10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs

Rob Clark, Hanna Silen, Tom Kenter, Ralph Leith

Google U.K.

rajclark@google.com, silen@google.com, tomkenter@google.com, leith@google.com

Abstract

Text-to-speech systems are typically evaluated on single sentences. When long-form content, such as data consisting of full paragraphs or dialogues is considered, evaluating sentences in isolation is not always appropriate as the context in which the sentences are synthesized is missing.

In this paper, we investigate three different ways of evaluating the naturalness of long-form text-to-speech synthesis. We compare the results obtained from evaluating sentences in isolation, evaluating whole paragraphs of speech, and presenting a selection of speech or text as context and evaluating the subsequent speech. We find that, even though these three evaluations are based upon the same material, the outcomes differ per setting, and moreover that these outcomes do not necessarily correlate with each other. We show that our findings are consistent between a single speaker setting of read paragraphs and a two-speaker dialogue scenario. We conclude that to evaluate the quality of long-form speech, the traditional way of evaluating sentences in isolation does not suffice, and that multiple evaluations are required.

1. Introduction

Traditionally, text-to-speech (TTS) systems are trained on corpora of isolated sentences. As such, their output is optimized, if only indirectly and inadvertently, for synthesizing isolated sentences. As the use of TTS proliferates and the application of TTS extends into domains where the required output is high quality discourse, long-form (multi-sentence) data is being used more frequently to build voices and to evaluate the quality of the long-form output.

The traditional evaluation approaches used in TTS are designed to assess the quality of synthesized sentences in isolation using metrics such as mean opinion score (MOS) [1] and side by side (SxS)[1] discriminative tasks. For long-form TTS, i.e., speech passages longer than one sentence, this evaluation scenario is limited in terms of what it can be used to evaluate; presenting sentences in isolation means that they are being evaluated out of their natural context. Long-form speech—which may consist of either single speaker data, such as an audio book, a news article, or a public speech; or multi-speaker data such as a conversation between multiple participants—should ideally be evaluated as a whole, because evaluating the quality of isolated sentences will not inform us of the overall quality of the discourse experience, which includes factors such as the appropriateness of prosody in context and fluency at paragraph-level.

The most obvious approach to evaluate long-form TTS is to use the existing standard evaluation techniques and simply present whole paragraphs or dialogues to raters. Doing so, however, raises questions about the impact of providing longer stimuli that vary in length, both from the perspective of increasing the cognitive load of the raters through presenting them with more material, as from the perspective of increasing the overall variability in the length of stimuli. Including paragraph length as a factor in any subsequent analysis is often impractical as it drastically increases the amount of evaluation material required to fully control for it and still obtain a meaningful result.

An additional scenario, which sits between evaluating isolated sentences and full long-form passages, would be to evaluate the quality of passages of speech in their immediate context. In this scenario, full long-form passages are divided into two parts to form a context part and a stimulus part. Raters are asked to evaluate the quality of the speech stimulus part as a continuation of a given context part, and are presented with the speech (or, potentially, just the text) of the context immediately before hearing the stimulus. Our hypothesis is that we can achieve a higher sentence-level precision in this scenario than we could if the sentences were presented in isolation—as listeners are explicitly asked to evaluate whether the stimulus is appropriate for a specific context rather than being allowed to hypothesize a context for which the stimulus would be appropriate—while keeping the cognitive load for raters low compared to presenting them with full paragraphs.

To develop a better understanding of the potentials of the methods described above:

- We analyze three different ways of evaluating long-form TTS speech. To the best of our knowledge this is the first time a formal comparison based on a multitude of experiments has been performed;
- We show that both evaluating long-form TTS speech as paragraphs and as context-stimulus pairs yields results distinctly different from the traditional single sentence evaluation approach, which is remarkable given that the evaluations in all settings are based on the same material;
- We propose to combine these evaluations to get the most complete picture of long-form TTS quality.

As we are interested primarily in the relative differences of results between the various evaluation scenarios, rather than the relative differences between the TTS systems used, we focus on MOS tasks in this paper, and leave out SxS evaluations.

