11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

# Factors Affecting the Evaluation of Synthetic Speech in Context

*Johannah O'Mahony, Pilar Oplustil-Gallegos, Catherine Lai, Simon King*

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
j.o'mahony-1@sms.ed.ac.uk

## Abstract

Text-to-Speech synthesis is approaching the limit of naturalness that is possible from an isolated sentence. The focus of research is shifting to modelling contextual information, typically with the goal of producing better prosodic realisations by accounting for longer-range text dependencies from preceding sentences. But current evaluation methods were developed for single sentences and it is not yet clear how the evaluation of longer texts should be approached. Previous work suggests that evaluation of utterances in context can lead to an increase in Mean Opinion Score ratings, even when the synthesis technique is not context-aware. We investigated several factors that might explain this increase. Three experiments manipulated: the wording of instructions that participants received; the textual characteristics of context-stimulus pairs; and the prosodic realisation of the synthetic speech. We found that the wording of instructions has an impact on listeners' ratings of stimuli presented in context. The between-sentence context dependency of stimulus text has no impact on ratings. Listeners are, however, sensitive to prosodic differences, both in context and in isolation.

**Index Terms**: long-form Text-to-Speech, Text-to-Speech evaluation, context-aware Text-to-Speech

## 1. Introduction

Recent improvements in Text-to-Speech (TTS) modelling have paved the way for approaches which can take textual [1, 2] or acoustic context beyond the current utterance into account [3, 4]. Accounting for context has the potential to capture long-range text dependencies, discourse [5] and paragraph information [6], which are known to affect the prosodic realisation of an utterance. Context-sensitive prosody should exhibit increased variation compared to 'default' prosody generated for isolated sentences, and thus better long-form TTS [7]. However, in previous work, TTS output has been almost exclusively generated utterance-by-utterance and has therefore also been evaluated using isolated utterances [8, 9]. For context-sensitive TTS, appropriate evaluation paradigms are not yet fully developed.

One difficulty when rating prosodic variability is that countless realisations may be equally valid given a specific context [9]. Rating an utterance in context is a fundamentally different task to rating an utterance in isolation: varying the context can change the rating of the utterance. Conversely, in isolation the listener does not have access to any contextual information [8] (although participants might be able to imagine it [10]). This could potentially cause marked prosodic forms, elicited by a very specific context, to be rated lower when presented out of context, where listeners would expect default prosody. The opposite could also be true: perfectly natural and well-spoken utterances are rated highly in isolation, but when heard in an infelicitous context they are rated lower.

Clark et al. [8] found that utterances presented in isolation vs. in context had significantly different Mean Opinion Scores (MOS), with those heard in context receiving a higher rating when both the context and target were synthetic speech. Importantly, the synthetic speech used in their study was *not* context-sensitive. This boost in MOS score calls into question whether the MOS paradigm is the right way to evaluate synthetic speech in context.

The goal of this study is to discover what factors lead to such differences in MOS ratings. We conducted three experiments investigating various factors of interest.

- In experiment one, we test whether the **instructions** have an effect on MOS ratings of utterances presented in context.
- In experiment two, we assess whether between-sentence **textual context dependency** has an effect on MOS ratings.
- In experiment three, we test whether the MOS paradigm is suitable for rating **prosodically varied** synthetic speech.

Although Clark et al. tested a range of presentation types, including paragraphs, we will focus on a comparison between isolated utterances and context-target pairs in which an utterance is presented after a single context utterance. Finally, although prosodic realisation changes as a function of much more than the preceding sentence, e.g., pragmatic context, emotional state of the speaker, etc [9], we will concentrate on prosodic realisations which are determined by the textual context alone.

## 2. Related Work

As Text-to-Speech synthesis approaches its limit of naturalness, there is more and more focus on prosodic variability [10, 11, 12, for example] including the use of surrounding context to condition the realisation of the current utterance [1, 2, 3]. There is, however, little agreement on the best method for evaluating such prosodically-varied synthetic speech.

