БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ • BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, **69**
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, **69**

София • 2018 • Sofia

Speech applications for human - robot interaction systems.

Matus Pleva, Stanislav Ondas

Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics,
Technical University of Kosice, Letna 9, Slovakia
Email: matus.pleva@tuke.sk

Abstract: *The speech communication is the most natural way of human-to-human-communication. The speech is a natural way of communication also for users of the robotic systems, especially for children, elderly and visually impaired people. This paper introduces recent progress in research and development of human robot interaction (HRI) based on speech enabled interfaces. Its main focus is to show the potential of the state-of-art automatic speech recognition techniques applied to modern HRI technologies in applications developed in Laboratory of speech and mobile technologies at the Technical university of Košice. Some of them were made together with international partners, like the English speech commands implemented to Jaguar robot with Center for Advanced Vehicular Technologies, Mississippi State University, US. Also the evaluation of the dialogue application with NAO robots will be presented with synthetized speech combined with gesture.*

Keywords: *speech recognition, HRI, personalization, application development.*

1. Introduction

The research group of the Laboratory of Speech and Mobile Technologies in Telecommunications (LSMTT - the Laboratory) is a part of Department of Electronics and Multimedia Communications which belongs to the Faculty of Electrical Engineering and Informatics of the Technical University of Košice (FEI TU). The young and productive team of around 20 members works continuously on automatic speech recognition and speech dialogue technologies for more than 30 years. During this period a lot of applications were developed thanks to numerous national projects, contracts with Slovak & foreign companies, 6th and 7th EU framework programme projects, international COST networking projects (cost.eu) and EU structural funds projects, which helps us to improve the technological facilities of the Laboratory. The research of the Laboratory is focused mainly on the Slovak language processing including automatic speech recognition [1] and synthesis [2]. A task for language independent voice search [3], English speech synthesis [4, 5] and English speech commands for noisy environment [6] were realized recently.

2. Speech application types

For speech communication with robots we have to distinguish between the main types of speech recognition technology used. We can divide the applications with automatic speech recognition (ASR) modules to 4 types:

1. Simple words – commands recognition. You can know these applications from multi-panel control in car with navigation and telephone dialing abilities, voice commands for mobile terminals / robots, etc. These systems can run locally on small embedded devices.
2. Sentences using fixed grammar (medium size vocabulary) – these applications usually provide simple dialogues. You can try them when using automatic telephone call center, interactive voice response (IVR) systems which can provide you some information or services only using your phone and speech input, etc. These dialogues could be implemented to robotic platform when there is an internet connection available.
3. Large Vocabulary Continuous Speech Recognition – LVCSR – these applications needs more computing power if they want to run locally on the device or robot. They are most commonly used for dictation software, voice search, etc. For running on robotic platform the internet connection to cloud or shared ASR modules are required if platform has lower computing resources.
4. Spontaneous speech recognition – this is still a challenge, especially if computer wants to understand the meaning of the recognized text. For spontaneous we can expect grammatically incorrect sentences, rude speaking, out of vocabulary (OOV) words like foreign names and companies, pauses during the sentence, etc.

3. Third party speech application programming interfaces (API)

When making decision about possible speech enabled API for the robotic platform the currently available third party API should be taken in the consideration. Many companies provide server based Large Vocabulary Continuous Speech Recognition services using their API. For example:

- ✓ https://cloud.google.com/speech/
  Google: Speech to text conversion powered by machine learning, for using the Speech API you have to use the Goggle Chrome web browser, or implement the required call to Google servers. The amount of requests is limited for free use.
- ✓ https://azure.microsoft.com/en-us/services/cognitive-services/speech/
  Microsoft: Convert spoken audio to text. It is a cloud service where you can request the transcription of the audio recording.
- ✓ https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API
  Mozilla: there is a CMUSphinx implementation in Gecko, but also implementation of Google Speech API could be executed in the browser

For using the online dictation mode there are more plugins for Google Chrome. For example https://dictation.io/ where you can choose a language from all Google Speech API supported languages and test the server platform before using it.
If a man wants to use paid online services there are also offerings from Amazon, Facebook and many others (vocapia.com, speechmatics.com, etc.).

Of course there are opportunities to use the local commercial apps, but they are standalone applications and for using the recognizer for a robotic platform you have to pay for a modified version and license every copy of it. The examples of standalone local applications are Dragon Naturally Speaking from Nuance, NEWTON Dictate, etc.