The remainder of this paper is organized as follows: Section 2 discusses related work and existing approaches to (long-form) TTS evaluation. Section 3 details the three ways of evaluating long-form TTS that we propose. Experimental details are presented in Section 4. Sections 5 and 6 present the results of the main and additional experiments, respectively. Section 7 concludes.

[1] Also referred to as AB tasks.

10.21437/SSW.2019-18

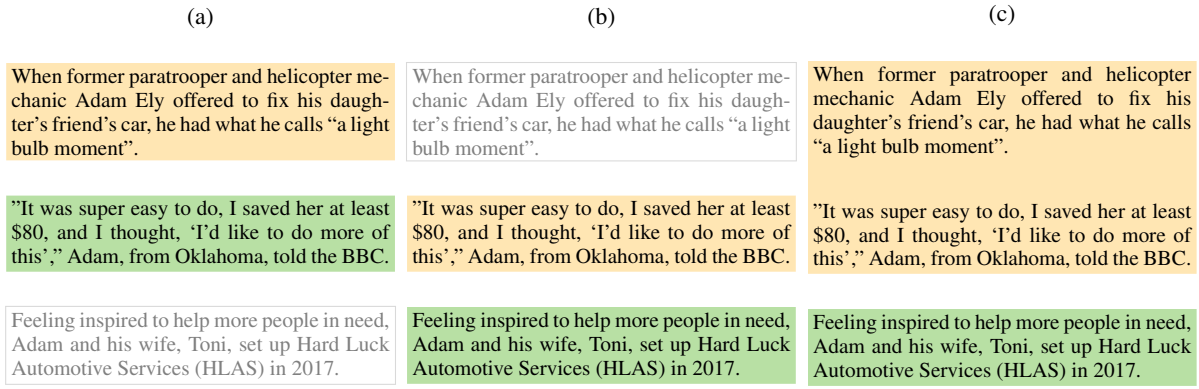

Figure 1: *Illustration of three ways to evaluate single sentences that are part of a three sentence paragraph, using other parts of the paragraph as context. Green boxes contain the audio to be evaluated. Yellow boxes are sentences presented as context (text and/or audio), not to be evaluated. White boxes show sentences of the paragraph not used in the rating task. (a) and (b) present the single previous sentence as context, while (c) presents two previous sentences in the paragraph.* *(Text courtesy of BBC News)*

2. Related Work

The currently used MOS [1] and SxS tasks for evaluating TTS naturalness were established in [2, 3]. Extensions and improvements to MOS evaluation have been made previously [4, 5, 6], but none of this work covers the long-form scenario.

In [7], the point is made that evaluating sentences in isolation when they are in fact part of a dialogue does not represent a real-world end-use scenario. An alternative evaluation setup is proposed in which raters interact with an avatar. The experiments on conversational data in Section 5.2 follow this work, in the sense that turns in the dialogue are presented in context rather than in isolation. A key difference is that we do not incorporate an interactive setting. This allows for comparison between the three different settings we propose, none of which involve interaction.

In [8] discourse structure is taken into account for improving prosody of longer passages of text. The focus in this work, however, is on the improvements of a supervised signal pertaining to rhetorical structure, rather than on the evaluation.

It is observed in [9] that evaluating sentences in isolation "may not be appropriate to measure the performance of intonation models." However, the objective in [9] is to show that when evaluating single sentences without providing context, multiple prosodic variants of the same sentence might be equally valid according to raters. No experiments were done to determine how those ratings change if a context is provided.

Lastly, an evaluation protocol for an audiobook reading task, adapted from the scales proposed by [10], is presented in [11]. The method is aimed at a fine-grained analysis of the audiobooks task in particular, and does not cover an analysis of different evaluation alternatives.

In short, to the best of our knowledge, no systematic analysis of the effect of different ways to evaluate long-form TTS context has been carried out before. The absence of such investigation is the primary motivation for this study.

