Toward More Expressive Speech Communication in Human-Robot Interaction

Vlado Delić[1](✉), Branislav Borovac[1], Milan Gnjatović[1], Jovica Tasevski[1],
Dragiša Mišković[1], Darko Pekar[2], and Milan Sečujski[1]

[1] University of Novi Sad Faculty of Technical Sciences, Novi Sad, Serbia
vlado.delic@uns.ac.rs
[2] AlfaNum – Speech Technologies, Novi Sad, Serbia

Abstract. It is well known that speech communication is a very important segment of human-robot interaction. The paper presents our experience from the project "Design of Robots as Assistive Technology for the Treatment of Children with Developmental Disorders", with focus on the development of more expressive dialogue systems based on automatic speech recognition (ASR) and text-to-speech synthesis (TTS) in South Slavic languages. The paper presents the most recent results of our research related to the development of expressive conversational human-robot interaction, specifically in the field of conversion of voice and style of synthesized speech based on a new generation of deep neural network (DNN) based speech synthesis algorithms, as well as the field of emotional speech recognition. The development of dialogue strategies is described in more details in the second part of the paper, as well as the experience in their clinical applications for treatment of children with cerebral palsy.

Keywords: Human-Robot interaction · Speech technology
Expressive communication · Dialogue systems

1 Introduction

Human-machine interaction is one of major challenges in the development of robots. Humanoid robots are usually capable of recognizing human speech with a certain degree of accuracy, as well as to synthesize human-like speech. The level of automatic speech recognition (ASR) and spoken language understanding (SLU) depends on many conditions: language, size of vocabulary, noise level, reverberation, microphone and its position, as well as speaking style. On the other hand, achieving a high level of naturalness and expressiveness of speech produced by TTS is also a great technological challenge, and it has been accomplished for still a relatively small number of languages, particularly those with large speaker bases and market potential. For these reasons, conversational human-robot interaction is still unavailable for a large majority of languages, and has reached various degrees of development.

Speech dialogue expressing emotions and attitudes is very important in human-robot communication and, in fact, more expressive robots have been proven to be preferable over more efficient ones [1]. Emotion based human-robot interaction can be considered from different points of view; apart from verbal communication, the use of non-verbal cues such

A. Ronzhin et al. (Eds.): ICR 2018, LNAI 11097, pp. 44–51, 2018.
https://doi.org/10.1007/978-3-319-99582-3_5

as mimics and body gestures can improve the understanding interlocutors' intentions [2]. More expressive verbal human-robot communication is considered in more details in this paper. Section 2 presents some research results in the development of emotional speech recognition and some new research results in the conversion of both voice and style of synthesized speech. The applications of speech technologies in the development of a humanoid robot and its dialogue system are described in more details in the Sect. 3, with focus on experience in applications of human-robot speech interaction tested in a hospital as assistive technology for the treatment of children with developmental disorders.

2 Development of ASR and TTS for More Expressive Speech Communication

Robot MARKO is able to speak in Serbian, based on ASR and TTS developed at the project "Development of Dialogue Systems for Serbian and Other South Slavic Languages". For the first time, these were small to medium size vocabulary ASR [3–5] and a concatenative TTS (a female voice) with intonation predicted by regression trees [6]. They provide from one side the recognition of speech commands in the domain of therapeutic work with children with developmental disorders, and from the other side synthesized speech that is comprehensible, reasonably natural-sounding, but always in a neutral speech style. In order to provide more expressive verbal communication, the activities on the project have included research on emotional speech recognition and synthesis of speech with emphasis on creating a framework which will allow easy incorporation of new speech styles/emotions and new speaker voices.

2.1 Emotional Speech Recognition

The research question of automatic emotional speech recognition integrates two issues: selection of an appropriate feature set and investigation of different classification techniques. Most speech emotion recognition systems were based on hidden Markov models (HMM) [7]. Recognising realistic emotions and affect in speech can be difficult, especially if emotions have a low level of arousal and valence [8].

