ARTICLE IN PRESS
JID: YCSLA [m3+;June 15, 2017;19:50]

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

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

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Committee (HHS, IRB 00003875, St. Petersburg State University) and written informed consent was obtained from parents of the child participant.

All children in the database were born (and lived) in the city of St. Petersburg, with parents who were also born in St. Petersburg, or have been living there for more than 10 years. The whole collection includes 20.340 utterances (more than 30 hours of speech). Recordings were made at home, in laboratory and at kindergarten. The three different recording conditions are playing with a standard set of toys, repetition of words from a toy-parrot in a game store setting, and watching a Russian cartoon called "Masha and bear" from iPad and the retelling of the story, respectively. Each experiment had a duration of 2 minutes (containing multiple utterances). Every record is accompanied by a protocol, which describes the recording conditions, and video recording of child's behavior in parallel. The speech materials are grouped based on the protocol and on the recording situation.

Model situations for provoking the child's emotional states were selected based on our previous experience - the supervision of children in various forms of interaction with adults (experimenters), and taking into account the response of children aged 4−7 years (Lyakso et al., 2010b). Using heart rate data, the child's emotional state was estimated during the preliminary experiments (which makes it possible to compare neutral and arousal states), and an additional video analysis was performed by experts. They annotated child behaviors, as well as facial expressions during these preliminary experiments.

The speech recordings were made with a "Marantz PMD660" digital recorder and with a single "SENNHEISER e835S" external microphone. The speech sounds were analyzed by experts, using Syntrillium's "Cool Edit Pro" sound editor. Speech files are stored in Windows PCM format, 22.050 Hz, 16 bits per sample. Stressed vowels were selected from stressed words of all phrases. Pitch and the vowel duration, as well as phrase prosody, were automatically calculated, based on the algorithms implemented in "Cool Edit Pro" sound editor. The waveform view was used for calculation of duration, and the spectral view was used to measure the pitch control for prosody. The corpus and the accompanying software package include the database, as well as a shell component to navigate, browse and search information in the database. So far, about 25% of the data are annotated for emotional states. The child's (ground truth) emotional state was determined based on the recording setting and by analysis of the video clips by five experts having professional experience of working with children and child speech. The database contains additional information about the child's psychophysiological condition before and after speech recording. Dialogue descriptions, speech developmental and cognitive scale data are included in the database, whenever available. We will focus on the acoustic information only for the purposes of this paper.

A software tool was developed in Microsoft Visual C# for enabling the experts to work with the "EmoChildRu" corpus under Windows OS. This software also allows choosing speech material using a query interface, along dimensions such as the type of emotional state, child age, and gender.

We performed a qualitative analysis of children's words reflecting different emotional states, using words from 4-years olds (4800 words from 39 children), 5-years olds (9030 words from 55 children), 6-years olds (4150 words from 26 children) and 7-years olds (846 words from 18 children). 4-year old children express themselves by antonyms (*yes − no; it is terrible − well; I'm afraid; I am glad; good − bad*). At the age of 7, the word range significantly expands (e.g., *very angry; angry; terrible; bad; not so good* for discomfort and *like; good; like more; like most; love; immense; wonderful; splendid* for comfort). In line with a recent study on a portion of this corpus (Lyakso et al., 2016), the number of words 7 year old children use to reflect a discomfort state ($n = 14 \pm 8$ words) is found higher than the number of words that reflect the comfort state ($n = 10 \pm 6$ words).

2.1. Dataset used for machine learning experiments

For automatic recognition experiments, we used a subset of the corpus (1.116 child speech utterances), where all speech files have 1 to 5 s of speech signal and all five child speech experts agree on the emotion annotation. Note that the annotation is done from the audio-visual data, including linguistic information, though for automatic classification, only acoustic features are used.

The subset contains data from 113 children, and the number of speech files per child ranges from 1 to 78 (mean = 9.9), and this imbalance makes automatic recognition more difficult. There are 54 boys and 59 girls in the dataset (mean ± std age is 5.3 ± 1.1 years). We use a different subset for the machine experiments, because the number of instances needed for a machine classifier is much higher than what can feasibly be set in a human perception study. Also, we split the data into training and test sets, which is not required in the human perception study. Class

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Table 1
Distribution of emotion and gender classes. #Inst. denotes number of utterances.

	Gender		Emotion		
Set	#M/#Inst	#F /#Inst.	Comfort	Discomfort	Neutral
Train	36 / 288	34 / 353	245	111	285
Test	18 / 209	25 / 266	189	94	192
Total	**54 / 497**	**59 / 619**	**434**	**205**	**477**

Table 2
Distribution of age group classes in train and test partitions.

Set	#Inst.	3−4 years	5 years	6−7 years
Train	641	130	168	343
Test	475	111	115	249
Total	**1116**	**241**	**283**	**592**

distribution for the three classification tasks over the speaker-independent training and test partitions are given in Tables 1 and 2.

2.2. Dataset used for human perception experiments

The dataset used for human perception experiments is a subset of the one used for the automatic recognition experiments. This is because we need a large number of instances to train an automatic recognizer as opposed to a naturally trained human.

30 children, aged from 3 to 7 years were selected for human perception study, and three test sequences were formed from the speech material of each child. These sequences were arranged such that they equally represent the three age groups (3−4 year old, 5 year old, 6−7 year old children, respectively) and that every test sequence includes 10 phrases uttered by children in a comfortable emotional state, 10 phrases in a discomfort state and 10 phrases in a neutral (calm) state. In total, we used 90 sequences for testing.

3. Human perceptual experiments

This section reports two human perceptual experiments to provide insight on the nature of the EmoChildRu Database. Listeners were Pediatric University Students 300 adults (age: 18.8 ± 2.2 years, median 18 years; 61 male, 239 female; 219 with the experience of interaction with children). Child interaction experience implies the presence of children in the family − younger brothers and sisters, communication with children of friends and relatives. Data about the listeners with experience and without experience of interaction with the children are presented together, as significant differences in the recognition of children were not found between these groups. The presentation of test sequences was carried out in an open field for 10-people groups (the listener location from the source of sounds had no effect on the recognition result). Each signal in the test (phrase) was presented one, the duration of pauses between the signals was 7 s, which allowed the listeners to fill in the forms with requested information. The 7-second interval is chosen experimentally.

3.1. Human perception and spectrographic analysis of emotional speech

The aim of the first study is to reveal how humans (Russian native speakers) can recognize emotional states in children via speech. Each of the test sequences was presented to 100 adults for perceptual analysis, and the ratio of listeners correctly recognizing the emotional state on the base of speech samples (perception rate) was calculated. Confusion matrices for perception experiments were prepared.