3. Evaluating Long-form TTS

We present three ways to evaluate long-form material: as single sentences in isolation, as full paragraphs, and as context-stimuli pairs. We should note that, even as the discussion below is presented in terms of sentences in a paragraph, it equally applies to turns in a dialogue. Furthermore, although the discussion is applied to MOS, it is independent of what type of evaluation is performed and applies equally to SxS tests as well as other varieties of evaluation such as MUSHRA [12].

3.1. Evaluating sentences in isolation

Firstly, we can use the traditional TTS approach and evaluate individual sentences separately as if they were isolated sentences. As mentioned in Section 1, the obvious disadvantage of this approach is that in evaluating isolated sentences, we are not considering the fact that these sentences are part of a larger discourse which may affect the way they should be synthesized. There are, however, advantages to this method of presentation. Raters, for example, are less likely to be able to infer the content based on context, in this setting, so lack of intelligibility is more likely to result in bad naturalness scores.

In the work presented here we treat this method of evaluation as a reference to compare other results to, which allows us to determine empirically whether we learn something different using alternative evaluation methods.

3.2. Evaluating full paragraphs

At the other end of the scale is the evaluation of full paragraphs. Evaluating full paragraphs imposes a higher cognitive load on raters which may impact the responses obtained. Paragraph length, becomes an issue in its own right, and we may get different results depending on how long the paragraphs are. An advantage of this setting, however, is that it is possible for raters to make judgments on the overall flow of the sentences in the paragraph, something they cannot do when they hear them in isolation.

3.3. Evaluating context-stimulus pairs

To compromise between evaluating isolated sentences and paragraphs we can present one or more sentences of the paragraph as context to the rater, and the subsequent sentence or sentences as the stimulus to be rated.

This approach raises questions regarding the amount of material that should be presented, both as context and as stimulus. Should we constrain the length of the context and stimulus in terms of the number of sentences or by overall length in words or syllables? E.g, a single long sentence may be longer than two short sentences. In the work presented here, we choose to control the variation in terms of number of sentences and length

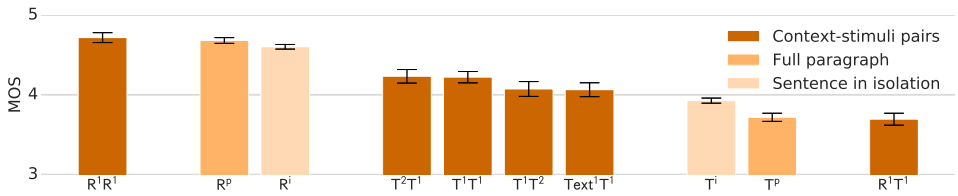

Figure 2: *MOS results on the news reading data set across evaluation strategies. 'R' refers to real speech, 'T' is for TTS (synthesized speech), 'Text' means no speech but text. For evaluations without context, superscript 'p' denotes a full paragraph and superscript 'i' denotes sentences in isolation. For evaluations with context, '**R¹R¹**' is a context-stimulus pair of one line of real speech context and one line of real speech stimulus, '**T²T¹**' is two lines of TTS context, one line of TTS stimulus.*

of paragraphs. We also evaluate whether paragraph length influences paragraph MOS scores (see Section 6.1).

Figure 1 shows various options for contexts. To keep the figure clear a single sentence stimulus is shown, but we note that multiple sentences can be presented as a stimulus too.

4. Experimental Setup

We compare the three approaches for long-form TTS evaluation outlined above: 1) sentences in isolation, 2) full paragraphs 3) context-stimulus pairs.

To test for consistency across different domains, we present results of evaluations in two distinctly different scenarios: news-reading and read conversations. We use a WaveNet [13] TTS voice that was built incorporating in-domain training data.

4.1. Data and TTS system

For our first series of evaluations we use a proprietary data set of read news articles. We select only paragraphs containing two sentences or more. Single sentence paragraphs are less interesting for our evaluations as any evaluation comparing single sentences to one-sentence paragraphs will come out even. In our final dataset, we have 103 paragraphs of news material. The longest paragraph contains 9 sentences, and the mean length is 3.0 sentences. To further compare the contexts of different lengths, we select a subset of paragraphs with a minimum length of three sentences resulting in a subset of 57 paragraphs with a mean length of 3.8 sentences.

