Back to the Future: Extending the Blizzard Challenge 2013

Sébastien Le Maguer\textsuperscript{1}, Simon King\textsuperscript{2}, Naomi Harte\textsuperscript{1}

\textsuperscript{1}Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
\textsuperscript{2}The Centre for Speech Technology Research, University of Edinburgh, UK
lemages@tcd.ie, Simon.King@ed.ac.uk, nharte@tcd.ie

Abstract

Nowadays, speech synthesis technology is synonymous with the use of Deep Learning. To understand more about how synthesis systems have progressed with the advent of Deep Learning requires open-sourced speech resources that connect past and present technologies. This would allow direct comparisons.

This paper presents such a resource by extending the 2013 edition of the Blizzard Challenge. Using this extension, we compare top-tier systems from the past to modern technologies in a controlled setting. From this edition, we selected the best representative of each historical synthesis technology, to which we added four systems representing combinations of modern acoustic models and neural vocoders. A large scale subjective evaluation was conducted to evaluate naturalness.

Our results show that, as expected, modern technologies generate more natural synthetic speech. However, these systems are still not perceived to be as natural as the human voice. Crucially, we also observed that the Mean Opinion Score (MOS) of the historical systems dropped a full MOS point from their scores in the original edition. This demonstrates the relative nature of MOS: it should generally not be reported as an absolute value despite its origin as an absolute category rating.

Index Terms: speech synthesis evaluation, Blizzard Challenge, reproducibility

1. Introduction

All the submissions to the most recent edition of the Blizzard Challenge\cite{1} were based on Deep Learning. This illustrates how ubiquitous this type of model has become in speech synthesis, which is the result of two breakthroughs in the last 6 years. The first\cite{2} was the development of neural vocoders, capable of rendering a signal almost indistinguishable from human speech. The second\cite{3} was the development of Sequence-To-Sequence (Seq2Seq) architectures which alleviated the need for expertly crafted features.

At the same time, large scale open-source datasets such as LibriTTS\cite{4} or CSS10\cite{5} have been developed. These models. The availability of these corpora, combined with the democratization of open-source implementations (e.g.\cite{6, 7, 8, 9}) of modern speech technologies, has enabled researchers to craft high-quality synthetic speech more easily. However, as a result, much of what was learned from using previous technologies on older datasets has become overshadowed. In order to understand which issues Deep Learning is able to address in comparison to previous technologies, we should not dismiss those past research efforts. In other words, we need anchors which connect past and present technologies.

A first attempt to address this was proposed by Cooper et al.\cite{10}. The primary goal of their study was to produce a dataset for training Deep Learning-based objective quality prediction systems. In order to generate this dataset, the authors conducted a large scale evaluation using samples from multiple editions of the Blizzard Challenge\cite{11} and voice conversion challenge\cite{12}. Because the speakers in these corpora vary between editions, and because multiple genres (e.g. news, novel or conversational) of sentences were evaluated at once, listeners had to assess a highly heterogeneous set of samples. Therefore, it is difficult to pinpoint how the listeners rated a specific sample. In addition to these issues, the measurement used in this study is the Mean Opinion Score (MOS).

The MOS is the result obtained after conducting an Absolute Category Rating (ACR) evaluation whose protocol is described in ITU-T recommendation P.800\cite{13}. Several studies\cite{14, 15, 16, 17, 18, 19} have already addressed the limitations of the MOS for various domains including speech synthesis. In speech synthesis, the main problem of the MOS is that the use of a 5-point Likert scale leads to a compression of the results\cite{16}. Consequently, some important differences are likely to be missed. Unlike the recommendation for the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)\cite{20}, P.800 does not impose the use of a hidden reference or anchors. Although many evaluations of speech synthesis include the natural voice as one of the “systems” being evaluated, MOS differs from MUSHRA in that the listener is not aware of the presence of this reference. Therefore, as\cite{15} points out, an additional problem is the interpretation of the MOS of a synthetic system.