Some opt to use a qualitative approach. After testing whether prosodic realisations were perceptually distinct using a discriminative task, Hodari et al. asked participants to judge what effect different prosodic renditions had on the interpretation of the sentence, i.e., subtle differences in meaning or intent [10]. They found that participants were able to describe different contexts or situations where the prosodic variant would be found. A different qualitative approach was taken by Xu et al. who constructed different textual contexts and used these to generate different prosodic realisations of a single sentence in order to determine what effect their BERT-based context-aware model had on the prosody of a sentence [2].

Others opt for quantitative subjective evaluation using a MUSHRA-like paradigm. For example, Tyagi et al. used linguistic information, such as syntactic information and word embeddings to generate richer prosodic variability and evaluated both isolated utterances and long-form material [12]. In order to assess the quality of the prosodic output of individual sentences, they asked ten linguists to judge the appropriateness of the prosody in isolation. They stated that judging prosody requires domain-specific knowledge. This raises an issue with


10.21437/SSW.2021-26

devising appropriate metrics for prosodic felicity, if using non-expert listeners requires them to have an awareness of this dimension of the speech signal. Even among experts, however, it has been shown that inter-annotator agreement can be quite low [13]. For long-form evaluation, Tyagi et al. used crowd-sourced listeners and asked them to rate whole news stories for the *suitability* of the speaker's style, which they said would assess naturalness. As we will see from the results of experiment one, changing just one word in the task instructions can lead to different ratings. Many studies have used instructions such as *suitability* and *appropriateness* as synonyms for naturalness when they are in fact asking something quite different [8, 12].

Another option is a preference test to determine which system or prosodic realisation listeners prefer. For example, Aubin et al. tested the difference between a TTS system using discourse relations and a baseline system, using a preference test in which the target sentence was presented in a natural speech context [5]. Oplustil et al. also used a preference test in order to evaluate whether systems which take acoustic context into account from the preceding sentence perform better than a non-context-aware baseline [3]. While preference tests and MUSHRA both ask participants to make *direct comparisons* of stimuli with differing prosodic renditions, MOS tests do not. By asking listeners to provide 'absolute' ratings, many stimuli could receive the same MOS score.

Clark et al. [8] were the first to systematically evaluate the use of MOS for long-form evaluation. They compared differences in MOS ratings for utterances presented in isolation, in a context-target pair, or in a paragraph. They asked participants to rate the *naturalness* of utterances presented in isolation, but for context-target pairs, they asked participants to rate *appropriateness* of the target utterance given the context. The type of context was also varied, being either text, synthetic speech, or natural speech. They found that target utterances presented in context were rated significantly higher than the same utterances presented in isolation, when the context was in the form of text or synthetic speech. It is important to re-iterate that the synthetic speech was *not* context-dependent.

Clark et al. postulated that the increase in rating might be due to the task specification, and indeed other work has found that instructions can have an impact on MOS rating [14]. They also suggested that this may be due to 'the fact that the content of a paragraph non-initial sentence sounds less natural when presented out of context.' [8, Section 5.1]. They found no increase in ratings when the preceding context utterance was (non-vocoded) natural speech, reasoning that mismatches in quality between natural and synthetic speech make the synthetic speech sound of lower quality.

In the study reported in this paper, we focus exclusively on the MOS paradigm and investigate what factors lead to differing MOS scores between utterances presented in isolation vs. in context. Clark et al. used different wording of instructions when presenting isolated utterances than when presenting them in context. One of our experiments investigates the effect of wording alone, to avoid this confound. We restrict the investigation to the case of both target and context being synthetic speech. We also investigate whether the paradigm is sensitive enough to differentiate prosodically-different renditions of a sentence by a single system, something that Clark et al. did not do.