The second data set consists of read conversations where two speakers take turns speaking. We use turns in conversation as the units making up our stimuli (similar to using sentences in paragraphs in the previous setting). An individual turn itself may consist of multiple sentences, which we keep together as a single turn. The conversation design determines the total amount of variation of length per turn and keeps the amount of material per turn reasonably balanced. We use two pairs of speakers. The first pair recorded 42 conversations and the second pair recorded 71. A key difference between this dataset and the news reading one is that the speaker changes between turns.

We should note that for the first dataset, a held-out set of passages was used for evaluation. In the conversation case we did not have sufficient data to do this, and the conversations used for evaluation were used as training example for the WaveNet voice as well. This is suboptimal, but as we are not trying to assess how well a particular TTS model can generalize, this should not affect the results presented here—we have seen little evidence that WaveNet models over-fit in such a way that any one utterance can have a significant impact on the resulting voice.

To synthesize speech we use a two-step approach where one model is trained to produce prosodic parameters (F_0, c_0 and duration) [14] to be used by a version of WaveNet [13], trained separately to produce speech from linguistic features and the predicted prosodic parameters. The model is not context-aware; it synthesizes speech sentence by sentence.

4.2. Rating task

We use a crowd-sourced MOS rating task for evaluation, where raters are asked to rate naturalness for the settings that do not include context, and appropriateness where the stimulus follows a context. Stimuli are rated on a scale of 1-5. The whole number points of the scale are labeled 'poor', 'bad', 'fair', 'good' and 'excellent'. Raters are allowed to rate at 0.5 increments of the scale, as we find this gives slightly finer resolution in MOS scores at the top end of the scale. Stimuli are presented to raters in blocks of 10, except for the full paragraphs, which are presented in blocks of 5. Each stimulus is presented 8 times per experiment to randomly assigned raters and the MOS results presented are calculated from the averages of those 8 ratings for each stimulus. Raters not using headphones are omitted from the analysis. The number of raters per task varies due to the overall number of stimuli in the task, with the lowest number of raters in a task being 35.

4.3. Evaluation tasks

For **news reading** the following evaluations are carried out:

1. **Sentences in isolation** Both real speech and TTS versions of each sentence are presented as stimulus. Below, these results are referred to as $\mathbf{R^i}$ (Real speech, individual sentences) and $\mathbf{T^i}$ (TTS, individual sentences).
2. **Full paragraphs** The same data is used as above, but presented as full paragraphs. Both real speech and TTS versions are presented to the raters. These results are labelled $\mathbf{R^p}$ and $\mathbf{T^p}$, respectively.
3. **Context-stimulus pairs** The first and second lines of paragraphs are presented, where the first line is the context and the second line is the stimulus to be rated. We experiment using a combination of real speech, TTS and text as the context, and both real speech and TTS as the stimulus. Additionally, to evaluate varying the length of the context, we provide two lines either as context or as stimulus. In this setting, only TTS is used as context:

   $\mathbf{R^1R^1}$ One sentence real speech as context, one sentence real speech as stimulus;

   $\mathbf{R^1T^1}$ One sentence real speech context, one sentence TTS as stimulus;

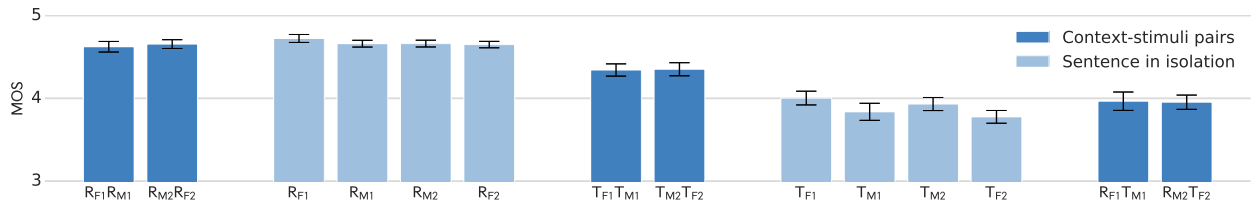

Figure 3: *MOS results on the conversational data set, presented with context (F1 M1 and M2 F2) and in isolation (F1, M1, M2, F2).*

T^1T^1 One sentence TTS context, one sentence TTS stimulus;

$Text^1T^1$ One sentence textual context, one sentence TTS stimulus.