Discrimination capability of the usually proposed feature set is compared with two feature sets (prosodic and spectral feature sets separately) in a five emotional states classification task (anger, joy, sadness, fear and neutral). Four different classifiers (linear Bayes classifier, perceptron rule, kNN classifier and multilayer perceptron) are trained and tested with observed three feature sets on the corpus of "Emotional and Attitude Expressive Speech" (GEES). A set of experiments with three feature sets: the prosodic, the spectral, and the combined one has been described in [9]. The linear Bayes, the perceptron rule and the kNN classifier were considered in all three experiments. The experimental results show that the highest recognition accuracy was obtained with the third feature set using the linear Bayes classifier.

Better recognition of emotional speech will provide more natural verbal human-machine communication. Based on recognized human emotions, a robot can correct the dialogue strategy. It should also be noted that a good dialogue strategy can sometimes be chosen solely on the basis of the decision whether the speaker emotion was identified as positive or negative.

2.2 Conversion of Voice and Speech Style

Recent technological development has enabled us to use deep neural networks (DNN) to model speech. A simple DNN-based TTS system is shown in Fig. 1. It consists of two neural networks, one predicting durations of phones, and the other predicting acoustic features. In our experiments all networks have 4 hidden layers with 1024 neurons, first three form a feed-forward network, while the output layer has neurons with long short-term memory (LSTM). The inputs are linguistic features extracted from annotated text. The acoustic features produced by the network are given to the vocoder from which the output speech signal is obtained. For producing intelligible and natural sounding speech, the model needs to be trained on several hours of speech, annotated phonetically and prosodically. While phonetic annotation is largely automatic, prosodic annotation requires significant human effort.

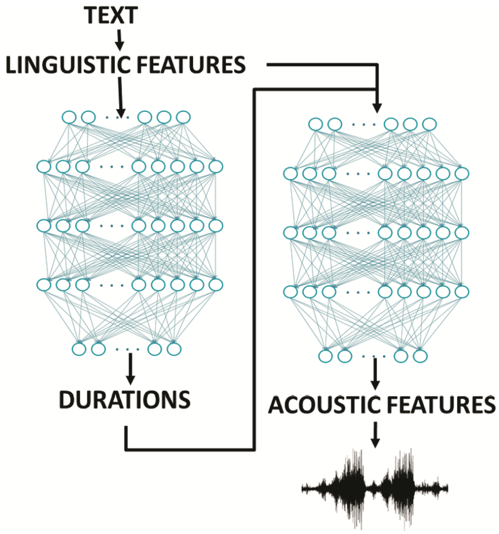

Fig. 1. Regular DNN based TTS.

In our recent research [10], we compared three different DNN-based methods for producing synthesized speech in multiple styles, originally proposed for modeling multiple speakers with a limited amount of data per speaker.

- **The method based on style code.** The idea is to have a database with one speaker, speaking in multiple styles. Besides standard linguistic feature inputs, another input is used, indicating the speech style. During training, this input is used to indicate the style of the speech data used, while in the synthesis phase, given a certain style code, the network will know which speech style to produce. The style is coded as one-hot vector.

- **Shared hidden layers.** This method is based on idea of having a separate output layer for each style, while sharing the hidden layers among different styles. In this way, all sections of the multi-style corpus are used to train the shared layers, but only a specific part of the corpus is used to train a certain output layer. In the synthesis stage, depending on which speaking style is to be produced, only a certain part of the entire model will be used.
- **Re-trained model.** The last model, maybe the most intuitive one, is based on re-training an already trained model. Namely, the idea is to train a model on a corpus of neutral speech in the regular way, and then to adapt it to a smaller speech corpus in a certain style. In such a case the number of models needed is equal to the number of styles to be produced.

Two listening tests (MOS and MUSHRA) with 20 amateur listeners, with neutral speech (3 h of training data) and angry, happy and sarcastic style of speech (5 min of training data per style) are conducted in [10]. It has been shown that intelligible and to some extent natural expressive speech can be produced by having only 5 min of speech in a certain style.

