Reproducing a Comparative Evaluation of German Text-to-Speech Systems

Manuela Hürlimann
Centre for Artificial Intelligence,
Zurich University of Applied Sciences,
Winterthur, Switzerland
manuela.huerlimann@zhaw.ch

Mark Cieliebak
Centre for Artificial Intelligence,
Zurich University of Applied Sciences,
Winterthur, Switzerland
mark.cieliebak@zhaw.ch

Abstract

This paper describes the reproduction of a human evaluation in *Language-Agnostic Meta-Learning for Low-Resource Text-to-Speech with Articulatory Features* reported in Lux and Vu (2022). It is a contribution to the ReproNLP 2023 Shared Task on Reproducibility of Evaluations in NLP. The original evaluation assessed the naturalness of audio generated by different Text-to-Speech (TTS) systems for German, and our goal was to repeat the experiment with a different set of evaluators.

We reproduced the evaluation based on data and instructions provided by the original authors, with some uncertainty concerning the randomisation of question order. Evaluators were recruited via email to relevant mailing lists and we received 157 responses over the course of three weeks. Our initial results show low reproducibility, but when we assume that the systems of the original and repeat evaluation experiment have been transposed, the reproducibility assessment improves markedly. We do not know if and at what point such a transposition happened; however, an initial analysis of our audio and video files provides some evidence that the system assignment in our repeat experiment is correct.

1 Introduction

The work reported in this paper has been carried out as part of a multi-lab, multi-test study in the context of the ReproHum project (Belz et al., 2023) and the ReproNLP shared task. The goal of the project is to assess the reproducibility of human evaluations in Natural Language Processing and to find out which factors contribute to making human evaluations more or less reproducible. Our contribution attempts to reproduce an evaluation in a paper from Track C of ReproNLP 2023, Lux and Vu (2022), which presents a language-agnostic low-resource approach for Text-to-Speech (TTS). The human evaluation is carried out on German audios generated with four different Text-to-Speech systems.

We first (Section 2) describe the approaches of the original experiment and our reproduction in detail.

In Section 3, we present the answer distribution of our results (Section 3.1) and the reproduction targets (Section 3.2). In Section 3.3 we then compare the results of both studies in terms of the scores obtained by each model and report the coefficients of variation (CV*), which quantify the variability of original-reproduced measurement pairs. We also report Pearson's correlation coefficients between the original and the reproduction system measurement sets. These results show very low reproducibility (large CV* and low Pearson correlations) and we notice a strong cross-similarity between the system results, meaning that the original results for one system are very similar to repeat results for the other, and vice versa. Therefore, in Section 3.4 we also re-evaluate the results with an assumed system transposition and find improved reproducibility (lower CV* and very high Pearson correlations).

In the light of these results, after ruling out some error sources (Section 4), we compare the Mel-Frequency Cepstral Coefficients (MFCC) of the audio and video files used in the repeat evaluation (Section 4.1). The results indicate that the system assignments in our repeat experiment are likely to be correct. In Section 5, we discuss our findings and in Section 6 we briefly compare our results with those of another reproduction submitted to ReproNLP 2023.

All our resources are publicly available.[1]

[1] https://github.com/manhue/repronlp2023_lux_and_vu

Proceedings of the Third Workshop on Human Evaluation of NLP Systems,
pages 136–144, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_012

2 Evaluation Experiments

In this section we first (Section 2.1) describe the original experiment and then (Section 2.2) our reproduction.

2.1 Original Evaluation

This section describes the original evaluation experiment as reported in Lux and Vu (2022), Section 4.2.2. The authors shared the details of their evaluation protocol with the ReproHum team in personal communication with the authors and the resources were subsequently provided to us.

Systems The original human evaluation was a preference study of four Text-to-Speech systems for German. The systems are based on two different models, FastSpeech 2 (Ren et al., 2021) and Tacotron 2 (Shen et al., 2018). For each model, there are two flavours: the baseline system (trained on 29 hours of German) and the proposed low-resource system (trained in a multilingual low-resource regime with 30 minutes of data for each of 8 languages,[2] then fine-tuned on 30 minutes of German). This results in a total of four different systems: FastSpeech-Baseline, FastSpeech-Proposed, Tacotron-Baseline and Tacotron-Proposed.