Nonetheless, while research on speech synthesis evaluation remains active\cite{21, 22, 23}, the ACR protocol remains stubbornly predominant in the face of alternatives such as the MUSHRA\cite{20}. The most likely reason behind the popularity of the ACR is convenience: it is easy to implement and is scalable for large subjective evaluation campaigns with several systems to evaluate. Unfortunately, this apparent simplicity also leads researchers to neglect the adequate design of their evaluation as demonstrated by Wester et al.\cite{24} in a meta-analysis of the Interspeech 2014 submissions.

To analyze the evolution of synthetic speech and what each technique improves, we need to develop better approaches. Towards this goal, we have developed a dataset dedicated to the evaluation of synthetic speech, whose core property is maintaining a connection between past and present technologies. The Blizzard Challenge has run every year since 2005 and provides the ideal source of material to build such a dataset. The current study extends the 2013 edition. We selected the top-tier submissions to the original challenge to which we added modern speech synthesis by training four modern open-source systems using Seq2Seq acoustic models and neural vocoders on the same data. We then conducted a subjective evaluation with native English listeners. The resulting dataset, freely available for research purposes\cite{1}, includes the models and the configurations to reproduce them as well as the subjective evaluation results.

\footnote{https://github.com/sigmedia/bc_2013_extension}
In addition to creating this valuable dataset, we use it to answer two critical questions about the evaluation of naturalness:

- How do past top-tier systems compare to present technologies? (Section 3.2 and Section 3.3)
- How reliable are the Mean Opinion Scores associated with synthesis systems? (Section 3.4)

2. Extending the Blizzard Challenge 2013

The Blizzard Challenge organizers have made numerous corpora freely available for research purposes. Among these corpora, the one provided in 2013 [25] has multiple advantages. It is in English and based on audiobooks with a more controlled prosody than the corpora in 2016–2018. This corpus is also the jointly most used (2011 and 2013 each have around 550 licenses issued).

2.1. Training dataset and selected systems

The 2013 edition is composed of 2 tasks for English [25]: 2013-EH1 and 2013-EH2. Both tasks used data from a single speaker: a female American professional narrator and actor. The data comprises the speaker reading audiobooks of various literary genres. The organizers provided an unsegmented dataset of 300h of speech for 2013-EH1, whereas for 2013-EH2 there was a sentence-level segmented subset of about 20 h of speech with corresponding transcriptions. For 2013-EH1, most audio files are only available in MP3 compressed form. For 2013-EH2, the audio files are noncompressed WAV files sampled at 44.1 kHz.

We chose to use 2013-EH2 because uncompressed audio was available and a segmentation into sentences is already provided. The resulting corpus comprises 9733 utterances whose durations vary between 0.5 s and 35.3 s.

2.1.1. Selected Systems

Ten systems were evaluated during the original 2013-EH2 task. These submissions belonged to three speech synthesis families: unit selection; statistical parametric HMM-based; and hybrid.

Our choice of submissions to include in the current work is determined by three criteria: exclude systems already rated with a low MOS; one representative system per family; avoid additional variables in our analysis.

Following these criteria, system K is selected as the representative of the hybrid synthesis family. For HMM synthesis, systems I and C are the available candidates. Given that system C is the HTS baseline for which the implementation is freely available, and was widely used by the community at the time, we select system C as the HMM synthesis system. Finally, for the unit selection family, the candidates are systems L and N. The sampling rate used by system L is 44.1 kHz. But, because the sampling rate used by systems K, C and N is 16 kHz, we exclude system L to satisfy the third criterion. For the same reason, we downsampled the natural speech to 16 kHz.

2.1.2. Test Utterances

Clark et al. [22] showed that the context of the utterances used during an evaluation influences its outcome. During the original listening test in 2013, sentence-sized and paragraph-sized speech samples were evaluated. However, the protocol [26] developed to evaluate the paragraphs is not widely used by the speech synthesis community. Because of this, we restrict our experiments to sentences. Two types of sentence were available: novel sentences taken from audiobooks read by the same speaker, but not in the training set; news sentences extracted from a newspaper.

We are focusing on the evaluation of naturalness of these two types of sentences, and this difference is expected to influence the ratings. Novel sentences are more expressive than news sentences. We do not have samples of the news sentences for the natural voice.

