Speech, Emotion, Age, Language, Task, and Typicality: Trying to Disentangle Performance and Feature Relevance

Erik Marchi[1], Anton Batliner[1], Björn Schuller[1], Shimrit Fridenzon[2], Shahar Tal[2], Ofer Golan[2]
[1]Institute for Human-Machine Communication, Technische Universität München, D-80333 München, Germany
[2]Department of Psychology, Bar-Ilan University, Ramat Gan, Israel
erik.marchi@tum.de, anton.batliner@lrz.uni-muenchen.de, schuller@tum.de
shimfri@gmail.com, shahar0190@gmail.com, ofer.golan@biu.ac.il

***Abstract*—The availability of speech corpora is positively correlated with typicality: The more typical the population is we draw our sample from, the easier it is to get enough data. The less typical the envisaged population is, the more difficult it is to get enough data. Children with Autism Spectrum Condition are atypical in several respect: They are children, they might have problems with an experimental setting where their speech should be recorded, and they belong to a specific subgroup of children. Thus we address two possible strategies: First, we analyse the feature relevance for samples taken from different populations; this is not directly improving performances but we found additional specific features within specific groups. Second, we perform cross-corpus experiments to evaluate if enriching the training data with data obtained from similar populations can increase classification performances. In this pilot study we therefore use four different samples of speakers, all of them producing one and the same emotion and in addition, the neutral state. We used two publicly available databases, the Berlin Emotional Speech database and the FAU Aibo Corpus, in addition to our own ASC-Inclusion database.**

Keywords–Autism Spectrum conditions, speech emotion recognition, cross-corpus evaluation, feature analysis.

I. Introduction

The modelling, generation and recognition of emotion has attracted more attention in recent years. Researchers first dealt with prototypical emotions (elicited, acted or prompted), then with real-life data with spontaneous emotional speech. In particular, children's emotional spontaneous speech has been investigated in [1]. However little is known about emotional speech of children with voice and language impairments and with Autism Spectrum Conditions (ASC).

Three decades of research have shown that children and adults with 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 online. The computerised environment is especially appealing for individuals with ASC, due to its predictable, controllable, and structured nature, which enables them to use their strong systemizing skills. Existing systems, such as the Rachel Embodied Conversational Agent (ECA) [2] and the Mind-Reading software [3], 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 past ICT solutions, the project will address the recognition and the expression of socio-emotional cues by providing an interactive game that scores the prototypicality and 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. In a previous study, we focused on the recognition of emotional vocal expressions and on feature 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 [4]. The importance of prosody with respect to several aspects of voice and language impairment in Autism Spectrum Conditions is addressed in [5], [6], [7], [8], [9], [10].

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 parameters 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.

The availability of speech corpora is positively correlated with typicality: The more typical the population is we draw our sample from, the easier it is to get enough data. The less typical the envisaged population is, the more difficult it is to get enough data. Children with Autism Spectrum Condition are atypical in several respect: They are children, they might have problems with an experimental setting where their speech should be recorded, and they belong to a specific subgroup of children.

It is therefore worth while to address two possible strategies: First, to try and enrich the training data with data obtained from similar populations that only differ 'slightly' from the target population; for instance, we can try and add speech data from children producing the same emotion but belonging to different languages. Our interest is whether by that, we can improve classification performance. Second, we can have a look at feature relevance for samples taken from different populations, while the speakers produce the same emotions. This will not directly improve performance but we might get some information on the factors that trigger the use of specific features within specific sub groups.

In this pilot study, we therefore want to use four different samples of speakers, all of them producing one and the same emotion and in addition, the neutral state. We are constrained to use those databases that are easily and freely available; moreover, these databases should contain at least one identical emotion – this state we want to keep constant across databases. We decided in favour of the following three databases: the Berlin Emotional Speech database (EMO-DB) [11], the FAU Aibo Corpus (FAU-AIBO) [1], and our own ASC-Inclusion database (ASC-DB).

