114); only few features are related to prosody, such as F0, root-mean-square energy, the sum of auditory spectrum with and without RASTA filtering, voicing probability and zero-crossing rate. The IS12-IG and PROS sets perform similarly: 76.8% and 77.5% UAR. The average mean correlation coefficients is equal to 0.44 (cf. Table 6), showing that the two feature sets are medium correlated also in the control group. The IS12-IG set contains again energy features related to auditory spectrum and 9 spectral features. In addition to the 99th percentile and the standard deviation of the sum of auditory spectrum that are found also in the PROS set, we observe the presence of further functionals, such as range and percentile range and mean peak absolute values. Thus, in the two groups, the arousal task can rely on prosodic features without losing performance.

4.2.3. Valence {2-class problem}

On the focus group data set, the influence of speaker normalisation improves UAR by over 8%, 7%, and 5% absolute, respectively, for IS12, IS12-CFS, and IS12-IG. Applying the full set of features (IS12), we obtain up to 82.1% UAR. IS12-CFS performs close to the baseline (IS12) (cf. Table 5) and consists of 41 features, including spectral features and voice quality features, such as jitter and shimmer and only 4 prosodic features such as energy and F0. We observe a very low average mean correlation coefficient (cf. Table 6), meaning that the two feature sets are not correlated; in fact, the IS12-IG comprises mainly voice quality features (3) and spectral (10) features; only two prosodic features related to F0 are comprised.

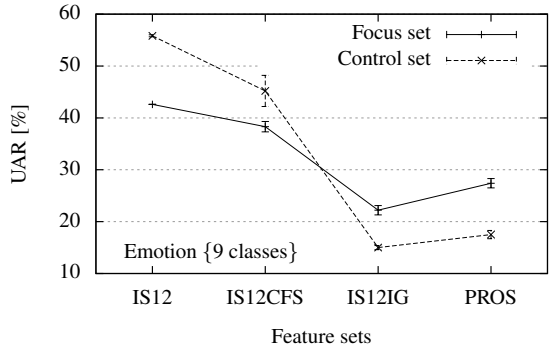
On the control group set, UAR is improved by speaker normalisation over 9% and 3% for the IS12 and the IS12-CFS feature sets. We obtain up to 81.8% UAR using the IS12 set, but reducing the feature space led to significant decrease in performances (cf. Figure 2c). IS12-CFS comprises 61 features, that are mainly spectral (58) and only 3 prosodic related to RMS energy and F0. As for the focus group, the average mean correlation coefficient (cf. Table 6) is low and, again, in IS12-IG the predominance of spectral features (15) is maintained. Thus, the valence task performs better with spectral and voice quality features than with prosodic features, achieving up to 72.0% and 64.4% UAR with the IS12-IG set, respectively, for the focus and the control group.

4.2.4. e-against-Neutral task {2-class problem}

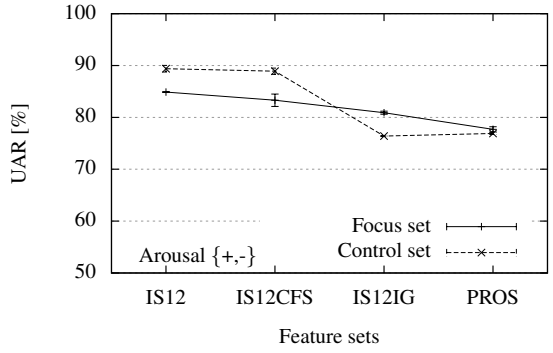
This task aims to classify each of the emotions (**Ha**ppy, **Su**rprised, **Pr**oud, **An**gry, **Af**raid, **Ca**lm, **Sa**d, **As**hamed) against the neutral emotional state. This is the heart of the matter because at the end of the day, we have to find out both classification performance and feature relevance for each emotion separately, and not for telling apart all emotions from neutral. For this task we only show results for the two smaller feature sets (IS12-IG and PROS). Now we want to find out if prosodic features alone can be employed for these tasks, and if not, what are the features to be taken from the automatically selected set. First we observe that speaker normalisation helps for almost all problems, on the two data sets (focus and control), and on both the IS12-IG and prosodic feature sets. UAR is improved up to 3.9% and 3.7% average absolute performance gain on the focus