4. Third party free local automatic speech recognition engines

The local ASR engines are necessary for robotic platforms with limited internet access or when the privacy issues do not allow the sharing of the audio data through the internet to third parties. There are many free decoding engines available for these purposes like Julius (C++, HMM based), Kaldi (NN based), HTK (C++, HMM based, only research purposes), Sphinx (Java, HMM based, pocket mobile version available), DeepSpeech (Mozilla open source Speech-To-Text engine uses uses Google's TensorFlow - open-source software library for Machine Intelligence and Deep Learning) etc.

For local ASR modules, the requirements for the developers are higher. These engines are provided mainly as decoders with no models for particular languages. Some of them are providing English general language and acoustic models. Sphinx is providing models also for German, French, Dutch, Spanish, Italian, Mandarin, Hindi, Kazakh, Russian, Greek and Indian English.

On the other hand, when the developer holds a control on full configuration of the engine he can use a fixed grammar for noisy environment and command control of the robot or he can add OOV words to the dictionary and language model too. Although these modifications are not trivial and requires an experienced user in speech applications. The specialized language models for particular domains as law enforcement, medical, business, banking, sports etc. could significantly increase the accuracy and usability of the whole system.

The main advantage is that no data needs to be shared to third parties and no internet connection is required for local ASR solutions. The end-user should always mind that when using the party server based engines he accept the conditions of storing and analyzing the audio data for future improvements of the systems.

5. SCORPIO robot speech commands application in Slovak language

The SCORPIO is a small-size mini-teleoperator mobile service robot for booby-trap disposal [7]. An operator can manually control it through the portable briefcase remote control device using joystick, keyboard and buttons [7]. ZTV-VVU Kosice Company manufactures this robot and thanks to joint projects, TUKE collaborated on speech interface.

As you can see on Figure 1. and 2. the robot is controlled from portable briefcase with limited number of buttons. As the functions of the robot was increasing in time (number of cameras, lights, etc.) the rebuilding of the briefcase was necessary. The optimal solution was using the joystick from the briefcase only for controlling the movements of the robot and control all additional functions with speech interface.

Fig. 1. Service robot SCORPIO.

Fig. 2. Portable briefcase used for remote control of the robot SCORPIO.

The SCORPIO robot vehicle contains currently:

- ✓ Black and white cameras
- ✓ One color camera
- ✓ 2 laser pointers
- ✓ 3 distance sensors
- ✓ 7 lights

The remote control briefcase contains a small embedded PC with Intel Pentium-III based 1GHz CPU, 1 core, 4GB of Flash card storage and 1GB of RAM. The user interface contains:

- ✓ 12” TFT display,
- ✓ Joystick with dead men button,
- ✓ keyboard,
- ✓ many buttons

The briefcase is easily extendable using the USB connector on the left side. For enhancing the system with speech, enabled application the external USB sound card was used. The Julius [8] recognizer was used with acoustic models trained on landline telephone speech database SpeechDatE-SK database [9] using reference recognizer training procedure from the COST-249 project [10]. The results during the testing in noisy environment are presented in the Table 1. For demonstration we used a head set microphone and loudspeakers.

The "dead man button" on the joystick was used for confirmation and execution of the recognized command as you can see on the demo video[1]. The built-in TFT 12" display is used for the chosen camera view and the system status windows. The system status windows are simpler when in production operations. This video was made during development and the robot vehicle is also uncovered.
For better feedback to the operator, the recognized command is synthetized using simple DB selection synthesizer.

Table 1. The Word Error Rates for recognition in noisy conditions.

SNR (dB)	**25**	**10**
WER (%)	2.76	9.81

6. Jaguar robot speech commands application in English language

The robotic platform, in this case was a Jaguar V4, a small unmanned ground vehicle, that acts as a forward member of the SWAT tactical team, searching dangerous areas without putting the officers in harm's way (see Fig. 3) [11]. There is no extra officer for the robot operation and one of the team member has to hold the gun and control the robot together. Dr. Pleva from TUKE visited the team from CAVS, Mississippi State University in order to embed the voice recognition system to the robotic platform.

The Julius recognizer and freely available TIMIT + WSJ acoustic models [12] for English language were used together with noise models from TUKE [13, 14]. The fixed grammar was designed for this purpose taking in the account the noisy environment and speech in the background [15].

