Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Emotion, age, and gender classification in children's speech by humans and machines[☆]

Q1 Heysem Kaya[*,a], Albert Ali Salah[b], Alexey Karpov[c,d], Olga Frolova[e], Aleksey Grigorev[e], Elena Lyakso[e]

[a] *Department of Computer Engineering, Namık Kemal University, Corlu, Tekirdag, Turkey*
[b] *Department of Computer Engineering, Bogazici University, Istanbul, Turkey*
[c] *Speech and Multimodal Interfaces Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia*
[d] *Department of Speech Information Systems, ITMO University, St. Petersburg, Russia*
[e] *Child Speech Research Group, St. Petersburg State University, St. Petersburg, Russia*

Received 1 May 2016; received in revised form 9 May 2017; accepted 8 June 2017

Abstract

In this article, we present the first child emotional speech corpus in Russian, called "EmoChildRu", collected from 3 to 7 years old children. The base corpus includes over 20 K recordings (approx. 30 h), collected from 120 children. Audio recordings are carried out in three controlled settings by creating different emotional states for children: playing with a standard set of toys; repetition of words from a toy-parrot in a game store setting; watching a cartoon and retelling of the story, respectively. This corpus is designed to study the reflection of the emotional state in the characteristics of voice and speech and for studies of the formation of emotional states in ontogenesis. A portion of the corpus is annotated for three emotional states (comfort, discomfort, neutral). Additional data include the results of the adult listeners' analysis of child speech, questionnaires, as well as annotation for gender and age in months. We also provide several baselines, comparing human and machine estimation on this corpus for prediction of age, gender and comfort state. While in age estimation, the acoustics-based automatic systems show higher performance, they do not reach human perception levels in comfort state and gender classification. The comparative results indicate the importance and necessity of developing further linguistic models for discrimination.

Keywords: Emotional child speech; Perception experiments; Spectrographic analysis; Emotional states; Age recognition; Gender recognition; Computational paralinguistics

1. Introduction and related work

Speech based communication contains both linguistic and paralinguistic information. The latter is particularly important in specifying factors of behavioral and functional status, and especially emotional states. For children's

[☆] This paper has been recommended for acceptance by Prof. R. K. Moore.

[*] Corresponding author.

E-mail address: hkaya@nku.edu.tr (H. Kaya), salah@boun.edu.tr (A. Ali Salah), karpov@iias.spb.su (A. Karpov), olchel@yandex.ru (O. Frolova), a.s.grigoriev89@gmail.com (A. Grigorev), lyakso@gmail.com (E. Lyakso).

http://dx.doi.org/10.1016/j.csl.2017.06.002
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

communications, self-reporting is not very reliable as a measure, and assessment of emotional speech becomes particularly valuable. There are two main approaches or the study of emotional speech. One approach focuses on the psychophysiological aspects of emotions, which can include studies of brain activity data (Lindquist et al., 2012; Watson et al., 2014), and cross-cultural investigation of emotional states in speech (Lyakso and Frolova, 2015; Rigoulot et al., 2013; Jürgens et al., 2013; Laukka et al., 2013). The second approach is focused on the physical speech signal and its analysis. Hence, it is geared towards software applications for human-computer interaction, such as automatic speech recognition (Fringi et al., 2015; Liao et al., 2015; Guo et al., 2015) and speech synthesis (Govender et al., 2015).

Adults perceive emotional states of infants in their vocalizations from the first months onwards (Lyakso and Frolova, 2015). For instance discomfort and comfort conditions of three months old infants are recognizable by people, but also via spectrographic analysis, which reveals differences in the pitch values and the duration of vocalizations. Crying and squeals of joy are indicative of emotional states, but acoustic features are not always significantly different. With child’s increasing age, lexical information acquires more discriminative power in the recognition of emotional states (Yildirim et al., 2011).

It is well known that acoustic and linguistic characteristics of child speech are essentially different from those of adult speech. The child speech is characterized by a higher pitch value, formant frequencies and specific indistinct articulation with respect to the adult speech. Recognition of child’s speech can be challenging. It was shown that adult Russians recognize between half and three quarters of 4−5 years old children’s words and phrases in calm and spontaneous conditions (Lyakso et al., 2006). The paralinguistic aspects, however, require more research, both from a human perceptual perspective, and from an automated speech processing perspective. This paper aims to address these points.

The first requirement for studying children’s emotional speech is the preparation of an adequate corpus (Ververidis and Kotropoulos, 2006). Creation of such a corpus is more difficult than the collection of emotional speech corpora of adults. In the case of adults, actors are often involved to portray the necessary emotional conditions (Engberg and Hansen, 1996; Burkhardt et al., 2005; Kaya et al., 2014; Lyakso and Frolova, 2015), or records of patients from a psychiatry clinic are used. Such approaches are not easily used for children. It is necessary to model communicative tasks in which the child is not conscious of being recorded to produce veridical emotional reactions. The creation of the corpus should be based on a verified and clear method of obtaining spontaneous speech manifestations of certain emotional reactions. By nature, collection of child emotional speech data should be under natural conditions that are not-controlled, not-induced (i.e., “spontaneous”).

At present there are a few spontaneous or emotional child speech databases available for the child speech research community. These include emotional and spontaneous corpora for Mexican Spanish (7−13 years old) (Pérez-Espinosa et al., 2011), British English (4−14 years old) (Batliner et al., 2005), and German (10−13 years old) (Batliner et al., 2005; Batliner, Steidl, Nöth, 2008. The SpontIt corpus is spontaneous child speech in Italian (8−12 years old) (Gerosa et al., 2007), and the NICE corpus is spontaneous child speech in Swedish, possibly emotional, but without emotion annotations (8−15 years old) (Bell et al., 2005). Recently, we have collected the first emotional child speech corpus in Russian, called “EmoChildRu”, and reported initial results (Lyakso et al., 2015). The present work greatly extends the scope of investigation on this corpus, doubling the annotated data, and providing age and gender estimation baselines for both machine classification and human perceptual tests.

The rest of the article is structured as follows: Section 2 introduces the Emotional Child Russian Speech Corpus “EmoChildRu”, including the recording setup and speech data analysis. Section 3 describes two separate human perception experiments, one on the recognition of emotional states and another for prediction of child’s age and gender by listeners, respectively. Section 4 presents baseline automatic classification systems for paralinguistic analysis, and reports extensive experimental results. Section 5 provides a discussion of the findings and conclusions.

2. Emotional Child Russian Speech Corpus

“EmoChildRu” is the first database containing emotional speech material from 3−7 year old Russian children. Three emotional states (discomfort, comfort, neutral) are used in the database. It is important to note that the “discomfort” state encapsulates a number of basic emotions, such as “sadness,” “fear,” and “anger,” but these emotional statements are not expressed strongly. It is not ethical to induce natural fear or anger in 3−7 year old children for the purposes of such a study. All procedures were approved by the Health and Human Research Ethics