Figure 2: *Classification of emotion, arousal, and valence: Mean and standard deviation of UAR by average of complexity for the four different feature sets with speaker normalisation.*

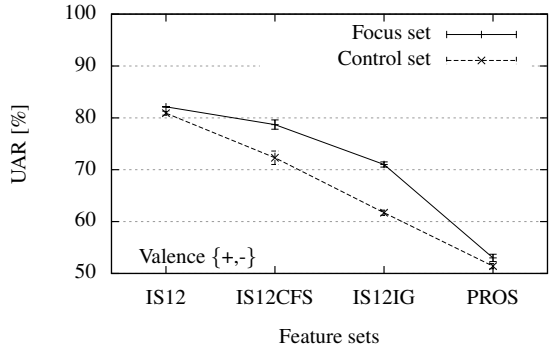
(a) Emotion {9-class}



(b) Arousal {+,-}



(c) Valence {+,-}



data set for IS12-IG and PROS, respectively (cf. Table 7). On the control data set, we achieve 3.3% and 5.8% average absolute performance gain for the two feature sets; note that due to the small size of the database, these differences are not significant and have to be corroborated in a later phase. Apparently, centring and scaling the feature space does not improve performance for the "Ashamed" task on the focus data set, and for the "Afraid" task on the control data set. Table 7 shows the best results obtained over the different tasks among the two feature sets on the focus and control groups. Applying the IS12-IG set, and by averaging the results over the eight emotions, we obtain 89.7% and 90.4% average UAR on the focus and control

Table 5: *Unweighted Average Recall for the 9 emotions task and for the arousal/valence tasks, for the focus group and the control group. Shown are the performances obtained with and without speaker z-normalisation (SN) using SVM with linear kernel.*

UAR[%]		**IS12**	**IS12-CFS**	**IS12-IG**	**PROS**
Focus group subset					
Emotion	–	39.0	31.1	24.3	21.0
{**9-classes**}	**SN**	**42.6**	41.4	23.9	28.9
Arousal	–	78.9	83.6	78.9	66.7
{**+,-**}	**SN**	84.9	**86.4**	81.4	78.8
Valence	–	73.8	72.7	67.4	57.4
{**+,-**}	**SN**	**82.1**	79.9	72.0	55.1
Control group subset					
Emotion	–	47.5	29.9	20.3	15.0
{**9-classes**}	**SN**	**55.9**	50.0	16.6	18.8
Arousal	–	85.1	82.3	61.1	63.9
{**+,-**}	**SN**	89.0	**90.0**	76.8	77.5
Valence	–	72.3	71.6	64.4	54.0
{**+,-**}	**SN**	**81.8**	74.2	62.4	52.4

Table 6: *Correlation of IS12-IG and PROS features for emotion, arousal, and valence tasks: average mean correlation coefficient ($\bar{r}$), standard deviation (stdev), and maximum absolute correlation coefficient (max).*

	$\bar{r}$	*stdev*	*max*
Focus group subset			
Emotion {**9-class**}	0.51	0.31	1.00
Arousal {**+,-**}	0.48	0.30	1.00
Valence {**+,-**}	0.15	0.14	0.54
Control group subset			
Emotion {**9-class**}	0.42	0.31	1.00
Arousal {**+,-**}	0.44	0.32	1.00
Valence {**+,-**}	0.16	0.12	0.42

group, respectively. Thus, on average this feature set led to similar performance for the two groups and it performs better than the PROS set on all the tasks. However, Figure 3a shows that the trends of the "Happy", "Surprised", "Angry" and "Afraid" tasks - which all belong to "high arousal" – are quite similar for the focus group across the two feature sets. We obtain up to 82%, 88%, 81% and 83.5% UAR with PROS for the above mentioned four tasks. The control group behaves in a similar way for three out of these four emotions, namely for "Happy", "Proud" and "Angry",(cf. Figure 3), displaying an UAR of up to 81.3%, 81.8%, and 84% with PROS. Thus we have seen, on the one hand, that prosodic features can be used to model at least those emotions which represent "high arousal", both for the focus and the control group. The low correlation between the PROS set and the automatically selected features for the typicality task in (Section 4.1), however, might indicate that the two groups indeed employ prosodic features in some different way – else, it probably would make no sense to speak about ASC children having problems in expressing emotions. In the following, we analyse the differences between the two feature sets (IS12-IG and PROS), separately for each emotion.

