Location, location

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
s1740779@sms.ed.ac.uk, p.s.oplustil-gallegos@sms.ed.ac.uk, c.lai@ed.ac.uk

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.

Index Terms: Speech Synthesis, TTS, TTS Evaluation, MOS, Prosody, Rapid Prosody Transcription, Speech Perception

1. Introduction

Modern text-to-speech (TTS) systems have attained a level of naturalness that is approaching human parity for isolated utterances [1]. 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 [2, 3, 4]. New prosodic control mechanisms have allowed TTS systems to produce more variable and expressive speech [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 [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 [8].

The appropriatess 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 [9, 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 [14, 15, 16]. Incorporating discourse relations has been shown to improve the perceived naturalness of synthesized speech [17], while incorporating information structure into generated speech has been shown to improve naturalness of automated task oriented dialogues [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 [2].

In order to evalute 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 [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 [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 distibution 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.

2. Background

TTS researchers have developed a wide range of methods to evaluate the quality of synthetic speech [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

(i.e., how 'humanlike' synthesized speech sounds). The most commonly used subjective evaluation type is the Mean Opinion Score (MOS) test: listeners are presented with a synthetic stimulus and asked about their overall impression of it, scoring the stimulus on a (usually 5-point) Likert scale [21].

Some of the advantages of MOS tests are that they are straightforward to set up, they are quicker and less cognitively taxing than ranking tasks like MUSHRA, and MOS test design choices have been well extensively investigated [6, 22]. Recent work has expanded to evaluation beyond the single sentence [6, 7]. However, these studies generally focus on holistic evaluations over multi-utterances segments, rather than the parts of utterances that may change listener perception. Meanwhile, the relationship between linguistic context and sub-utterance prosody has been extensively studied, particularly in English, in terms of information structure [15, 16, 23], i.e. how information is organised in an utterance. This is usually cast in terms of given/new information (similarly topic/focus). These constraints are usually demonstrated using question-answer constructions: For example, 'ALEX ate the brownies' is an appropriate answer to 'Who ate the brownies?' because 'Alex' is the new information, while 'Alex ate the BROWNIES' is infelicitous because 'brownies' is contextually given. Thus, use of stimuli with clear information structure provides a precise way of probing whether prosody is context appropriate or not.

Though Information Structure theory can give us an idea of prosodic expectations, prosody perception is known for high inter-listener variability [19, 20] even with expert training [24]. So, the fact that the RPT framework was specifically developed to capture variability in prosody perception from non-expert listeners makes it a natural choice for exploring the perception of synthesized speech. In RPT, the perceptual salience of a prosodic features (e.g. prominence) is determined by the number of listeners who mark a specific segment (e.g. word) with that feature. Because RPT is a task that is conducted in real-time, responses are more sensitive to subtle local changes in quality than offline ones [21, 25].

While RPT has been extensively used to study perception of prosody in human speech [26], there has been little empirical work investigating within utterance prosody for TTS. The closest related work is Edlund et al.'s Audience Response System [27], where a group of listeners judged a synthetic sample simultaneously, pressing a button whenever they hear 'oddities'. This was used to identify common error regions and find the average response latency of listeners when marking errors. However, the definition of a perceptual error was left (deliberately) unclear. Edlund et al. used a single long-form stimulus of about 3 minutes in length. In contrast, our study compares different types of stimuli across multiple TTS systems and focuses on prosody.

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 audiobook 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 [16] 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 (LMEDS) [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).

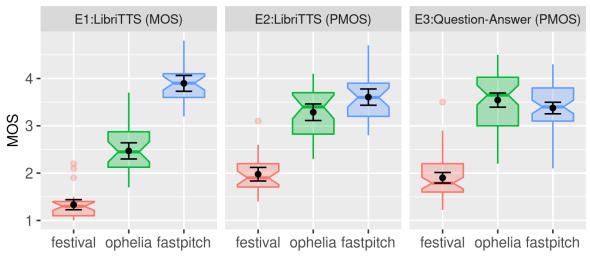

Figure 1: *Distribution of Mean Opinion Scores per experiment (boxplots and means with 95% confidence intervals in black).*

The augmented MOS tests (E2, E3) included the additional RPT-based error marking task, the MOS slider, and a further error type survey. The error marking task appeared first: participants were asked to listen to the stimulus and to click on any words in the transcript where the intonation did not sound correct (possibly none), highlighting them in red. For E3, participants were told to read the context question before marking errors on the answer stimulus. Participants were allowed to replay the stimulus up to 3 times and change their error marks. The MOS slider was positioned after the error marking task, rating *'How natural is the speaker's intonation?'* on a 5-point scale (PMOS). Finally, participants were asked to select which error types they noticed out of: 'Abrupt change in pitch', 'Awkward pause', 'Unexpected intonation' and 'Lacking intonation'. These choices were based on our initial impressions of potentially common errors. Participants also had access to an 'Other' box to enter additional comments or a custom response. In E2 and E3, participants were initially shown 3 examples of stimuli with prosodic errors along with an explanations of why they were considered odd or unnatural. Once this familiarisation phase was complete, participants moved on to the main task. In all three experiments, the audio stimuli were presented in a random order.