Our extension of the Blizzard Challenge 2013 included seven systems. The established Blizzard Challenge protocol requires that listeners rate only one sample per system for each section. To satisfy this constraint, given that we are evaluating fewer systems than in 2013, we had to select, per section, a subset of 8 sentences among the 11 test sentences previously used. In the analysis section, we therefore only analyze the ratings for the same subset of utterances from the original challenge.

2.2. Extension with Modern Systems

We built four additional systems, covering the combinations of two acoustic modelling architectures with two neural vocoders, to represent modern technologies. For the acoustic models, we used Tacotron [3] as implemented by [9] and FastPitch [27, 28]. Tacotron is the original Sequence-To-Sequence architecture. It uses an attention mechanism both to learn phone-spectrogram alignment during training and to predict duration during inference. FastPitch, on the other hand, is fully-supervised regarding phone-spectrogram alignment during training, and performs explicit duration prediction during inference. Externally-provided F0 provides additional supervision and is likewise explicitly predicted during inference. FastPitch is therefore significantly faster to train, taking one day compared to Tacotron’s three and a half weeks (on a single NVIDIA GeForce RTX 2080 Ti).

As the corpus contains utterances with long pauses which can derail the training of attention, we used guided-attention for Tacotron. This, and FastPitch, both require a phone-spectrogram alignment, which we obtained using Montreal Forced Aligner (MFA) [29].

The neural vocoders we selected are WaveNet [2, 6] and Parallel WaveGAN [30, 7]. WaveNet is the original neural vocoder and is therefore an important milestone in the evolution of speech synthesis. While the quality produced by WaveNet is known to be outstanding, training and inference are slow due to its auto-regressive architecture. Parallel WaveGAN uses the Generative Adversarial Network (GAN) paradigm, with a generator based on components developed for WaveNet. The main difference with WaveNet is that Parallel WaveGAN is not auto-regressive and the loss function is a multi-resolution STFT. This allows Parallel WaveGAN to be trained faster; it is also fast at inference time. We use the following identifiers for each system: FastPitch/WaveNet (F-N), FastPitch/Parallel WaveGAN (F-G), Tacotron/WaveNet (T-N) and Tacotron/Parallel WaveGAN (T-G).

The input to these models is a phonetic sequence, obtained using MaryTTS [31], enriched by appending elementary prosodic information (consistent with what was standard in 2013). For each vowel, we provide the lexical stress of the containing syllable. We associate the punctuation type to each pause label, if available.

When training these systems, operating at 16 kHz is a necessary compromise in order to compare to the selected historical systems. Therefore, we selected these implementations as each of them provides a training recipe for either the Blizzard Chal-
lenge 2013 dataset or cmu-slt-artic which is another standard corpus with waveforms sampled at 16 kHz. In addition, the acoustic parameters are those proposed for Parallel WaveGAN: a mel-spectrogram using 80 filters with cutoff frequencies from 50 Hz for the lower bound and 7600 kHz for the upper bound. We experimented with multiple normalization strategies which have been proposed for training neural vocoders [6, 8, 7, 9]. The results of these experiments showed that using a z-score normalization of the mel-spectrograms led to more stable results for both neural vocoders.

2.3. Recruitment and evaluation protocol

For the current study, we are only concerned with evaluation of naturalness by native listeners. In 2013, these listeners (labelled EE in [25]) were recruited and the evaluation was run in-person at the University of Edinburgh’s perceptual testing laboratory. Here, we recruited listeners via Prolific [32] and they performed the evaluation online. We pre-screened participants to balance their sex and English accent (UK/US). The total duration of the evaluation for each listener was calibrated to take typically 20 minutes and no more than 30 minutes, as recommended by the ITU-T [13]. Each participant was paid £3.75, a rate suggested as good remuneration by Prolific.

We wished to deviate as little as possible from the subjective evaluation protocol used in 2013 to enable meaningful comparisons. Besides evaluating fewer samples, we split the post-test questionnaire used in the original challenge into pre- and post-test questionnaires. For the pre-test questionnaire, the listeners were asked about their age, accent (identified as “dialect” in the questionnaire), and listening conditions. For the post-test questionnaire, the listeners were asked about their familiarity with speech technologies. The questions were the same as those used in the original edition.

