Lithuanian speech in Humanoid robots

L. Aidokas, G. Tamulevičius

Vilnius University, Institute of Data Science and Technologies

E-mail

Introduction

Automated speech recognition systems have been an established research topic for decades now. There has been some major breakthroughs recently, mostly because of various neural network techniques used in various ways. Social robots are robots that exploit ASR systems for communication purposes. As social robots are a rapidly growing area, they will have a huge impact on our everyday life in the future.

Problem statement

Challenges with Lithuanian language ASR systems:

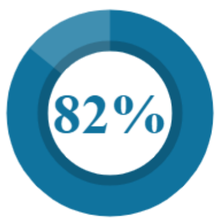
- Small population of native speakers;
- The training should include more data on young and elderly voices;
- Different Lithuanian language dialect and non-native speakers experience severely deteriorated language recognition performance;
- The current Lithuanian ASR do not consider the speaker's emotions;

Research goal

The goals of the research are:

- Collect and transcribe Lithuanian language speech examples from various places in Lithuania and train new Lithuanian ASR models, that would be more familiar with different Lithuanian language dialects;
- Develop solutions for interacting in HRI when atypical language problem arises, especially with non-locals;
- Increase the Lithuanian ASR model reliability in noisy, reverberating or crowded spaces;
- Improve Voice Activity Detection in situations with music in the background or when a person is talking and the speech is not directed towards the robot;
- To improve the extraction of suprasegmental properties, emotional and semantic content for more natural and immersive HRI process;

Collected data



Male voices

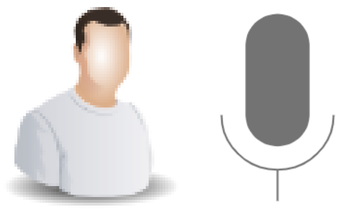
82% of the collected voices are done by males, majority of the voices are younger than 23 years old

10 hours of recorded data

The collected data was done using NAO V6 robot. The recorded speech is read speech. Topics are ranging from commands to articles, fantasy literature, detective fiction and more.

4 channel recordings

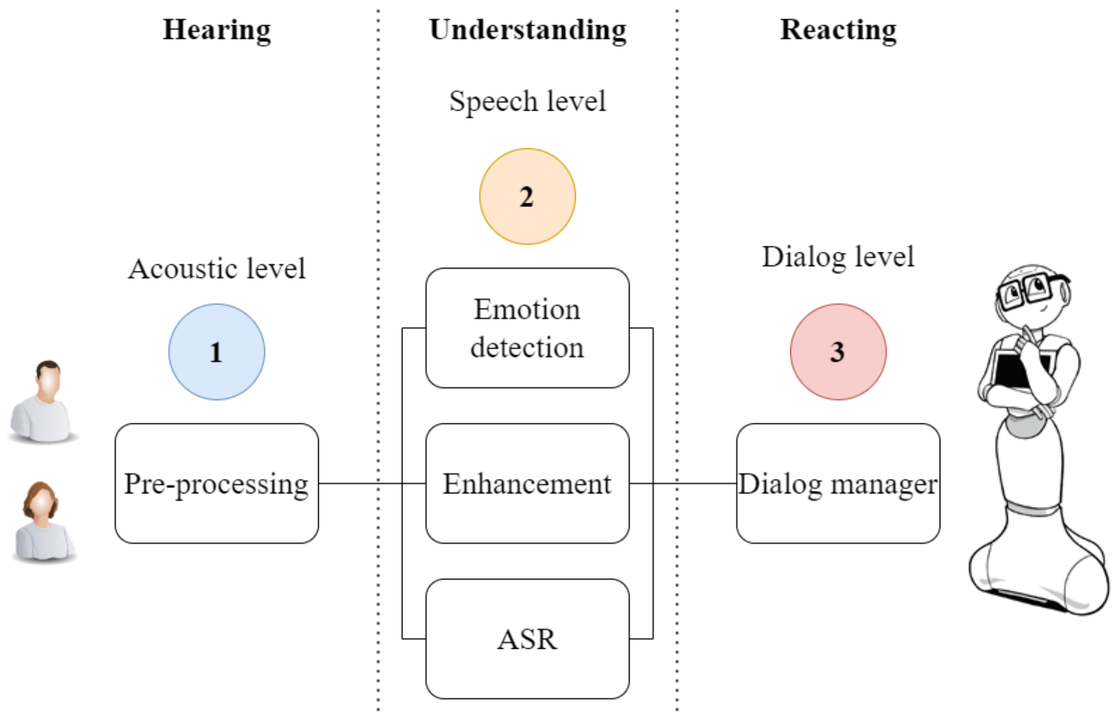
The recordings are done in .wav file format, cut into 1 minute audio chunks, using 4 different audio channels.



Lithuanian speaking robot

Above QR codes are links to the videos of a robot, that can communicate in Lithuanian language as an example. Both the Lithuanian language ASR and TTS were developed during "LIEPA-2" project in Vilnius University. The presented robot uses ASR and TTS engines locally in its system for instant reaction in HRI communication.

HRI in practice



In human-human interaction speech is being used to convey information. Just as human-human interaction, HRI also can be split into 3 stages: Hearing, Understanding what is being said and Reacting to the information.

1. Speech signal enhancement, Speech sources localization, Voice activity detection
2. Emotion/stress recognition, speech recognition, Prosody recongnition, Keywords
3. Dialog control, Summarization, Keyword extraction

In reality the speech is very complex and is open to multiple interpretations. In normal conversation between humans a subtle shift on emphasis or intonation can switch the meaning of a sentence drastically. Normal verbal communication is also usually enriched by paralinguistic information, such as: prosody or nonverbal behavior like facial expressions, gaze, various types of gestures. Humans, unlike robots, also use conversational fillers, that make up a part of the speech without directly relating to the specific type of information.

Conclusions

- One of the most obvious forms of communication between humans is speech. Therefore it is necessary to develop robots with Lithuanian language for rich and diverse forms of communication for achieving HRI goals.
- Open-ended, natural-language conversations are not yet possible, but it is possible for the robot to interact using and reading emotions, which increases the HRI experience.
- For more pleasant HRI experience, it is necessary to develop local Lithuanian dialog managers, that can deal with incomplete sentences, unknown states from the spoken utterances and have an ability to complete or fill the missing information if the speaker is omitting words.