# 3. Research Questions

## 3.1. Effect of instructions

As noted in [8], the increase in MOS rating between the isolated condition and the TTS context condition was rather unexpected, given that the TTS model in question was not context-aware. One factor that might have influenced MOS was the task specification. Specifically, participants were asked to rate the *naturalness* of isolated utterances but the *appropriateness* of utterances presented in context. By wording the instructions to ask for either naturalness or appropriateness ratings, our first experiment tests whether this difference leads to changes in rating, independent of how the stimuli are presented.

## 3.2. Effect of between-sentence textual context dependency

Although [8] suggested that the increase in MOS rating may have been due to the task, they also suggested that utterances from non-paragraph-initial position may benefit from being presented with a preceding context. This is because non-initial sentences more often contain anaphoric references, such as pronouns, and therefore need a context in order to be fully understood. In experiment two, we manipulate the context-dependency of the target sentence text to test whether sentences containing anaphora receive higher MOS ratings when presented in a context that provides the referent, than non-anaphoric versions that do not need context in order to be fully understood.

## 3.3. Sensitivity of MOS to prosodic differences

While [8] investigated the effect of synthetic spoken context, natural spoken context and text context, they did not investigate whether participants are sensitive to changes in prosodic realisation when both context and target are synthetic and differ only in their prosody. [12] suggests that rating speech in context is difficult because there is no *correct* realisation and multiple variations will be equally acceptable. Therefore, in experiment three, we make one stimulus obviously non-canonical and ill-fitting to the context, in order to evaluate whether such a mismatch is salient for participants. If participants rate both the non-canonical and canonical highly in context, that would be evidence that this task is ill-suited to evaluating prosodic variation.

# 4. Methods

## 4.1. Data and models

We used the LJ Speech corpus, which consists of roughly 13 000 sentences read by a female speaker [15], for training all models. The model used in all experiments was the Ophelia implementation [16] of DC-TTS [17] . For experiment 3, we needed to manipulate prosody. We used the publicly-available training data used in [18] which is the LJ Speech corpus marked up with prosodic labels automatically generated using continuous wavelet transform (CWT) features which correlate with prosodic attributes such as prominence and boundaries. By marking up the training data with these labels, we obtained a model that offered control over prosody during inference, simply by changing the labels. Suni et al. used three strength levels of both accent and boundary labels, with accent level 0 signifying a de-accented word and boundary level 0 signifying no prosodic boundary. Level 2 accent signifies an emphasised word and level 2 boundary is roughly equivalent to an intona-

| | Condition | |
|---|---|---|
| | Context-dependent | Context-independent |
| Context | **Storms** have been named in the US since the 1700s, for the UK it's a relatively new thing. | **Storms** have been named in the US since the 1700s, for the UK it's a relatively new thing. |
| Target | The first **one** to receive a name in the UK was storm Abigail in 2015. | The first **storm** to receive a name in the UK was storm Abigail in 2015. |

Table 1: *Example of context-dependent (left column) and context-independent (right column) sentence pairs.*

Please, read the instructions carefully:
- You will be presented with **one sentence at a time**.
- We want you to rate how **natural** the sentence **sounds**.

(a) *Rating naturalness of utterances presented in isolation*

Please, read the instructions carefully:
- You will listen to **two sentences**.
- The second sentence will be highlighted in **bold text**.
- We want you to rate how **natural** the second sentence *sounds*, given the first sentence.

(b) *Rating naturalness of target utterances presented in context*

Please, read the instructions carefully:
- You will listen to **two sentences**.
- The second sentence will be highlighted in **bold text**.
- We want you to rate how **appropriate** the second sentence *sounds*, given the first sentence.

(c) *Rating appropriateness of target utterances presented in context*

Table 2: *Participant instructions.*

tional phrase boundary [18]. The LJ Speech recordings contain some background noise and reverberation, which we mitigated by post-processing all generated synthetic speech with the Automatic Sound Engineer (ASE) [19].

### 4.2. Stimuli

We created 110 pairs[1] of sentences each comprising a context sentence followed by a target sentence, using facts from Wikipedia. An example of two context-target pairs is given in Table 1. The same sentences were used in all experiments. We did not create test material using held out utterances from LJ Speech because this was too restrictive for carefully crafting suitable sentence pairs. All sentences were phonetised using [20] manually corrected, then synthesised.

