h_da@ReproHum – Reproduction of Human Evaluation and Technical Pipeline

Margot Mieskes
University of Applied Sciences
Darmstadt, Germany
margot.mieskes@h-da.de

Jacob Benz
University of Applied Sciences
Darmstadt, Germany
jacob.benz@stud.h-da.de

Abstract

How reliable are human evaluation results? Is it possible to replicate human evaluation? This work takes a closer look at the evaluation of the output of a Text-to-Speech (TTS) system. Unfortunately, our results indicate that human evaluation is not as straightforward to replicate as expected. Additionally, we additionally present results on reproducing the technical background of the TTS system and discuss potential reasons for the reproduction failure.

1 Introduction

Replication of research results in Natural Language Processing (NLP) has gained considerable attention in the past years. While quite some progress has been achieved with initiatives such as the Responsible Research Checklist[1] and the Reproduction Checklist [2] (Dodge et al., 2019), the question about the reproduction of human evaluation is widely unanswered. The work presented here is part of the ReproHum Project[3], which aimes to reproduce human evaluation. In our experiment, we tried to reproduce the evaluation of a low-resource Text-to-Speech (TTS) system for German. As the results of our reproduction indicated that we were unsuccessful, we also had a closer look at the technical aspects of the work and attempted to reproduce those elements for our study as well.

Our major contributions are therefore: 1) the results on the reproduction of the human evaluation of the TTS output, 2) the results of the reconstruction of the language data required for the TTS system and 3) the results of the reconstruction of the TTS model required to create the TTS output, which is then judged during the human evaluation.

[1] https://aclrollingreview.org/responsibleNLPresearch/

[2] https://2021.aclweb.org/calls/reproducibility-checklist/

[3] https://reprohum.github.io/

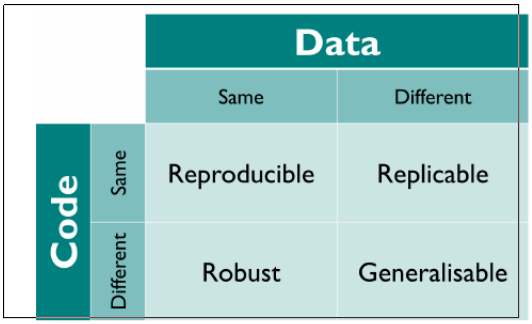

Figure 1: Dimensions of Reproducibility according to (Whitaker, 2017)

2 Background and Related Work

Replication is a topic that is being discussed in a wide range of fields. In NLP the primary focus so far has been on the technical reproduction – i.e. reproducing results based on quantitative evaluation. (Cohen et al., 2018) presented three dimensions of reproduction:

- Reproduction of a Conclusion
- Reproduction of Results
- Reproduction of a Value

But their focus has been on the technical reproduction.

Figure 1 shows another set of parameters for the reproduction: Whether the Code and the Data are the same or different allows for different conclusions with respect to Reproducibility, Replicability, Robustness and Generlizability.

This is also clear from the reproducibilty spectrum according to (Peng, 2011), which focuses heavily on code and data (see Figure 2, similar to (Whitaker, 2017)).

There are major differences between the technical reproduction and the reproduction of human evaluation results, although initially, the aim is also to reproduce a certain value, a certain result or a

Proceedings of the Third Workshop on Human Evaluation of NLP Systems,
pages 130–135, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_011

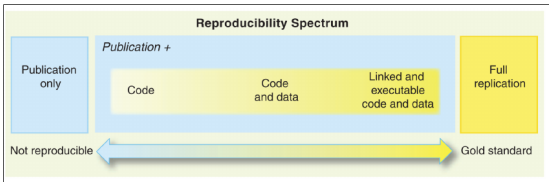

Figure 2: Spectrum of Reproducibility according to (Peng, 2011)

certain conclusion. But a look at other fields, where the reproduction of human input has been already evaluated, such as Psychology and Social Sciences, reveals that this is also far from straightforward. For Psychology it seems that only between 36 % and 68 % of the results were reproducible by an independent researcher (Open Science Collaboration, 2015), while in Social Sciences between 57 % and 67 % of the studies were reproducible (Camerer et al., 2018). Although what dimension of reproduction has been aimed for, is open.