[1] http://speetis.fei.tuke.sk/video/scor2012.wmv

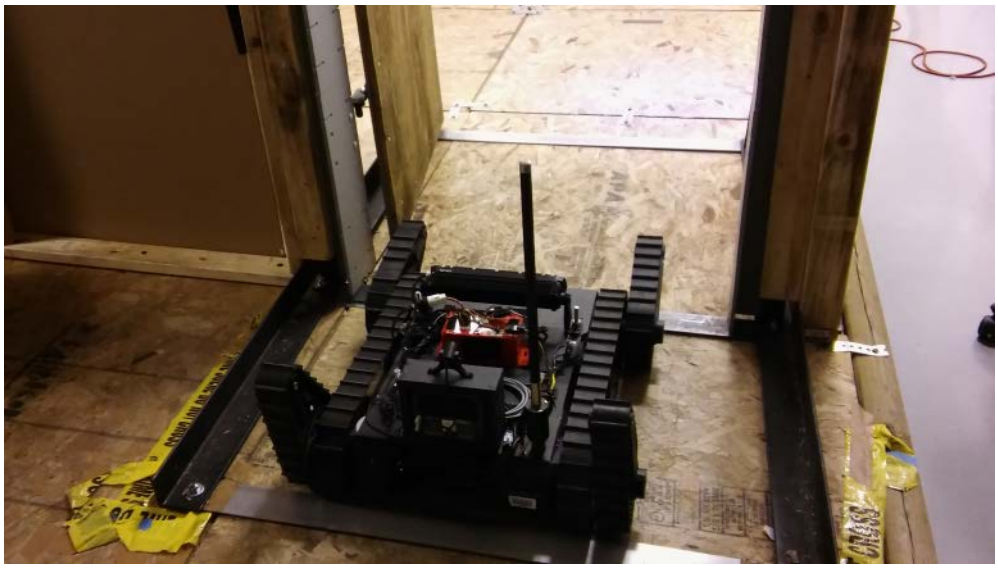

Fig. 3. Extended Jaguar V4 robotic platform on CAVS – Center for Advanced Vehicular Systems.

The solution presented in the demo video[2] is based on wireless Nunchuck Wii gamepad controller for robot movement and steering control on Fig. 4. The video from the chosen camera is displayed on smart glasses of the officer. The voice recognition module was used for controlling of the additional functions as lights, strobe light, cameras or playback of the chosen file in the loudspeaker of the robot.

Fig. 4. Wii Nunchuck controller with joystick and 2 buttons for robot movements control.

[2] http://bit.ly/1sxaqcK

7. NAO dialogue application based on VoiceXML interpreter in Slovak and English language with evaluation results and discussion

Humanoid robot NAO (see Fig. 5) is an autonomous programable robot developed by Softbank Robotics. NAO can be considered as a great tool to prepare a multimodal dialogue system due to its support of vision, hearing, gesture production and body language. Moreover, robot can run in an autonomous mode to produce human-like movements. The pilot version of the multimodal dialogue system for NAO enables multimodal interaction with the user in such manner that it takes a speech input from the user and it answers by a combination of synthetic speech and gestures [16]. Interaction with the user is managed by an external dialogue manager, which interprets VoiceXML language [17]. The architecture of the system is shown in Fig. 5.

Expect NAO, the basic system components are built-in automatic speech recognition system and text-to-speech system. Using of built-in speech technologies brings to the system the support of 19 languages that are supported by NAO itself. The next important component is the VoiceON dialogue manager, which manages the interaction by VoiceXML scripts interpretation. The communication between

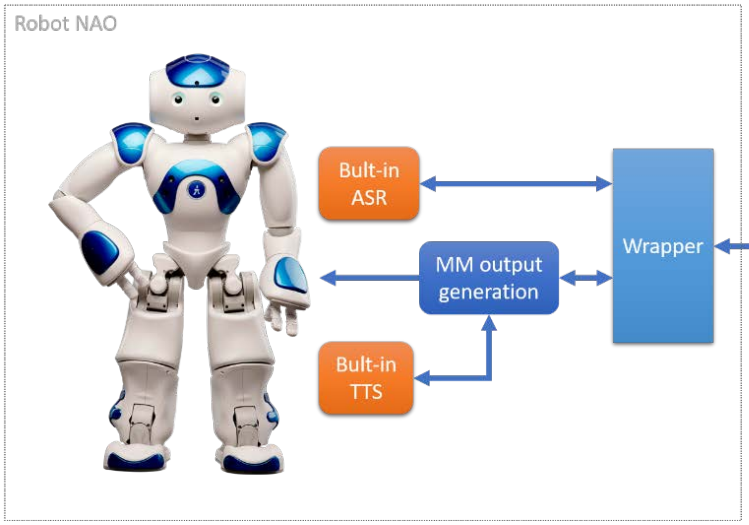

Fig. 5. NAO multimodal dialogue system.