T^2T^1 Two sentence TTS context, one sentence TTS stimulus;

T^1T^2 One sentence TTS context, two sentence TTS stimulus;

In the news reading tasks, real speech samples $\mathbf{R^i}$ are cleaned from sentence-initial breathing noise and $\mathbf{R^p}$ from paragraph-initial breathing noise. All real speech samples are downsampled to match the TTS sampling rate.

The **conversational** data includes two pairs of speakers: **F1** paired with **M1**, and **M2** paired with **F2**, where **F** and **M** denote female and male speakers, respectively. The evaluations use TTS samples and real speech samples from all four speakers: $\mathbf{T_{F1}}$, $\mathbf{T_{F2}}$, $\mathbf{T_{M1}}$, $\mathbf{T_{M2}}$ and $\mathbf{R_{F1}}$, $\mathbf{R_{F2}}$, $\mathbf{R_{M1}}$, $\mathbf{R_{M2}}$, respectively.

WaveNet voices were built for each of these speakers.

5. Results

In this section we discuss the results of the two sets of experiments performed. We use a two-tailed independent t-test with $\alpha = 0.05$ for calculating significance between results.

5.1. News reading

Figure 2 shows the results for all MOS evaluations, ordered from high to low.

The first block of results confirms the intuition that real speech scores higher than all settings involving TTS. The highest ratings are for appropriateness of a real speech stimulus in a real speech context ($\mathbf{R^1R^1}$). The scores are slightly higher than for naturalness ratings of both real speech paragraphs ($\mathbf{R^p}$) and real speech isolated sentences ($\mathbf{R^i}$). Real speech paragraphs ($\mathbf{R^p}$) themselves are rated slightly higher than real speech isolated sentences ($\mathbf{R^i}$). Within this grouping of real speech results, there are significant differences between all three conditions. These results alone already indicate that there is a difference between evaluating sentences in isolation and in context, even when only real speech is involved.

The next block of results in Figure 2 shows the results for context-stimuli pair evaluations. Presenting two sentences as context, while rating one follow-on sentence ($\mathbf{T^2T^1}$) scores highest, followed by one sentence as context with one sentence rated ($\mathbf{T^1T^1}$). The lowest scores are obtained when one sentence is presented as context followed by two sentences being rated ($\mathbf{T^1T^2}$). $\mathbf{T^1T^2}$ is found to be significantly different from $\mathbf{T^1T^1}$and $\mathbf{T^2T^1}$. The final bar in this block shows the result of presenting the context as text rather than speech ($\mathbf{Text^1T^1}$) and this gives a score not significantly different from $\mathbf{T^1T^2}$. These results indicate that the length of the context presented does not appear to have a significant effect on the MOS results, but increasing the length of the stimulus lowers the MOS result.

The next block holds the results for evaluating TTS sentence ($\mathbf{T^i}$) and paragraph ($\mathbf{T^p}$) naturalness in isolation. These results are significantly lower than the ones in the previous blocks. One potential explanation why raters would rate paragraphs lower than their individual sentences, is that ratings are strongly influenced by the worst thing they hear in the stimulus and thus as the stimulus becomes longer the rating is likely to be lower. This interpretation is consistent with the result above where a lower MOS was found when increasing the stimulus length in a context. It could suggest a (weak) correlation between the minimum sentence MOS and the paragraph MOS (cf. Table 1 discussed in Section 6.1). Alternatively, it may be that higher cognitive load simply results in lower ratings.

It is interesting to see that sentences with context are rated higher than when presented in isolation. As noted in Section 4.1, the TTS model used is not taking any paragraph level context into consideration, so the difference has to be attributed either to the task itself, or the fact that the content of a paragraph non-initial sentence sounds less natural when presented out of context.