Data and Task The evaluation was done via a comparative evaluation of generated audio. There were six text prompts, which were chosen to be phonetically balanced. Each of these six prompts was synthesised using each of the four systems. In each judgement, evaluators were presented with two synthesised audio files, one from the baseline and one from the low-resource flavour of the same model. They then had to choose one of the following three responses:

- Audio 1 is significantly better than Audio 2
- Audio 2 is significantly better than Audio 1
- Audio 1 and Audio 2 are about equally good

Evaluators were not informed of the number or type of systems that were used to generate the audios but were simply asked to make a preference judgement as outlined above for each audio pair. As far as we can tell from the provided materials, "naturalness" was not mentioned to evaluators as an explicit criterion.

[2]English, Greek, Spanish, Finnish, Russian, Hungarian, Dutch and French

Survey Form The authors of the original paper conducted the evaluation using a Google Form survey[3]. Since Google Forms do not have any functionality to embed audio directly, they converted the audio files to videos with a black image as visual. They then uploaded these videos to YouTube and embedded them in the Google Form.

Not Reproducible: Randomised Question Order The original authors reported that they had randomised the order of the questions in the Google Form. When working on the repeat evaluation experiment, the authors of the current work and the ReproHum project team were not able to reproduce this functionality: there was no option to randomise the order of Google Form questions which preserved the video-response pairs. A randomisation option was available in the current version of Google Forms but its functionality proved unsuitable for the proposed setting since it jumbled all elements of the questionnaire, breaking the link between videos and questions. It remains unclear whether this feature has changed since the original authors did their evaluation or whether they in fact proceeded differently from what they reported. In Section 2.2 below, we describe how this was handled.

Evaluators The original survey was sent via email to students in speech-related courses at the original authors' university. 34 evaluators who self-identified as native speakers of German participated in the evaluation, leading to a total of 408 human judgements (6 prompts x 2 systems x 34 evaluators = 408 judgements).

Results The authors of the original evaluation aggregated the survey responses per system and found the preference distributions in Table 1 (from Figure 3 in Lux and Vu (2022)[4]).

Their results show a clear preference for the proposed low-resource system for the Tacotron model. For FastSpeech, the most frequently chosen option

[3]https://www.google.com/intl/en/forms/about/

[4]The numbers in Figure 3 of (Lux and Vu, 2022) do not agree completely with the text. In Section 4.2.2, the authors write "In 56% of the cases, the [Tacotron] model fine-tuned on 30 minutes of data was perceived to be as good or better than the model trained on 29 hours." During correspondence with the ReproHum project team, they said that this number should in fact be 69% (=52% + 37%) as in the figure.
Also note that the caption of Figure 3 in (Lux and Vu, 2022) mentions 102 judgements per system, but this number should be 204.

Label	%
Fastspeech-baseline	31%
Fastspeech-proposed	25%
Fastspeech-equal	43%
Tacotron-baseline	11%
Tacotron-proposed	52%
Tacotron-equal	37%

Table 1: Percentages of answers reported in the original study, from Figure 3 of (Lux and Vu, 2022). The number in each row indicates the proportion of responses for a specific option; for example, Tacotron-baseline was preferred over Tacotron-proposed in 11% of the cases and Tacotron-proposed over Tacotron-baseline in 52%.

was that both audios are equal, and the baseline was preferred more frequently than the proposed low-resource system.

2.2 Repeat Evaluation

For the repeat evaluation, the authors of the original paper provided us with the following:

- The introductory text, instructions, and set of answer options for the survey.
- The 24 audios that were presented to evaluators.
- An explanation of how they had created the survey.

We added a short consent screen, which evaluators saw first and had to agree to. We do not know if the original study also had a consent screen but we assume that it did not since this information was not provided to us. We then used the provided introductory text and instructions[5] and the provided answer options to create the survey.[6]

As explained in Section 2.1, it was not possible to reproduce the randomised order of the questions that the original authors reported. To standardise the question order of the different repeat evaluations, the ReproHum project team created a randomly shuffled order to be used in each repeat experiment. They used a Python script, *random_videos.py*[7] to shuffle the questions.