NAO functionalities and the dialogue manager is transformed in wrapper module. To produce multimodal output of the robot the new simple module was designed – MultiModal output generation module, which join synthesized speech and gestures into one output stream. Module analysis sentences, which should be synthesized by

NAO's TTS module and it search for keywords, which relates to particular movements and head/hand gestures.

Because NAO originally does not support Slovak language, we use Czech language pack to enable Slovak dialogues. This cross lingual using is possible thanks to similar phonetic set of both languages. In case of speech recognition, ne special changes need to be done. In case of speech synthesis, Slovak text need to be modified to overcome of differences.

A pilot speech communication application was designed, and a preliminary evaluation was performed using subjective methods[3]. Evaluation was done by thirty test subjects – students, which interacted with robot. They were split into three distinct groups:

- First group interacted with the robot, which produced only speech without gestures and autonomous movements.
- Second group interacted with the robot, which produced speech together with gestures, but autonomous behavior was switched off.
- Third group interacted with the robot, when the autonomous behavior was switched on and the robot produced speech and gestures, which resulted from text processing.

The goal of evaluation was to examine the impact of different modes of gesture production on the perception of naturalness and smoothness of the interaction. Test subjects fill the questionnaire after the interaction with robot. Results of the preliminary evaluation (Fig.6) highlights importance of involving gestures into communication exchange. The interaction was realized also in English language for demonstration purposes of easy adaptation of the robotic multimodal platform to different languages[4].

[3] https://goo.gl/9I8xsS
[4] https://goo.gl/nVjBgj

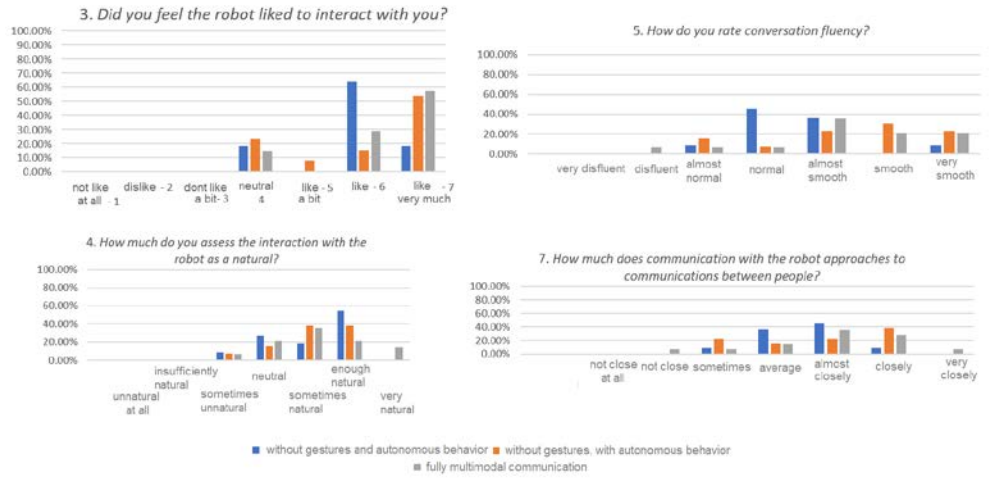

Fig. 6. Results of subjective evaluation.

Conclusion

The obtained results show the potential of using different speech modules for task of automatic speech recognition and synthesis. The evaluation of the multimodal platform shows that the automatic gesture synthesis allows a more natural human - robotic interaction and dialogue with the user. The presented modules was realized in offline mode with no internet connection needed so they provide better privacy of the user and better independence on the network infrastructure in rescue or law-enforcement operations.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under the research projects APVV-15-0517, APVV-15-0731 & APVV-14-0894 and by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the projects VEGA 1/0075/15 \& KEGA 055TUKE-4/2016.