The final and lowest result in Figure 2 is for TTS stimuli with real speech context ($\mathbf{R^1T^1}$). A key observation to point out is that these results are considerably lower than the ones where *the same stimuli* are presented with a TTS context. This seems to indicate an anchoring effect of the real speech lowering the perceived quality of the TTS, suggesting that when rating appropriateness in context, raters pay particular attention to whether the quality of the stimulus matches the quality of the context.

Lastly, the fact that cases where the TTS context was used score higher than when sentences are rated in isolation suggests that part of the appropriateness judgment relates to similarity in quality compared to the context, and the rating does not just relate to overall naturalness and how well the prosody is suited in context to the paragraph. The implication here is that the context-stimuli setting cannot be considered to be an alternative to the sentence-in-isolation naturalness MOS task, because it will produce varying results depending on the quality of the context. The MOS result a context-stimulus evaluation yields can be substantially higher than one obtained for a sentence in isolation when there is a quality match between the context and stimulus, or lower it when the quality of the context is higher than that of the stimulus.

5.2. Conversations

To determine if the differences observed between ratings for sentences presented in isolation versus sentences presented in context are consistent across domains, we perform evaluations on a distinctly different dataset that consists of conversations.

Table 1: *Correlations of sentence MOS scores with paragraph MOS (news reading data).*

Correlate	**Mean sentence MOS**	**First Sentence MOS**	**Second Sentence MOS**	**Last sentence MOS**	**Min. sentence MOS**	**Max. sentence MOS**	**Paragraph no. of sentences**	**Paragraph no. of words**
r	**0.296**	0.087	0.114	**0.268**	**0.234**	**0.345**	-0.020	0.029
p	$<$ **0.05**	$>$ 0.05	$>$ 0.05	$<$ **0.05**	$<$ **0.05**	$<$ **0.01**	$<$ **0.05**	$>$ 0.05

Table 2: *Regression model coefficients for predicting the paragraph MOS from individual sentence MOS for paragraphs of lengths two, three and four sentences long (news reading data).*

Model for paragraphs of two sentences

Num. of paragraphs 46

$R^2 = 0.04, \quad (F = 0.95,\ p > 0.05)$

	coef	std err	t	$P>\|t\|$
intercept	2.50	0.92	2.74	0.01
s1	0.15	0.15	0.99	0.33
s2	0.17	0.15	1.13	0.26

Model for paragraphs of three sentences

Num. of paragraphs 31

$R^2 = 0.27, \quad (F = 3.35,\ \mathbf{p < 0.05})$

	coef	std err	t	$P>\|t\|$
intercept	1.30	0.95	1.37	0.18
s1	0.39	0.14	2.77	0.01
s2	0.01	0.12	0.11	0.92
s3	0.22	0.14	1.63	0.17

Model for paragraphs of four sentences

Num. of paragraphs 15

$R^2 = 0.54, \quad (F = 2.97,\ p > 0.05)$

	coef	std err	t	$P>\|t\|$
intercept	4.23	2.03	2.08	0.06
s1	-0.53	0.25	-2.14	0.06
s2	0.12	0.16	0.73	0.48
s3	0.17	0.17	1.00	0.34
s4	0.11	0.22	0.51	0.62

We restrict the evaluation to using only the first and second turns of the dialogues, as we saw previously that amount of context presented did not greatly affect the results.

We both evaluate the first and second turns in isolation, and we evaluate the second turns using the first turns as context. Note that, different from the news data, the context in this scenario is uttered by a different speaker. In two separate tasks, we present the context either as real speech or as TTS.