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 text 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 complete on average, while the augmented MOS tests (E2, E3) took 15 minutes.[2]

4. Results

Figure 1 shows the distribution of mean Likert scale ratings per stimuli for the three experimental conditions. For E1 this is the classic 'naturalness' (MOS), while for E2 and E3 this is a prosodic naturalness rating (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.

Table 1: *Mean / IQR per stimulus mean MOS*

System	E1 (MOS)	E2 (PMOS)	E3 (PMOS)
Festival	1.33 / 0.30	1.98 / 0.49	1.90 / 0.60
Ophelia	2.47 / *0.75*	3.29 / *0.88*	**3.54** / *1.03*
FastPitch	**3.90** / 0.50	**3.61** / 0.70	3.38 / 0.70

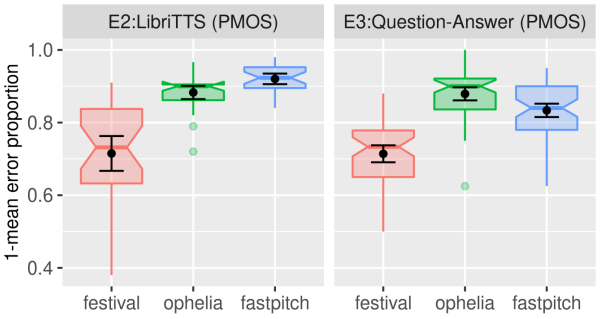

Figure 2: *Distribution of mean error rates (per stimulus).*

These distributional differences indicate that shifting participants focus to prosodic errors changed how they rated the stimuli. This also suggests that lower ratings for Festival and Ophelia in E1 were due to non-prosodic issues. Conversely, the higher ratings for FastPitch are for overall better synthesis quality, but not necessarily for more natural prosodic realization. As we shift to test stimuli with clearer prosodic expectations, the gap between systems in terms of prosodic naturalness is reduced and sometimes reversed relative to what we'd expect given only a standard MOS naturalness test.

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 for question-answer stimuli Ophelia performs significantly better than FastPitch. This indicates that the fine-grained evaluation has better ability to differentiate the

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

[2] Further details/stimuli: http://sweb.inf.ed.ac.uk/clai/tts-rpt

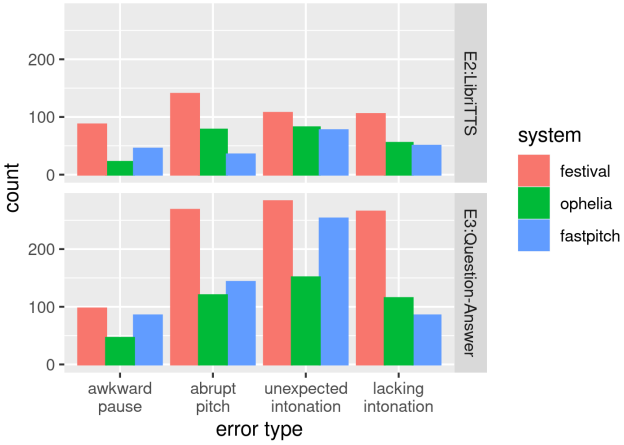

Figure 3: *Counts of error types per system for E2, E3.*

Test set	system	α	α_p	N_p
LibriTTS	festival	0.12	0.18	7.67
	ophelia	0.16	0.26	5.60
	fastpitch	**0.24**	**0.28**	**4.90**
Question-Answer	festival	0.10	**0.24**	6.53
	ophelia	**0.28**	0.18	**3.90**
	fastpitch	0.18	**0.24**	5.20

Table 2: *Mean interannotator agreement: Krippendorff's α, Krippendorff's α restricted to participants that marked at least one error in the stimulus (α_p), the number of participants who marked an 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 α [34] was calculated across the annotations for each stimulus in E2 and E3. We used the coding error=1, no error=0 for the annotation. Since participants could mark no errors in a stimulus, we added 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, we also calculated agreement per stimulus discarding annotations with no error marks (α_p). We also count the number of participants who marked any error in a stimulus (N_p).