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

In terms of spectrographic analysis, we analyzed and compared pitch values, max and min values of pitch, pitch range, energy and duration. The vowel duration and pitch values were determined based on the analysis of the dynamic spectrogram. Spectral analysis was performed by fast Fourier transformation weighted using a Hamming window, with a window length of 512 samples. To consider word stress development, the vowel duration and its stationary part duration were compared in the stressed versus the unstressed vowels, as well as the pitch and formant values in the stationary parts.

Stressed and unstressed vowels in the words are defined initially by experts via auditory perception of the word meaning in the Russian language. A secondary (instrumental) spectrographic analysis was performed on stressed vowel duration and pitch, because a former study found that the stressed vowels in the child speech are characterized by higher values of the pitch and duration (Lyakso and Gromova, 2005), unlike the stressed vowels in Russian adult speech that only exhibit longer duration compared to the unstressed.

The average recognition accuracy was 56% for the 3−4 year old group, 66% for the 5 year old group, and 73% for 6−7 year old group. The comparison of perception rate for emotional speech samples showed that the amount of samples attributed by most of listeners (0.75−1.0) as the state of comfort was higher than the amount of signals attributed as neutral and discomfort (Fig. 1-A). The agreement among the listeners in determining neutral signals and discomfort was lower: majority of neutral signals were recognized with a ratio of 0.5−0.74 (Fig. 1-A). The state of comfort is more clearly reflected in the child speech compared to the state of discomfort. Often 3−6 year old children spoke excitedly, smiled and laughed in comfortable emotional state during recording. At the same time, children of pre-school age do not express intense discomfort, only a slight discomfort is considered admissible in recording situation in the presence of an unfamiliar adult. The slight discomfort was recorded when child spoke about something unpleasant, manifested disgust or anger. Consequently, listeners recognized slight discomfort of child with a smaller ratio compared to comfort. The neutral state is recognized by listeners with ratio of 0.5−0.74, because they attributed the part of neutral speech samples to comfort and discomfort samples.

Amount of emotional samples correctly recognized by most of listeners (i.e., perception rate of 0.75−1.0) increased with child age. The emotion perception rate for the 6−7 year old children was higher than that of 3−4 year old children, as shown in Fig. 1−B ($p < 0.01$, Mann−Whitney U test).

Errors in emotional state classification were associated primarily with the allocation of discomfort and comfort speech samples to the neutral state class. Speech samples that reflected a neutral state were classified more often as discomfort rather than comfort. The corresponding confusion matrices are presented in Table 3.

Spectrographic analysis also revealed that speech samples interpreted by humans as discomfort, neutral and comfort are characterized by specific sets of acoustic features (Fig. 2). Discomfort speech samples are characterized by phrase duration lengthening; highest pitch values (comparatively) in phrase and in stressed vowels selected from stressed words in the phrase ($p < 0.05$ − vs. neutral state, Mann−Whitney U test); an increase of minimum pitch value in phrases; an increase of the pitch range in stressed vowels of stressed words in the phrase ($p < 0.05$ − vs. neutral state); falling pitch contour of stressed vowels of stressed words in the phrase. Comfort speech phrases have short

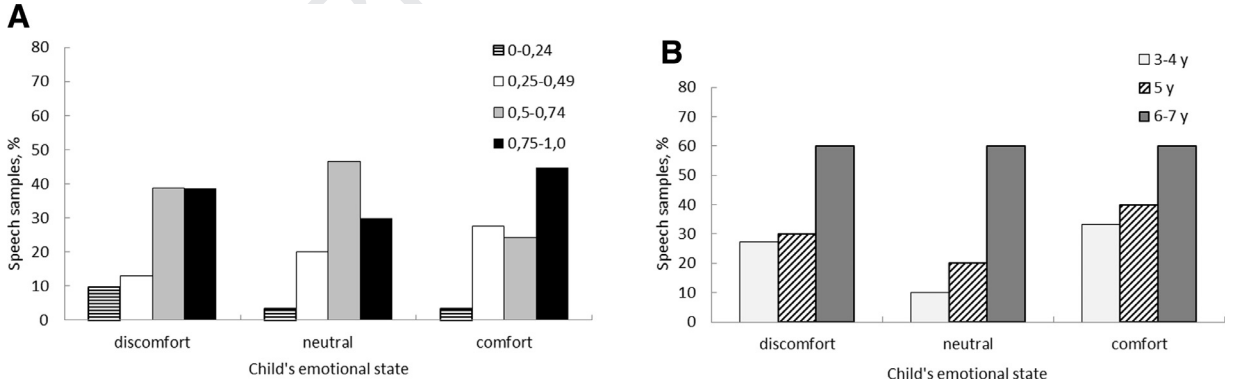

Fig. 1. Percentages of emotional child speech samples perceived by the listeners: (A) correctly recognized states as 'discomfort', 'neutral' and 'comfort' with the perception rate of 0−0.24 (horizontal hatch), with the rate of 0.25−0.49 (white color bar), with the rate of 0.5−0.74 (light gray) and with the rate of 0.75−1 (black); (B) correct recognition with the rate of 0.75−1.0 for different age groups: 3−4 years (light gray), 5 years (sloping hatch), 6−7 years old (gray).

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Table 3
Confusion matrices for emotion recognition by humans with age group breakdown.

	Child's age								
	3−4 years			5 years			6−7 years		
State	Disc	Neut	Comf	Disc	Neut	Comf	Disc	Neut	Comf
Discomfort	**64**	22	14	**68**	23	9	**69**	25	6
Neutral	23	**56**	21	24	**65**	11	18	**70**	12
Comfort	13	33	**54**	9	26	**65**	4	18	**78**

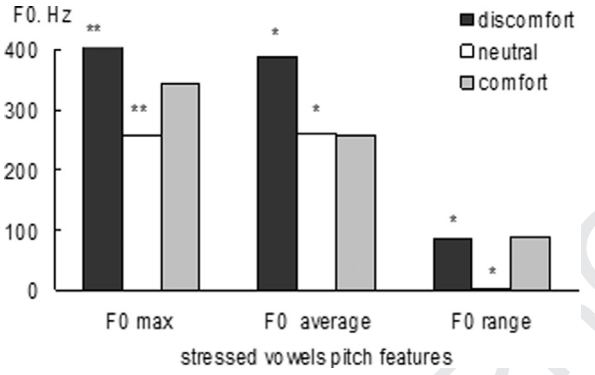

Fig. 2. Pitch features of stressed vowels (medians)− F0 max: maximum pitch value, Hz; F0 average: vowel's average pitch value, Hz; F0 range: vowel's pitch range value (F0max−F0min); * − $p <$ 0.05, **− $p <$ 0.01 Mann−Whitney U test.

duration, together with long stressed vowel duration; pitch values of phrases are increased, but less so compared to discomfort samples; pitch range in the stressed vowels is similar to discomfort samples; pitch contours of stressed vowels are rising. Neutral speech samples are characterized by lowest values of vowel duration, stressed vowels' pitch and pitch range (Fig. 2). Flat pitch contours are observed for 3−4 and 6−7 years old children (Table 4).