Stimuli comprising a context-target pair were created by synthesising the two sentences separately then concatenating them into a single audio file separated by a 400 ms pause, a duration chosen through informal listening. This differs from [8], who asked listeners to click separate buttons to play context and target utterances.

#### *4.2.1. Text manipulation*

Each stimulus is the synthesised speech of a context sentence followed by one of two possible sentences: either the context-dependent follow-up (CD) or context-independent follow-up (CI). Table 1 provides an example. The CD target sentence needs the context sentence for the listener to resolve the anaphoric reference, such as *it* or *they*. In the CI condition, the target sentence has the referent filled in. The only difference between CI and CD conditions is the referent. Any two-word referents were matched with a two-word anaphoric reference so that the number of words in both conditions is the same.

#### *4.2.2. Prosodic manipulation*

To achieve prosodic manipulation, we manually modified the CWT labels on the input to the TTS model in order to create a *canonical* and a *non-canonical* rendition of each target sentence. Non-canonical renditions (as judged by one of the authors) were created by changing the accent and phrase boundary structure of the target utterances such that accents were placed on unexpected words (e.g., function words) or placing prosodic phrase boundaries in unexpected places. Figure 1 provides an example: *first* is de-accented in the non-canonical renditions, but accented in the canonical renditions; *and* receives a strong emphasis in the non-canonical renditions, but is de-accented in the canonical renditions. The creation of prosodic variants was constrained by the ability of the model, which did not render intelligible speech for every possible combination of accents and boundaries.

### 4.3. Participants

Listeners who self-reported to have no hearing impairment, be resident in the United States and have English as their first language were recruited through Prolific.[2] No other demographic information was asked for. None were allowed to participate more than once within this study. They received monetary compensation for taking part. Participants were asked whether they were using headphones. The responses from anyone who answered *no* were removed from analysis, following [8], as were those from participants who took less than 10 minutes (the minimum time required to listen to all stimuli).

### 4.4. MOS task

We implemented the MOS task in Qualtrics.[3] Following [8], participants were asked to rate stimuli on a scale of 1-5 in 0.5 increments (i.e., a 9-point scale). Points 1 to 5 were labelled as *poor*, *bad*, *fair*, *good* and *excellent*.

#### *4.4.1. Experiment 1 - Effect of instructions*

Each participant was assigned to one of 3 conditions. All participants in any given condition rated the same stimuli.

[1] Stimuli can be found: `https://johannahom.github.io/SSW-samples/index.html`

[2] https://www.prolific.co

[3] https://www.qualtrics.com/

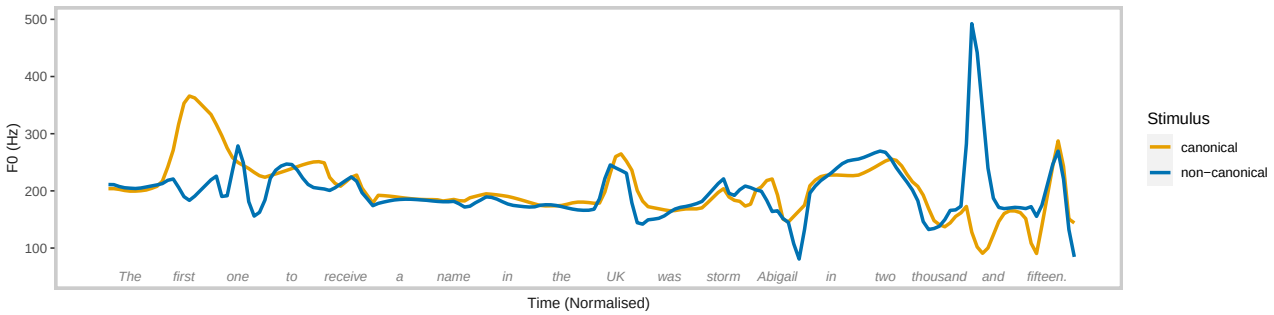


Figure 1: *Time-normalised F0 contour of a canonical and non-canonical stimulus from experiment 3.*

Condition 1: each participant was given the instructions in Table 2a, then rated 110 isolated sentences comprising all 55 unique context sentences, and 55 target sentences (a mixture of CI and CD) presented in randomised order. Condition 2: each participant was given the instructions in Table 2b then rated 55 context-target pairs presented in a random order. Condition 3: identical to condition 2, except using the instructions in Table 2c.