The results of this experiment are shown in Figure 3. The MOS for the evaluations involving only recorded voices $\mathbf{R_{F1}}$, $\mathbf{R_{F2}}$, $\mathbf{R_{M1}}$ and $\mathbf{R_{M2}}$ range between 4.6 – 4.7 with no statistically significant difference between the scores for the second turns presented in isolation or in their recorded context—the only statistically significant difference in this group of evaluations is observed between $\mathbf{R_{F1}}$ and $\mathbf{R_{F1}R_{M2}}$ or $\mathbf{R_{F2}}$. Furthermore, MOS scores for the synthesized turns in isolation range from 3.8 for voices $\mathbf{T_{M1}}$, $\mathbf{T_{M2}}$ and $\mathbf{T_{F2}}$ and to 4.0 for voice $\mathbf{T_{F1}}$. Conversely, when the second turns of the dialogues are preceded by their context, the MOS for the TTS voices rises to the 4.3 – 4.4 range, mirroring the effect we saw for the read news data. Furthermore, using real speech as context ($\mathbf{R_{F1}T_{M1}}$ and $\mathbf{R_{M2}T_{F2}}$) decreases the resulting MOS for TTS stimuli as again the raters appear to consider the quality of the context as an anchor. However in this case these ratings do not drop below the ratings of the turns in isolation. We attribute this to the fact that, even if the context is presented as the natural speech of a different speaker, this still acts as an anchor, but a weaker one than the natural speech of the same speaker would be.

6. Further analysis

The results presented in the previous section show that rating a full paragraph gives different results then rating sentences in isolation does, regardless of how the task is set up. To gain more insight into this observation, we analyze correlations between full paragraph and sentence ratings. These tests are carried out on the news reading data set.

6.1. Correlating ratings of full paragraphs and single sentences

Table 1 shows the correlations (Pearson's r) between paragraph MOS scores and various sentence MOS ratings. We see significant correlations, at the 5% level, of around 0.3 between the MOS rating of the full paragraph and the 'Mean sentence MOS', 'The last sentence MOS' and the 'Minimum sentence MOS'. Furthermore, there is a significant correlation of 0.345, at 1% level, between paragraph MOS and the maximum MOS of the sentences the paragraphs consists of. All of these r values are small, however, and only 12% of the variance can be accounted by the maximum sentence MOS of r = 0.345. These correlations show that paragraph MOS is influenced by the individual sentences, both collectively through the means and individually through the extremes, yet only less than half the variance can be accounted for this way. We conclude that, while paragraph and individual sentence ratings cannot be considered to be independent, the majority of the variance seen in paragraph MOS scores is not accounted for by the MOS ratings of the individual sentences.

Lastly, the rightmost columns of Table 1 shows the correlations between the paragraph length (measured in sentences or words) and the paragraph MOS rating. There is a correlation, supporting the intuition that MOS ratings go down as paragraphs get longer in terms of sentences, but the r value is so small that it does not appear to be meaningful.

6.2. Correlating ratings of full paragraphs and single sentences and their positions

An alternative hypothesis is that, even if little correlation between the MOS ratings of paragraphs and the MOS ratings of each individual constituent sentence is found when these are analyzed all together, perhaps the latter can be inferred from the former if the order of sentences is taken into account. To test this hypothesis we create linear regression models predicting the paragraph MOS from the individual sentence MOS values, depending on their position in the paragraph. We restrict these experiments to paragraphs of length two, three and four sentences, as we do not have sufficient data in the current experiments to analyze longer paragraphs, and it is not immediately

clear how models allowing for variable paragraph length should be designed.

The results of the regression experiments are shown in Table 2. First, we note that the only model with a significant R-squared value, i.e., that can account for a variance in a significant way, is the model for paragraphs of three sentences. For this model the only significantly non-zero contribution is made by the MOS of the first sentence. That trend is not repeated for the other models.

For the two sentence model only the constant term contributes in a non-zero way, i.e., the paragraph MOS is the same for all paragraphs under this model. Hence, it is unsurprising that this has an insignificant R-squared value.

For paragraphs of four sentences there are non-zero contributions from both the constant term and the first sentence in the paragraph, but the low and non-significant R-squared value for this model means this model does not fit the data well.

In short, we conclude from these results that the individual MOS ratings of sentences are bad predictors of paragraph MOS, and if a MOS which reflects the overall quality of the paragraph is required, it needs to be obtained directly.

7. Conclusions

Now that the performance of TTS systems has come to a level where voice quality itself is close to human level, interesting and challenging new tasks are being undertaken, like synthesizing speech for an entire audio book or in a multi-turn conversation. The experiments presented here suggest that, as these new tasks go beyond the scope of traditional TTS, new ways of evaluation should be considered including task based evaluations.