Most speech samples that were correctly recognized by humans in the experiment have a complex shape of phrase pitch contours (> 70% samples). The analysis of features of all stressed vowels from stressed words revealed that discomfort speech samples have mainly a falling shape, while comfort speech samples have a rising shape.

Almost all neutral speech samples have flat, falling and bell-shaped pitch contours, and the first two patterns are the most common. U-shaped pitch contour is revealed in comfort speech samples only. The variety of pitch contour shapes in stressed vowels increases by 6−7 years, compared to younger children. With increasing age, the duration of phrases increases, and the duration of stressed vowels and pitch values decreases. The differences in the acoustic characteristics of speech samples correctly recognized as discomfort, neutral and comfort are more expressed at the age of 3−5 years. Correctly recognized speech samples of 6−7 years old children do not differ significantly in

Table 4
Distribution of pitch contour shapes for speech samples with correctly recognized emotional states.

		Pitch contour shape (%)				
State	Age (y)	Flat	Rising	Falling	U-shaped	Bell-shaped
Discomfort	3−4	0	33	**67**	0	0
	5	0	0	**100**	0	0
	6−7	33	0	**67**	0	0
Neutral	3−4	100	0	0	0	0
	5	0	0	100	0	0
	6−7	67	0	16.5	0	16.5
Comfort	3−4	0	**67**	0	33	0
	5	0	**75**	25	0	0
	6−7	17	**50**	0	33	0

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

acoustic features. Adult listeners mostly rely on the meaning of the phrase. Analysis of speech samples correctly recognized by listeners revealed that detection of word meaning improved with increase of child's age up to 100% at the age of 6−7 years.

3.2. Estimation/perception of child's age/gender by humans

The aim of the second study was to investigate the possibility of child's age and gender recognition by listeners. Three test sequences were formed from the speech material of 45 children. The sequences contain speech data uttered by children in a discomfortable emotional state (Test 1 − discomfort), in a neutral/calm state (Test 2 − neutral) and in a comfortable state (Test 3 − comfort). Every test sequence includes 30 phrases: 10 speech samples for each age group (3-4- years, 5 years, and 6−7 years; five speech samples per gender). Thus, each test sequence includes 15 speech samples uttered by boys and 15 samples by girls. For testing, we used 90 sequences in total; each speech signal was included in the test sequence only once; the time interval between the signals was seven seconds.

It should be noted that Russian is a gender-dependent language and most of the verbs in the past tense form and some adjectives have both masculine and feminine gender word-forms. We have excluded such phrases from the test set for the human perception experiments, so that Russian speaking listeners could not easily predict child's gender using linguistic knowledge. In the gender and age prediction tasks, each test sequence was evaluated by 100 adult listeners (a total of 300 listeners), who were asked to select the gender (male or female) and age group (3−4, 5, or 6−7 years) of each child in a questionnaire.

For gender prediction, the average recognition accuracies were 66%, 64% and 71% for speech samples uttered by children in the discomfort state (Test-1), neutral state (Test-2), and comfort state (Test-3), respectively. Most of listeners (0.75−1.0) correctly recognized the gender of child on the base of child speech samples reflected neutral, discomfort, and especially comfort state (see Fig. 3). The amount of comfort samples on which most of listeners (0.75−1.0) correctly recognized child's gender was higher than the amount of discomfort and neutral samples. The agreement among the listeners in determining child's gender on the base of neutral signals was the lowest. Child gender was recognized with ratio 0−0.5 in more than 30% of neutral samples. We can conclude that the emotional child speech includes more acoustic and linguistic information for gender recognition as opposed to neutral speech.

Percentages of correct and incorrect gender estimation are reported individually for comfort, neutral and discomfort states with confusion matrices in Table 5. The least number of errors was made for male speakers at the discomfort state, and the most error-prone classification was female speaker recognition at the comfort state. Human recognition of male speakers was better than female speakers for all emotional states.

In the age prediction task, the average recognition accuracy was 50%, 52%, and 51% of speech samples uttered by children in the discomfort state, neutral state, and comfort state, respectively. The listeners recognized child's age mainly with a perception rate of 0.5−0.74 (Fig. 4). Generally the agreement among the listeners in determining child's age was less than in gender and emotional state recognition. The amount of neutral samples on which most of listeners (0.75−1.0) correctly recognized child's age was higher than comfort and discomfort samples. This fact

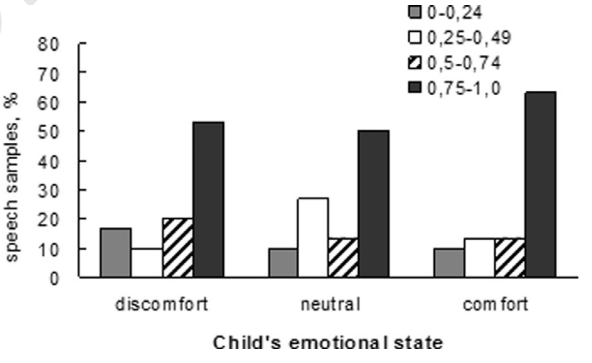

Fig. 3. Percentages of emotional child speech samples perceived by the listeners: correctly recognized gender with the perception rate of 0−0.24 (gray), with the rate of 0.25−0.49 (white color), with the rate of 0.5−0.74 (sloping hatch) and with the rate of 0.75−1 (black).

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Table 5
Confusion matrices for gender prediction in three emotional states.

	Emotional state					
	Discomfort		Neutral		Comfort	
Gender	Male	Female	Male	Female	Male	Female
Male	**80**	20	**68**	32	**76**	24
Female	48	**52**	40	**60**	34	**66**

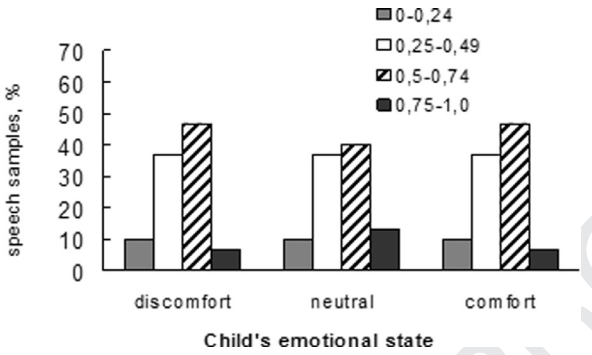

Fig. 4. Percentages of emotional child speech samples perceived by the listeners: correctly recognized age with the perception rate of 0−0.24 (gray), with the rate of 0.25−0.49 (white color), with the rate of 0.5−0.74 (sloping hatch) and with the rate of 0.75−1.0 (black).

can be explained from the point of view of the pitch values. Decrease of pitch values from 3 to 7 years is shown. We can assume that listeners first of all lean on the pitch values in the determining the age of the child. Pitch values in emotional speech higher than in neutral speech that can confuse listeners.