There are various reasons for the lack of reproducibility of human generated results. One element is the lack in objectivity in humans and their individuality, as each human has individual experiences and opinions. Another element is the language, the instructions are presented in. Some languages distinguish between a formal address and an informal address. A person used to being addressed formally, might react negatively to an informal address and the other way around. When performing an evaluation using online tools or any form of technical equipment, this too can affect the results. A high-resolution screen will represent colours differently to a smartphone screen. When dealing with acoustical data, using a headset or speakers can make a vast difference and the quality of each can also influence the results, when asked to evaluate the quality of the presented sound.

3 The Original TTS Experiment

The basis for our work is the paper by (Lux and Vu, 2022). Its aim is to present the possibility to create TTS systems with little training data and reduced training time. This is achieved by using a large multilingual model, which is then fine-tuned towards the target language based on the reduced training data and reduced training time. A specific focus is put to model articulatory features of the language.

The technical basis for the model is Tacotron2 (Shen and Pang) and FastSpeech2 (Ren et al., 2020). Where Tacotron2 is based on a recurrent sequence-to-sequence network, FastSpeech2 is based on a Feed-Forward Transformer network.

The basis for the multilingual model is data from English, Greek, Spanish, Finnish, Russian, Hungarian, Dutch and French. The German data is derived from the HUI corpus (see Section 6 below).

While the multi-lingual model required lots of resources, both in time and hardware, the adaptation to German was performed using 30 minutes of speech and training for about 2 hours. In order to allow for a comparison and to verify the low-resource approach, the authors also trained both FastSpeech2 and Tacotron2 exclusively on German, using 29 hours of recorded speech.

4 Reproduction – Experimental Setup

Following the original study, we set up a Google Form survey, where each participant is presented with two stimuli and asked to judge, which of the two sounds more natural. Figure 3 shows the interface we used to conduct the survey. As we were dealing with German speech output and German students were asked to judge the TTS output, we also addressed participants in German. Participants could choose from three different options: Either one of the outputs is better than the other, or both are equally good.

Prior to starting the evaluation, we submitted all relevant information to the University of Aberdeen Ethics Board for evaluation, which approved of our experimental setup, the way we dealt with the data and the personal information collected from the participants.

The participants were recruited by email from our university. Other than sending out an email via a central email address, we did not collect any personal data from our participants.

5 Reproduction – Results

In the end, 37 participants took part in our experiment, which is comparable to the original study. In general, the output from the proposed Fastspeech2 model is considered better than the baseline system in 41 % of the cases, while the baseline system is considered better in 13 % of the cases. When comparing the two Fastspeech2 versions, 46 % of the participants did not hear any noticeable difference. This is comparable to the original evaluation, where 43 % of the participants did not hear a difference. See also Figure 4.