We applied the suggested process to create the videos that could be embedded in the Google Form and extended the *random_videos.py* script to generate a unique four-character identifier for each video in order not to reveal the system type.

We then sent the survey via email to different mailing lists within and outside our university. These included staff mailing lists for institutes and communities, as well as a dedicated mailing list of students who had consented to participate in research surveys. In the email, which was written in German, potential evaluators were told that they needed to speak German as their native language in order to participate.

3 Results

Below, we first (Section 3.1) present the results obtained in our reproduction study. We then show the reproduction targets (Section 3.2) and compare our results to the original study, assessing their reproducibility (Section 3.3). Since we find that the original results for FastSpeech are very similar to the repeat results for Tacotron and vice versa, we add Section 3.4, where we redo the comparisons and reproducibility assessments after *transposing* the system labels of our results. Note that we cannot be certain that such a transposition happened.

3.1 Results Obtained in the Reproduction Study

In this section, we present the results that we obtained in the repeat experiment. We show the distribution of answers and calculate the interrater agreement. We also run a Logistic Random-Effects Model to assess the preferences between the two systems, FastSpeech and Tacotron. Finally, we aggregate the preferences per evaluator per system, creating Per-Person Preference Data (PPPD), which allows to run a binomial test, testing against the mean preferences obtained in the original study - see Sections 3.3 and 3.1 for the tests and results.

Answer Distribution A total of 157 evaluators participated in our survey over the course of three weeks, creating 1878 individual judgements (6 prompts x 2 systems x 157 evaluators - 6 skipped questions[8] = 1878 judgements). Table 2 shows the distribution of the obtained answers.

[5] We only removed the final sentence from the original instructions which said that the order of answer options and audios could vary, since this was not the case in our survey.

[6] A PDF version of the Google Form survey is available in the project documentation: https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/google_form_pdf/GoogleForm_Evaluation%20von%20Text-zu-Sprache-Systemen%20-%20Google%20Formulare.pdf

[7] https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/random_videos.py

[8] Since the questions in our survey were not mandatory, it was possible to skip.

Label	n	%
Fastspeech-baseline	113	12%
Fastspeech-proposed	471	50%
Fastspeech-equal	358	38%
Fastspeech-skipped	0	-
Tacotron-baseline	274	29%
Tacotron-proposed	271	29%
Tacotron-equal	391	41 %
Tacotron-skipped	6	<1%

Table 2: Distribution of answers obtained in the reproduction study.

Interrater Agreement In order to assess the interrater agreement, we calculate Krippendorff's alpha on the evaluator judgements. We find rather low agreement: 0.12 overall, 0.18 for the FastSpeech questions, and 0.055 for the Tacotron questions.

Within-Rater Variability By survey design, our 157 evaluators rated both systems several times. The data therefore contains a between-rater variability (difference in judgements between the evaluators) as well as a within-rater variability (difference of an individual evaluator's judgement of the same system). There are several ways to address the within-rater variability, e.g., as a random effect in a mixed model or aggregating the data to obtain one judgement per person and system. We describe both below.

Logistic Random Effects Model We run a logistic random effects regression model with a random effect for person. The results show that the odds of the proposed model being perceived as better than the baseline is 0.385 times lower (95% confidence interval [0.315, 0.468]) for the Tacotron answers than the corresponding odds for the FastSpeech answers. In percentages, this means it is 61.5% less likely that Tacotron is perceived as better than the baseline in comparison to the same judgement for FastSpeech (95% CI [51.5%, 68.5%]). This constrasts with the results of Lux and Vu (2022), who found a much higher preference for Tacotron as opposed to FastSpeech.

Per-Person Preference Data (PPPD) If the data are aggregated per-label as in Table 2, we brush over potential effects of individual annotators. We therefore additionally create per-person preference data (we will refer to this as PPPD in the remainder of this paper). For this, we aggregate the raw counts from the survey into agreement ratios per system and per person, i.e. we count in how many questions about system X did person Y perceive the proposed system as better than the baseline. The PPPD will be used for binomial tests comparing against the mean preferences found in the original study in Sections 3.3 and 3.4 below.

3.2 Reproduction Targets

In line with the ReproNLP shared task guidelines, we attempt to reproduce the following type (i) and type (ii) results from Lux and Vu (2022).