The main sources of error in the age recognition task were the following: the listeners confused the 5 year old group with other ages (close ages were confused); speech samples of 5 year old children uttered in the discomfort and comfort states were often attributed to 3−4 year old kids. Confusion matrices are presented in Table 6. Based on these results, we can conclude that the age prediction task is more difficult for humans than the gender prediction task in our database.

Recently, two studies on human and machine prediction of gender and age of children are presented by Safavi et al. (2013); 2014). These studies use material from an age range of 5 to 13 years, which does not entirely overlap with the child age range in our data. Safavi et al. (2013) analyze the spectrum (24 filtered frequencies) obtained from sliding windows of length 20 ms shifted with 10 ms. It was shown that the frequencies below 1.8 kHz and above 3.8 kHz are most useful for gender identification for older children (13−16 years), and the frequencies above 1.4 kHz are most useful for the youngest children (5−9 years). The frequencies above 5.5 kHz are the least useful for age identification (Safavi et al., 2014). Results show that in both cases, machine identification is more accurate compared to humans.

Table 6
Confusion matrices for child's age prediction in three emotional states.

	Emotional state								
	Discomfort			Neutral			Comfort		
Age (y)	3−4 y	5 y	6−7 y	3−4 y	5 y	6−7 y	3−4 y	5 y	6−7 y
3−4	**53**	39	8	**65**	29	6	**69**	28	3
5	35	**49**	16	22	**50**	28	31	**44**	25
6−7	16	36	**48**	13	44	**43**	17	43	**40**

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

In the following machine learning experiments, we use a standard set of supra-segmental features without focus on feature or frequency band selection. Finding the most potent spectrum/cepstrum bands and most predictive features for these three classification tasks are left as future work.

4. Automatic classification systems for paralinguistic analysis

In this section, we investigate machine classification of the emotion, age, and gender of the child from speech segments. While there are several studies for the automatic processing of child speech (e.g., (Potamianos et al., 2011; Meinedo and Trancoso, 2011; Bolaños et al., 2011; Safavi et al., 2013; 2014), etc.), automatic detection and classification of emotional states of speech of children in natural conditions is a new direction of research (Batliner et al., 2008; Lyakso et al., 2015). As our previous perception analysis reveals, the recognition of children's emotions from speech is hard, even though some prosodic patterns can be discerned. The overall human recognition accuracy for a balanced three-class problem (i.e., discomfort, comfort, neutral) is found as 65%.

In the following subsections, we provide a brief overview of paralinguistic analysis and major elements of its pipeline, followed by the experimental results.

4.1. Background on automatic paralinguistic speech analysis

Paralinguistics (meaning alongside linguistics), studies short term states and long term traits of the speaker(s), particularly focusing on non-verbal aspects of speech that convey emotions. It deals with how the words are spoken, rather than what is being spoken. Speech Emotion Recognition (SER) is a branch of paralinguistics, related to speaker state and trait recognition (SSTR).

A general pipeline for SSTR is shown in Fig. 5 (Schuller, 2011). The speech signal is processed to extract informative features, which are fed to machine learning modules for acoustic- and/or language-based classification. In paralinguistic analysis, acoustic models refer to affect classification models trained on features derived from acoustic/prosodic Low Level Descriptors (LLD) such as pitch, energy, jitter, shimmer and MFCCs (Schuller, 2011). On the other hand, language-based classification employs linguistic information provided by automatic speech recognition (ASR), which for instance can be represented as a bag-of-words. While emotion recognition is a worthy goal on its own, it can also help ASR through emotion-specific models. In turn, emotion recognition performance can be improved by good ASR, by providing robust linguistic features to be fused with acoustic features (Schuller et al., 2004). However, this is not always possible, since the recognition of affective (emotional) speech is itself very challenging (Schuller, 2011). In the present study, we use only acoustic models due to lack of Russian ASR trained

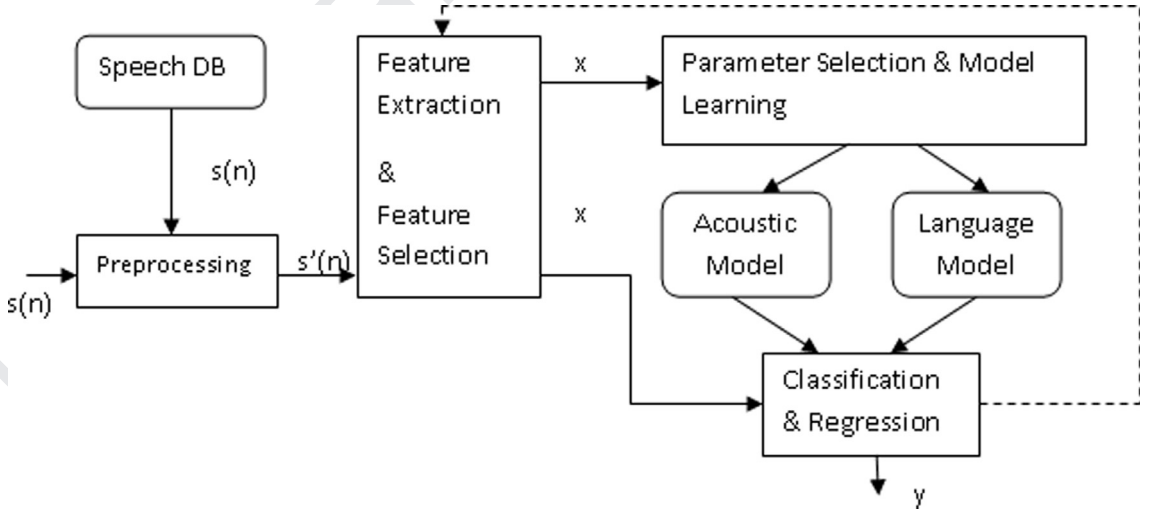

Fig. 5. General speaker state and trait recognition pipeline.