#### *4.4.2. Experiment 2 - Effect of between-sentence textual context dependency*

Each participant rated one of 4 sets of stimuli: Set 1: each participant was given the instructions in Table 2a, then rated 110 isolated sentences comprising all 55 unique context sentences, and 55 target sentences (a mixture of CI and CD) presented in randomised order. (*Since this is identical to experiment 1 condition 1, the same participant responses were re-used.*) Set 2: identical to set 1, and also using all 55 unique context sentences, except now using the remaining 55 target sentences not presented in condition 1 (also a mixture of CI and CD), to counterbalance. Set 3: each participant was given the instructions in Table 2c then rated all 55 context-target pairs presented in a random order. (*Since this is identical to experiment 1 condition 3, the same participant responses were re-used.*) Set 4: identical to set 3, except using the remaining 55 sentence pairs not presented in set 3, to counterbalance.

#### *4.4.3. Experiment 3 - Sensitivity of MOS to prosodic differences*

Each participant rated one of 4 sets of stimuli: Set 1: each participant was given the instructions in Table 2a, then rated 110 isolated sentences comprising all 55 unique context sentences rendered canonically, and 55 target sentences of which around half were rendered canonically and the rest rendered non-canonically, all presented in randomised order. Set 2: identical to set 1, with the same canonical renditions of all 55 unique context sentences, except with the canonical vs. non-canonical renditions of the target sentences swapped, to counterbalance. Set 3: each participant was given the instructions in Table 2c then rated 55 context-target pairs presented in a random order. Context sentences were always rendered canonically. Around half the target sentences were rendered canonically and the rest rendered non-canonically. Set 4: identical to set 3, except with the canonical vs. non-canonical renditions of the target sentences swapped, to counterbalance.

## 5. Results

All analyses were done in a by-items fashion such that, for each stimulus, the MOS rating is the mean of all participants' ratings for that stimulus. All data were found to be normally distributed following an insignificant Sharpiro-Wilk test and we therefore used two-tailed paired t-tests. Whenever making multiple pairwise comparisons, p-values were adjusted with Bonferroni coefficients.

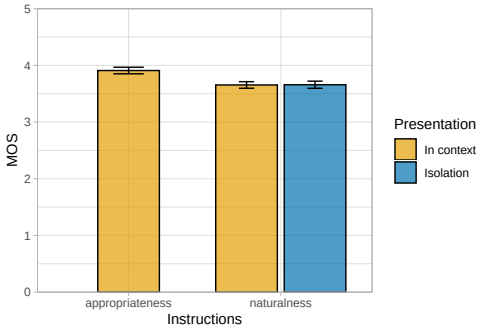


Figure 2: *Results for experiment one: MOS ratings of appropriateness and naturalness for utterances presented in isolation and in context.*

### 5.1. Experiment one

The experiment tests whether the instruction to listeners affects their ratings. A total of 108 participants took part of which 8 (7.4%) were removed using exclusion criteria from Section 4.3. As we see in Figure 2, stimuli were rated lower on the 5-point MOS scale when presented in isolation (M = 3.66 SD = 0.239) than in context. However this is only the case when using the instructions in Table 2c which asked them to rate how *appropriate* they sounded in context (M = 3.91 SD = 0.220) but *not* when using the instructions in Table 2b which asked them to rate how *natural* they sounded in context (M = 3.65 SD = 0.219). Ratings obtained with the 'how appropriate' instructions were significantly higher than those obtained with the 'how natural' instructions: t(54) = 9.94, $p < 0.001$. When using the 'how natural' instructions, there is no significant difference in ratings for stimuli presented in isolation vs. in context: t(54) = -0.16, $p = 1$. This refutes Clark et al.'s [8] hypothesis that it is the quality of the context and the match in quality (i.e., both context and target are synthetic speech) which leads to an increase in MOS rating.