(i) Single numeric values, i.e., the overall number of times each label was chosen.

(ii) Sets of related numeric values, i.e. sets of label counts per system.

Note that we cannot assess the reproducibility of type (iii) results since we do not have these from the original study. We reported our own type (iii) results (Krippendorff's alpha) above. The sets of labels are *Fastspeech-baseline*, *Fastspeech-proposed* and *Fastspeech-equal* for the FastSpeech system and *Tacotron-baseline*, *Tacotron-proposed* and *Tacotron-equal* for the Tacotron sytem.

3.3 Comparison to Original Study

Type (i) results In Table 3 we show the raw counts[9], the percentages of each answer category and the coefficient of variation (CV*) computed on the percentages for the original study and our reproduction.

The CV* in each row provides a measure of the dispersion of the original versus repeat percentages. A lower value means that the repeat result matches the original one more closely. The values in Table 3 show that the judgements of equality are more easily reproducible than the preference judgements for the baseline or proposed systems. Overall, the CV* values are rather high, indicating that the repeat results diverge from the original ones.

Type (ii) results In order to compare the full sets of results of the two studies, i.e. the sets of counts per label, we calculate Pearson's r. The results are shown in Table 4. The observed Pearson correlations are very low and none of them are significant, meaning that our repeat experiment does not confirm the original results.

[9] Lux and Vu (2022) (Figure 3) provide percentages but not raw counts per answer, so we calculated these. For FastSpeech, the counts add up to 202 instead of the expected 204, which could mean that two answers were skipped, or perhaps this is due to rounding the percentages.

	(Lux and Vu, 2022)		Current work		*CV**
	n	%	n	%	
Fastspeech-baseline	63	31%	113	12%	88.1
Fastspeech-proposed	51	25%	471	50%	66.5
Fastspeech-equal	88	43%	358	38%	12.3
Tacotron-baseline	22	11%	274	29%	89.7
Tacotron-proposed	106	52%	271	29%	56.6
Tacotron-equal	76	37%	391	41%	10.2

Table 3: Comparison of original and repeat evaluation. The Coefficient of Variation (*CV**) is calculated on the percentages.

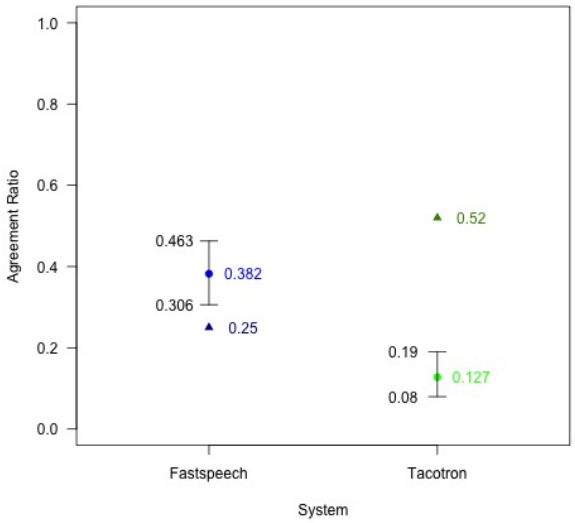


Figure 1: True preference confidence intervals from a binomial test on PPPD. The y-axis represents the percent of questions in which an evaluator agreed that the proposed system is superior to the baseline. The circles mark the estimated mean preference values from the binomial test and the whiskers show the 95% confidence intervals. The triangles indicate the values from the original study.

Comparison	Pearson's r	p-value
All labels	0.0019	0.997
Fastspeech	-0.113	0.928
Tacotron	0.141	0.910

Table 4: Pearson's r for label counts.

Binomial Test on PPPD We run a binomial test on the PPPD and test against the original study's reported preference outcomes. For both systems, we can reject the null hypothesis that our preference data leads to the preference outcomes reported in Lux and Vu (2022) (FastSpeech preferred in 25% of cases, Tacotron in 52%) with p-values < 0.05 for both systems (FastSpeech=0.00029, Tacotron<2.2e-16). This is visualised in Figure 1.