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

on child speech, and since ASR trained on adult speech does not work on children's speech. Working on Russian child ASR is one of our future research directions.

The state-of-the-art computational paralinguistics systems use large scale suprasegmental feature extraction via passing a set of summarizing statistical functionals (such as moments, extremes) over LLD contours (Eyben et al., 2010). Pitch, Formants (resonant frequencies of vocal tract filter), Mel Frequency Cepstral Coefficients (MFCC), Modulation Spectrum, Relative Spectral Transform − Perceptual Linear Prediction (RASTA-PLP), Energy and variation features (i.e., Shimmer and Jitter) are frequently used as LLDs (Schuller, 2011). Among these, MFCC and RASTA-PLP are the most widely used. In line with the state-of-the-art, we extract acoustic features in this work using the freely available openSMILE tool (Eyben et al., 2010), using a standard configuration file.

The most commonly employed classifiers in paralinguistics are Support Vector Machines (SVM), Artificial Neural Networks (ANN), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM). The state-of-the-art models of SER for the current databases are those trained with Support Vector Machines (SVMs) and Deep Neural Networks (DNN) (Schuller, 2011). From the ANN family, Extreme Learning Machines (ELM), which combine fast model learning with accurate prediction capability, are recently applied to multi-modal emotion recognition and computational paralinguistics, obtaining state-of-the-art results with modest computational resources (Kaya et al., 2015a; 2015b; Kaya and Salah, 2016). Consequently, we employ Kernel ELMs (Huang et al., 2012) in this work, as well as a fast and robust classifier based on Partial Least Squares (PLS) regression (Wold, 1985). In the next subsection, we give a brief summary of these two classification approaches. As a further baseline, we use SVMs.

4.2. Background on least squares regression based classifiers

To learn a classification model, we employ kernel extreme learning machine (ELM) and Partial Least Squares (PLS) regression due to their fast and accurate learning capability (Wold, 1985; Huang et al., 2012) and state-of-the-art achievements in recent audio and video based challenges (Kaya et al., 2015b; Kaya and Karpov, 2016; Kaya et al., 2017).

ELM proposes a Single Layer Feedforward Network (SLFN) architecture, but unsupervised, even random generation of the hidden node output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$, where N and h denote the number of data samples and the hidden neurons, respectively. The hidden node output matrix $\mathbf{H}$ is obtained by projecting the data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ using a randomly generated first layer weight matrix $\mathbf{W} \in \mathbb{R}^{d \times h}$, and subsequently applying an infinitely differentiable non-linear activation function (e. g., logistic sigmoid). The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T} \in \mathbb{R}^{N \times L}$, where L is the number of classes. Let y^t denote the class label of t^{th} data instance, in the case of L-class classification, $\mathbf{T}$ is represented in one vs. all coding as follows:

$$\mathbf{T}_{t,l} = \begin{cases} +1 & \text{if } y^t = l, \\ -1 & \text{if } y^t \neq l. \end{cases} \tag{1}$$

In case of regression, $\mathbf{T} \in \mathbb{R}^{N \times 1}$ is the continuous target variable.

The second level weights $\beta \in \mathbb{R}^{h \times L}$ are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. The output weights can be learned via:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{2}$$

where $\mathbf{H}^{\dagger}$ is the Moore−Penrose generalized inverse (Rao and Mitra, 1971) that gives the minimum L_2 norm solution to $\| \mathbf{H}\beta - \mathbf{T} \|$, simultaneously minimizing the norm of $||\beta||$. To increase the robustness and generalization capability, the optimization problem of ELM is reformulated using a regularization coefficient on the residual error $\| \mathbf{H}\beta - \mathbf{T} \|$. The learning rule of this alternative ELM is related to Least Square SVMs (LSSVM) via the following output weight learning formulation:

$$\beta = \mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T \right)^{-1} \mathbf{T}, \tag{3}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, and C, which is used to regularize the linear kernel $\mathbf{H}\mathbf{H}^T$, corresponds to the complexity parameter of LSSVM (Suykens and Vandewalle, 1999). This formulation is further simplified by noting

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

that the hidden layer matrix need not be generated explicitly given a kernel $\mathbf{K}$, which can be seen identical to Kernel Regularized Least Squares (Huang et al., 2012; Rifkin et al., 2003):

$$\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{K}\right)^{-1} \mathbf{T}. \tag{4}$$

The second approach we use for classification is partial least squares (PLS) regression. PLS regression between two sets of variables $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times p}$ is based on decomposing the matrices as $\mathbf{X} = \mathbf{U}_x \mathbf{V}_x + r_x$, $\mathbf{Y} = \mathbf{U}_y \mathbf{V}_y + r_y$, where $\mathbf{U}$ denotes the latent factors, $\mathbf{V}$ denotes the loadings and r stands for the residuals. The decomposition is done by finding projection weights $\mathbf{W}_x$, $\mathbf{W}_y$ that jointly maximize the covariance of corresponding columns of $\mathbf{U}_x = \mathbf{X}\mathbf{W}_x$ and $\mathbf{U}_y = \mathbf{Y}\mathbf{W}_y$. For further details of PLS regression, the reader is referred to (Wold, 1985). When PLS is applied to the classification problem in a one-versus-all setting, it learns the regression function between the feature matrix $\mathbf{X}$ and the binary label vector $\mathbf{Y}$, and the class giving the highest regression score is taken as prediction. The number of latent factors is a hyper-parameter to tune via cross-validation.

4.3. Features and performance measures used in machine classification

We extract openSMILE (Eyben et al., 2010) features with a configuration file used in the INTERSPEECH 2010 Computational Paralinguistics Challenge (ComParE) as baseline set (Schuller et al., 2010a). This feature set contains 1.582 suprasegmental features obtained by passing 21 descriptive functionals (e.g., moments, percentiles, regression coefficients) on 38 Low Level Descriptors (LLD) extracted from the speech signal (see Table 7 for details). This configuration file is preferred over the one used in the 2015 edition of the ComParE Challenge, since in our recent work on a subset of this corpus (Lyakso et al., 2015), ComParE 2010 baseline set gave better results compared to the 2015 version, which is a 6.373-dimensional acoustic feature set.

In all classification experiments reported below, the acoustic features are first normalized and kernelized using Linear and Radial Basis Function (RBF) kernels before classification. Kernelization refers to obtaining an instance similarity matrix dubbed *kernel* $\mathbf{K} \in \mathbb{R}^{N \times N}$ from the data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$. It is popularly employed in linear classifiers to avoid the curse of dimensionality (especially when $d \gg N$) and to allow non-linear separability in an implicitly mapped hyper-space by means of non-linear kernels, such as RBF. As preprocessing, we apply z-normalization (i.e., standardization to zero-mean, unit variance) or min-max normalization to [0,1] range. The hyper parameters of classifiers are optimized using a two-fold, speaker independent cross-validation within the training set. The optimal parameters are finally used for model training and predicting the labels of the test set, which is only used once for reporting the results.