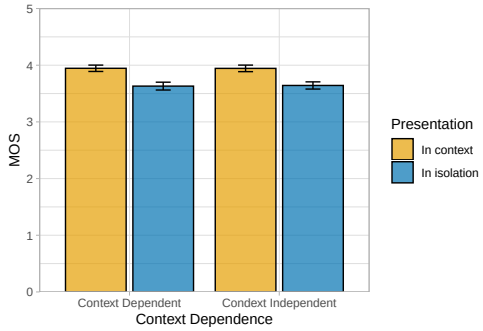


Figure 3: *Results for experiment two: MOS ratings for context-dependent and context-independent utterances presented in isolation and in context.*

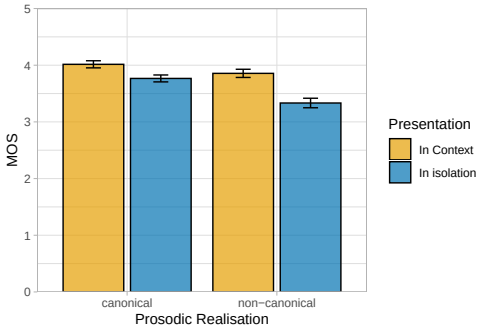


Figure 4: *Results for experiment three: MOS ratings for prosodically canonical and non-canonical renditions, presented in isolation and in context.*

A better explanation, also mentioned in [8], is that differences in ratings arise because participants interpret 'appropriate' differently to 'natural'. This implies that, in the condition from [8] where a synthetic utterance is presented after a natural spoken context utterance, listeners were rating the target as less *appropriate* rather than less natural: it is not appropriate for speech to change from natural to synthetic. We conclude that asking for ratings of *appropriateness* is different to asking for ratings of *naturalness*, for stimuli presented in context.

### 5.2. Experiment two

This experiment tests whether ratings of appropriateness are affected by textual dependence between the target and its context. A total of 144 participants took part of which 10 (6.9%) were removed using the exclusion criteria in Section 4.3. Results are shown in Figure 3. First, for utterances presented in isolation, there is no significant difference in ratings of naturalness for context-dependent (M = 3.63 SD = 0.262) and context-independent (M = 3.64 SD = 0.240) sentences (t(54) = -0.34, $p = 1$). When rated in context, there is no significant difference in ratings of appropriateness between context-dependent (M = 3.95 SD = 0.212) and context-independent utterances (M = 3.94 SD = 0.219): t(54) = 0.048, $p = 1$. Finally, consistent with the results from experiment 1, there is a significant difference between ratings of isolated utterances and utterances presented in context. This is true regardless of whether the utterance is context-dependent or is context-independent: t(54) = -8.30, $p < 0.001$ and t(54) = -10.48, $p < 0.001$ respectively. We conclude that textual context dependence does not affect listeners' ratings. However, as in experiment one, ratings of appropriateness for utterances presented in context are higher than ratings of naturalness for utterances presented in isolation.