We demonstrated that long-form evaluation can be improved beyond evaluating isolated sentences by showing that different results are obtained when the material is presented in different ways. Asking raters to rate the paragraph as a whole does not give the same results as asking raters to rate the constituent sentences in isolation or asking raters to rate using the previous parts of the paragraph as context. Additionally, we proved that it is difficult and inconclusive to try to predict paragraph MOS from the MOS of the individual sentences in it, which suggests that raters do pay attention to contextual cues when performing these different tasks.

We conclude, therefore, that to fully evaluate long-form paragraphs or dialogues, a combination of tests is necessary. In some circumstances it may be sufficient to only evaluate the paragraphs as a whole, and this is probably what should be done if resources are limited and the paragraphs are not too long. Yet, as observed above, this method gives lower scores than the scores for individual sentences when rated either in isolation or with discourse context. One potential reason is that although our TTS training data consists of multi-sentence data, no significant effort has been made to model paragraph level structures in a TTS system, for example varying the prosody of a sentence based on the content or realization of the previous sentence, and it will be interesting to see if successfully doing so can close the gap between the rating for paragraphs and sentences.

One shortcoming of the three different approaches of evaluating long-form TTS we presented is that they do not consider unbalanced numbers of sentences per paragraph in the data. That is, we have a lot more second sentences in a paragraph than we do fifth sentences in a paragraph. Future work could investigate how to handle unbalanced data in a rigorous way.

Lastly, evaluating sentences in context produced interesting results with higher scores in general, specifically when the context was also TTS: with the same voice in the case of the read news experiments, but also when the context was a different TTS speaker, in case of the conversation experiments. We attribute this effect to the raters including a similarity judgment between the quality of the context and stimulus in their scores. This is corroborated by the experiments with real speech context, which yielded lower ratings. Evaluating in context is therefore our recommended way to evaluate long-form material as it allows sentences to be presented individually, while paragraph effect judgments can be considered in the rating.

8. Acknowledgments

We would like to acknowledge contributions to this work from the wider TTS research community within Google AI and DeepMind, your thirst for understanding lead to this study. Specific thanks to Xingyang Cai, Anna Greenwood, Mateusz Westa, Dina Kelesi and Leilani Kurtak-McDonald for help with evaluation tools and voice building.

9. References

[1] ITU-T P.800.1, "Mean opinion score (MOS) terminology," *International Telecommunication Union*, 2016.

[2] V. J. van Heuven and R. van Bezooijen, "Quality evaluation of synthesized speech." in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995.

[3] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of text and speech systems*. Springer, 2007.

[4] M. Viswanathan and M. Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

[5] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Interspeech*, 2015.

[6] S. Shirali-Shahreza and G. Penn, "MOS Naturalness and the Quest for Human-Like Speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[7] J. Mendelson and M. P. Aylett, "Beyond the Listening Test: An Interactive Approach to TTS Evaluation," in *Interspeech*, 2017.

[8] N. Hu, P. Shao, Y. Zu, Z. Wang, W. Huang, and S. Wang, "Discourse prosody and its application to speech synthesis," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016.

[9] J. Latorre, K. Yanagisawa, V. Wan, B. Kolluru, and M. J. Gales, "Speech intonation for TTS: Study on evaluation methodology," in *Interspeech*, 2014.

[10] ITU-T Rec. P.85, "A method for subjective performance assessment of the quality of speech voice output devices," *International Telecommunication Union*, 1985.

[11] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An Evaluation Protocol for the Subjective Assessment of Text-to-speech in Audiobook Reading Tasks," in *Proceedings of the Blizzard Challenge Workshop (ISCA)*, 2011.

[12] ITU-R BS. 1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[13] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 3915–3923.

[14] V. Wan, C. Chan, T. Kenter, J. Vit, and R. Clark, "CHiVE: Varying Prosody in Speech Synthesis with a Linguistically Driven Dynamic Hierarchical Conditional Variational Network," in *International Conference on Machine Learning*, 2019, pp. 3331–3340.