"Happy" and "Angry" – We grouped these two tasks together since they perform similarly with both the IS12-IG and PROS feature sets (cf. Figure 3a, Figure 3b, and Table 7). For

the "Happy" task, we achieve up to 90.4% and 82.3% UAR for the focus group, and up to 88.3% and 81.3% UAR for the IS12-IG and PROS sets, respectively. We observe a low to medium average mean correlation coefficient of 0.39 (cf. Table 8); for the focus group, the IS12-IG comprises mainly spectral features (12), and only three prosodic features related to energy and voice probability. The maximum absolute value of 0.91 is found for the correlation between the 3rd quartile and the arithmetic mean of the sum of auditory spectrum. For the control group, we observe a medium average mean correlation coefficient of 0.52, showing that the two feature sets comprise correlated features; in fact, 11 out of the 15 features model energy, in particular the sum of auditory spectrum with functionals such as percentile range, peak range, peak mean, 2nd and 3rd quartile, and arithmetic mean. The remaining 4 features consist of spectral features. For this task, prosodic features seem to be relevant, yielding comparable performance with respect to the automatically selected features. For the "Angry" task, we achieve up to 86.3% and 81.3% UAR for the focus group, and up to 89.8% and 84.1% UAR for the two feature sets, showing similar average mean correlation coefficients of 0.42 and 0.47 for the focus and the control group (cf. Table 8). For the focus group, the IS12-IG set mainly consists of spectral features with only three prosodic features (F0, energy and zero crossing rate). For the control group, it comprises nine energy features such as RMS energy and auditory spectrum features, and spectral features (6). In particular the 99th percentile and the standard deviation of the sum of auditory spectrum and the 99th percentile of RMS energy are found in both the IS12-IG and PROS feature sets. Thus, also in this task the prosodic features seem to perform quite close to the IS12-IG.

"Surprised" – We observe that for the focus group, the performance is quite similar across the two feature sets (cf. Figure 3a); IS12-IG consists of 11 spectral features, one voice quality feature, and three prosodic features related to F0 and energy. The average mean correlation coefficient shows a low level of correlation of 0.33, and the maximum absolute correlation values are found for the three prosodic features that IS12-IG comprises. We achieve up to 88.2% UAR with PROS. However, for the control group, PROS does not perform as well as IG12-IG (cf. Figure 3b). It also seems that speaker normalisation is not effective for the task; the two feature sets are not much correlated, and the IS12-IG set comprises spectral features, along with five prosodic features such as statistical functionals of F0 and energy. For this task, prosodic features are relevant only for the focus group, while for the control group, spectral and cepstral features perform better.

"Proud" – We achieve up to 89.9% and 72.4% for the focus group, and up to 84.5% and 81.8% UAR for IS12-IG and PROS, respectively. As shown in Figure 3a, the prosodic features yield lower results in comparison to those obtained with IS12-IG. However, for the control group, the prosodic set performs better leading to an UAR comparable to the automatically selected feature set; here, IS12-IG is slightly correlated to the prosodic set, comprising five prosodic features (mainly energy functionals), and 10 spectral features. For the focus group, the IS12-IG set consists of ten spectral features and four prosodic features related to RMS energy and the sum of auditory spectrum; the average mean correlation coefficient shows a very low correlation for the two sets. For this task prosodic features seem to be more relevant for the control group than for the focus group.

