Naturalness and Intelligibility Monitoring for Text-to-Speech Evaluation

Alex Peiró-Lilja1, Guillermo Cámara1,2, Mireia Farrús3, Jordi Luque2

1TALN Research Group, Universitat Pompeu Fabra, Barcelona, Spain
2Telefónica I+D, Research, Barcelona, Spain
3Language and Computation Centre, UBICS, Universitat de Barcelona, Spain

alex.peiro@upf.edu, guillermo.cambara@upf.edu, mfarrus@ub.edu, jordi.luque@telefonica.com

Abstract

Current text-to-speech (TTS) systems are deep learning-based models capable of learning phonetic articulation and intelligibility, as well as prosodic attributes that model speaking style, providing naturalness to synthetic voices. However, the performance of these models highly depends on their training of hyper-parameters and iterations. Besides, a conventional loss function does not reflect a correct voice modeling; thus, we believe a dedicated training assessment on TTS is needed. To this end, we monitor intelligibility and naturalness during training of Tacotron2 model in a 2-step process. First, we report the analysis of a method to follow up the intelligibility of the TTS in terms of character-token level error rate (TER) by using five different automatic speech recognition (ASR) systems. Second, we extend this work with a recently published TTS naturalness predictor that estimates this aspect in terms of mean opinion scores (MOS). Finally, we unify predicted MOS with TER measurements to return, over each training checkpoint, a single score that we name Full Assessment Score (FAS). We report the relevant preference of our listeners on the checkpoint with maximum FAS rather than the one with minimum validation loss, both in intelligibility and naturalness —up to 62.3% in the latter.

Index Terms: naturalness, prosody, intelligibility, assessment, speech synthesis, automatic speech recognition

1. Introduction

Nowadays, with the use of deep learning (DL) technologies, text-to-speech (TTS) systems are closer than ever to human-like voices. Most of the known architectures are based on a concatenation of sequence-to-sequence (seq2seq) models that map input text to acoustic features [1, 2, 3, 4, 5, 6, 7] followed by a neural vocoder to produce waveforms [8, 9]. Others can already tackle the whole pipeline in a single model [10]. DL-based TTS models are complex enough to catch a wide variety of speech features and speech styles, but they still rely on a common training objective function that does not consider any of the specific aspects of human speech. It has been demonstrated that lower values from such types of loss functions do not necessarily correspond to an improvement of TTS performance [11]. Being aware that the majority of these models require a high number of training steps, if we cannot fully rely on the training and validation loss behaviors, the decision of where to stop training becomes hard and uncertain. For this reason, we believe a dedicated assessment is crucial for the best model selection. In general, we can highlight two main quality aspects of synthesized speech [12]: 1) **Intelligibility**, which defines how well the words of a spoken message are understood, and 2) **Naturalness** of speaker voice, that is to say, the correct use of emphasis, intonation, pitch, intensity, and pauses according to the message and intention [13, 14]. Also other properties such as semantics, comprehension or preference have been defined and studied. Until now, these properties have been quantified by subjective tests with different purposes. The most known is the Mean Opinion Score (MOS), used to evaluate global quality, originally used for phone speech, and later extended to the speech synthesis field. Currently, MOS is widely used to score naturalness, intelligibility, and quality, among other aspects. But also other less frequent tests were designed, such as the diagnostic rhyme test (DRT) with the use of CVC-words (e.g. dune-tune) or the semantically unpredictable sentences (SUS) [15]. However, the main disadvantages of carrying out subjective tests are the human cost and performance time.

In order to cover the necessities and drawbacks commented, our work aims to provide evidence that a complete TTS assessment could be beneficial in obtaining an optimal performance. We present two separated methods that return metrics proportional to the two main aspects of speech: intelligibility and naturalness. The former, related to the correct pronunciation and construction of words, is obtained by averaging the Token Error Rate (TER) measured from five different neural ASR architectures, while the latter is represented by MOS scores predicted by a neural model trained with manually annotated MOS ratings in terms of naturalness from synthetic speech samples. Later, we unify both metrics into a single score, obtaining thus a single curve throughout the whole training. We call this score Full Assessment Score (FAS). Then, we choose the training checkpoint with the highest FAS and the one with the lowest validation loss and observe that, according to the preference test presented, the checkpoint with highest FAS presents a better performance both in intelligibility and naturalness.

The structure of this paper unfolds as follows: Section 2 describes the related work previously published, in Section 3 we describe models and measurements performed for the TTS assessment, Section 4 shows the results observed and the discussion and Section 5 the conclusions and the future work.

2. Related work

In terms of intelligibility assessment, we took inspiration from [16], where the authors used the phone error rate (PER) to select the best training model, using a hybrid hidden Markov model (HMM) with deep neural network (DNN) ASR model from Kaldi Toolkit. The use of an ASR to measure synthetic speech intelligibility comes from years before. In [17] an ASR model was presented as a substitute for the Perceptual Evaluation of Speech Quality (PESQ) in order to evaluate intelligibility without the need of human reference, also such as [18] and [19]. On the other hand, naturalness has been included for years in automatic TTS assessment among other aspects (i.e. effort, overall impression, or fluency) [20]. But more recently, some
deep learning-based models for naturalness assessment have been developed. Motivated by the Voice Conversion Challenge (VCC), MOSNet was presented as MOS human rating predictor [21] presenting a high performance at system-level. VCC has its own large-scale public dataset with manually annotated MOS ratings. A year later, the implementation of a global quality token and an encoding layer to the MOSNet brought an improvement both at system and utterance frame-level [22]. In parallel, another DL-based architecture was retrieved and trained with signal quality data scoring to later use Blizzard and VCC datasets as a transfer learning approach [23]. The original model was presented by the same authors in 2016 [24] as an automatic non-intrusive quality assessment. For this work we used the latter naturalness assessment pre-trained model, which is detailed in the next section.