The article is structured as follows: First, a detailed description of the three databases is given (Section II); then we define experimental tasks, features and set-up (Section III). We next present evaluation results (Section IV) before concluding the paper in Section V.

II. Databases

We decided to adopt three databases (EMO-DB, FAU-AIBO and ASC-DB). They contain at least one identical emotion, namely Anger, along with Neutral. They feature differences with respect to contents, population, and type. EMOD-DB contains acted emotion recordings, performed by adults in a studio environment in German. The FAU-AIBO comprises spontaneous emotional speech, recorded while children are interacting with a pet robot. The ASC-DB contains acted emotional speech of two groups of children: typically developing (control group: ASC-C) and children with ASC (focus group: ASC-F). In this section we describe the three databases used for our evaluations: the ASC-Inclusion database (Section II-A), the FAU Aibo Corpus (Section II-B), and the Berlin Emotional Speech database (Section II-C).

A. 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 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 5 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 five "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 (later dowsampled to 16kHz) 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 [3]. 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!"*.

TABLE I: ***Age/Group*** *(adults or children and typical or atypical).* ***Content*** *of speech (fixed/variable). Number of utterances per emotion category (**# Emotion**); Emotion classes: angry (**An**), neutral (**Ne**). Overall number of turns (**# All**). Number of subjects (**# Sub**), number of female (f), number of male(m) subjects.* ***Type*** *of material (acted/natural) and recording conditions (studio/normal/noisy). Sampling* ***Rate****.*

Database	Age Group	Content	# Emotion An	# Emotion Ne	# All	# Sub	Type	Rate kHz
EMO-DB	adults typical	German fixed	127	78	205	6f 4m	acted studio	16
FAU-AIBO	children typical	German variable	165	230	395	6f 5m	natural normal	16
ASC-C	children typical	Hebrew fixed	38	40	78	5f 5m	acted noisy	16
ASC-F	children atypical	Hebrew fixed	16	16	32	1f 3m	acted noisy	16

Sad - Afterwards he entered the kitchen. He noticed his mother was preparing a simple breakfast for him and 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 with 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 only 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 developing 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 produced, respectively, by children with Asperger syndrome and high-functioning diagnosis. The remaining 351 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.

For our experiments we only used the utterances related to Anger and Neutral. Morover, we left out three speakers in the focus group since they only produced less than three angry utterances and no neutral utterance. Table I shows the number of utterances for the classification task.

B. FAU Aibo Emotion Corpus

The second database that we employ for our experiments is the FAU Aibo Emotion Corpus, a corpus with recordings of children interacting with Sonys pet robot Aibo. The corpus consists of spontaneous, German speech that is emotionally coloured. The children were led to believe that the Aibo was responding to their commands, whereas the robot was actually controlled by a human operator. The wizard caused Aibo to perform a fixed, predetermined sequence of actions;

sometimes Aibo behaved disobediently, thereby provoking emotional reactions. The data were collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). Speech was transmitted with a high quality wireless head set and recorded with a DAT-recorder (16 bit, 48 kHz down-sampled to 16 kHz). The recordings were segmented automatically into turns using a pause threshold of 1 s. Five labellers (advanced students of linguistics) listened to the turns in sequential order and annotated each word independently from each other as neutral (default) or as belonging to one of ten other classes. Since many utterances are only short commands and rather long pauses can occur between words due to Aibos reaction time, the emotional/emotion-related state of the child can change also within turns. Hence, the data are labelled on the word level. We resort to majority voting (MV): If three or more labellers agreed, the label was attributed to the word. In the following, the number of cases with MV is given in parentheses: joyful (101), surprised (0), emphatic (2,528), helpless (3), touchy, i. e. irritated (225), angry (84), motherese (1,260), bored (11), reprimanding (310), rest, i. e. non-neutral, but not belonging to the other categories (3), neutral (39,169); 4,707 words had no MV; all in all, there were 48,401 words. Classication experiments on a subset of the corpus [1] showed that the best unit of analysis is neither the word nor the turn, but some intermediate chunk being the best compromise between the length of the unit of analysis and the homogeneity of the different emotional/emotion-related states within one unit. Hence, manually defined chunks based on syntactic-prosodic criteria [1] are used here.