We report classification results in terms of accuracy and Unweighted Average Recall (UAR), which is introduced as performance measure in the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009). UAR is used to overcome the biased calculation of accuracy towards the majority class. It also gives a chance-level baseline performance as $1/K$, where K is the number of classes. Therefore, in a 3-class problem, we have 33.3% chance-level UAR.

Table 7
The openSMILE feature set used in the study with a standard configuration from INTERSPEECH 2010 ComParE Challenge (Schuller et al., 2010a). DDP: difference of difference of periods; LSP: line spectral pairs.

Descriptors	Functionals
F0 by sub-harmonic sum.	Arithmetic mean, standard deviation
F0 envelope	Linear regression coefficients 1/2
Jitter DDP	Linear regression error (quadratic/absolute)
Jitter local	Percentile 1/99
Log mel freq. band [0−7]	Percentile range 99−1
LSP frequency [0−7]	Quartile 1/2/3
MFCC [0−14]	Quartile range 2-1/3-2/3-1
PCM loudness	Relative position of minimum/maximum
Probability of voicing	Skewness, kurtosis
Shimmer local	Up-level time 75/90

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Table 8
Test set automatic classification results (%) for three emotional states.

Preprocess		UAR			Accuracy		
Normalization	Kernel	PLS	ELM	SVM	PLS	ELM	SVM
z-norm	Linear	47.5	**51.5**	48.6	52.0	52.0	53.7
Min−max	Linear	47.2	49.1	48.2	51.8	53.9	53.3
z-norm	RBF	50.7	50.9	49.6	56.0	56.2	55.6
Min−max	RBF	48.1	48.9	49.6	52.0	53.9	54.9

Table 9
Row normalized (%) confusion matrices for automatic emotional state recognition giving the highest UAR (51.5% using z-norm, linear kernel, ELM) with age-group breakdown.

	Child's age								
	3−4 years			5 years			6−7 years		
State	Disc	Neut	Comf	Disc	Neut	Comf	Disc	Neut	Comf
Discomfort	**51**	15	34	**64**	24	12	**27**	59	14
Neutral	14	**69**	17	34	**48**	18	17	**58**	25
Comfort	34	20	**46**	41	10	**48**	19	30	**51**

4.4. Automatic classification of child emotional states

In affective computing, *arousal* and *valence* are the two main dimensions along which continuous and dimensional affect is measured (Russell, 1980). Arousal is defined as physiological/psychological state of being (re-)active, while valence is the feeling of positiveness (Schuller, 2011). The comfort classification can be thought as a three-state valence classification problem. It is well known that valence classification from acoustics is poorer compared to arousal classification, and is almost at chance level in challenging conditions (i.e., without adaptation to cross-corpus settings) (Schuller et al., 2010b). The test set classification results (in terms of both accuracy and UAR) obtained from two normalization, two kernel and three classifier alternatives are presented in Table 8. The confusion matrices for the predictions with the highest UAR (z-norm, Linear kernel, ELM) are shown in Table 9.

From Table 8, we observe that using only the acoustic features, it is possible to get higher than chance-level UAR scores. However, the best accuracy (56%) is much below the gold standard obtained from human perception experiments (66%). Note that the human perception experiments described in Section 3 were done on a subset of the test set with 90 instances, and the decisions of 300 human listeners were fused. Here, we report the performance of individual classifiers over 475 test set instances. Moreover, the classifiers do not benefit from linguistic information that might have been useful for human discrimination.

The results in Table 9 are quite different than their human perception counterpart. Interestingly, the automatic emotion recognition performances are found as 55.3% and 45.5% UAR for 3−4 years group and 6−7 years group, respectively. We observe that while the UAR performance of human perception of affective states improves with increasing age, acoustics-based automatic classification gives higher performance with younger children. This may imply that the human listeners make implicit use of linguistic information that develops with age, such as the expanding vocabulary of the child.

4.5. Automatic classification of child age group and gender

The results for automatic three-level age group classification and the confusion matrices corresponding to the best UAR performance (z-norm, Linear kernel, PLS) are given in Tables 10 and 11, respectively. Comparing the UAR performances against the chance level, both in human and in machine classification this task is found to be the hardest among the three paralinguistic tasks dealt with in this paper. On the other hand, this is the only task where the best overall UAR score outperforms the one found in the human perception tests.

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Table 10
Test set automatic classification results (%) for three-class age group estimation.

Preprocess		UAR			Accuracy		
Normalization	Kernel	PLS	ELM	SVM	PLS	ELM	SVM
z-norm	Linear	**54.2**	53.0	52.5	51.8	48.2	53.3
Min−max	Linear	49.1	48.7	52.4	46.7	47.2	53.7
z-norm	RBF	52.0	48.9	52.4	48.0	48.0	49.9
Min−max	RBF	48.4	50.0	51.5	46.1	48.2	48.6

Table 11
Confusion matrices for child age classification giving the highest test set UAR (54.2% using z-norm, linear kernel and PLS) with comfort state breakdown.

	Emotional state								
	Discomfort			Neutral			Comfort		
Age (y)	3−4 y	5 y	6−7 y	3−4 y	5 y	6−7 y	3−4 y	5 y	6−7 y
3−4	**64**	25	11	**48**	45	7	**63**	31	6
5	12	**72**	16	8	**64**	28	12	**57**	31
6−7	0	73	**27**	8	54	**38**	13	41	**46**

Considering the confusion matrices in Table 11, we observe that similar to the case with humans, confusions arise between the middle group (5 years) and the other two groups, which can be attributed to the narrow age span of this class. Confusion between 3−4 and 6−7 years is low in general. On the other hand, it is interesting to see that higher UAR performance of age prediction is observed with comfort and discomfort states (55%) compared to the neutral state (50%). Ordinarily, one would expect the contrary, as both in speech recognition and in recognition of speaker traits, emotional speech generally gives lower performance compared to neutral speech (Schuller, 2011).

Finally, the automatic classification results for child gender and the confusion matrices corresponding to the best UAR performance prediction are listed in Tables 12 and 13, respectively. For gender classification, the best machine UAR performance is found lower than the UAR obtained from the human perception test. Comparing the confusion matrices of machine and human classification, we see a similar pattern in the discomfort state: recall of the male

Table 12
Test set automatic classification results (%) for child's gender.