3.4 Comparison to Original Study - Transposed Systems

Since the analysis in Section 3.3 show a large similarity between the FastSpeech results of the original study and our Tacotron results, and vice versa. Therefore, in this section, we re-run the comparisons after transposing the labels of the two systems in our results. We do not know where the transposition happened, so this should not be taken as a statement regarding which system label corresponds to which set of results. The goal at this point is to see how the reproducibility assessment changes after the transposition.

Type (i) results In Table 5 we show the raw counts, the percentages of each answer category and the coefficient of variation computed on the percentages when the repeat results are transposed. We can see that the coefficients of variation are much lower for each original-repeated value pair than in Table 3. The Tacotron-Equal outcome is the easiest to reproduce and FastSpeech-proposed the most difficult.

Type (ii) results We also repeat the comparisons of type (ii) results with transposed labels. Table 6 shows the Pearson's r values. They show very high correlations of at least 0.95; the correlation for the combined set of labels (FastSpeech and Tacotron) as well as for FastSpeech on its own are significant, but not for Tacotron on its own. This indicates that our results broadly reproduce those of the original study when we transpose the system labels.

Binomial Test on PPPD We re-run the binomial test with transposed system labels on our PPPD. We find that also in the transposed scenario, we can reject the null hypotheses that we reproduce the mean preference of the original study with p-values < 0.05 for both systems (FastSpeech=0.00057, Tacotron=0.0002).

The identified 95% confidence intervals for the

	(Lux and Vu, 2022)		Current work **transposed**		*CV**
	n	%	n	%	
Fastspeech-baseline	63	31%	274	29%	6.7
Fastspeech-proposed	51	25%	271	29%	14.8
Fastspeech-equal	88	43%	391	41%	4.8
Tacotron-baseline	22	11%	113	12%	8.7
Tacotron-proposed	106	52%	471	50%	3.9
Tacotron-equal	76	37%	358	38%	2.7

Table 5: Comparison of original and **transposed** repeat evaluation. The Coefficient of Variation (*CV**) is calculated on the percentages.

Comparison	Pearson's r	p-value
All labels	0.991	0.00012
Fastspeech	0.999	0.0295
Tacotron	0.955	0.192

Table 6: Pearson's r for label counts with **transposition**.

true preference are [31%, 46%] for FastSpeech and [8%, 19%] for Tacotron. Note that our values of 50% and 29% also lie outside these intervals. All the per-label aggregated values are beyond the upper bound of the 95% confidence intervals on the PPPD. It thus appears that the per-label aggregation overestimates the preferences due to some effects of the evaluators.

Figure 2 visualises the outcomes of the binomial tests with transposed repeat results. We can see that the mean values from the original study now match the distributions better, but, as discussed above, they do not fall within the 95% confidence intervals of the PPPD.

4 Analysing Potential Error Sources

Our analysis show a more positive reproducibility assessment for system-transposed results. The current section is an attempt to assess potential sources of this supposed transposition error.

We were able to verify the following:

- The files provided to us match the corresponding ones in the possession of the original authors in terms of file size.
- The order of the videos in the Google Form[10] corresponds to the order created by the *random_videos.py* script, which is stored in *video2id.csv*.[11]

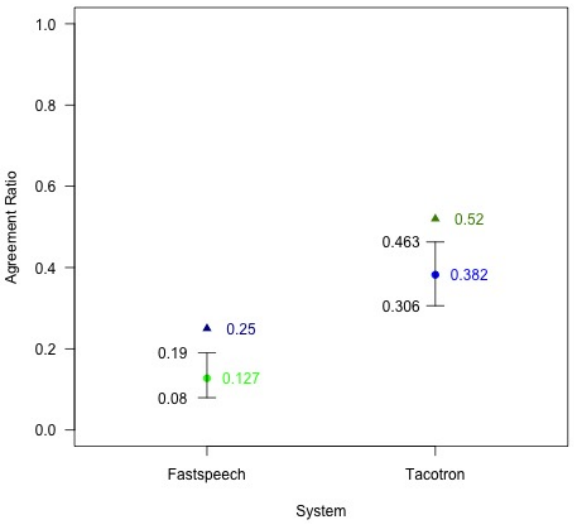


Figure 2: True preference confidence intervals from a binomial test on PPPD with **transposed** values from the repeat evaluation. The y-axis represents the percent of questions in which an evaluator agreed that the proposed system is superior to the baseline. The circles mark the estimated preference values from the binomial test and the whiskers show the 95% confidence intervals. The triangles indicate the values from the original study.

