1 Conclusion and future works

In this thesis we presented the design and implementation of a remote real time noise robust end of turn detection system. This system was developed to be used as part of software architecture of the Abel android, a hyperrealistic humanoid robot, used as a research platform in various AI application in the E. Piaggio research lab of the University of Pisa. Before this system Abel used a silence-based turn-taking module that behaved only based on silence and not on the absence of speech, having bad performances in noisy scenarios. The idea was to develop a turn-taking module robust to noisy scenarios and that could be deployed remotely to be integrated with the rest of the software architecture of the android.

First we presented the current literature review on the main methods and models used in the end of turn detection tasks and their pros and cons. Then we presented the implementation of our silence-based solution. Lastly, we validated our solution with the creation of a novel dataset that could meet our test requirements, and we compared the performances of the system using another vad module, the one integrated with WebRTC.

Our solution uses the state-of-the-art WebRTC technology to reliably transmit in real time the audio from the speaker to the remote server. The Python server thanks to the aiortc library processes the incoming audio frames that will be analyzed by the Silero VAD voice activity detection model to detect the presence of voice in the audio. Here a turn based system manages the status of the current turn and updates in real time the current turn status using websockets to the external environment and in particular to the speech elaboration module of the Abel android.

In the Experiments and Results section we showed an analysis of the performances of the system both in terms of inference and end of turn detection latency, and in terms of accuracy of the system and in particular of the Silero

VAD model that was compared with the WebRTC vad. The first analysis showed that the system has a detection latency compatible with the use case application, while the last analysis showed that the system has good detection capabilities even in noisy scenarios and that the Silero VAD model outperform the WebRTC vad showing better performances with the diminishing of the Signal to Noise Ratio (SNR), so in more noisy scenarios.

Further work on this system could include a speaker diarization module, capable of detecting multiple source of voices from different speakers, in order to correctly keep track of multiple conversations concurrently.