3. Results and analysis

3.1. Listeners

68 native speakers of English participated in the listening test. Even though it was explicitly stated that listeners should use headphones, 8 of them indicated that they did not. While their responses are included in the released dataset, we removed them for the analysis presented in this paper.

Of the 60 remaining listeners, 30 are native British English speakers, 29 native US English speakers and one of an unknown accent. This last listener is excluded from the analysis. There-fore, our analysis is based on 59 listeners’ responses of which 28 are from females and 31 from males.

3.2. How do modern systems compare to past top-tier ones?

The overall results are presented in a standard boxplot (Figure 1) and a summary table (Table 1). As observed in [15], comparing the average MOS provides little information about the statistical significance between the systems. Thus as recommended by [15], we conducted pairwise Wilcoxon signed-rank tests between systems’ MOS to test the ordering significance. Table 2 presents the result of this test. In all the figures and tables, we use the suffix “-E” to identify the new results from our listening test. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS.

Figure 1: Listening test results. The yellow ones are those obtained in the 2013 challenge and the green ones with the suffix “-E” are the new results from our listening test. The median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers are represented as circles. The systems are ordered using the mean MOS.

<table>
<thead>
<tr>
<th>System</th>
<th>Mean</th>
<th>Median</th>
<th>Std</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.74</td>
<td>5.00</td>
<td>0.64</td>
<td>118</td>
</tr>
<tr>
<td>A-16</td>
<td>4.39</td>
<td>5.00</td>
<td>0.89</td>
<td>118</td>
</tr>
<tr>
<td>K</td>
<td>3.81</td>
<td>4.00</td>
<td>0.95</td>
<td>347</td>
</tr>
<tr>
<td>T-N</td>
<td>3.69</td>
<td>4.00</td>
<td>1.08</td>
<td>236</td>
</tr>
<tr>
<td>F-N</td>
<td>3.65</td>
<td>4.00</td>
<td>1.11</td>
<td>236</td>
</tr>
<tr>
<td>T-G</td>
<td>3.44</td>
<td>4.00</td>
<td>1.22</td>
<td>236</td>
</tr>
<tr>
<td>F-G</td>
<td>3.30</td>
<td>3.00</td>
<td>1.13</td>
<td>236</td>
</tr>
<tr>
<td>N</td>
<td>3.22</td>
<td>3.00</td>
<td>1.04</td>
<td>387</td>
</tr>
<tr>
<td>C</td>
<td>2.88</td>
<td>3.00</td>
<td>1.01</td>
<td>380</td>
</tr>
<tr>
<td>K-E</td>
<td>2.62</td>
<td>3.00</td>
<td>1.24</td>
<td>236</td>
</tr>
<tr>
<td>N-E</td>
<td>2.00</td>
<td>2.00</td>
<td>1.12</td>
<td>236</td>
</tr>
<tr>
<td>C-E</td>
<td>1.96</td>
<td>2.00</td>
<td>1.13</td>
<td>236</td>
</tr>
</tbody>
</table>

Table 1: Mean Opinion Score for Naturalness. MAD stands for Mean Absolute Deviation; STD stands for Standard Deviation. Highlighted rows are the results from the original challenge. The suffix “-E” are the new results of the systems submitted in the 2013 challenge from our listening test.

The Seq2Seq systems with median MOS varying from 4 for Tacotron/WaveNet to 3 for FastPitch/WaveGAN. Finally, the historical systems are rated as the least natural: the median MOS for K, N and C are varying from 3 to 2. The results presented in table 2 confirm the previous ordering with the following additional information: F-G differs significantly from T-N only for the news sentences; systems N and C do not differ significantly from each other.

As expected, modern systems produce more natural-sounding synthetic speech than all 2013 state-of-the-art technologies. The difference between F-G and T-N/F-N is also consistent with the results reported in [27, 30]. Our results also...
show that, despite this improvement, the latest technologies still do not typically generate speech that is considered as natural as recorded human speech.

### 3.3. Investigating the influence of each balanced dimension

Table 2 shows that the pattern is stable across the two types of sentences with the only discrepancy being that, for news sentences, F-G is considered less natural than T-N. This difference is likely due to the absence of samples of news sentences for the natural voice; the natural voice being rated significantly higher for the novel sentences.