### 5.3. Experiment three

This experiment tests whether MOS rating is sensitive to differences in prosodic realisation. A total of 144 participants took part of which 13 (9.0%) were removed using the exclusion criteria in Section 4.3. Results are shown in Figure 4. When presented in isolation, naturalness ratings of non-canonical renditions (M = 3.33, SD = 0.318) were significantly lower than of canonical renditions (M = 3.77 , SD = 0.231), t(54) = 9.41, $p < 0.0001$. This also holds true when these stimuli were presented in context and rated for appropriateness, although ratings of non-canonical (M= 3.86 SD= 0.273) and canonical (M = 4.02 SD = 0.237) are closer: t(54) = 3.86, $p = 0.001$. Both canonical renditions and non-canonical renditions received higher appropriateness ratings when presented in context than naturalness ratings when presented in isolation: t(54) = -7.29, $p < 0.001$ and t(54) = -18.33, $p < 0.001$ respectively. This is consistent with the findings reported in [8] and our results in experiments one and two. We conclude that MOS is sensitive enough to measure prosodic differences. As in experiments one and two, we once again conclude that appropriateness ratings for utterances presented in context are higher than naturalness ratings for utterances presented in isolation.

## 6. Discussion

Like Clark et al. [8], we found that utterances presented in context receive higher ratings of appropriateness than when presented in isolation, across all three experiments. In experiment one, we concluded asking whether an utterance sounds *appropriate* in context is not the same as asking whether it sounds *natural*. We believe the boost in rating is caused by the task specification, as Clark et al. suggested. This could be because the term *appropriate* is open to interpretation by listeners as textual appropriateness or prosodic appropriateness.

We tested whether context dependent targets received a boost in rating when their context was provided. The results from experiment two suggest this is not the case: this context-dependency of text does not play a significant role in listeners' ratings. This does not mean that participants were not taking the text into account at all. All our sentence pairs (an example is in Table 1) fitted together contextually, whether the target contained anaphoric reference or not: so all target sentences were appropriate in context, and listeners' ratings may reflect that. Of course, if they were *only* rating the text, we would expect the same high MOS across all stimuli, which was not the case: the speech did also matter. A future experiment could manipulate semantic or syntactic mismatch between context and target.

In experiment three, we tested whether MOS is sufficiently sensitive to measure differences in prosodic realisation. Clark et al [8] showed that varying the contexts between natural speech, synthetic speech and just text led to changes in MOS rating. They postulated that this was due to quality mismatches, with natural speech lowering the perceived quality of the following synthetic target. Our experiments exclusively used synthetic speech and did not vary the context utterance, so we can rule out any effects caused by differing contexts. We found that participants rated prosodically non-canonical targets as significantly less natural in isolation than canonical targets: so MOS

is sensitive to the difference. Our stimuli generally had substantial prosodic differences (the non-canonical renditions were very different to the canonical ones), so we are unable to say whether MOS would be sensitive to more subtle differences.

But, unexpectedly, *both* non-canonical and canonical target utterance received significantly higher ratings for appropriateness when presented in context, than identical utterances presented in isolation and rated for naturalness. Sometimes, a non-canonical form may indeed sound unnatural if heard in isolation, unless a very specific context is provided in which it sounds felicitous. Our stimuli, however, were constructed to ensure that the non-canonical renditions were *infelicitous* to their contexts, which is why we did not expect ratings of appropriateness to still be higher.

We would also like to extend this study to implicit measures of speech processing, such as reaction time word monitoring tasks, for example to evaluate which prosodic realisation leads fastest processing.

## 7. Conclusions

We replicated the most interesting finding in [8]: that synthetic speech is rated more highly in context. We investigated the source of this effect, considering the instructions to listeners, textual context-dependence and prosodic felicity. We found that the wording of instructions had a significant effect on the final MOS score. Instructions that asked listeners to rate *naturalness* resulted in the same rating regardless of whether utterances were presented in isolation or in context. In contrast, asking listeners to rate *appropriateness* of utterances presented in context resulted in a rating higher than the naturalness score, as in [8]. Naturalness and appropriateness are fundamentally different things. It is important, when reporting listening test results, to also report the exact wording of instructions to listeners.

To understand how listeners are interpreting appropriateness, we manipulated the target sentence text. We found no significant difference in the ratings of context-dependent and context-independent text. This does not mean that text plays no role in appropriateness rating. Future research could manipulate semantic and syntactic factors to gain a better understanding.

