Corpus of Children Voices for Mid-level Markers and Affect Bursts Analysis

Marie Tahon, Agnes Delaborde, Laurence Devillers
Department of Human Communication, LIMSI-CNRS,
91 403 Orsay, France
Department of Computer Sciences, University Paris 11,
91 405 Orsay, France
E-mail: mtahon@limsi.fr, agdelabo@limsi.fr, devil@limsi.fr

Abstract

This article presents a corpus featuring children playing games in interaction with the humanoid robot Nao: children have to express emotions in the course of a storytelling by the robot. This corpus was collected to design an affective interactive system driven by an interactional and emotional representation of the user. We evaluate here some mid-level markers used in our system: reaction time, speech duration and intensity level. We also question the presence of affect bursts, which are quite numerous in our corpus, probably because of the young age of the children and the absence of predefined lexical content.

Keywords: Audio Signal Processing, Emotion Detection, Human-Robot Interaction

1. Introduction

In the context of Human-Robot Interaction, the robot usually evolves in real-life conditions and then faces a rich multimodal contextual environment. While spoken language constitutes a very strong communication channel in interaction, it is known that lots of information is conveyed nonverbally simultaneously to spoken words (Campbell, 2007). Experimental evidence shows that many of our social behaviours and actions are mostly determined by the display and interpretation of nonverbal cues without relying on speech understanding. Among social markers, we can consider three main kinds of markers: interactional, emotional and personality markers. Generally-speaking, social markers are computed as long-term markers which include a memory management of the multi-level markers during interaction. In this paper, we focus on specific mid-level and short-time acoustic markers: affect bursts, speech duration, reaction time and intensity level which can be used for computing the interactional and emotional profile of the user.
In a previous study, we have collected a realistic corpus (Delaborde, 2010a) of children interacting with the robot Nao (called NAO-HR1). In order to study social markers, we have recorded a second corpus (called NAO-HR2), featuring children playing an emotion game with the robot Nao. The game is called interactive story game (Delaborde, 2010b). So far, there exist few realistic children voices corpora. The best known being the AIBO corpus (Batliner, 2004), in which children give orders to the Sony’s pet robot Aibo. Two corpora were collected for studying speech disorders in impaired communication children (Ringeval, 2008). In both studies, there are no spoken dialogs with robots; only the children are speaking.

Many previous studies focus on one of the three social markers. Interactional markers can be prosodic as in (Breazeal, 2002): five different pitch contours (praise, prohibition, comfort and attentional bids and neutral) learnt from infant-mother interaction are recognised by the Kismet robot. Mental state markers can also be only linguistic as the number of words, the speech rate (Kalman, 2010). Personality markers can be linguistic and prosodic cues (Mairesse, 2007). Emotional markers can be prosodic, affect bursts and also linguistic. The concept of “affect bursts” has been introduced by Scherer. He defines them as “very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events” (Scherer, 1994). Affect bursts are very important for real-life interactions but they are not well recognized by emotion detection systems because of their particular temporal pattern. Schröder (2003) shows that affect bursts have a meaningful emotional content. Our hypothesis is that non verbal events and specific affect bursts production are important social cues during a spontaneous Human-Robot Interaction and probably even more with young children.

Section 2 presents the protocol for collecting our second children emotional voices corpus. The content of the corpus NAO-HR2 is described in Section 3: affect bursts, speakers and other interactional information. Section 4 summarizes the values we can expect for some mid-level social cues. Finally, Section 5 presents our conclusion and future work.

2. Data collection

2.1 Interactive Story Game

We have collected the voices of children playing with the robot Nao and recorded with lapel-microphone. Nao told a story, and two children in front of it where supposed to act the expected emotions in the course of the story.
A game session consists in 3 phases: first the robot explains the rules and suggests some examples, the second part is the game itself, and the last part is a questionnaire proposed by an experimenter. The children are presented a board, on which words or concepts are drawn and written (such as “house”, or “poverty”). Emotion tags are written in correspondence for each of this word. The player

number one knows that, for example, if the notion "poverty" occurs in the course of the story, he will have to express sadness. He can express it the way he wants: he can speak sadly, or do as though he was weeping; children were free to interpret the rules as they wanted to. Once the rules are understood by the two players, Nao starts to tell the story. When it stops speaking, one of the players is supposed to have spotted a concept in the previous sentence, and is expected to play the corresponding emotion. If the robot detects the right emotion, the child wins one point.

