Location, location

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm

Elijah Gutierrez\(^1\), Pilar Oplustil-Gallegos\(^2\), Catherine Lai\(^1,2\)

\(^1\)Linguistics and English Language, University of Edinburgh, United Kingdom
\(^2\)The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

\texttt{s1740779@sms.ed.ac.uk, p.s.oplustil-gallegos@sms.ed.ac.uk, c.lai@ed.ac.uk}

\begin{abstract}
Text-to-Speech synthesis systems are generally evaluated using Mean Opinion Score (MOS) tests, where listeners score samples of synthetic speech on a Likert scale. A major drawback of MOS tests is that they only offer a general measure of overall quality—i.e., the naturalness of an utterance—and so cannot tell us where exactly synthesis errors occur. This can make evaluation of the appropriateness of prosodic variation within utterances inconclusive. To address this, we propose a novel evaluation method based on the Rapid Prosody Transcription paradigm. This allows listeners to mark the locations of errors in an utterance in real-time, providing a probabilistic representation of the perceptual errors that occur in the synthetic signal. We conduct experiments that confirm that the fine-grained evaluation can be mapped to system rankings of standard MOS tests, but the error marking gives a much more comprehensive assessment of synthesized prosody. In particular, for standard audiobook test set samples, we see that error marks consistently cluster around words at major prosodic boundaries indicated by punctuation. However, for question-answer based stimuli, where we control information structure, we see differences emerge in the ability of neural TTS systems to generate context-appropriate prosodic prominence.

\textbf{Index Terms}: Speech Synthesis, TTS, TTS Evaluation, MOS, Prosody, Rapid Prosody Transcription, Speech Perception
\end{abstract}

\section{1. Introduction}

Modern text-to-speech (TTS) systems have attained a level of naturalness that is approaching human parity for isolated utterances \cite{11}. This progress is in large part due to the rise of neural network based machine learning methods, which have drastically improved the overall quality of synthetic speech and enabled researchers to focus more attention on generating natural sounding prosodic variation. In recent years, there has been substantial research on achieving fine-grained control over synthetic prosody \cite{8,9,4}. New prosodic control mechanisms have allowed TTS systems to produce more variable and expressive speech \cite{5}. However, there has been relatively little work determining whether the prosody that is assigned to an utterance is actually licensed by a given context \cite{6,7}, and it is not clear whether current subjective evaluation methods, such as Mean Opinion Score (MOS) tests, provide enough information to determine the contextual appropriateness \cite{8}.

The appropriateness of utterance prosody—which broadly includes pitch, energy, timing and other suprasegmental characteristics of speech—can vary greatly depending on context. In fact, prosodic differences can help disambiguate many aspects of discourse and dialogue structure \cite{2,10,11,12,13}. Many studies have also shown the close relationship between context-induced expectations about the prosodic form of utterances and information structural notions like newness and givenness \cite{14,15,16}. Incorporating discourse relations has been shown to improve the perceived naturalness of synthesized speech \cite{17}, while incorporating information structure into generated speech has been shown to improve naturalness of automated task oriented dialogues \cite{18}. As neural TTS models continue to improve in their ability to generate variable prosody, it is important to note that not all variation is appropriate in all contexts and increased variation within an utterance is not always perceived as natural \cite{2}.

In order to evaluate how and where TTS systems are really improving in terms of prosody, we need methods that give us a clearer view of what sort of prosodic patterns they generate, and how their appropriateness changes with context. To do this, we propose a new evaluation method that augments traditional MOS-based listening tests with finer-grained error annotations. Specifically, we draw on the Rapid Prosody Transcription (RPT) framework \cite{19,20} to obtain information about the location of perceived errors in the prosody of synthesized speech. In RPT, non-expert listeners mark the presence of prosodic phenomena (e.g. prominence or boundary placement) in real time. This approach allows us to more precisely identify contextual/linguistic sources of prosodic errors.

Much of the current work on TTS in context has focused on monologue or narrative style generation, where information structural relationships are generally unclear \cite{6,7}, and prosodic expectations may not be strong. To address this, we created a schema for generating question-answer pairs with well defined information structure, which in turn project clear prosodic expectations for synthesized answers. Combined with word-level error annotation, this allows us to identify cases of contextually inappropriate prosodic variation.