Preprocess		UAR			Accuracy		
Normalization	Kernel	PLS	ELM	SVM	PLS	ELM	SVM
z-norm	Linear	54.6	49.5	52.4	56.6	49.1	52.6
Min−max	Linear	**57.0**	50.2	51.7	58.5	50.5	52.0
z-norm	RBF	48.7	49.5	48.6	48.6	49.7	48.2
Min−max	RBF	56.7	52.5	50.6	58.5	53.3	50.5

Table 13
Confusion matrices for automatic gender classification giving the highest UAR (57.0% using min−max normalization, linear kernel and PLS) with emotional state breakdown.

	Emotional state					
	Discomfort		Neutral		Comfort	
Gender	Male	Female	Male	Female	Male	Female
Male	**60**	40	**37**	63	**44**	56
Female	51	**49**	13	**87**	35	**65**

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

class is markedly higher than the recall of the female class. On the other hand, the recall patterns are different in the other two emotional states: the recall of female class is higher than the recall of male class in machine classification, while it is the opposite case for human perception. In the special case of discomfort, females are highly confused as male (51% by machine, 48% by humans). It is widely known that children's acoustic features are very similar in two genders and thus discrimination is hard. Our experimental results indicate that discomfort vocalizations of female children resemble those of male children.

Mann−Whitney U tests are administered on predictions to measure statistical significance of gender recall performance under different emotional states. The test results indicate that female recall performances are statistically different between emotional states ($p < 10^{-6}$, $p < 10^{-4}$, $p < 0.05$ for neutral vs. discomfort, neutral vs. comfort and comfort vs. discomfort, respectively). Male recall is found significantly different between comfort and discomfort as well as between neutral and discomfort states ($p < 0.05$), while no statistical difference is observed between male recall performances under neutral and comfort states.

4.6. Reproducibility

The MATLAB scripts and extracted openSMILE features used in automatic classification experiments are available over the following GitHub repository https://github.com/hysmky/emochildru. The codes can be used to reproduce the results given in this paper.

5. Discussion and conclusions

The present work is part of an emotional development study, which investigates emotional states in verbal and non-verbal behavior of kids during the first seven years of life. Choosing the age range as 3−7 years is due to the evolution of the grammatical skills of speech at 4 years, and the ability of effective communication of a child with an adult. In this age range, regulation of emotional expressions is not fully developed yet, and the emotional expressions are purer, as the contribution of society in the organization of the child's behavior is comparatively small. The upper bound age of seven years is associated with the end of the preschool time of children in the Russian Federation. There are only a few databases with emotional speech of children before 4 years of age in general (Lyakso et al., 2010a).

The presented experimental results show that lexical information has more discriminative power in recognition of comfort compared to acoustic features for speech samples of 6−7 year old children. The database contains instances (especially for smaller infants) that are difficult to annotate, even by the parents. Older infants speak more clearly. Human perception of emotion is higher in the speech of older children, which has several implications. It is harder to recognize sentiment with younger children, since linguistic and acoustic control skills are not mature enough. Furthermore, the comparative experiments in age and gender classification tasks also reveal the importance of linguistic models for better discrimination. There is no ASR solution for Russian child speech yet, which suggests this would be good future research direction, both as an independent study, and for multi-modal paralinguistic analysis. Analyzing/monitoring child emotion in early ages is important not only for linguistics, but also for analysis of neurological development/disorders. Moreover, all three paralinguistic tasks targeted here are important for developing intelligent tutoring systems for pre-school education.

In age group classification, the automatic classification study reveals higher performance compared to human perception, which can be taken as a gold standard, and falls below human performance in the other two. Especially for emotional state classification, the human listeners may be using linguistic content to their advantage, and it is difficult to quantify this. Further research is necessary to achieve and to outperform human performance. Multimodal fusion from linguistic and visual cues seems a promising step in this direction. Another important research direction is cross-corpus, cross-language analysis in child paralinguistics, as it is the key to leveraging additional data sources in training automatic systems.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grant nos. 16-06-00024, 15-06-07852, and 16-37-60100), Russian Foundation for Basic Research − DHSS (grant No 17-06-00503), by the grant of the

President of Russia (project No MD-254.2017.8), by the Government of Russia (grant No 074-U01), by Boğaziçi University (project BAP 16A01P4) and by the BAGEP Award of the Science Academy.

References

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M.J., Steidl, S., Wong, M., 2005. The PF_STAR children's speech corpus. In: Proceedings of INTERSPEECH, pp. 2761–2764.

Batliner, A., Steidl, S., Nöth, E., 2008. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In: Proceedings of the LREC-2008 Workshop of on Corpora for Research on Emotion and Affect, pp. 28–31.

Bell, L., Boye, J., Gustafson, J., Heldner, M., Lindström, A., Wirén, M., 2005. The Swedish NICE Corpus−spoken dialogues between children and embodied characters in a computer game scenario. In: Proceedings of EUROSPEECH. ISCA, pp. 2765–2768.

Bolaños, D., Cole, R.A., Ward, W., Borts, E., Svirsky, E., 2011. FLORA: fluent oral reading assessment of children's speech. ACM Trans. Speech Lang. Process. (TSLP) 7, 16.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of German emotional speech. In: Proceedings of INTERSPEECH, pp. 1517–1520.

Engberg, I.S., Hansen, A.V., 1996. Documentation of the danish emotional speech database DES. Center for Person Kommunikation, Denmark, pp. 1–22. Internal AAU report.

Eyben, F., Wöllmer, M., Schuller, B., 2010. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the Eighteenth ACM International Conference on Multimedia. ACM, pp. 1459–1462.

Fringi, E., Lehman, J.F., Russell, M., 2015. Evidence of phonological processes in automatic recognition of children's speech. In: Proceedings of INTERSPEECH, pp. 1621–1624.

Gerosa, M., Giuliani, D., Brugnara, F., 2007. Acoustic variability and automatic recognition of children's speech. Speech Commun. 49, 847–860.

Govender, A., Wet, F.d., Tapamo, J.R., 2015. HMM adaptation for child speech synthesis. In: Proceedings of INTERSPEECH, pp. 1640–1644.

Guo, J., Paturi, R., Yeung, G., Lulich, S.M., Arsikere, H., Alwan, A., 2015. Age-dependent height estimation and speaker normalization for children's speech using the first three subglottal resonances. In: Proceedings of INTERSPEECH, pp. 1665–1669.

Huang, G.B., Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42, 513–529.

Jürgens, R., Drolet, M., Pirow, R., Scheiner, E., Fischer, J., 2013. Encoding conditions affect recognition of vocally expressed emotions across cultures. Front. Psychol. 4, 111.