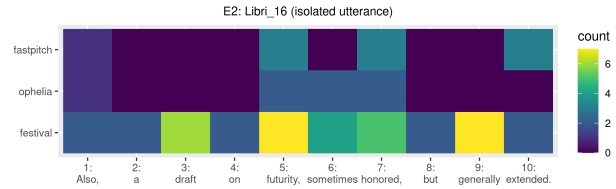

Figure 4: *Error Heatmap (Libri16), PMOS: Festival=1.7, Ophelia=3.6, Fastpitch=2.90.*

System	E2: LibriTTS	E3: Question-Answer
Festival	0.50	0.52
Ophelia	0.67	0.45
FastPitch	0.73	0.60

Table 3: *Proportion of time the most error marked word in a stimulus preceded punctuation. Note, LibriTTS includes much more within utterances punctuation and punctuation variation than the Question-Answer set.*

Agreement statistics are shown in Table 2. The LibriTTS results were as expected: as the lowest quality system, Festival, displays the lowest inter-annotator agreement, while Ophelia and FastPitch exhibit greater agreement for both α and α_p. The mean values for N_p also align with the PMOS ranking. For the question-answer test set, α and N_p also reflects the PMOS ranking. However, α_p, is higher for Festival and FastPitch, indicating that, while there was less agreement on whether there was an error in a stimuli: when participants marked an error they were more likely to choose the same word in the FastPitch and Festival cases. However, we note that both types of α value are still in the low agreement range, so other types of errors likely came into play (cf. Figure 3).

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.

Similarly, Figure 6 shows error distributions for a contrastive focus example. Figure 7 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)

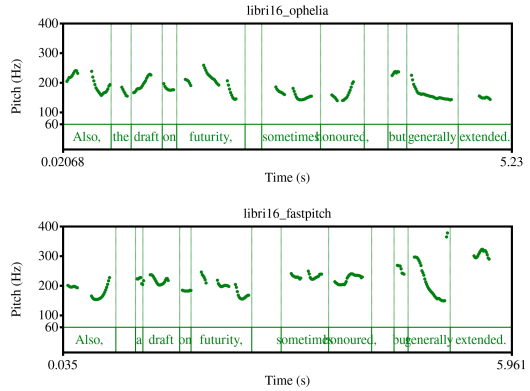

Figure 5: *F0 differences for Libri16: FastPitch has unexpected pitch accents on before punctuation.*

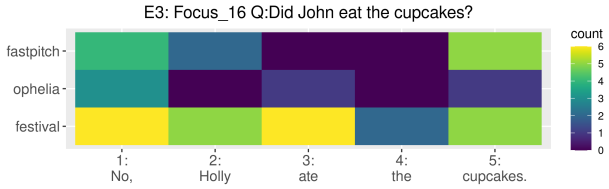

Figure 6: *Error Heatmap (focus16); PMOS: Festival=1.5 Fastpitch=3.2 Ophelia=3.0*

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-

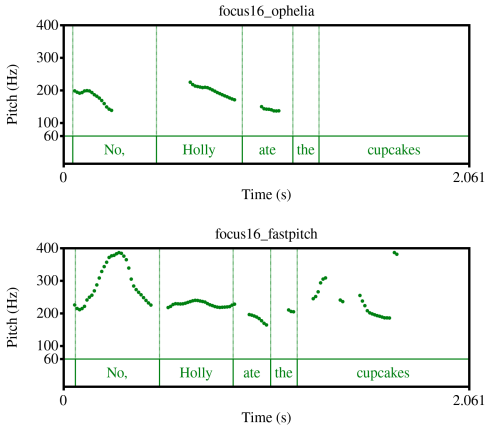

Figure 7: *F0 differences (focus16): FastPitch produces an extra prominence on 'cupcakes' (cf. Figure 6)*

uli, where factors contributing to prosodic expectations are well understood. A key factor for English prosody is information structure, but other factors will likely be important for other languages. In general, there are many paths which might lead listeners to give similar a (P)MOS, especially if the stimuli contains other types of errors (cf. relatively low interannotator agreement). This indicates that simply asking more specific MOS questions isn't enough to pinpoint differences in systems. The results of the current study support the case for more fine-grained error analysis.

The real-time marking used in this study can help TTS researchers and designers understand what non-expert listeners pay attention to when they evaluate and perceive synthetic speech. For example, our study suggests that listeners have specific expectations about what should happen around prosodic boundaries signalled by punctuation. Similarly, the results from the question-answer testset provide evidence for an expectation-driven view of prosody perception [16, 23]. Further analysis of the acoustic properties around error marks will help improve our understanding of these expectations in future work. Similarly, this method may shed light on cases where additional context actually allows for greater prosodic variability than in the isolated case [6].