In the following, we show that there is a strong negative correlation between measures based on error marking and MOS, from which we can recreate MOS based system rankings. Moreover, our question-answer stimuli can be used to induce stronger expectations about prosody than classic audiobook style test utterances, and so better highlights differences in the system prosodies. In general, inspection of the distribution of errors across systems for specific stimuli can lead to better understanding of the sources of system differences, which may otherwise be obscured by MOS alone.

\section{2. Background}

TTS researchers have developed a wide range of methods to evaluate the quality of synthetic speech \cite{8}. However, subjective methods are still considered to be the gold standard in TTS evaluation. These generally involve asking listeners to rate speech samples on a specified dimension, usually naturalness...
3. Experimental Setup

3.1. Experiments and Hypotheses

We perform three listening tests to probe the usefulness of our proposed evaluation method. Experiment condition 1 (E1) is a standard MOS test, while Experiment condition 2 (E2) is a MOS test augmented with the RPT-based error marking task. We compare the results of these two tests to see whether orienting the evaluation to prosody and adding the error marking task affects MOS results.

In E1 and E2 listeners rate single utterances taken from the widely used LibriTTS test set [28]. In Experiment condition 3 (E3), listeners completed the augmented MOS test on question-answer stimuli designed to evoke specific information structural expectations. The goal here was to determine if the error marking would bring out listener expectations about utterance prosody, and hence allow us to distinguish between the appropriateness of prosodic renditions more precisely. This also allows us compare the types of error marks between the two types of test data (audiobook vs dialogue).

3.2. TTS systems

In each of the 3 listening tests, we compare three TTS systems: the Festival [29], Ophelia [30], and FastPitch [31].

Festival is a standard toolkit for building synthetic voices with unit selection. For these experiments, the ‘SLT’ voice distributed by FestVox was used, i.e. a female voice with a General American accent, built from the Arctic A corpus. We note that this voice is far from the current state-of-the-art in TTS, and so we use it as a baseline to see if listeners would ignore other signal naturalness issues when asked to attend to prosodic errors.

We use two neural TTS models as representative of the current state of the art in TTS. These were both trained on the Linda Johnson (LJ) Speech dataset [32], which consists of 13,100 recorded utterances from 7 non-fiction books. Ophelia models were trained using the default recipe (500 epochs for Text2Mel, 250 epoch for SSRN). FastPitch stimuli were synthesised using character (rather than phone) inputs via a pre-trained sequence-to-sequence model that was trained for 1000 epochs.

3.3. Stimuli

For E1 and E2, 30 sentences were sampled randomly from the evaluation set of the LibriTTS corpus [28], a popular audio-book corpus specially designed for TTS research. The maximum stimulus length was controlled to be 15 words to mitigate listener boredom and fatigue.

For E3, contexts and stimuli were generated in a similar manner to those used by [10] for their study of the acoustic correlates of information structure. We used a template-based approach, involving simple Subject Verb Object sentences, generating two types of question-answer pairs:

- Informational Focus: SVO
e.g., Q: What did Mary eat?
A: Mary ate the cake.
- Corrective Focus: No, SVO
e.g., Q: Did Mary buy the cookies?
A: No, John bought the cookies.

Questions were generated to change which constituent represented the new information/correction in the answer, which in English determines the appropriate prominence placement in the response stimuli. We created 10 stimuli per prominence position. Since there were two stimulus structures, this resulted in 10*3*2=60 stimuli in total.

3.4. Evaluation Tasks

The experiments were designed and distributed remotely using a customized version of the Language Markup and Experimental Design Software (LMEDES) [33]. Each stimulus was presented on its own page as follows.

For the standard MOS test (E1), a transcript of the audio stimulus was presented with a ‘Play’ button. Participants were asked to answer the question ‘How natural does the speaker sound?’ on a 5-point Likert scale via a scale slider (MOS).
3.5. Participants and Groups

English-speaking participants were recruited with the crowd sourcing platform Prolific Academic\(^1\). Each participant was paid £2 for their participation in the study. Participants were assigned a random group via Prolific and directed to a listening test based on a Latin square design for each evaluation condition (E1, E2: 3 groups of 10, E3: 6 groups of 10). Participants evaluated stimuli from every system, but did not evaluate the same stimulus more than once.