The approaches described above have originally been used for producing synthetic speech in different voices [11, 12]. However, in this case the amount of training material per speaker is higher, around 20 min. The ultimate aim of our research is to develop a framework able to produce synthetic speech in an arbitrary voice and speaking style, contributing to more expressive human-robot voice communication.

3 Evaluating the Robot Dialogue Behavior

The conversational human-like robot MARKO was developed as assistive tool for robot-supported therapy of children with cerebral palsy and similar movement disorders. One of the crucial functionalities of this robot is that it can engage in three-party natural language interaction with the child and the therapist [13–16, 18].

The dialogue behavior of the robot MARKO was experimentally evaluated in a therapeutic settings at the Clinic of Pediatric Rehabilitation in Novi Sad, Serbia (cf. Figure 2). The children who participated in this study were selected by qualified therapists. Their parents gave written permission for the children to participate, and, when possible, the children above age five gave assent. Due to the sensitivity of the research, all robot actions and speech acts were controlled by a human operator, under supervision of a therapist (cf. [18]).

The evaluation was performed in two phases. The aim of the first phase was to assess the children's receptivity to the robot MARKO at the first encounter. Thus, the experimental settings were not strictly therapeutic, but rather designed to engage a child in the interaction with the robot, and to motivate it to perform nonverbal actions on request. For example, MARKO says that it has lost a toy (i.e., confronting the child with a simple discourse), mimicking a sad facial expression (non-verbal expression of emotions can improve intention recognition, cf. [2]), and asks the child for help to find it. Twenty-nine children were involved in the first experimental phase (13 f, 16 m, avg. age 9.1, st. dev. 3.54). Twelve of them were healthy, and 17 were recruited from among patients

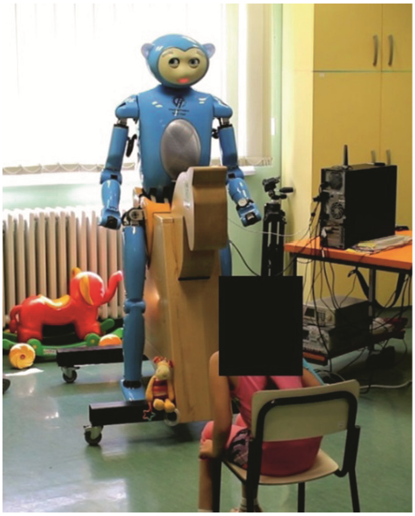

Fig. 2. The robot MARKO in the experimental setting.

with cerebral palsy and other movement disorders. In addition, the control group of patients not exposed to the robot contained 15 children (6 f, 9 m, avg. age 6.8, st. dev. 3.19). The first experimental phase validated that children positively respond to MARKO at first encounter, engage in interaction, and accept and perform given instructions. More detailed information on subjects, experimental settings, and produced corpus of child-robot interaction is given in [17, 18].

Table 1. Subjects.

ID	Age	Sex	Height [cm]	Width [kg]	Diagnosis	Mobility
s_1	15	F	159	46	hemiparesis, right sided weakness	can stand, can walk
s_2	14	F	164	50	hemiparesis, right arm disorder	can stand, can walk
s_3	9	F	129	27	paralysis cerebralis infantilis, vision problems	sitting, can stand, can walk a little with assistance
s_4	9	M	127	32	hemiparesis, left sided weakness, brain hemorrhage	can stand, can walk
s_5	15	F	150	50	paralysis cerebralis infantilis, spinal surgery	can stand, can walk
s_6	11	F	155	45	paralysis cerebralis infantilis (spastic diplegia), difficulty with speaking, vision problems, sensorimotor integration disorder	can stand, can walk
s_7	5	M	124	34	paralysis cerebralis infantilis	can stand, can walk with assistance

The second experimental phase was conducted as part of work reported in this paper. Seven children with cerebral palsy and similar movement disorders participated in this study (5 f, 2 m). The basic information on the children is given in Table 1.