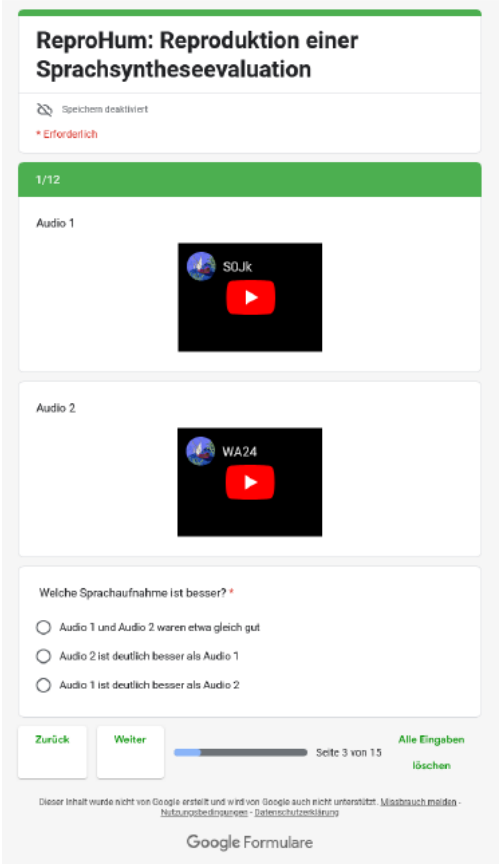

Figure 3: The survey interface.

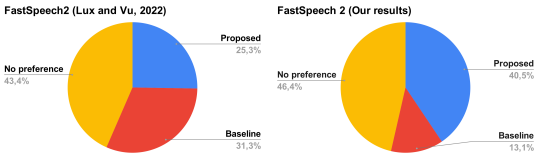

Figure 4: Human Evaluation for FastSpeech2 Low Resource and Baseline.

When evaluating Tacotron2, 26 % of the participants preferred the low-resource model, while 23 % preferred the original version. But, 51 % of the participants did not hear a difference between the two versions. Compared to the original evaluation, where 52 % of the participants preferred the low-resource version, while 11 % preferred the original system and only 37 % did not hear a difference. The results are also shown in Figure 5.

As shown in table 1, the coefficient of variation values for the pair-wise comparisons between the original results and our reproduction are with the exception of one value always in the double digits, further indicating that our reproduction resulted not only in rather different values but different results as well.

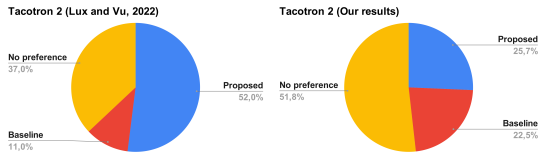

Figure 5: Human Evaluation for Tacotron2 Low Resource and Baseline.

6 Technical Reproduction

In light of these results for the reproduction of the human evaluation we had a closer look at the background of the TTS system. First, we tried to reproduce the data and then we aimed to reproduce the TTS model.

6.1 Reproducing the Data

The original corpus project, as presented in (Puchtler et al., 2021). The Hof Universität – Institut für Informationsysteme (HUI) Audio Corpus German aimed to create a high-quality, open source dataset for German TTS systems. Figure 6 schematically describes the approach.

The authors originally defined a range of parameters for choosing data for their speech synthesis system:

- at least 20 hours of audio per speaker
- minimal sampling rate of 22 kHz
- normalization of textual data
- normalization of loudness
- audios of between 5 and 10 seconds of length
- recording of punctuation

In the end, the original study had collected 326 hours of audio and processed them according to their pipeline in Figure 6. This included five speakers with between 32 and 96 hours of audio and another set of 97 hours of audio by 117 other speakers.

We tried to be very accurate with our reproduction, documenting all steps. Unfortunately, due to a range of errors described below, this reproduction proved to be unsuccessful in the limited time. Initially, the link for the German Deep Speech Model was faulty. Luckily, the original authors reacted quickly and fixed this.

Next, the textual representation of the spoken data had to be downloaded. This referred to a Gutenberg repository, where the mirror was hardcoded, but not valid anymore. Additionally, the URI was automatically created, but again, in the

Model	(Lux and Vu, 2022)	Our Reproduction	Coefficient of Variation
Tacotron2 Proposed preferred	52 %	25,7 %	33,9 %
Tacotron2 Baseline preferred	11 %	22,5 %	34,4 %
Tacotron2 No preference	37 %	51,8 %	16,6 %
FastSpeech2 Proposed preferred	25,3 %	40,5 %	23,7 %
FastSpeech2 Baseline preferred	31,3 %	13,1 %	40,7 %
FastSpeech2 No preference	43,4 %	46,4 %	3,8 %

Table 1: Comparison of the results of the original evaluation and our reproduction.