After consenting to participate in the study, participants were instructed to wear headphones for best audio quality, to ensure they had a stable connection to the server, and to focus their attention on the evaluation task. A brief explanation of what was meant by intonation was also given for E2 and E3. Each participant rated 30 stimuli via the LMEDS interface described above. The standard MOS (E1) test took 8 minutes to calculate the error marking rate (number errors/number of words) per stimuli and participant. Figure 1 shows the mean Likert scale ratings per stimuli for the three experimental conditions. For E1 this is the classic ‘naturality’ (MOS), while for E2 and E3 this is a prosodic naturalness test (PMOS). Comparing the results for E1 and E2 (LibriTTS), we see that the MOS and PMOS scores show the same overall ranking of systems. However, the absolute difference between the system means is reduced in E2, with a marked increase for the scores for Festival and Ophelia.

Table 1 shows the means and Interquartile ranges (IQR) for the 3 tests (IQR is reported as a measure of dispersion for consistency instead of standard deviation as system distributions were skewed). The overall mean MOS is significantly different for all systems in E1 (paired t-test, \(p < 0.01\) with Bonferroni correction), resulting in the ranking Fastpitch > Ophelia > Festival. However, in the PMOS conditions (E2, E3), the difference between FastPitch and Ophelia is no longer significant at the same level (i.e., \(p > 0.01\)). Ratings of Ophelia-produced stimuli were the most variable for all conditions, with the greatest dispersion shown for the question-answer condition.

To see how the error marking task relates to PMOS, we calculated the error marking rate (number errors/number of words) per stimuli and participant. Figure 2 shows the distribution of the mean error rate per stimuli (shown as 1-mean error rate to mirror PMOS ranking). We see that the overall system rankings are the same as that shown in Figure 1 for PMOS. Unsurprisingly, the correlation between stimulus PMOS and error rate is strongly negative when we pool data across all conditions (Pearson’s \(R = -0.75\)). All differences in mean error rate are significant (paired t-tests, \(p < 0.01\), Bonferroni correction) except between Ophelia and FastPitch in E2, i.e., when we look at the word level errors in the question-answer stimuli Ophelia performs significantly better than FastPitch. This indicates that the fine-grained evaluation has better ability to differentiate the
tions with no error marks (\(\alpha\)). We also calculated agreement per stimulus discarding annotations with no error marks (\(\alpha_p\)). We also count the number of participants who marked any error in a stimulus (\(N_p\), max 10).

System prosody when prosodic expectations are designed to be stronger (E3), but this difference may not be apparent for classic narrative style test sets.

Figure 3 shows the distribution of error types selected per experimental condition. This again supports the idea that prosodic expectations had a larger role when evaluating question-answer pairs. Overall, we see a lower number of error types selected for the LibriTTS set than the question-answer set. In particular, participants seemed less likely to detect abrupt pitch changes in FastPitch compared to Ophelia in the LibriTTS stimuli. However, we see a marked increase in unexpected intonation errors for the FastPitch question-answer set. In contrast, FastPitch garnered slightly less ‘lacking intonation’ errors than Ophelia for the question-answer stimuli. This suggests that issues with FastPitch came from an excess of prosodic variation, which was more salient for the question-answer set.

We originally expected that the question-answer pairs would lead to greater interannotator agreement on error locations compared to the LibriTTS data due to stronger prosodic expectations. To investigate this, Krippendorff’s \(\alpha\) was calculated across the annotations for each stimulus in E2 and E3. We used the coding error=1, no error=0 for the annotation. This ensured that an additional ‘word’ to each annotation marked 1 if the participant marked no other errors, and 0 otherwise. This ensured that ‘no error’ annotations would be counted as agreeing. To see more clearly if errors tended to be marked on the same words, ‘no error’ annotations would be counted as agreeing. To see 'no error' annotations would be counted as agreeing. To see whether this occurred more generally, we calculated the proportion of times that the most error marked word in a stimulus preceded punctuation. The results in Table 3 indicate that punctuation produces specific prosodic expectations which were violated by the high level of prosodic variability (i.e., expressiveness) of FastPitch.