The aim of this phase was to assess the children's longer-term motivation to undergo therapy in a therapeutic context. Therefore, compared to the first phase, the following aspects of the experimental settings were modified in the second phase:

(i) *Therapeutic settings.* The interaction was strictly focused on therapeutic exercises. The children were verbally instructed to perform ten Frenkel's exercises [19] selected by their respective therapists, including, but not limited to stretching in standing and sitting positions, walking along a line and between two parallel lines, lateral walking over obstacles, etc.

(ii) *Multiple sessions per child.* Each child was participating in a series of therapeutic sessions – one session per day, in a sequence of working days. Forty-four sessions were recorded. The average duration of a session was approximately 16 min, with a standard deviation of approximately 4 min. The additional information on the recorded sessions is given in Table 2.

Table 2. Sessions.

Subject ID	# Sessions	Average duration of session [s]	Standard deviation [s]
s_1	7	758	135
s_2	5	869	310
s_3	13	1104	286
s_4	4	949	237
s_5	3	933	22
s_6	7	1015	120
s_7	5	834	242
Total:	44	952	249

(iii) *Three-party interaction:* The interaction was evolving between the child, the therapist and the robot. For each child separately, in the first session, the therapist verbally instructed the child to perform exercises, while MARKO engaged in conversation either to encourage the child (e.g. by commending it), or to draw the child's attention to some aspects of the current exercise (e.g., warning the child to straighten the spine or knees, to drop the heel, etc.). From the second session, MARKO was taking the initiative in interaction from the therapist, following the dialogue strategy introduced in [18]. The robot was primarily instructing the child, while the role of the therapist was corrective.

A qualitative insight into the children's motivation to undergo robot-supported therapy was provided by their long-term therapists. The positive motivation was observed in all subjects. With respect to the level of motivation, the subjects can be classified in two groups. Subjects s_1, s_2 and s_7 were accepting the robot's instructions, and expressed steady-state motivation that was maintained but not increased through the sessions. In the remaining subjects, the motivation to undergo the therapy was established and then increased through the sessions. These subjects were not only accepting the robot's instructions, but also expressing higher engagement, e.g., they verbally interacted with MARKO, memorized the exercises, and tended to perform them in advance of the robot's instruction.

4 Conclusion

Experience in the development of a dialogue system for a humanoid robot is presented in the paper with focus on possible progress based on expressive speech recognition and synthesis. Speech technology development has evolved from small vocabulary HMM-based speech command recognition and neutral-style speech synthesized by concatenation, toward large vocabulary ASR including emotive speech and multi-style TTS – based on advantages of deep neural networks.

The reported dialogue system integrated with the robot MARKO has been tested in a clinical context. Although it includes the functionality of concatenative TTS, the children liked to participate in the dialogue and were motivated to repeat exercises that they otherwise find hard and boring (e.g., subject s_3 asked her parents to exercise on her own between sessions in order to be better during the next session with MARKO). According to the qualitative assessment provided by the involved therapists, introduction of emotions in human-robot interaction either by mimics or voice has contributed to the effects of dialogue. It is expected that a more expressive voice of the robot will further increase the positive effects, but real benefits will be assessed in future work.

Acknowledgments. Research was supported in part by the Ministry of Education, Science and Technological Development of Serbia (grants TR32035 and III44008).

References

1. Hamacher, A., Bianchi-Berthouze, N., Pipe, A.G., Eder, K.: Believing in BERT: using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction. In: 25th IEEE International Symposium on Robot and Human Interactive Communication, 26–31 August 2016, 8 pages (2016). https://doi.org/10.1109/roman.2016.7745163
2. Berns, K., Zafar, Z.: Emotion based human-robot interaction. In: Ronzhin, A., Shishlakov, V. (eds.) 13th International Scientific-Technical Conference on Electromechanics and Robotics "Zavalishin's Readings", St. Petersburg, Russia, 18–21 April 2018, MATEC Web of Conferences, vol. 161, Article 01001, 7 pages (2018). https://doi.org/10.1051/matecconf/201816101001
3. Popović, B., et al.: A novel split-and-merge algorithm for hierarchical clustering of Gaussian mixture models. Appl. Intell. **37**(3), 377–389 (2012). https://doi.org/10.1007/s10489-011-0333-9

4. Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., Delić, V.: Deep Neural Network based continuous speech recognition for Serbian Using the Kaldi Toolkit. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) SPECOM 2015. LNCS (LNAI), vol. 9319, pp. 186–192. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23132-7_23
5. Pakoci, E., Popović, B., Pekar, D.: Language model optimization for a deep neural network based speech recognition system for Serbian. In: Karpov, A., Potapova, R., Mporas, I. (eds.) SPECOM 2017. LNCS (LNAI), vol. 10458, pp. 483–492. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66429-3_48
6. Sečujski, M., Pekar, D., Knežević, D., Svrkota V.: Prosody prediction in speech synthesis based on regression trees. In: Halupka-Rešetar, S., et al. (eds.) The 3rd International Conference of Syntax, Phonology and Language Analysis, pp. 224–236. Cambridge Scholar Publishing (2012)
7. Nwe, T., Foo, S., De Silva, L.: Speech emotion recognition using hidden Markov models. Speech. **41**, 603–623 (2003)
8. Schüller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**, 1062–1087 (2011)
9. Delić, V., Bojanić, M., Gnjatović, M., Sečujski, M., Jovičić, S.: Discrimination capability of prosodic and spectral features for emotional speech recognition. Elektronika ir Elektrotechnika **18**(9), 51–54 (2012). https://doi.org/10.5755/j01.eee.18.9.2806
10. Suzić, S., Delić, T., Jovanović, V., Sečujski, M., Pekar D., Delić, V.: A comparison of multi-style DNN-based TTS approaches using small datasets. In: 13th International Scientific-Technical Conference on Electromechanics and Robotics "Zavalishin's Readings", St. Petersburg, Russia, April 2018, MATEC Web Conference, vol. 161, 6 pages (2018). https://doi.org/10.1051/matecconf/201816103005
11. Fan, Y., Qian, Y., Soong, F. K., He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, April 2015. https://doi.org/10.1109/icassp.2015.7178817
12. Hojo, N., Ijima, Y., Mizuno, H.: An investigation of DNN-based speech synthesis using speaker codes. In: Interspeech, San Francisco, USA. https://doi.org/10.21437/interspeech.2016-589
13. Gnjatović, M.: Therapist-centered design of a robot's dialogue behavior. Cogn. Comput. **6**(4), 775–788 (2014)
14. Gnjatović, M., Delić, V.: Cognitively-inspired representational approach to meaning in machine dialogue. Knowl. Based Syst. **71**, 25–33 (2014)
15. Gnjatović, M., Janev, M., Delić, V.: Focus tree: modeling attentional information in task-oriented human-machine interaction. Appl. Intell. **37**(3), 305–320 (2012)
16. Mišković, D., Gnjatović, M., Štrbac, P., Trenkić, B., Jakovljević, N., Delić, V.: Hybrid methodological approach to context-dependent speech recognition. Int. J. Adv. Robot. Syst. **14**(1), 12 (2017)
17. Gnjatović, M., et al.: Pilot corpus of child-robot interaction in therapeutic settings. In: Proceedings of the 8th IEEE International Conference on Cognitive Infocom. (CogInfoCom), Debrecen, Hungary, pp. 253–257 (2017)
18. Tasevski, J., Gnjatović, M., Borovac, B.: Assessing the Children's Receptivity to the Robot MARKO. Acta Polytechnica Hungarica, Special Issue on Cognitive Infocommunications (in press)
19. Zwecker, M., Zeilig, G., Ohry, A.: Professor Heinrich Sebastian Frenkel: a forgotten founder of rehabilitation medicine. Spinal Cord **42**, 55–56 (2004)