Microphone Test

Introduction

We tested the ***VideoMic NTG*** by ***RØDE*** as it is to be used as the audio input device for the robot.

The Microphone comes with multiple gestures and as such, we decided on testing the audio quality when using different of these features. In the following section, we describe the process by which we tested the different features of the microphone. For now, we will shortly describe the features we utilized:

- **High-pass Filters:**There are two high-pass filters available. These will cut lower frequencies to help mitigate wind noise and other low-frequency rumble
  - 75 Hz high-pass filter
  - 150 Hz high-pass filter
- **Padding:** To make sure audio is clear and distortion-free when recording loud sound sources
  - -20dB pad

Set-Up

We decided to record three different sentences (all relevant to RoboCup challenges) several times, each time under different conditions. See both the sentences and conditions below:

Sentences

- **NameAndDrink:** "My Name is Sarah and my favourite drink is Coffee"
- **Order:** "I would like to have 2 steaks, fries and cola"
- **Hobby:** "I like to play video games"

Conditions

- **Condition 1:** Use no features of the microphone
- **Condition 2:** Use the 75 Hz high-pass filter
- **Condition 3:** Use the 150 Hz high-pass filter
- **Condition 4:** Use the -20dB pad
- **Condition 5**: Use the 75 Hz high-pass filter and the -20dB pad
- **Condition 6**: Use the 150 Hz high-pass filter and the -20dB pad

Noise

Each recording was captured while also playing an ambient sound of a large crowd of people talking (see ▶ 10 Full Hours Of People Talking). The sound was played using a **JBL GO 2** bluetooth speaker. This was done because we expect a lot of noise at the RoboCup.

Physical Set-Up

The microphone was placed on a around 76cm high table. The speakers were placed around 60 cm far away from the microphone. The speaker was also placed on the table, around 90 cm away from the microphone, to the opposite direction of the speaker. The speaker faced into the direction of the microphone.

Storing

The different audio files are stored in different directories, corresponding with the condition (see above) under which they have been captured. The naming convention goes as follows:

{Name of the task}{Version of the taks}.{Number of recording of the version}

For example: ***NameAndDrink1.1.wav*** refers to a sentence where a name is said and a favourite drink. The first **1** refers to which exact sentence is said, for example "My name is Sarah and my favourite drink is coffee", while the **.1** refers to a different recording of this sentence. So the file ***NameAndDrink1.1.wav*** is a recording of the sentence "My name is Sarah and my favourite drink is coffee", while ***NameAndDrink1.2.wav*** is a recording of the same sentence but recorded by a different person.

Audio File Specs

The audio files have been captured using Audacity and exported with the following specs:

- **Format:** WAV
- **Channels:** Stereo
- **Sample Rate:** 44100 Hz
- **Encoding:** Signed 16-Bit PCM

Test Phase One

In the first test phase 3 different speakers recorded the different sentences under each condition as described before, accumulating a total of 54 voice recordings, 9 for each condition.

These audio files have been used as input for the tests.py script, where we compare the transcribed output for each audio file to a ground truth. At this point we only focused on the actual transcription result and not the accuracy for extracted entities for example. This decision is based on the assumption that an inaccurate transcription result will likely result in a wrong final output.

Results

Figure 1 shows the amount of correctly transcribed sentences per condition, where we can see that except for condition 1 and condition 4, most conditions result in a 66.66% percent accuracy.

