Thomas Peterson Thpeter@kth.se
Yonk Shi Pyshi@kth.se
Clément Morin Clemor@kth.se

Attended Feedback

This document contains the feedback we received during peer-review and comments regarding how it was attended. In the beginning of the project we were planning to conduct transfer learning from synthetic data(A google translate and wavenet synthesised dataset that we would create on our own) to the VCTK dataset(A dataset of spoken audio with labels). The idea was to reimplement wavenet but train it to for ASR instead of audiogeneration. However, we realized that this task was way to big for the time we had in the course. After a meeting with the teacher it was concluded that we should only focus on the wavenet for ASR task. Because of the size of the project, our draft did not yet contain any results and lacked in several aspects. Additionally, at the time of the creation of the draft report, our primary focus still lied in exploring transfer learning for ASR. Thus, a small portion of the feedback was no longer relevant for us anymore(for example “missing reference in the section about transfer learning” since this section would be removed). For clarity all relevant comments received follow down-below.

Scores

15/30

The score is low as the course instructor said that the project draft should be like a final report with some missing results and the reader should be able to reproduce the experiments just by reading the project draft but as explained earlier there are several factors which can restrict the reader.

15/30

The paper is not ready for draft submission. Two main parts (result and conclusion) are missing. I believe that in draft submission, we must submit the almost complete version of the report. Minor missing result is acceptable. However, when reviewing this draft, I have difficulties to evaluate the real experiment done by the authors and the correctness of the analysis because the methodology is not detailed and there is no result/expected-result written in the report.

16.5/30

[No comments were given by this peer-review:er regarding the overall quality of the report]

Peer Review 1

“*Suggestion:* Author needs to define and structure the experiments and can take the previous work’s results for comparison (as a reference). If there is no previous work done in this specific domain (transfer learning), he can take the results published from the research work done on the similar datasets with different method.”

The “Discussion and conclusions” section now mentions the difference between our results and the results of the independent python implementation we took most inspiration form.

Thomas Peterson Thpeter@kth.se
Yonk Shi Pyshi@kth.se
Clément Morin Clemor@kth.se

“.. explain the speech related features of the experiment”
The speech related features of the experiment have now been explained in the final version of the report.

“The author must explain, how he is planning and conducting the experiments and how he is going to evaluate the system. Without evaluation and results it is very difficult to assess the correctness of the stated method.”
The report now explains how the experiments were conducted and how the system was evaluated.

“There is no information about the data preprocessing of the dataset in the report even author mentioned that they are doing LMFCC. Please explain how you are applying LMFCC as there are many variables involved in this process.”
It has now been detailed how MFCCs(We are not using LMFCC but MFCC) are calculated and used.

“Several methodologies are explained in background which is inappropriate. For example, Wavenet’s architecture and Connectionist Temporal Classification are explained in background. These should be moved to methodology.”
The sections have been moved to the methodology chapter

“It is quite common to illustrate the architecture and loss with the help of diagrams. I would like to have an illustration of Wavenet’s architecture and Connectionist Temporal Classification loss diagram in the report.”
More figures have been added to illustrate the architecture and the loss.

“There is insufficient information in the draft regarding methodologies. The author just wrote that “While the initial idea behind the paper was to make inference from raw audio, the authors of this (referenced github) implementation have decided to use Liftered Mel-Frequency Cepstrum Coefficients (LMFCC)”. You need to motivate or compare why you choose LMFCC over raw audio.”
It has been clarified why we are using MFCCs instead of raw audio.

“Please explain the input and output of the Wavenet architecture along with output of individual layers (in the diagram if you wish) so that it would be easy for a reader to reimplement the architecture.”
The complete WaveNet architecture is now described in the report and an overview figure helps the reader understand the input and output of the model.

Peer Review 2

“The use of CSTR VCTK Corpus’ is not consistent in the report. In other sentences, the authors only wrote ‘VCTK Corpus’ rather than its complete term. It made the reader confused of ‘CSTR’ term that is written in the abstract.”
We didn’t use the VCTK dataset but switched to the LibreSpeech dataset(Not to avoid this issue of course but because of implementational convinience). Thus this comment is not relevant anymore.

“There is no literature study about LMFCC. The background theory of LMFCC is important since the experiment will use LMFCC feature for the input of the network. It is also important to state why LMFCC can represent the voice speech property that will be recognized by the system.“
A section describing MFCC in relatively much detail has been added. It explains why certain steps of the procedure is conducted.

“The author did not mention the detail of the LMFCC feature extraction. We need to know the settings of the LMFCC feature and the motivation of doing it. One of the reason might be to reduce the number of parameter. However, we also need to justify the sampling period and its shifting time to make sure how the information of the audio is preserved during LMFCC feature extraction.”
The MFCC feature extraction details have now been added to the report.

“The implementation of WaveNet for ASR was not clear. We know that WaveNet is a text-to-speech system. The authors mentioned that their reference of WaveNet did not give accurate details on how to use it for ASR, but the authors also did not explain their strategy to use WaveNet as ASR system.“
The report now explains how wavenet is adapted for ASR.

“There is no figure or table in the report. Figure and table will help to explain the experiment/result better for the reader.**“**
Figures have been added to the report. However, tables were not added since they did not have any relevant use-cases in the context of our work.

“The reference is not numbered. The author also did not manage to write the correct reference in the report. In every sentence with reference, there is a mistype of the sentence and the reference. The last word in the sentence is always mixed up with the last name of the reference. The author must only write the number and point to the correct reference number.”
The report has been changed to use numbered references. The missing of a space has been corrected.

“There is a mistype word in section 2.4 line no.4. The word ‘red’ must be replaced by ‘read’ as ‘read’ is the correct past tense.“
This minor spelling mistake was corrected

Peer Review 3

Thomas Peterson Thpeter@kth.se
Yonk Shi Pyshi@kth.se
Clément Morin Clemor@kth.se

“Fill in missing citations in subsections ‘Transfer learning’ and ‘Connectionist Temporal Classification’.”
The transfer learning subsection has been removed completely since it was no longer relevant after the scope reduction of the project. However, the missing citations for the CTC section has been fixed.

“More specifically, subsection ‘wavenet for ASR’ needs improvement. You state: “the authors of this implementation have decided to use Liftered Mel-Frequency Cepstrum Coefficients (LMFCC)”. How? What led them to take this decision? Elaborate. “
The subsection ‘WaveNet for ASR’ was merged into the experiment chapter and the ‘WaveNet’ subsection. We have clarified why they chose this approach and elaborated.

“Try explaining some concepts that might contain specialized terminology, in the simplest way possible e.g. in subsection ‘Wavenet’ terms like receptive field could be unfamiliar to someone.”
We have now clarified complicated terms and concepts to make the text easier to read.

“Fix:
• section 1 – missing citation ‘Bishop - pattern recognition’
• section 1 – ‘using WaveNet and famous translating service’ add ‘the’ before famous
• subsection 2.1 – ‘that has been previously been solved’ use instead ‘that has been previously solved’”
This misses have been corrected