WIYE: building a corpus of children's audio and video recordings with a story-based app

Micol Spitale
micol.spitale@polimi.it
Politecnico di Milano
Milan, Italy

Fabio Catania
fabio.catania@polimi.it
Politecnico di Milano
Milan, Italy

Giulia Cosentino
giulia.cosentino@polimi.it
Politecnico di Milano
Milan, Italy

Mirko Gelsomini
mirko.gelsomini@polimi.it
Politecnico di Milano
Milan, Italy

Franca Garzotto
franca.garzotto@polimi.it
Politecnico di Milano
Milan, Italy

ABSTRACT

This paper describes the procedure to create an emotional data-set, named WIYE (What Is Your Emotion?), composed by semantic contents, audio and video recordings of children. Data have been collected using an interactive storytelling application, leading children aged 4 years to 12 years to discover about their emotional sphere and emotion expression skills. During the story, every episode is dedicated and focused on a specific emotion. The investigated emotional states are sadness, anger, fear, surprise and joy. This corpus can be exploited by Conversational Technologies: it can be used with Machine Learning classification algorithms to train models to recognize emotions expressed by children starting from the pitch of their voice, their facial expression and the contents of their conversations.

CCS CONCEPTS

• **Human-centered computing** → *Natural language interfaces.*

KEYWORDS

Emotional speech and facial corpus; Emotional database; Affective Computing; Emotion Recognition; Emotion classification.

ACM Reference Format:
Micol Spitale, Fabio Catania, Giulia Cosentino, Mirko Gelsomini, and Franca Garzotto. 2019. WIYE: building a corpus of children's audio and video recordings with a story-based app. In *24th International Conference on Intelligent User Interfaces (IUI '19 Companion), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3308557.3308684

1 INTRODUCTION

Communication consists of verbal and non-verbal information transmission. According to [2], facial expressions are responsible for about 55% of the message perception, voice intonation for about 38% and actual words just for 7%. This emphasizes the importance of the ability to express and to understand information not openly communicated as a content. Emotions are usually among this kind of info. An emotion is *an episode of coordinated changes in several components including at least neuro-physiological activation, motor expression and subjective feeling but possibly also action tendencies and cognitive processes in response to external or internal events of major significance to the organism* [7]. In psychology, there are a lot of theories providing different ways to organize emotions in classes. One of the most popular is the one by Ekmann [3], who identified six basic emotions that can be universally recognized: joy, sadness, surprise, fear, anger and disgust. These are named the *Big Six*.
The impact of emotion theories on other areas of research, like affective computing in computer science, is extraordinarily high. For this reason, in the last years it has been researched a lot to make it possible to realize spoken and visual emotional human-machine interaction, to develop security systems, to raise driver safety and to investigate in the field of health-care. To achieve all such ambitious goals, an emotional database containing labelled audio and video recordings is a prerequisite. Despite their usefulness, currently available emotional data-sets are limited as most of them only capture the emotional expressions (with the face and the voice) in adults.
In this paper, we describe the developmental process of WIYE, our original emotional database that stores audio and video files, pictures and utterances by Italian children aged 4 to 12 years. WIYE is the first Italian corpus collecting multimedia files exploiting visual, oral as well as semantic communication channels. In addition, it is the first Italian data-set of emotional speech by children and the first collection of utterances with emotional content produced with the direct contribution by youngsters. This corpus has a wide range of possible applications: for instance, it can be used to train Machine Learning (ML) algorithms in Speech-to-Text from emotional child-voice and in automatic emotion recognition and synthesis starting both from the voice and the facial expression. This kind of systems has been recently investigated a lot as didactic supporting tool and in the study of NeuroDevelopmental Disorders (NDD) such as Autism Spectrum Disorder [8], [4].