For the INTERSPEECH 2009 Emotion Challenge [12] the whole corpus consisting of 18,216 chunks was mapped onto five category labels: Anger (A), which included angry, touchy and reprimanding; Emphatic (E); Neutral (N); Positive (P), which included motherese and joyful; and Rest (R).

For our experiments, we selected utterances in the Anger category that had a rating better than 0.7 and equal to 1 for the Neutral emotional state. According to this first filtering, we selected only those speakers that produced at least 7 Anger utterances and less than 40 Neutral utterances, in order to avoid high unbalanced class distribution. Table I shows the number of utterances for the classification task.

C. Berlin emotional speech database

The third database chosen for our evaluations is the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [11]. It covers *anger, boredom, disgust, fear, joy, sadness and neutral* as emotions. Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances like *"Der Lappen liegt auf dem Eisschrank"* (*"The cloth is lying on the fridge"*) in all seven emotional states. The recordings took place in the anechoic chamber of the Technical University of Berlin, using a Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder. Recordings were done with a sampling frequency of 48kHz and later downsampled to 16kHz. The whole set comprises around 800 phrases. 494 utterances with a recognition rate better than 80% and a naturalness over 60% were selected through a listening experiment at which 20 subjects took part.

For our experiments, we selected utterances in the Anger and Neutral categories. Table I shows the number of utterances for the classification task.

III. Experiments

In this part we describe the classification tasks in Section III-A, the feature sets in Section III-B, the experimental set-up in Section III-C, and our evaluation and analysis criteria in Section III-D.

A. Tasks

One task was evaluated on each of the databases separately. The task concerns the classification of Anger against Neutral. We choose anger since it is the only emotional state that can be found in all the three databases along with Neutral. Thus, we analyse the differences in classification performances and in the feature sets across the different content types and populations. The task was performed on the selected sets shown in Table I.

Then we perform cross-corpus evaluation in order to analyse the behaviour of the different population samples to see if we could enrich the training data with data obtained from similar groups that only differ 'slightly' from the target population.

B. 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** comprising 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 two feature sets: IS12-IG and PROS. The **IS12** features set, from the INTERSPEECH 2012 Speaker Trait Challenge [13], contains 6128 features (84.6% spectral, 9.4% prosodic and 6% voice quality) and is taken as large feature set on which we perform feature selection since it contains a great variety of functionals and low level descriptors. We applied feature selection to IS12 by measuring the information gain (**IS12-IG**) and 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.

While for the within-corpus experiments, we perform feature selection separately on each database, for the cross-corpus evaluations, we merged all the three databases and then we perform feature selection on the whole dataset obtaining the **IS12-IGA** feature set. In this way we obtained a unique feature set to be used for the comparison of cross-corpus performances across the automatically selected features and our prosody feature set.

C. Setup

Since some of the data sets (EMO-DB and FAU-AIBO) 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 [14], 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 $C \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. For the within-corpus experiments, in order to ensure speaker independent evaluations, we performed Leave-One-Speaker-Out (LOSO) cross-validation.

For the cross-corpus evaluations, we train on one database and test on the left out ones. 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 [15]. With such a method, the feature values are normalised to a mean of zero and a standard deviation of one for each speaker.

D. Evaluation

In the within-corpus scenario, we first perform classification experiments using 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 negative and positive linear relationships between the features.

In the cross-corpus scenario, we perform classification experiments using the IS12-IGA and PROS feature sets, and we describe the difference across the feature sets.

IV. Results

This section shows evaluation and feature analysis for the targeted task in the two scenarios: within-corpus (Section IV-A) and cross-corpus (Section IV-B).