While we have not done a comprehensive usability study of this method, many participants reported in the post-survey feedback form that they didn't find the experiment tiring and were even entertained by the experiment. This feedback suggests that the methodology is feasible and may help mitigate loss of attention in evaluating long-form TTS [35]. The fact that non-expert listeners can be used for the evaluation means that the methodology is scalable and gives a more realistic account of how a synthetic voice is perceived than a method using expert prosodic labelling.

6. Conclusion

This study introduced a novel evaluation paradigm that augments the standard MOS test with an RPT-based error marking task. Our experiments showed how this fine-grained error marking can uncover differences in systems in prosody generation. We confirmed that our error marking method can be used to distinguish prosodic quality of different TTS systems with a greater degree of precision than MOS-only tests. The experiments highlighted the usefulness of including question-answer

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, nº 72190135.

7. References

[1] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: A discussion and an evaluation," in *Proceedings of ICPhS 2019*, 2019.

[2] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proceedings Interspeech 2015*, 2015.

[3] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0," in *Proceedings of Speech Prosody 2020*, 2020.

[4] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proceedings of ICASSP 2019*. IEEE, 2019.

[5] Z. Hodari, O. Watts, and S. King, "Using generative modelling to produce varied intonation for speech synthesis," in *Proceedings of SSW 2019*, 2019.

[6] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs Rob," in *Proceedings of SSW 2019*, 2019.

[7] 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, China: ISCA, 2020, pp. 4407–4411.

[8] 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 *Proceedings of SSW 2019*, 2019, pp. 105–110.

[9] C. Lai, M. Farrús, and J. D. Moore, "Integrating lexical and prosodic features for automatic paragraph segmentation," *Speech Communication*, vol. 121, pp. 44–57, 2020.

[10] J. Kleinhans, M. Farrús, A. Gravano, J. M. Pérez, C. Lai, and L. Wanner, "Using prosody to classify discourse relations," in *Proceedings of Interspeech 2017*, 2017.

[11] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[12] T. Tran, "Neural models for integrating prosody in spoken language understanding," Ph.D. dissertation, University of Washington, 2020.

[13] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[14] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, 2000.

[15] S. Calhoun, "The centrality of metrical structure in signaling information structure: A probabilistic perspective," *Language*, pp. 1–42, 2010.

[16] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, no. 7, pp. 1044–1098, 2010.

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

[18] M. White, R. A. Clark, and J. D. Moore, "Generating tailored, comparative descriptions with contextually appropriate intonation," *Computational Linguistics*, vol. 36, no. 2, pp. 159–201, 2010.

[19] Y. Mo, J. Cole, and E.-K. Lee, "Naïve listeners' prominence and boundary perception," *Proc. Speech Prosody 2008*, 2008.

[20] J. Cole and S. Shattuck-Hufnagel, "New methods for prosodic transcription: Capturing variability as a source of information," *Laboratory Phonology*, vol. 7, no. 1, 2016.

[21] V. J. van Heuven and R. van Bezooijen, "Quality evaluation of synthesized speech," in *Speech coding and synthesis*, 1995, no. 21, pp. 707–738.

[22] 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 *Proceedings of Interspeech 2015*, 2015.

[23] T. B. Roettger, T. Mahrt, and J. Cole, "Mapping prosody onto meaning–the case of information structure in american english," *Language, Cognition and Neuroscience*, vol. 34, no. 7, pp. 841–860, 2019.

[24] A. K. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of ICSLP 2000*, 2000.

[25] J. V. Ralston, D. B. Pisoni, S. E. Lively, B. G. Greene, and J. W. Mullennix, "Comprehension of synthetic speech produced by rule: Word monitoring and sentence-by-sentence listening times," *Human factors*, vol. 33, no. 4, pp. 471–491, 1991.

[26] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015.

[27] J. Edlund, C. Tånnander, and J. Gustafson, "Audience response system-based assessment for analysis-by-synthesis." in *Proceedings of ICPhS*, 2015.

[28] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *Proc. Interspeech 2019*, 2019.

[29] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

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

[31] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," *arXiv preprint arXiv:2006.06873*, 2020.

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

[33] T. Mahrt, "LMEDS: A Platform for Collecting Prosodic Annotations Online," 2016. [Online]. Available: https://github.com/timmahrt/LMEDS

[34] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011. [Online]. Available: https://repository.upenn.edu/asc_papers/43/

[35] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," in *Proceedings of Interspeech 2018*, 2018.