We investigated whether MOS is sensitive to prosody, which will be the main difference between the output of a context-aware model and a context-independent one. We found that, for utterances presented in isolation, participants exhibited a greater preference for canonical renditions, a preference that was maintained for utterances presented in context. MOS is an appropriate paradigm for evaluating prosodic differences. This increase in MOS was also found for non-canonical items, although they were constructed to be less felicitous in context. It is therefore still unclear what is exactly taken into account in the appropriateness rating. We would like to extend this work by including other variations in the instructions to participants, such as attempting to focus their attention on prosody.

**Acknowledgements:** we thank Rob Clark for providing additional details about the work in [8]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588 and was supported in part by: ANID, Becas Chile, nº 72190135.

## 8. References

[1] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational End-to-End TTS for Voice Agents," in *IEEE Spoken Language Technology Workshop (SLT)*, vol. 2, 2021, pp. 403–409.

[2] G. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, "Improving Prosody Modelling with Cross-Utterance BERT Embeddings for End-to-end Speech Synthesis," 2020, pp. 2–6. [Online]. Available: http://arxiv.org/abs/2011.05161

[3] P. Oplustil-Gallegos and S. King, "Using previous acoustic context to improve Text-to-Speech synthesis," in *arXiv preprint*, arXiv:2012.03763, 2020.

[4] P. Oplustil-Gallegos, J. O'Mahony, and S. King, "Comparing acoustic and textual representations of previous linguistic context for improving Text-to-Speech," in *SSW 2021*, 2021.

[5] A. Aubin, A. Cervone, O. Watts, and S. King, "Improving speech synthesis with discourse relations," in *Interspeech*. Graz: ISCA, 2019, pp. 4470–4474.

[6] M. Farrús, C. Lai, and J. D. Moore, "Paragraph-based prosodic cues for speech synthesis applications," in *Proceedings of the International Conference on Speech Prosody*, 2016, pp. 1143–1147.

[7] S. Prevost and M. Steedman, "Specifying intonation from context for speech synthesis," *Speech Communication*, vol. 15, no. 1, pp. 139–153, 1994.

[8] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs Rob," in *The 10th ISCA Speech Synthesis Workshop*, Vienna, 2019.

[9] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, E. Szekely, C. Tannander, and J. Vosse, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *The 10th ISCA Speech Synthesis Workshop*, 2019, pp. 105–110.

[10] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0," in *Proceedings of the International Conference on Speech Prosody*, 2020, pp. 965–969.

[11] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading TTS: Exploiting text structure information," in *Interspeech*. Stockholm: ISCA, 2017, pp. 1064–1068.

[12] S. Tyagi, M. Nicolis, J. Rohnke, T. Drugman, and J. Lorenzo-Trueba, "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection," in *Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds. Shanghai: ISCA, 2020, pp. 4407–4411.

[13] A. K. Syrdal and J. T. McGory, "Inter-transcriber reliability of toBI prosodic labeling," in *Sixth International Conference on Spoken Language Processing*. Beijing, China,: ISCA, 2000, pp. 235–238.

[14] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech Synthesis: The effect of style and expectation," *Proceedings of the International Conference on Speech Prosody*, pp. 1012–1016, 2014.

[15] I. Keith and J. Linda, "The LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[16] CSTR-Edinburgh, "Ophelia," 2018. [Online]. Available: https://github.com/CSTR-Edinburgh/ophelia

[17] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2018, pp. 4784–4788.

[18] A. Suni, S. Kakouros, M. Vainio, and J. Šimko, "Prosodic Prominence and Boundaries in Sequence-to-Sequence Speech Synthesis," in *Proceedings of the International Conference on Speech Prosody*, Tokyo, 2020, pp. 940–944.

[19] C. Chermaz and S. King, "A sound engineering approach to near end listening enhancement," in *Interspeech*. Shanghai: ISCA, 2020, pp. 1356–1360.

[20] K. Park and J. Kim, "g2pE," 2019. [Online]. Available: https://github.com/Kyubyong/g2p