2.2 Semi-automatic Human-Robot Interaction System

The behaviour of the robot changes in the course of the game. It can be neutral, just saying "Your answer is correct", or "not correct". It can also be empathic "I know this is a hard task", etc. Fuzzy logic rules select the most desirable behaviour for the robot, according to the emotional and interactional profile of each child, and their sex. This profile is built according to another set of fuzzy logic rules which process the emotional cues provided manually by the Wizard experimenter. The latter provides the system with the emotion expressed by the child (a label such as "Happiness", "Anger", "Sadness", etc.), the strength of the emotion (low, average or high activation), the elapsed time between the moment when the child is expected to speak and the time he starts speaking, and the duration of the speaking turn (both in seconds). From these manually captured cues, the Human-Robot Interaction system builds automatically an emotional and interactional representation of each child, and the behaviour of the robot changes according to this representation.

The dynamic adaptation of the behaviour of the robot and the design of the profile, based on a multi-level processing of the emotional audio cues, are explained in (Delaborde, 2010b). Table 1 gives an overview of the different level of processing of the emotional audio signal: from low level cues computed from the audio signal, to high level markers such as emotions, emotional tendencies, and interactional tendencies.

Low-level Cues	Mid-level Cues	High Level Social Markers
• Intensity level • Prosody • Spectral envelope	• Affect bursts (Laughs, hesitation, 'grr') • Speech duration • Reaction Time • Speaking rate	• Emotion (label, dimension) • Interactional tendencies (e.g. dominance) • Emotional tendencies (e.g. extraversion)

Table 1: Multi-level cues and social markers

The collected audio data is subsequently processed by expert labellers. On each speaker's track, we define speaker turns called instances. The annotation protocol is described in detail in (Delaborde, 2010b). The annotation scheme consists in emotional information (labels, dimensions and affect bursts), but also mental-state and personality information based on different time windows. In this paper, we focus on the study of affect bursts and others mid-level markers such as reaction time, duration but also the low-level marker intensity.

3. Contents of NAO-HR2 corpus

3.1 Description of the corpus

The NAO-HR2 corpus is made up of 603 emotional segments for a total amount of 21mn 16s. Twelve children (from six to eleven years old) and four adults have been recorded (five boys, seven girls, one woman and three men).

For this study, we have selected only the speech instances which occur during the story game (not during the questionnaire). In consequence, we obtain 20 emotional answers per gaming session: 10 emotional answers for each speaker. In that way the number of speaker turns is quite similar from one speaker to another.

3.2 Affect bursts

An annotation tag indicates the presence or absence of an affect burst in the instances. We notice that a large majority of the corpus is made up of affect bursts.

Table 2 summarizes the number of affect bursts (AB) over the total number of instances (TT) for each group of speaker. We have separated the children in two groups of 5 according to their age: the younger are from 6 to 7 years old, the older over 8 year old.

	# AB (TT)	Mean AB (TT) per speaker
Adults	12 (114)	3.0/17.3
Children (6-7 y.o.)	30 (85)	6.0/17.0
Children (8-11 y.o.)	19 (80)	3.8/16.0

Table 2: Affect bursts (AB) compared to the total (TT) number of instances

From these results we can conclude that asking a participant to express an emotion without any predefined lexical content leads to a high number of affect bursts. Children seem to use more often affect bursts than adults and young children even more. It seems that they are not at ease with finding words to express an emotion. Both children and adults express happiness laughing, but only children use "grr" affect bursts for anger in our corpora. Expressions of fear are usually more affect bursts for children than for adults. Affect bursts usually contain only a single phoneme; it is not possible to compute easily a speaking rate.