A. Within-corpus

For the classification of Anger-against-Neutral within the databases, we perform the task on the selected data sets (cf. Table I). Table II shows the best results obtained over the different complexities among the two feature sets on each data set. The table contains results obtained with speaker normalisation since it performs better on all the datasets. Applying the IS12-IG set, we obtain up to 99.8%, 86.7%, 90.4% and 95.3% UAR for EMO, AIBO, ASC-C and ASC-F datasets, respectively. We observe that the PROS set led to very similar performances for the EMO and AIBO data sets, while for the focus and control group data sets, the results are lower, in particular for the focus group. Figure 1 displays differences and performance trends over the four data sets. In order to investigate and explain the performance differences, we analyse the relationship between the feature sets, looking at feature relevance among different populations. In the following subsections we focus on each data set separately: EMO-DB (Section IV-A1), FAU-AIBO (Section IV-A2), and ASC-Inclusion control (Section IV-A3) and focus (Section IV-A3) data sets.

1) EMO-DB data set: Within the EMO-DB data set, the IS12-IG set comprises mainly spectral features (12) and only three prosodic features such as first, second and third quartile of F0 contour. However, we observe a medium average mean correlation coefficient of 0.52 (cf. Table III) showing that the two feature sets comprise correlated features. In particular we obtain the maximum absolute correlation value of 0.99 between the above mentioned F0 quartiles and F0 arithmetic mean; spectral features such as spectral roll off and harmonicity show medium-high correlation values with

TABLE II: *Unweighted Average Recall for Anger-against-Neutral task, on the four selected datasets: EMO-DB, FAU-AIBO, ASC-F (focus group) and ASC-C (control group). Shown is performance obtained using SVMs with linear kernel.*

UAR[%] Task {An,Ne}	**EMO**	**AIBO**	**ASC-C**	**ASC-F**
IS12-IG	99.8	86.7	90.4	95.3
PROS	99.3	87.1	86.6	78.1

TABLE III: *Correlation of IS12-IG and PROS features for Anger-against-Neutral task: average mean correlation coefficient ($\bar{r}$), standard deviation (stdev) and maximum absolute correlation coefficient (max).*

	$\bar{r}$	*stdev*	*max*
EMO	0.52	0.25	0.99
AIBO	0.55	0.33	1.00
ASC-C	0.47	0.34	1.00
ASC-F	0.37	0.24	0.96

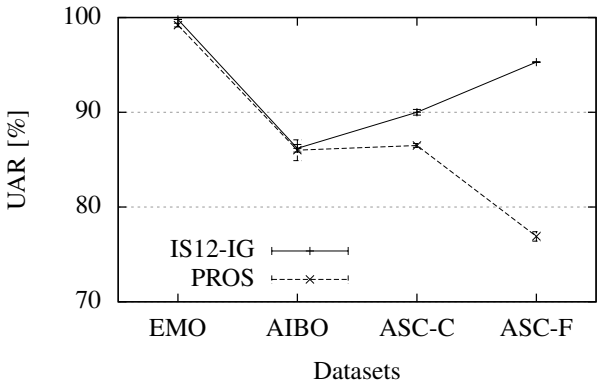
F0 and Energy functionals. Moreover, the correlation values corroborate the performance trends across the IS12-IG and PROS sets, showing a very low absolute difference of 0.5% UAR (cf. Table II). Thus, we observe that anger-against-neutral discrimination performs similarly on the two set (cf. Figure 1), showing the relevance of pitch and energy features.

2) AIBO data set: For the AIBO data set, the IS12-IG set comprises unexpectedly only prosodic features such as mean, standard deviation, percentiles and flatness of RMS energy and sum of auditory spectrum. Thus intensity and loudness are very relevant for discriminating Anger within this data set. In fact, we observe a medium to high average mean correlation coefficient of 0.55 (cf. Table III) showing that the two feature sets comprise highly correlated features. In particular, we observe identical features in the two feature sets such as standard deviation, arithmetic mean, 1st and 99th percentile of RMS energy, and standard deviation of the sum of auditory spectrum; this is confirmed by a absolute maximum correlation coefficient of 1.0 (cf. Table III). The correlation analysis follows the performance trends across the IS12-IG and PROS sets (cf. Figure 1), showing that PROS performs a little bit better than IS12-IG (cf. Table II). Thus, we observe that the task discrimination can obviously rely only on prosodic features; energy and loudness are the most relevant features to look at. This is important since this population includes **children's** spontaneous speech and it confirms, for this task, that only prosodic features can be used for classification and as potential consistent parameters to convey to the children.