"Afraid" – This is the last task in which prosodic features perform quite close to the IS12-IG for the focus group (cf. Figure 3a). Moreover, Afraid is the last remaining emotion mapped onto positive arousal. For the focus group, we observe an average mean correlation coefficient of 0.4; the IS12-IG set comprises 11 spectral features, in addition to four prosodic features such as functionals of fundamental frequency contour and 3rd quartile of the sum of auditory spectrum. For the control group, the average mean correlation coefficient is lower (cf. Table 8); the IS12-IG set consists of voice quality (2) and spectral (7) features. Six prosodic features such as the 2nd quartile of F0 contour and five energy functionals are found in the feature set.

"Calm", "Sad" and "Ashamed" – These last there tasks show significant performance differences across the two feature sets. Morover, Calm, Sad and Ashamed are the three classes mapped onto low arousal. The IS12-IG set performs better for both the focus group and the control group (cf. Figure 3a, Figure 3b). We observe very low average mean correlation coefficients, and the automatically selected sets comprise only spectral, cepstral and voice quality features.

Table 7: *Unweighted Average Recall for "e-against-Neutral" classification task on both the focus and control group subsets. Shown are the performances obtained with and without speaker z-normalisation (SN) using SVM with linear kernel. Arousal (A) and Valence (V) are indicated according to the mentioned mapping.*

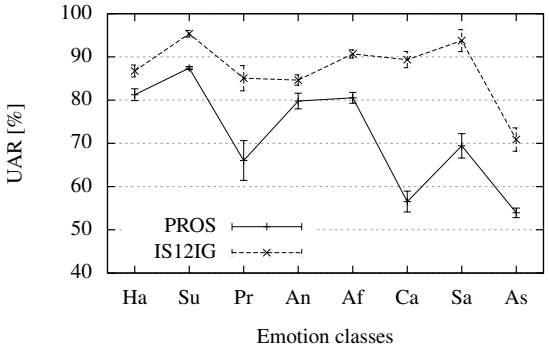
UAR[%]				**Focus**		**Control**	
	A	**V**	**features**	**–**	**SN**	**–**	**SN**
Happy	+	+	IS12-IG	87.0	**90.4**	82.1	**88.3**
			PROS	82.5	82.3	65.9	81.3
Surprised	+	+	IS12-IG	89.0	**96.1**	88.0	**88.5**
			PROS	81.9	88.2	72.2	68.9
Proud	+	+	IS12-IG	85.4	**89.9**	77.7	**84.5**
			PROS	53.8	72.4	71.1	81.8
Angry	+	-	IS12-IG	83.8	**86.3**	87.2	**89.8**
			PROS	74.1	81.3	73.2	84.1
Afraid	+	-	IS12-IG	85.4	**91.3**	**94.9**	88.6
			PROS	80.7	83.5	65.1	71.4
Calm	-	+	IS12-IG	82.5	**91.8**	83.5	**96.6**
			PROS	61.5	61.8	59.1	57.1
Sad	-	-	IS12-IG	89.7	**96.1**	**94.9**	94.3
			PROS	74.7	74.0	53.2	56.4
Ashamed	-	-	IS12-IG	**83.5**	75.9	88.6	**93.0**
			PROS	59.3	55.0	51.6	56.6

Table 8: *Correlation of IS12-IG and PROS features for the e-against-Neutral tasks: average mean correlation coefficient ($\bar{r}$), standard deviation (stdev) and maximum absolute correlation coefficient (max).*

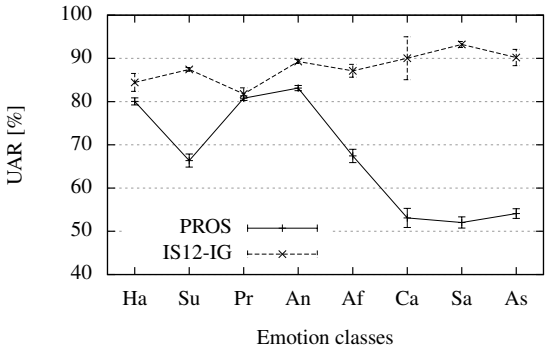
	Focus			**Control**		
	$\bar{r}$	*stdev*	*max*	$\bar{r}$	*stdev*	*max*
Happy	0.39	0.22	0.91	0.52	0.32	1.00
Surprised	0.33	0.22	0.96	0.35	0.29	1.00
Proud	0.29	0.22	0.97	0.41	0.30	1.00
Angry	0.42	0.31	0.92	0.47	0.34	1.00
Afraid	0.40	0.25	0.96	0.29	0.27	1.00
Calm	0.17	0.14	0.51	0.14	0.10	0.45
Sad	0.17	0.14	0.68	0.14	0.10	0.50
Ashamed	0.20	0.14	0.59	0.11	0.09	0.49