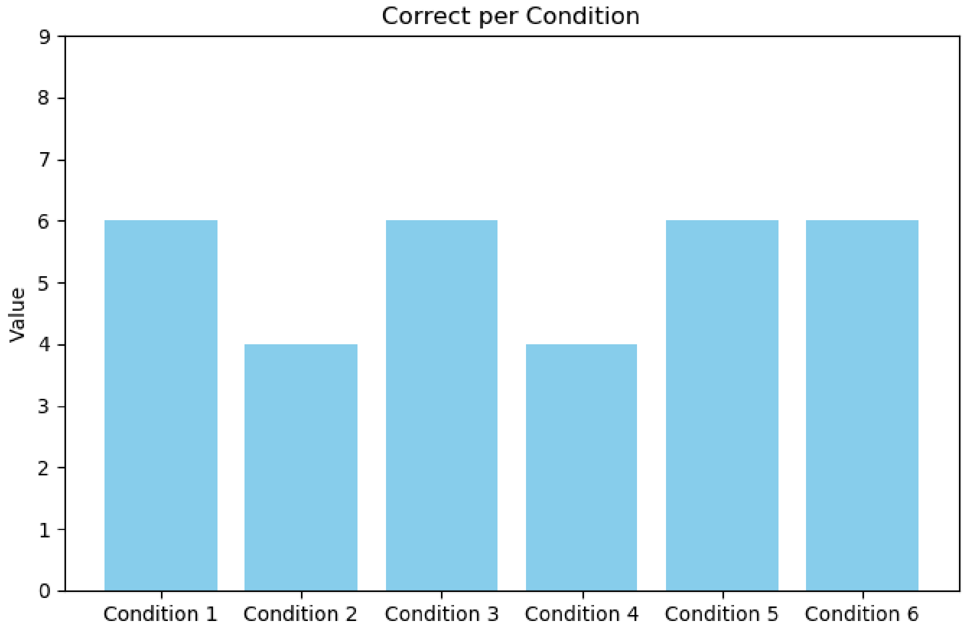

Figure 1

Furthermore, Figure 2 shows the number of correctly transcribed sentences of each phrase overall. Of the 54 recordings, 32 have been transcribed correctly, where Phrase 1 and Phrase 2 have been transcribed 8 times correctly and Phrase 3 has been transcribed accurately 16 times.

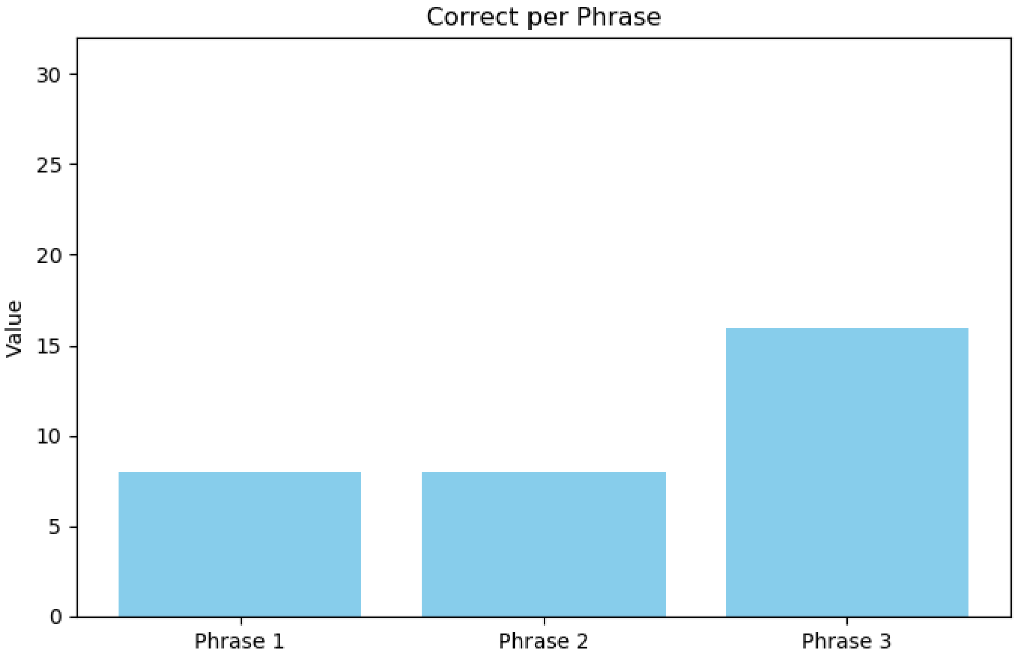

Figure 2

Test Phase Two

In the second test phase 3 different speakers recorded the different sentences under each condition as described before, accumulating a total of 54 voice recordings, 9 for each condition. This time the background noise has been around double in volume compared to test phase one. Otherwise the same conditions as in test phase one hold true.

Results

Figure 3 shows the amount of correctly transcribed sentences per condition, where we can see that condition 2 and condition 6 performed the best overall (77.77% accuracy).