Modell	Hardware	Duration Preprocessing	Iterations	Time/Iteration	Total Duration
Tacotron2 Low Resource	GPU	1:13 min	10,020	1.25 It/sec	2:25 hrs
Tacotron2 full	GPU	50:32 min	100,224	1.4 It/sec	19:54 hrs
Tacotron2 Low Resource	CPU	NA	925	22 sec/It	6 hrs
FastSpeech2 Low Resource	GPU	NA	100,071	4.4 It/sec	6:27 hrs

Table 2: Retraining of the Low Resource and Full Models according to the specifications given in (Lux and Vu, 2022)

wrong format for the mirror we chose instead of the original one.

The next problem was linked to FFMPEG and NLTK packages that had to be added to the original installation.

Finally, we had to remove one speaker completely from the data set, as several files associated with that speaker could not be processed and this error could not be eliminated.

This resulted in the abortion of the replication attempt, as removing one of the five major speakers from the data set did not allow for a plausible further result.

6.2 Reproducing the TTS Model

Furthermore, we tried to replicate the initial speech synthesis model, as described by (Lux and Vu, 2022). Figure 7 represents the pipeline to create the TTS model, including the technical packages used. Theoretically, this reproduction attempt should have been straightforward, as most research artifacts have been made available to the research community. Unfortunately, the resulting model has not been provided and the TTS outputs are also only available in the context of this project.

Despite the seemingly straightforward problem, the availability of the research artifacts and an extensive `Readme` file, we came across a range of issues in the process. First of all, not all required packages are listed in the `requirements.txt` file. The biggest issue was a `Invalid render options` error during the data pre-processing, which occurred multiple times and only with some files, but not all. Identifying the specific files which caused issues, was quite time-consuming. It turned out, that the original problem is the `unsilence` package, that is used to skip over longer period of silence in the recorded data. With some of those, a parameter required for `ffmpeg` is set to an invalid value, which results in the `invalid render option` error. We extended the code to check for invalid values and set them to a default value, in cases where an invalid value was reached.

Another issue is the fact that the HUI-corpus is available in two versions: *clean* and *full*. Unfortunately, the authors did not report which version of the data has been used for the original experiments, so we decided to use the *full* version.

Finally, the number of training iterations has not been reported. We assume that the figures set in the original code represent these numbers, but it is unsure, if those are actually the figures used in the original experiments.

Table 2 shows the duration of training for the reproduced models. We retrained both the Low Resource models for Tacotron2 and FastSpeech2 and the full Tacotron model. As a proof-of-concept, we also retrained the Tacotron2 Low Resource model on a CPU rather than a GPU. Retraining the FastSpeech2 Full model was beyond the scope of our work. We can support the results from previous work, that indeed, low-resource models can be quickly trained. But we observed a notable difference in the sound quality, pronunciation and the prosody of the resulting output, leading to the conclusion that despite not changing any of the given parameters, the reproduction of the final results was only partially successful.

7 Discussion

Table 3 shows a summary of the different reproductions we attempted and the respective results.

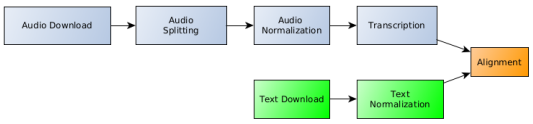

Figure 6: Pipeline for creating the Audio-Transcript Data according to (Puchtler et al., 2021)

Reproduction	Reproducibility	Remarks
Data set	Reproduction had to be abandoned	Mirrors unavailable, software issues
TTS Model	Partially, conclusions were reproduced	Different results, conclusion can be supported
Human Evaluation	Values and results not reproducible	Overall conclusion reproducible

Table 3: List of our attempted reproductions and the respective results.