Kaya, H., Gürpınar, F., Afshar, S., Salah, A. A., 2015a. Contrasting and combining least squares based learners for emotion recognition in the wild. In: Proceedings of the ACM on International Conference on Multimodal Interaction, ACM. 459−466,

Kaya, H., Gürpınar, F., Salah, A.A., 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis. Comput. http://dx.doi.org/10.1016/j.imavis.2017.01.012.

Kaya, H., Karpov, A.A., 2016. Fusing acoustic feature representations for computational paralinguistics tasks. In: Proceedings of INTERSPEECH. San Francisco, USA, pp. 2046–2050.

Kaya, H., Karpov, A.A., Salah, A.A., 2015. Fisher vectors with cascaded normalization for paralinguistic analysis. In: Proceedings of INTERSPEECH. Dresden, Germany, pp. 909–913.

Kaya, H., Salah, A.A., 2016. Combining modality-specific extreme learning machines for emotion recognition in the wild. J. Multimodal User Interfaces 10, 139–149. http://dx.doi.org/10.1007/s12193-015-0175-6.

Kaya, H., Salah, A.A., Gurgen, S.F., Ekenel, H., 2014. Protocol and baseline for experiments on Bogazici University Turkish emotional speech corpus. In: Proceedings of the Twenty-Second IEEE Signal Processing and Communications Applications Conference (SIU), pp. 1698–1701.

Laukka, P., Elfenbein, H.A., Söder, N., Nordström, H., Althoff, J., Chui, W., Iraki, F.K., Rockstuhl, T., Thingujam, N.S., 2013. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. Front. Psychol. 4, 353.

Liao, H., Pundak, G., Siohan, O., Carroll, M.K., Coccaro, N., Jiang, Q.M., Sainath, T.N., Senior, A., Beaufays, F., Bacchiani, M., 2015. Large vocabulary automatic speech recognition for children. In: Proceedings of INTERSPEECH, pp. 1611–1615.

Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F., 2012. The brain basis of emotion: a meta-analytic review. Behav. Brain Sci. 35, 121–143.

Lyakso, E., Frolova, O., 2015. Emotion state manifestation in voice features: chimpanzees, human infants, children, adults. In: Proceedings of the International Conference on Speech and Computer (SPECOM). Springer, pp. 201–208.

Lyakso, E., Frolova, O., Dmitrieva, E., Grigorev, A., Kaya, H., Salah, A.A., Karpov, A., 2015. EmoChildRu: emotional child russian speech corpus. In: Proceedings of the International Conference on Speech and Computer (SPECOM). Springer, pp. 144–152.

Lyakso, E., Gromova, A., 2005. The acoustic characteristics of russian vowels in children of 4−5 years of age. Psychol. Lang. Commun. 9, 5–14.

Lyakso, E., Kurazova, A., Gromova, A., Ostrouxov, A., 2006. Recognition of words and phrases of 4-5-years-old children by adults. In: Proceedings of the International Conference on Speech and Computer (SPECOM), pp. 567–570.

Lyakso, E.E., Frolova, O.V., Grigorev, A.S., Sokolova, V.D., Yarotskaya, K.A., 2016. Recognition of adults emotional state of typically developing children and children with autism spectrum disorders. Neurosci. Behav. Physiol. 102, 729–741.

Lyakso, E.E., Frolova, O.V., Kurazhova, A.V., Gaikova, J.S., 2010a. Russian infants and children's sounds and speech corpuses for language acquisition studies. In: Proceedings of INTERSPEECH, pp. 1981–1988.

Lyakso, E.E., Ushakova, T.N., Frolova, O.V., Kurazhova, A.V., Bednaya, E.D., Gaikova, J.S., Grigoriev, A.S., Soloviev, A.N., Ostrouchov, A.V., 2010b. Russian children's vocabulary, speech imitation and reading skills mastery. Int. J. Psychophysiol. 77, 310.

Meinedo, H., Trancoso, I., 2011. Age and gender detection in the I-DASH project. ACM Trans. Speech Lang. Process. (TSLP) 7, 13.

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002

Pérez-Espinosa, H., Reyes-García, C.A., Villaseñor Pineda, L., 2011. EmoWisconsin: an emotional children speech database in Mexican Spanish. In: Proceedings of the Affective Computing and Intelligent Interaction, LNCS, 6975. Springer, pp. 62–71.
Potamianos, A., Giuliani, D., Narayanan, S.S., Berkling, K., 2011. Introduction to the special issue on speech and language processing of children's speech for child-machine interaction applications. ACM Trans. Speech Lang. Process. (TSLP) 7, 11.
Rao, C.R., Mitra, S.K., 1971. Generalized inverse of matrices and its applications. 7, Wiley, New York.
Rifkin, R., Yeo, G., Poggio, T., 2003. Regularized least-squares classification. Nato Sci. Ser. Sub Ser. III Comput. Syst. Sci. 190, 131–154.
Rigoulot, S., Wassiliwizky, E., Pell, M., 2013. Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition. Front. Psychol. 4, 367.
Russell, J.A., 1980. A circumplex model of affect. J. Personal. Soc. Psychol. 39, 1161–1178.
Safavi, S., Jancovic, P., Russell, M.J., Carey, M.J., 2013. Identification of gender from children's speech by computers and humans. In: Proceedings of INTERSPEECH. Lyon, France, pp. 2440–2444.
Safavi, S., Russell, M.J., Jancovic, P., 2014. Identification of age-group from children's speech by computers and humans. In: Proceedings of INTERSPEECH. Singapore, pp. 243–247.
Schuller, B., 2011. Voice and speech analysis in search of states and traits. Computer Anxalysis of Human Behavior. Springer, pp. 227–253.
Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), IEEE, Montreal, Canada, 577−580,
Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. In: Proceedings of INTERSPEECH, pp. 312–315.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., 2010a. The INTERSPEECH 2010 paralinguistic challenge. In: Proceedings of INTERSPEECH, pp. 2795–2798.
Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010b. Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. 1, 119–131.
Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9, 293–300.
Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun. 48, 1162–1181.
Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., Belin, P., 2014. Crossmodal adaptation in right posterior superior temporal sulcus during face−voice emotional integration. J. Neurosci. 34, 6813–6821.
Wold, H., 1985. Partial least squares. In: Kotz, S., Johnson, N. (Eds.), Encyclopedia of Statistical Sciences. Wiley, New York, pp. 581–591.
Yildirim, S., Narayanan, S., Potamianos, A., 2011. Detecting emotional state of a child in a conversational computer game. Comput. Speech Lang. 25, 29–44.

Please cite this article as: H. Kaya et al., Emotion, age, and gender classification in children's speech by humans and machines, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.06.002