4. Results on Social Markers

In this section, we have manually measured the different markers in all game sessions.

An example is shown in Figure 1. Nao says: "*a lot of sadness*", the word "sadness" is one of the keywords written on the board and the child has to express the corresponding emotional state which is sadness. The four social markers we are studying, are represented in red: reaction time is 4.42s, speech duration is 2.17s, mean intensity is 52.83dB (after normalization: 28.67dB) and mean Harmonics-to-noise Ratio is 10.95dB. Reaction Time is important for this turn; the mean value of this 10 year old boy is 3.07s. Intensity and HNR are also lower than the mean values obtained on his whole session (Intensity mean is 32.43dB and HNR mean is 12.56dB). Intensity and HNR values correspond to what is expected when acting sadness; a high reaction time probably means that the boy was not at ease with this specific turn.

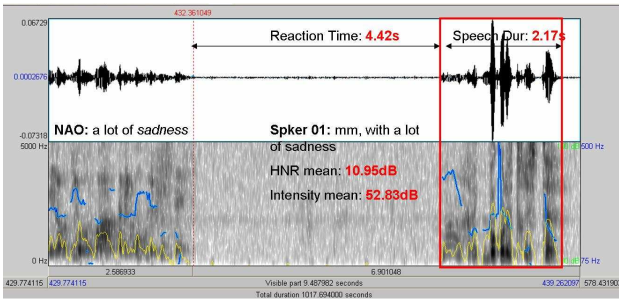

Figure 1: An example of social markers during the story game, the markers are collected with Praat

4.1 Reaction Time

The reaction time (RT) represents the interval between the time when the speaker is expected to speak (when Nao stops telling the story), and the time he indeed starts to speak. In the context of our game, the children were not supposed to call up their knowledge, or to think about the best answer. They were supposed to act the emotion written on the board. The longer the reaction time, the more the speaker postpones the time of his oral production. This parameter is one of the parameters used for the definition of the dimension "self-confidence" of the emotional profile. The shorter the reaction time, the more the speaker tends to be self-confident. Table 3 presents the mean and standard deviation of mean reaction times for each child.

Mean RT (s)	Std RT (s)
4.62	2.00

Table 3: Reaction Time

Some children are not at ease with the game, and their RT is much more important than the other (RT = 7.73 for children n°12, 6 year old). When the RT value is so high it often means that the children did not find any answer to give to NAO in the time he has to (if the child did not answer after 12.5s, the robot continues the story). Hesitation is quite used by children who have an important RT.

4.2 Estimation of Speech Duration

The speech duration (SD) is another parameter used for the emotional profile of the speaker. It corresponds to the duration of speech of the speaker, for each speaking turn. Children included small pauses (from 850ms to 1.40s) in their speech. These short silences are not considered as ends of speaking turn: it can be breathing, hesitating, thinking, and the speaker resumes speaking.

Mean SD (s)	Std SD (s)
2.01	1.30

Table 4: Speech Duration for each turn

We notice in table 4 that the mean SD is generally quite short. The turns are mostly composed of one single syllable. As we have seen before the proportion of affect bursts is quite important and most of them have short durations. As the players do not have any lexical support except what Nao have just said, they are not simulated to speak a lot.

4.3 Estimation of Intensity

For each session, both children were recorded with separate microphones which have their own gain. We compute the mean intensity (Int) normalized to the noise value for each session. It is also possible to estimate the HNR value on voiced parts only.

Hesitation is often expressed with a lower intensity: on hesitation turns, mean intensity is from 45% to 70% lower than the mean intensity for the same child.