[10]Google Form: https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/google_form_pdf/GoogleForm_Evaluation%20von%20Text-zu-Sprache-Systemen%20-%20Google%20Formulare.pdf

[11]https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/

- The same order from *video2id.csv* is used to evaluate the results of the form and calculate the scores.[12]

This leaves us with the following potential sources of error:

1. The systems were transposed when creating the videos from the audio files
   (a) in the original experiment
   (b) in the repeat experiment
2. The results of the original survey were transposed when they were reported (due to the validation of the video order with *video2id.csv*, we can exclude this option for the repeat experiment.)

We cannot assess potential error sources (1a) and (2), since we do not have access to the required materials from the original study. Therefore, below we analyse the likelihood of option (1b) by comparing the audio files with the generated videos.

4.1 Audio Features Analysis

It is possible that systems were transposed when we create the videos from the audio files in order to embed them in the Google Form (option 1b above). We therefore want to verify if the created videos are similar to the audio files that they should correspond to. For this comparison, we use the Mel-Frequency Cepstral Coefficient (MFCC) audio features and cross-compare the audios and videos that correspond to each of the six text prompts.

We first generate the MFCC features of the audio and video using the Librosa Python library.[13]

For each audio-video pair with the same prompt (4x4=16 pairs per prompt) we then truncate the longer MFCC to the length of the shorter MFCC[14] and calculate the L2-norm of the difference between two MFCC-vectors as follows: $distance = \sqrt{\sum_1^n (a_i - b_i)^2}$, where a and b are the two vectors, x_i the element of vector x at index i and n is the length of the shorter MFCC-vector.

We visualise the resulting values as heatmaps in Figure 3: the x-axis shows the audios and the y-axis shows the videos that we presume to correspond to each audio. If our audio-video assignment is correct, the diagonal should display the lowest values. Indeed this is what we find: the diagonal is zero for all prompts, which makes it appear unlikely that there is a mistake in the audio-video assignment of our repeat experiment. Unfortunately, we cannot compare this to the audio-video assignment of Lux and Vu (2022) since their videos are no longer available.

5 Discussion and Conclusions

The positive aspects of this evaluation were that the original authors were able to provide the exact prompts, instructions, and questions used for the evaluation as well as information on how they set up the evaluation, so the setup was relatively straightforward. However, the question randomisation in the survey form could not be reproduced.

As for reproducibility, our initial assessment completely fails to confirm the results of the original study (see Section 3.3). Once we assume a transposition of systems (Section 3.4), we can paint a more positive picture with strong positive correlations and agreement. However, even in the transposed scenario, the per-label aggregation does not fully agree with the per-user preference data (PPPD): in a binomial test, we reject the null hypothesis that the per-label aggregated means could be drawn from the per-user preference data distribution. It appears that, when we aggregate on the question level, as opposed to the user level, we smooth over some within-rater variability. The low inter-annotator agreement (Krippendorff's alpha) further underscores that there are disagreements between the different evaluators. For both these assessments, the binomial tests and the inter-annotator agreement, we do not have any comparison to the original study since these data were not reported.

Table 7 summarises the findings of our repeat experiment for the originally obtained and transposed results.