References

1. Juhar et al., Recent progress in development of language model for Slovak large vocabulary continuous speech recognition. *New Technologies*. InTech, 261-276, 2012
2. Sulír, M., Juhár, J., Rusko, M., Development of the Slovak HMM-based tts system and evaluation of voices in respect to the used vocoding techniques, *Computing and Informatics*, 35 (6), pp. 1467-1490, 2016
3. Vavrek, J., Viszlay, P., Kiktova, E., Lojka, M., Juhar, J., Cizmar, A., Query-by-example retrieval via fast sequential dynamic time warping algorithm, *2015 38th International Conference on Telecommunications and Signal Processing, TSP 2015*, art. no. 7296440, IEEE, 2015

4. Sulir M et al. Speech Synthesis Evaluation for Robotic Interface. *Complex Control Systems*. 11 (1), 64-69, 2012
5. Ondas S et al., Speech technologies for advanced applications in service robotics. *Acta Polytechnica Hungarica*. 10 (5), 45-61, 2013
6. Pleva, M., Juhar, J., Cizmar, A., Hudson, C., Carruth, D.W., Bethel, C.L., *Implementing English speech interface to Jaguar robot for SWAT training*, SAMI 2017 - IEEE 15th International Symposium on Applied Machine Intelligence and Informatics, Proceedings, art. no. 7880284, IEEE, pp. 105-110, 2017
7. Ondas, S., Juhar, J., Pleva, M., Cizmar, A., Holcer, R. Service robot SCORPIO with robust speech interface, *International Journal of Advanced Robotic Systems*, SAGE, 10, art. no. 3, 2013
8. Lee A., Kawahara T. and Shikano K. Julius - an open source real-time large vocabulary recognition engine. *EUROSPEECH*, Aalborg, ISCA, pp. 1691-1694, 2001
9. Pollak P., et al. SpeechDat(E) Eastern European Telephone Speech Databases. Proc. of *LREC Satellite workshop XLDB*, Athens, Greece, ELRA, pp. 20-25, 2000
10. Johansen F.T., et al. The COST 249 SpeechDat Multilingual Reference Recogniser. *LREC*, Athens, Vol. 3, ELRA, pp. 1351-1355, 2000
11. C. Hudson, C. L. Bethel, D. W. Carruth, M. Pleva, J. Juhar and S. Ondas, A training tool for speech driven human-robot interaction applications, *2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Stary Smokovec, IEEE, pp. 1-6, 2017
12. Vertanen K., *Baseline WSJ acoustic models for HTK and Sphinx*: Training recipes and recognition experiments, Technical report, Cambridge, United Kingdom: Cavendish Laboratory, 2006
13. Pleva, M. Juhar, J., TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation, In: *LREC 2014*, Ninth International Conference on Language Resources and Evaluation, ELRA, Reykjavik, pp. 1709–1713, 2014
14. Pleva, M., Vozarikova, E., Dobos, L., Cizmar, A. The Joint Database ofAudio Events and Backgrounds for Monitoring of Urban Areas, *Journal of Electrical and Electronics Engineering*, 4 (1), pp. 185–188, 2011
15. M. Pleva, J. Juhar, A. Cizmar, C. Hudson, D. W. Carruth and C. L. Bethel, Implementing English speech interface to Jaguar robot for SWAT training, *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Herlany, IEEE, pp. 105-110, 2017
16. S. Ondáš, M. Pleva, J. Juhár and R. Husovský, Preliminary evaluation of the multimodal interactive system for NAO robot, *2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Stary Smokovec, IEEE, pp. 1-6, 2017
17. Juhár J, Ondas S, Cizmar A, Rusko M, Rozinaj G, Jarina R. Development of Slovak GALAXY/VoiceXML based spoken language dialogue system to retrieve information from the internet. In: *Ninth International Conference on Spoken Language Processing* ICSLP. - Bonn : Universität Bonn, pp. 485-488, 2006.

Речевые приложения для систем взаимодействия человека и робота.

Матуш Плева, Станислав Ондас
Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics,
Technical University of Kosice, Letna 9, Slovakia
Email: matus.pleva@tuke.sk

Аннотация: *Речевое общение является наиболее естественным способом общения человека и человека. Речь является естественным способом общения также для пользователей роботизированных систем, особенно для детей плоховидящих и пожильек людей. В настоящем документе представлен недавний прогресс в исследованиях и разработках человеческого робота (HRI) на основе интерфейсов с поддержкой речи. Основное внимание в нем - показать возможности современных автоматических методов распознавания речи, применяемых к современным технологиям HRI, в приложениях, разработанных в Лаборатории речевых и мобильных технологий в Техническом университете Кошице. Некоторые из них были сделаны вместе с международными партнерами, такими как английские речевые команды, реализованные для робота Jaguar с Центром усовершенствованных автомобильных технологий, Университет штата Миссисипи, США. Также оценка применения диалога с роботами NAO будет представлена синтезированной речью в сочетании с жестом.*

Ключевые слова: *распознавание речи, HRI, персонализация, разработка приложений.*