3) ASC-C data set: On the control group data set, we obtain up to 90.4% and 86.6% UAR with IS12-IG and PROS (cf. Table II), respectively. Here, the IS12-IG set comprises a balanced number of spectral (7) and prosodic (8) features, such as arithmetic mean, standard deviation, 99th percentile, percentile range and quadratic regression error (quadratic error between contour and quadratic regression line) of RMS energy, and range of sum of auditory spectrum. Three of them (RMS energy mean, standard deviation and 99th percentile) are present in the two feature set and thus for them we observe absolute maximum correlation coefficient of 1.0. A low to medium average mean correlation coefficient of 0.47 (cf. Table III) is observed, showing that for this task the IS12-IG that combines spectral and prosodic features led to better performance trend (cf. Figure 1). However, the PROS set performs quite close to the IS12-IG and it can be enriched with the prosodic features that have been automatically selected.

4) ASC-F data set: On the focus group data set, we observe a marked difference in performance between IS12-IG and PROS, which led up to 95.3% and 78.1% UAR, respectively (cf. Table II). The IS12-IG mainly consists of spectral features (10) and includes one voice quality (shimmer) and four prosodic features such as quadratic regression offset of F0 contour, linear regression coefficient of RMS energy, and linear regression coefficient and mean peak distance of the sum of auditory spectrum. The low average mean correlation coefficient of 0.37 (cf. Table III) shows that IS12-IG and PROS are not highly correlated; this corroborates the marked difference in performance trends (cf. Figure 1). On this data set, the task classification seems to rely more on spectral features, but the above mentioned prosodic functionals can be employed to improve the PROS set performances.

Fig. 1: *Classification of Anger-against-Neutral: Mean and standard deviation of UAR by average of complexity for the four different datasets.*



B. Cross-corpus

For the classification of Anger-against-Neutral across the databases, we perform the task training on one of the selected

data sets (cf. Table I), and test on the left out ones. Table IV shows the performance obtained with respect to the different combination of train and test sets. We show evaluation performed with and without speaker normalisation along with the absolute difference of UAR between the PROS set and the IS12-IGA set, in order to gain more insight into the use of prosodic features within different populations. The best performances within the feature sets are highlighted in bold.

Testing with the Emo-DB data set shows that better performances are obtained without speaker normalisation when using IS12-IGA. This can be related to the fact that Emo-DB is the most dissimilar population with respect to age (adults) and type of content (acted/studio). For this reason, centring and scaling the feature space can flatten the differences between adult speech and children's speech resulting in a decrease of performances. Applying the PROS set we observe that the performances go down to 62.4%, 65.0% and 67.1% UAR when training with AIBO, ASC-C and ASC-F, respectively. This can be explained considering that IS12-IGA comprises mainly spectral features (only one is prosodic), and the two feature set are not highly correlated.

Testing with the AIBO data set shows quite similar performances among the two feature sets, even if IS12-IGA again performs better. We observe that the lowest results are obtained when training with Emo-DB, achieving 83.5% and 77.1% for IS12-IG and PROS, respectively. This confirms the dissimilarities across Emo-DB and the other data set. Moreover, speaker normalisation is effective only when training with ASC-C and ASC-F.

Using ASC-C as test set shows again the dissimilarity of performance between IS12-IGA and PROS when training with Emo-DB. However, we observe similar performances when training with AIBO and ASC-F.

Testing on the ASC-F set with the IS12-IGA feature set, we achieve up to 84.3% and 85.6% UAR when training with AIBO and ASC-C, respectively. Very similar results are obtained using the PROS set, and we outperform the performance achieved on the within-corpus scenario that led up to 78.% UAR. In fact, with the PROS feature set, we achieve up to 84.4% and 84.6% when training with AIBO and ASC-C, respectively. This seems to be promising for simply enriching training databases with more data from different but similar databases.