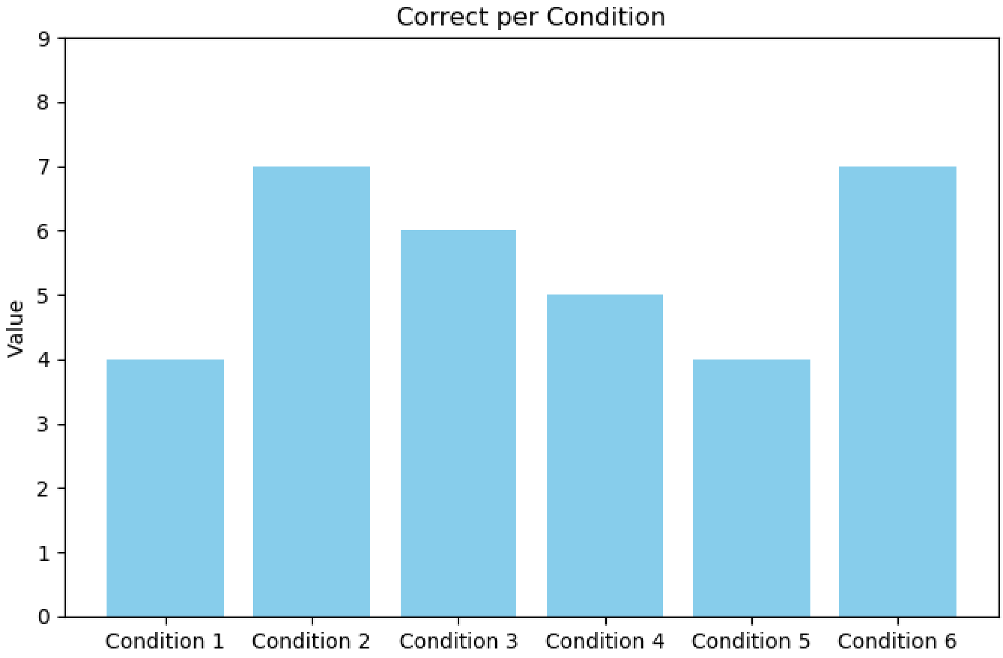

Figure 3

Furthermore, Figure 4 shows the number of correctly transcribed sentences of each phrase overall. Of the 54 recordings, 33 have been transcribed correctly, where once again Phrase 3 performed noticeably better than Phrase 1 and 2.

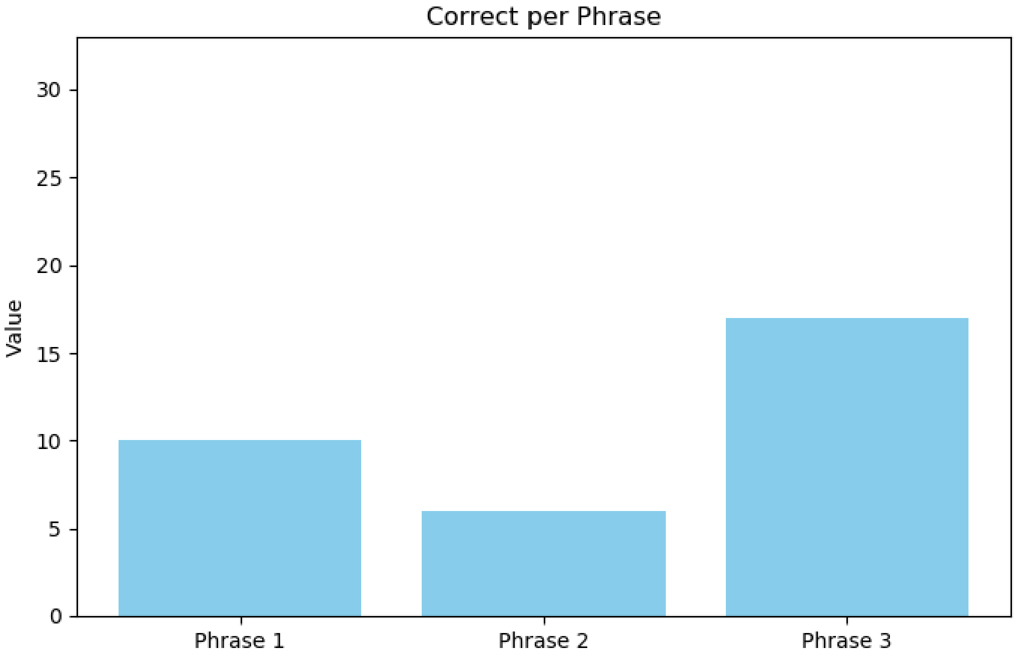

Figure 4

Discussion

The current test results (status 2025-02-03) show differentiating results for test phase one and test phase 2. One can assume that the high pass filters may only perform well under the condition that noise exists. Overall it is still hard to tell what condition may perform best and further testing is required, where even more training data is needed. Furthermore do the current results not take false negatives into account, which is something that may be improved upon later.
One noticeable factor the test results have shown, is that shorter, less complex phrases perform noticeably better than longer ones. This should motivate us to find solutions for accurately transcribing complex phrases more reliably.