IUI '19 Companion, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6673-1/19/03…$15.00
https://doi.org/10.1145/3308557.3308684

2 WIYE: THE CORPUS

The WIYE corpus is an emotional database storing audio and video recordings, pictures and utterances by Italian children aged 4 to 12 years. The current version of the data-set counts 710 audio and video recordings and 710 original utterances with emotional semantics by 142 children equally divided between males and females. Recordings and utterances have been collected following a systematic procedure within a game by asking the children to act expressing emotions with their voice and with their face and then to invent sentences with emotional contents. The emotions we investigated are *the Big Six* excluding *disgust*, as *indico.io* does [1], which performs emotion recognition from text and facial expression. The whole user experience has been thought as a game to let the children feel comfortable during the recordings. In particular, the game is a story-based application exploiting Conversational Technology: it is a Conversational Agent (CA), or dialogue system, namely a software program able to interact with human beings through natural language. The application has been designed according to the educational pattern thanks to the cooperation with psychologists and child development experts who created the content. Indeed, the narration is inspired to a classical novel told in primary school [5] adapted to a space environment and structured into different scenes.

2.1 The setting

Before starting the game, the parents of the participants gave us the authorization to record their children and we received the IRB approval from university. The little participants were invited to enter one by one in a space-tent and immerse themselves in a different planet. The setting was composed by some cushions where they could sit on and a tablet, earphones and a microphone to isolate them from the noise.

2.2 The story-telling session

The main character of the story is an alien called *Boo* who asks to the user to teach him how to correctly express emotions both with the voice and with the facial expressions. Indeed *Boo* can communicate only with neutral intonation and no facial emotion manifestation. The story is divided into five different scenes each of which corresponds to one emotion. During each scene, the user is invited to accomplish three different tasks:

1) to repeat a specific sentence, already spelled by the alien during the story-telling. The utterance needs to be expressed with the proper, specified emotion. The sentences assigned have been chosen with an emotional semantics in order to let the children get in the same mood as the main character;
2) to invent a sentence linked to the current mood of the story. To do so, they are asked to use their imagination and emotional intelligence as well as their ability to understand the context;
3) to show her/his facial expression fitting with that emotion.

For all these tasks, on the advice of the experts we relied on the children' ability of self-induction by imagining or remembering a situation when the desired emotion had been felt strongly, which is known as the Stanislavski method [6].

2.3 The storage method

We collected one video recording and one picture for each emotion expressed with the face, and then ten audio files with the voice of each participant: five for the sentences to be repeated and five for the invented ones. In the database, no stored data is directly attributable to a specific user.

3 CONCLUSION AND FUTURE WORKS

This paper has presented a detailed description of the emotional corpus WIYE, which stores audio and video recordings, pictures and utterances by Italian, acting children aged 4 years to 12 years. Currently, the data-set counts 710 audio and 710 video recordings and 710 utterances with emotional semantics by 142 children equally divided between males and females.

Although acted expressions have intrinsic limitations related to their naturalness with respect to spontaneous ones, we believe that WIYE is suitable for use in Speech-to-Text from emotional child-voice and in automatic emotion recognition and synthesis. Indeed, the database is foreseen to be used for the development of new algorithms for the recognition of vocal and facial emotional expressions by children and for cross-modal emotional analysis, comparing the audio pitch, the semantics of the speech, the facial expression and biological data. A limitation of the storytelling method used to collect emotional sentences produced by the users is that the children were asked to invent sentences linked to predefined contexts and this produced a topically homogeneous set of utterances. This has huge, negative model training implications, since the model would easily be over-fitted if the corpus is not diverse enough.

The natural follow up of this work is the enlargement of the first preliminary data-set within an already planned elementary school experimentation. Indeed, we are working on refining the story-based app to ensure an immersing and more engaging game-experience. This can be exploited for collecting a larger amount of data, useful for many applications, including children-oriented researches to better investigate children' verbal and non-verbal communication skills and methods. In parallel, we are also working on the validation process via crowd-sourcing in order to increase the quality of the provided recordings and utterances.

ACKNOWLEDGMENTS

We are very grateful for the cooperation to the psychologists, the caregivers, the children and their parents.

REFERENCES

[1] 2016. Indico - Emotion recognition. *https://indico.io* (2016).
[2] Claude C Chibelushi et al. 2003. Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision* (2003).
[3] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* (1992).
[4] Liliana Laranjo et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* (2018).
[5] David McKee et al. 1990. *Elmer: l'elefante variopinto.* Mondadori.
[6] Perviz Sawoski. 2010. The Stanislavski System: growth and methodology. *Santa Monica College* (2010).
[7] Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* (2003).
[8] Hiroki Tanaka et al. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one* (2017).