As the difference in required expressivity could also play a role, a more dedicated evaluation would be required to explore this. In a similar way, we compared significant differences across the sex of the listeners. The results showed that F-G does not differ significantly from the other Seq2Seq systems and K does not differ significantly from C for male listeners. The data showed this could not be attributed to the listeners’ declared familiarity with Text-To-Speech (TTS). This difference for males is potentially due to additional factors e.g. linguistic skills or other demographic factors, but our listener tests were not designed specifically to investigate gender differences and thus this result should not be over-interpreted [33]. Finally, when comparing the significance patterns regarding the accent of the listeners, we did not observe any differences between British and American listeners.

Our results are consistent with concerns over MOS already raised by several studies [15, 24, 19]. At 60 listeners, we can adequately rank the systems for naturalness, but the scores do not permit further analysis of the differences between systems. Even when controlling the balance of factors such as sex and accent of English, a typical ACR is not sufficient to give insights into how these factors influence perception. Alternative protocols, such as MUSHRA [20] and Prosody-MOS[23], may reveal more but new protocols are clearly needed, if we wish to properly evaluate modern synthesis systems.

### 3.4. How reliable is MOS?

When comparing the results of the submissions from the original 2013 edition, we observe the following. From Table 1, we see that downsampling the natural recorded speech does not significantly impact the perception of its naturalness, which is reassuring: recorded speech should indeed be rated as perfectly natural regardless of its bandwidth (within reasonable limits).

However, from the same table, the most striking point is the drop in MOS (both the average and the median) for all the historical systems, when comparing their scores from the 2013 listening test with the current one. Specifically, system K average MOS drops from 3.81 to 2.62; system N average MOS from 3.22 to 2.00 and system C average MOS from 2.88 to 1.96. Ex- actively the same stimuli were presented to listeners in both listening tests, yet the MOS scores given by listeners to each and every one of these systems dropped by a full point.

Despite this drop in the absolute scores, the ranking of systems is maintained. System K is still considered as more natural than C and N with these two systems no longer found to be significantly different. In other words, the scores have been compressed. This is likely to be the consequence of the presence of the neural voices in this listening test. It might also be due to a change in the listener population, who are likely to be more exposed to higher-quality synthetic speech than the 2013 listeners.

The fact remains that MOS is not absolute. Researchers should never report an isolated MOS out of context and must be cautious about their conclusions when using this metric.

Our results have immediate implications for MOS prediction models. Their central tenet is that MOS is an absolute category rating. We have shown that MOS is highly sensitive to which systems are included in a particular listening test: it is in fact partially or perhaps entirely a relative score. It is certainly not valid to simply pool MOS obtained from multiple listening tests, to form a training set for a MOS prediction model.

### 4. Conclusions

By extending the Blizzard Challenge 2013, we created a dataset connecting past and present synthesis technologies. This dataset uses open-source tools and other resources that are free for research purposes. Using this dataset, we investigated the evolution of MOS. As expected, we showed that modern synthesis technologies are perceived as more natural. Yet we also found that the natural voice remains distinguishable from synthetic voices. Crucially, we observed that the MOS of the historical systems dropped a full point compared to the original 2013 listening test. This demonstrates the relative nature of MOS despite its origin as an absolute category rating. An ACR is deceptively easy to conduct: the difficulty lies in the interpretation and reliability of the results.

Further replication studies are necessary to determine the extent to which we can rely on the results of the highly subjective evaluation protocols used in our field. This also applies to other protocols such as the MUSHRA [20] even if it is a hybrid rating/ranking task which removes the temptation to quote absolute scores. In the meantime, authors should not draw conclusions about the performance of their system solely based on its MOS value.

### 5. Acknowledgements

The authors would like to thank Rob Clark, Rasmus Dall, Cassandra Valentini-Boitinhao, Erica Cooper, Xing Wang and Junichi Yamagishi for the fruitful discussions on the topic presented in this study. This research was conducted with the financial support of Irish Research Council (IRC) under Grant Agreement No. 208222/15425 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme underGrant No. 13/RC/2106_P2.
6. References