Lastly, computing the average of UAR, we observe that speaker normalisation led to better performances both for IS12-IGA and PROS (cf. Table IV), and that the loss of performance across databases with children's speech is low.

TABLE IV: *Unweighted Average Recall for "Anger-against-Neutral" classification task on all the combination of test and train datasets. Shown are the performances obtained with and without speaker z-normalisation (SN) using SVM with linear kernel. The absolute difference (Δ) of UAR across the IS12-IGA and PROS feature sets.*

UAR[%]		IS12-IGA		PROS		Δ	
Test set	**Train set**	**-**	**SN**	**-**	**SN**	**-**	**SN**
EMO	**AIBO**	**96.8**	89.0	50	**62.4**	-46.8	-26.6
	ASC-C	**99.4**	88.5	**65.0**	60.1	-34.4	-28.4
	ASC-F	**98.72**	88.3	53.8	**67.1**	-45.0	-21.3
AIBO	**EMO**	**83.5**	72.5	**77.1**	67.3	-6.4	-5.2
	ASC-C	**85.6**	85.2	70.2	**83.7**	-15.4	-1.5
	ASC-F	83.5	**84.7**	63.5	**78.3**	-20.0	-6.4
ASC-C	**EMO**	75	**78.1**	62.5	**62.5**	-12.5	-15.6
	AIBO	75	**81.3**	62.5	**81.3**	-12.5	0.0
	ASC-F	**81.3**	71.9	59.4	**75**	-21.9	3.1
ASC-F	**EMO**	64.8	**83.6**	57.8	**62.0**	-7.0	-21.6
	AIBO	70.2	**84.3**	64.7	**84.4**	-5.5	0.1
	ASC-C	66.3	**85.6**	74.4	**84.6**	8.1	-1.0
	Average	81.7	**82.8**	63.4	**72.4**	-18.3	**-10.4**

V. Conclusion

We investigated the classification of Anger-against-Neutral, evaluating the task on two different scenarios: within-corpus and cross-corpus. 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 almost 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, we found some additional prosodic features in the IS12-IG set that we did not envision in our manually selected prosodic feature set, gaining more insight into the use of acoustic and prosodic parameters within different populations. Moreover, the results corroborate common wisdom, for instance, that prosody is relevant if it comes to modelling Anger. ASC children seem to employ prosodic features, albeit in a different way. Lastly, we found that training on similar datasets (AIBO and ASC-C) can outperform the results obtained on the ASC-F data set. This seems to be promising for simply enriching training databases with more data from different but similar corpora.

Coming back to the title of this paper: We are yet far away from effectively disentangling the possibly intervening factors mentioned in the title: "Speech, Emotion, Age, Language, Task, and Typicality". Emo-DB differs most from all other databases, cf. Table 1, and yields pronunced performance differences, cf. the absolute differences displayed in Table 4, both for employing EMO-DB as train or test set. In contrast,

FAU-Aibo used as either train or test for ASC-C and ASC-F does not result in markedly lower performance. This might indicate that it will be more promising to use children's speech than adults' speech for enriching training databases.

ACKNOWLEDGMENT

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).

REFERENCES

[1] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," 2008. [Online]. Available: http://d-nb.info/992551641/04

[2] 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.

[3] 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.

[4] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, Shahar, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: Prosody and everything else," in *Workshop on Child, Computer and Interaction (WOCCI), Satellite Event of INTERSPEECH*, 2012, to appear.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] D. V. Lancker, C. Cornelius, and J. Kreiman, "Recognition of emotional-prosodic meanings in speech by autistic, schizophrenic, and normal children," *Developmental Neuropsychology*, vol. 5, pp. 207–226, 1989.

[10] R. Paul, A. Augustyn, A. Klin, , and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 35, pp. 205–220, 2005.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH*. ISCA, 2005, pp. 1517–1520.

[12] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing*, vol. 53, no. 9/10, pp. 1062–1087, November/December 2011.

[13] 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.

[14] 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

[15] 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.