A benefit of the error annotation is that we can visualize the distribution of errors across systems to direct further investigation. For example, Figure 4 shows the error heatmap for a LibriTTS stimulus where FastPitch was rated lower than Ophelia in PMOS. This shows that error markings for FastPitch tended to occur on words attached to punctuation marks. To check whether this occurred more generally, we calculated the proportion of times that the most error marked word per stimulus preceded punctuation. The results in Table 3 indicate that punctuation was a more salient issue for FastPitch than for Ophelia.

Figure 5 shows F0 contours corresponding to the heatmap in Figure 4. Out of the 11 error types checked for the FastPitch version, 5 were for ‘awkward pause’, 2 for ‘abrupt pitch’ and 4 for ‘unexpected prosody’, while for Ophelia 3/5 votes were for ‘lacking intonation’. On the FastPitch version we observe unexpected *H* like pitch accents on ‘honoured,’ and ‘extended’, while the Ophelia rendition has a continuation rise on ‘honoured,’ and a fall to low pitch through ‘extended’. This supports the idea that punctuation produces specific prosodic expectations which were violated by the high level of prosodic variability (i.e., expressiveness) of FastPitch.

Similarily, Figure 6 shows error distributions for a contrastive focus example. Figure 4 indicates the error marks on ‘cupcakes’ in the FastPitch version are due to an unexpected pitch accent: ‘cupcakes’ is given relative to the context question and so should be deaccented. Interestingly, pitch tracking for the Ophelia version fails on ‘cupcakes’ due to issues in the signal quality, resulting in creaky-sounding (i.e., low pitched)
Figure 5: F0 differences for Libri16: FastPitch has unexpected pitch accents on before punctuation.

Figure 6: Error Heatmap (focus16); PMOS: Festival=1.5 Fastpitch=3.2 Ophelia=3.0

voice. This is in line with information structure expectations but still may have reduced overall stimuli PMOS. We can also see that the large pitch excursion on the FastPitch ‘No’ was perceived as an error. While this doesn’t produce an information structural clash, it does present an unexpected level of emphasis without further contextual information to justify it.

5. Discussion

Our results indicate that error markings are consistent with rankings from MOS tests. However, differences between systems changed when participants were primed to focus on prosodic issues rather than naturalness in general. This means that the large lead FastPitch had over Ophelia in the naturalness (E1) is likely due to improvements in speech quality separate to prosody. It appears participants did separate out prosody and other quality issues, even for Festival which exhibited much lower naturalness than our neural TTS models. In fact, the error rate measure was better than PMOS at discriminating system prosody when combined with the question-answer test set, where prosodic expectations are more constrained.

It’s important to note that neither FastPitch or Ophelia take into account preceding context in their generation processes. The lower ranking of FastPitch in the question-answer test is likely due to overly-variable (unexpected) prosody, rather than Ophelia being intrinsically better in context. MOS scores for Ophelia were generally more variable, especially in E3. So, it is likely that Ophelia generates a more typical ‘reading style’ intonation, which works well for some question-answer pairs, but not for others.

Default ‘reading’ intonation can work well for narrative-style (e.g., LibriTTS), but can be problematic when prosodic expectations are stronger, such as in task-oriented dialogues. This motivates more design and use of context sensitive stim-

Figure 7: F0 differences (focus16): FastPitch produces an extra prominence on ‘cupcakes’ (cf. Figure 9)
test materials, and more generally stimuli which induce clear prosodic expectations. This new test set provided evidence for an expectation-driven model of prosody perception in TTS. This highlighted the fact that the high prosodic variability, often associated with expressive TTS, may be perceived as errors when it doesn’t match prosodic expectations induced by the context.

Future work will involve a more detailed study of the acoustic properties of the error markings, and the priming effect of RPT-based error marking on PMOS scores. We would also like to extend this work to evaluate other long-form synthesis, e.g. narrative and conversational TTS, to better understand when contexts admits prosodic variation. We would also like to extend the paradigm to evaluate, for example, speaker intent and speaker stance.

Acknowledgements. This work was supported in part by: ANID, Becas Chile, n° 72190135.

7. References