Figure 3: *Classification of "e-against-Neutral" task: Mean and standard deviation of UAR by average of complexity for the four different feature sets with speaker normalisation. Emotion classes: happy (Ha), surprised (Su), proud (Pr), angry (An), afraid (Af), calm (Ca), sad (Sa), ashamed (As).*

(a) "e vs. Neutral" $\{e$, Neutral$\}$ (focus group)



(b) "e vs. Neutral" $\{e$, Neutral$\}$ (control group)



5. Conclusions

Summing up, we first described the speech emotion database that has been used for evaluation; it is unique in composing speech data of children on the spectrum and a control group under the same conditions. Then, we discussed results concerning the classification of typicality and diagnosis, and of ASC children's emotional expressions, evaluating the 9-emotions task and the binary arousal/valence discrimination task. In addition, we classified each of the emotions against the neutral emotional state. Together with the classification evaluation, we analyse how prosodic features behave in the tasks. We focus on mainly three prosodic low level descriptors (energy, pitch and duration) with their basic functionals (mean, standard deviation, 1st percentile and 99th percentile), as these can be easily conveyed to the children and modified by them during the game. For example, the child can modulate his/her pitch in order to reach a target, or he/she has to increase or decrease energy to jump over an obstacle. Such intuitive and easy interaction would be hardly possible for spectral and cepstral features. Speaker normalisation increases performance for all the emotion related tasks, and this technique will be adopted also in the prototype of the ASC-Inclusion platform since we will incrementally collect more speech material from the same subject throughout the game. The caveat has to be made that this is a pilot study, with a rather small number of cases per class; the results will be reviewed, verified or falsified, with larger databases collected in the future. However, so far the results corroborate common wisdom, for instance, that prosody is more relevant if it comes to modelling arousal, and less relevant for modelling valence. ASC children seem to employ prosodic features, albeit in a different way. The correlation between the prosodic and the automatically selected feature sets is not very high but not low, either. Moreover, we can expect that by intentionally modulating and manipulating prosodic features, other acoustic parameters will change accordingly.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion).

7. References

[1] E. Mower, M. P. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *International Conference on Multimedia Computing and Systems/International Conference on Multimedia and Expo*, 2011, pp. 1–6.

[2] O. Golan and S. Baron-Cohen, "Systemizing empathy: Teaching adults with asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia," *Development and Psychopathology*, vol. 18, no. 02, pp. 591–617, 2006.

[3] J. Demouy, M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Périsse, D. Chauvin, S. Viaux, B. Golse, D. Cohen, and L. Robel, "Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment," *Research in Autism Spectrum Disorders*, vol. 5, no. 4, pp. 1402–1412, 2011.

[4] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal speech spectrum and increased pitch variability in young autistic children," *Frontiers in Human Neuroscience*, vol. 4, 2011, 7 pages.

[5] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: a critical review," *International Journal of Language & Communication Disorders*, vol. 38, pp. 325–350, 2003.

[6] N. Russo, C. Larson, and N. Kraus, "Audio-vocal system regulation in children with autism spectrum disorders," *Experimental Brain Research*, vol. 188, pp. 111–124, 2008.

[7] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA. Portland, OR: ISCA, September 2012, to appear.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 9th ACM International Conference on Multimedia, MM 2010*, ACM. Florence, Italy: ACM, October 2010, pp. 1459–1462.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[10] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-December 2010.