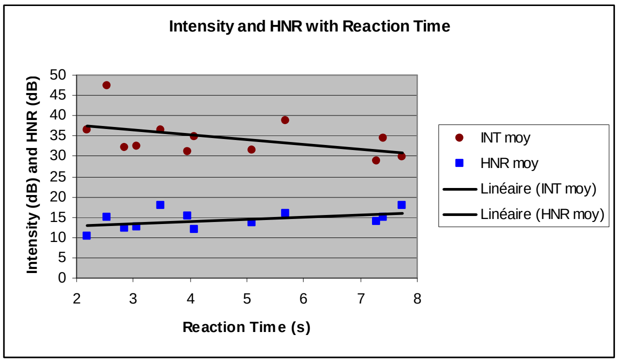

Figure 2: Intensity and HNR in function of the reaction time for the 12 children

Figure 2 shows that mean Intensity seems to decrease with RT and HNR to increase with RT. As we have said, a small RT generally signifies a good self-confidence; our data show that it is correlated with a high Intensity and a small HNR. When the child is at ease, he will speak loud. The correlation with HNR value is less evident. More data could help us to generalize this information.

Mean Int (dB)	Std Int (dB)	Mean HNR (dB)	Std HNR (dB)
34.46	5.01	14.25	2.35

Table 5: Intensity and HNR means and std

5. Conclusion and Future Works

The NAO-HR2 children voices corpus is composed of

French emotional speech collected in the course of a game between two children and the robot Nao. A semi-automatic Human-Robot Interaction system built the emotional and interactional representation of each child and selected the behaviour of the robot, based on the emotions captured manually by an experimenter. The data we collected allow us to study some parameters which take part in the setting up of the emotional and interactional profile.

We have analysed some of the mid-level cues which are used in our Human-Robot Interaction system. Among those cues, reaction time, intensity level and speech duration do make sense in our child-robot interaction game, but speaking rate does not seem to be relevant in that particular context. Indeed, as the children are quite young (from six to eleven years old), and as they are not given any predefined lexical content, they usually express their emotions with affect bursts. The younger the child, the more he/she will use affect bursts.

In a future work, we will also study the speaking rate in longer turns of child speech. For the needs of our data collection, the affective interactive system was used in Wizard-of-Oz (an experimenter captured manually the emotional inputs); in a next collection, we will use it with automatic detection of the emotions in speech, and then collect more data to confirm our analysis.

6. Acknowledgement

This work is financed by national funds FUI6 under the French ROMEO project labelled by CAP DIGITAL competitive centre (Paris Region).

7. References

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D’Arcy, S., Russell, M., Wong, M. (2004) *“You stupid tin box” – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus*. In proc. of the 4th International Conference of Language Resources and Evaluation, pp. 171–174.

Breazeal, C. and Aryananda, L. (2002) Recognition of affective communicative intent in Robot-Directed speech, Autonomous Robots 12, pp 83-104.

Campbell, N. (2007). *On the use of nonverbal speech sounds in human communication,* in proc. of the COST 2102 Workshop on Verbal and Nonverbal Communication Behaviours, Mar..

Delaborde, A., Tahon, M., Barras, C., Devillers, L. (2010a) *Affecive links in a child-robot interaction*. LREC 2010, Malte.

Delaborde, A., Devillers, L. (2010b) *Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers,* AFFIN’10, Firenze, Italy.

Kalman, Y. (2011) *HCI markers: A conceptual framework for using human-computer interaction data to detect disease processes*, 6th Conference on Information Systems (MCIS), Cyprus, 2011.

Mairesse, F., A. Walker, M., R. Mehl Matthias and K. Moore, R. (2007) *Using linguistic cues for the automatic recognition of personality in conversation and text,* in Journal of Artificial Intelligence Research 30, pp 457-500.

Ringeval, F., Sztaho, D., Chetouani, M. and Visci, K. (2008) *Automatic prosodic disorders analysis for impaired communication children,* 1st Workshop on Child, Computer and Interaction (WOCCI), IEEE International Conference on Multimodal Interfaces.

Scherer, K.R. (1994). *Affect Bursts*, in Emotions (S.H. M. van Goozen, N.E. van de Poll, & J.A. Sergeant, eds), p. 161-193. Hillsdale, NJ: Lawrence Erlbaum.

Schröder, M., (2003) *Experimental study of affect bursts*, Speech Communication – Special session on speech and emotion, vol. 40, Issue 1-2.