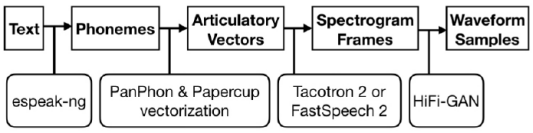

Figure 7: Pipline to create the TTS model according to (Lux and Vu, 2022)

These are quite baffling, since none of our approaches reached the same values or results. There are a number of potential reasons for this:

The differences in results when reproducing the TTS models could be explainable by different hardware or slightly different software versions, especially since we faced issues that the original authors obviously did not encounter.

Regarding the different results for the reproduction of the human evaluation, one reason could be the different group of people. While both studies employed students to evaluate the synthesis output, in the original study, the students are from the field of computational linguistics and natural language processing and as such more used to hearing and judging synthetic speech. In our study, the students did not have any particular training in judging synthetic speech.

Another reason could be that the stimuli were somehow mixed up. If that would be case, we would have to transpose the results and would have results that are more comparable to the original study.

The problem might be related to the problems with reproducing the original data set and/or the original TTS models, since the stimuli were recreated for the purpose of this study[4], which could have lead to a variance in sound quality compared to the original stimuli.

Comparing our results to the results of (Hürlimann and Cieliebak, 2023), who ran the exact same experiment, the chances that the stimuli were transposed somewhere in the process are increasing, as their results also indicate low reproducibility, except if a transposition is assumed. As their results are based on a larger number of participants, they are more pronounced than ours and statistically more reliable. The authors state a range of other potential error sources, which have to be taken into account in addition to our experiments. Additionally, it is certainly remarkable that in both reproductions the lowest coefficients of variation were achieved for the ”no preference” option.

8 Conclusion

In general, we can support the conclusion of the previous study, that the low-resource speech synthesis (both using Tacotron2 and Fastspeech2) are viable approaches to produce reasonable TTS output based on limited resources (time, computing and available speech data). Our results also show, that the reproduction of human evaluation and possibly human annotation as well are important research areas. As quantitative results can only give so much information, while human evaluation in various domains (i.e. synthetic speech, but also text quality in Natural Language Generation) can provide a more detailed insight into the data.

Unfortunately, the way human evaluation is currently reported, the reproduction of human evaluation has not been successful.

With respect to the whole pipeline, of a technical reproduction based on which a human evaluation can take place, it is important to make sure, that research artifacts are stored properly, documented thoroughly and potential pitfalls (i.e. dying links) are noted.

Our results indicate that more research is necessary into the issue of human evaluation. Related to

[4]Florian Lux personal communication.

this, it would be interesting to study human annotation tasks, which are related to human evaluation and are the basis of a wide range of models built in the context of NLP.

Acknowledgments

We would like to thank Jonathan Baum for his experiments on the replication of the TTS model and Christian Stute for his support in the replication of the human evaluation replication study.

References

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. 2(9):637–644.

K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurelie Neveol, Cyril Grouin, and Lawrence E Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 156–65.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Manuela Hürlimann and Mark Cieliebak. 2023. Reproducing a comparative evaluation of german text-to-speech systems. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems (HumEval'23)*.

Florian Lux and Thang Vu. 2022. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long Papers*, pages 6858–6868.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Roger D. Peng. 2011. Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Pascal Puchtler, Johannes Wirth, and René Peinl. 2021. HUI-audio-corpus-german. a high quality TTS dataset.

Yi Ren, Chenxu Hu, Xu Tan, and Tao Qin. 2020. FastSpeech 2. fast and high-quality end-to-end text to speech. arXiv.

Jonathan Shen and Ruoming Pang. Tacotron 2: Generating human-like speech from text.

Kirstie Whitaker. 2017. Showing your working. a how to guide to reproducible research.

A Human Evaluation Datasheet (HEDS)

The Human Evaluation Datasheet (HEDS) is part of the supplemental material.

B Spreadsheet Results Evaluation

The spreadsheet that we used for analysing the results of our human evaluation is part of the supplemental material.