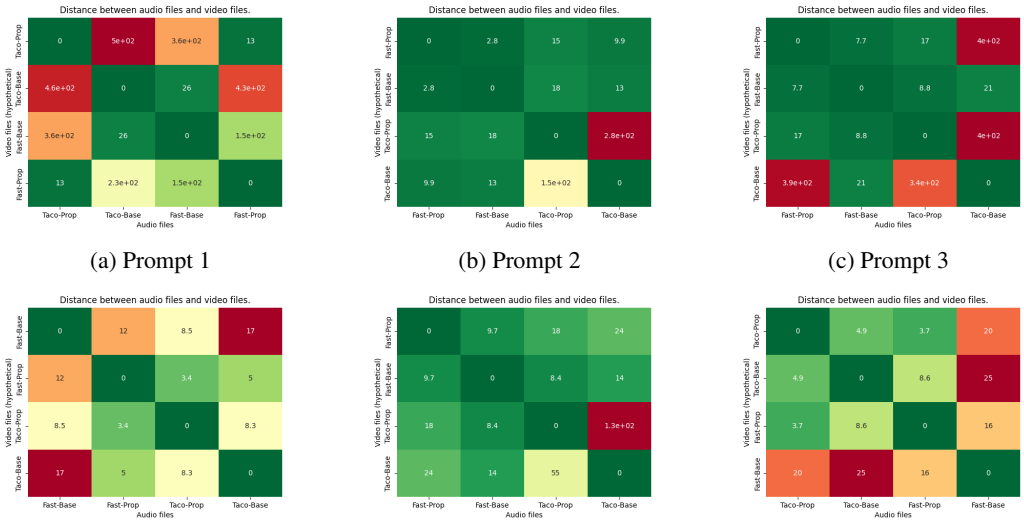
Finally, it is unclear in which study the hypothesised transposition happened. We can only confirm that for one of the systems there is a relatively clear preference for the proposed low-resource model (as opposed to the baseline), but we do not know for certain whether this is FastSpeech or Tacotron.

video2id.csv

[12] See script https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/get_label_counts_from_raw_results.py

[13] https://librosa.org/doc/latest/generated/librosa.feature.mfcc.html

[14] This is necessary because we do not exactly truncate the videos to the audio length and there can be trailing silence



(a) Prompt 1 (b) Prompt 2 (c) Prompt 3

(d) Prompt 4 (e) Prompt 5 (f) Prompt 6

Figure 3: Heatmaps showing distance between audio and video files for each text prompt. Video labels are hypothetical and correspond to the ones used in the current study.

Test	Outcome	Reproducibility	Table/Figure Ref.
Type (i) results	High CV* values	Not reproduced	Table 3
Type (ii) results	Low Pearson correlations, not significant	Not reproduced	Table 4
Binomial Test	Reject null hypothesis	Not congruent	Figure 1

(a) Findings of repeat experiment

Test	Outcome	Reproducibility	Table/Figure Ref.
Type (i) results	Lower CV* values	Reproduced	Table 5
Type (ii) results	High Pearson correlations, some significant	Reproduced	Table 6
Binomial Test	Reject null hypothesis	Not congruent	Figure 2

(b) Findings of **transposed** repeat experiment

Table 7: Summary of findings and reproducibility assessment.

6 Post-reporting Comparison Between Reproductions

The ReproHum team gave us access to another study which reproduced the same evaluation after finalising our report. Here, we briefly comment on their approach and findings. Mieskes and Benz (2023) also reproduced the human evaluation from Lux and Vu (2022). As far as we can see, there are two differences between their reproduction and ours: they randomised the order of answer options for each survey (whereas we always had the same order) and they informed participants that the study is a reproduction (whereas we did not). They collected a somewhat smaller set of responses (n=37) and their results also show high Coefficients of Variation. This finding provides further evidence (in addition to our audio/video features comparison in Section 4.1) that the label transposition happened in the original paper, either when creating the videos or when reporting the results (cp. Section 4). Therefore, if one wanted to interpret the results of the human evaluation with respect to the two systems, one should likely use the system label assignment from our study. The conclusion would then be that there is a preference for the proposed low-resource model (as opposed to the baseline) for FastSpeech, while for Tacotron, there is no clear preference.

Acknowledgments

We would like to thank Jan Deriu for his careful proofreading of the manuscript and his helpful suggestions.

References

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP. In *Proceedings of The Fourth Workshop on Insights from Negative Results in NLP*.

Florian Lux and Ngoc 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, Dublin, Ireland. Association for Computational Linguistics.

Margot Mieskes and Jacob Benz. 2023. h_da@ReproHum – Reproduction of Human Evaluation and Technical Pipeline. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems (HumEval'23)*.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

A Human Evaluation Datasheet (HEDS)

The Human Evaluation Datasheet (HEDS) for our evaluation can be accessed at `https://github.com/manhue/repronlp2023_lux_and_vu/blob/main/HEDS/datacard.json` and is also included in the supplementary materials of this paper.