## 3. TTS assessment

In this section we describe the models and how we proceeded to perform monitoring and assessment.

### 3.1. Resources

#### 3.1.1. Automatic speech recognition systems

State-of-the-art end-to-end (E2E) ASR models are capable of providing high quality transcriptions without the need of language models (LM) [25, 26], which tend to correct defects in the transcriptions caused by any source such as mispronunciations, poor clarity in speech or speech and audio distortions. Although these corrections caused by LMs are an advantage for speech recognition tasks, such effects are not desired when assessing the quality of a synthesized sample. There is an important factor that tends to be uncovered, which is the correct pronunciation of the spoken text or, in other words, the amount of mispronunciations in the utterance. For instance, the words “west” and “waste” sound in a similar manner, but have totally different meanings. A TTS might pronounce “waste” as “west”, even in a natural and intelligible way, but still the utterance would not be correct at all. Thus, this is a variable that is not necessarily correlated with these two main aspects, and also affects the overall quality of the synthesizer. From now on, we will refer to such quality as correctness.

We want the ASR to accurately transcribe synthesized speech when pronunciation is correct and clear, while providing wrong transcriptions whenever the mispronunciations and unclear speech occur, therefore yielding a higher TER denoting worse correctness and/or intelligibility. This balance between accuracy and sensitivity to speech quality is well matched by the mentioned E2E ASR systems without any LM. However, there is still a chance that an ASR wrongly transcribes good quality speech, as well as it is able to infer the correct transcription from a mispronounced or distorted utterance, by using inherently learned linguistic knowledge, even without an LM. For that reason, we use 5 different ASR models for evaluating all the validation samples generated by a TTS checkpoint, averaging the total TERS across such models. This sort of machine inter-annotator agreement is done to regularize the errors across models.

For all 5 ASR models are systems trained with the wav2letter toolkit [27], all of them available at the framework’s recipe repository. These are all Transformer-based models following the proposed architecture in [26], which mainly consists of a convolutional front-end using gated linear units (GLUs) [28], followed by several Transformer blocks [29]. The differences between the models rely on the training dataset, the amount of attention layers, the size of the self-attention dimension or the use of Conformer (CFR) blocks [30] instead of Transformer (TR) ones. The used datasets are the Multilingual LibriSpeech (MLS) database, and a combination of several datasets to cover different domains for robust ASR (RASR), as described in [31]. This combination of datasets is constituted by the following ones: LibriSpeech [32], WSJ [33], Switchboard & Fisher [34, 35], Common Voice [36] and a Facebook’s in-house video dataset [31]. We have publicly released our recipe for TTS evaluation with the 5 ASR models, so it is available for the community.

### 3.1.3. Tacotron2 seq2seq for evaluation

We implemented our modified version of Tacotron2 NVIDIA. The main changes were: (1) the replacement of dropout in the prenet module, as it introduced randomness in the inference, by batch normalization, (2) implementation of a guided attention mask as in [37] to enforce the first steps to pay attention into the diagonal, and (3), in order to improve performance in larger synthesis, we added an extra decoder in the training process to

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Type</th>
<th>Dim</th>
<th>Layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLS_T</td>
<td>MLS</td>
<td>TR</td>
<td>768</td>
<td>36</td>
</tr>
<tr>
<td>RASR_T,70M</td>
<td>RASR</td>
<td>384</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>RASR_T,300M</td>
<td>”</td>
<td>768</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>RASR_C,28M</td>
<td>CFR</td>
<td>256</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>RASR_C,57M</td>
<td>”</td>
<td>512</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

### 3.1.2. MOS prediction model

For this work, we used the publicly available and trained Non-Intrusive Quality Assessment (NISQA) model, presented originally in [24] as a model able to automatically evaluate super-wideband speech quality from communication networks without the need of a clean reference signal. The model is based on a convolutional neural network (CNN) followed by a recurrent neural network (RNN) architecture trained in two stages: first, the CNN to predict per-frame quality/similarity scores, and later, the RNN long short-term memory (LSTM) module to predict the MOS scores from per-frame scores and mel-frequency cepstral coefficients (MFCC). More recently, same authors trained NISQA first with quality assessment scores to later train it with other datasets such as Blizzard and Voice Conversion Challenge (VCC). The model was able to predict naturalness ratings of TTS samples [23], in which we assume prosodic attributes are implicitly scored. It is language independent and, according to the authors, NISQA-TTS model performs better at system level assessment. The model is publicly available for research purposes.

3.1.3. Tacotron2 seq2seq for evaluation

We implemented our modified version of Tacotron2 NVIDIA. The main changes were: (1) the replacement of dropout in the prenet module, as it introduced randomness in the inference, by batch normalization, (2) implementation of a guided attention mask as in [37] to enforce the first steps to pay attention into the diagonal, and (3), in order to improve performance in larger synthesis, we added an extra decoder in the training process to
parallel decode same frames but with a higher step size. This technique is termed by its author Double Decoder Consistency (DDC). In our model, we kept the batch size fixed to 32, learning rate to 1e-3, coarse decoder step was 7 and the fine decoder step 2. We used the LJSpeech dataset for training, splitting the total number of samples (13100) into training (12843), validation (128) and test (129).

3.2. Methodology

3.2.1. Experimental setup

Since the seq2seq part of a neural TTS system is the one that generates all speech aspects on the spectrogram, the assessment was centered on Tacotron2 model. Thus, during its training, we saved a checkpoint of the model every 2 epochs and ran each of them with the test set together with a pre-trained neural vocoder. Regarding to the vocoder election, we tested two recent generative adversarial network (GAN) models, which present remarkable results while performing with very low timings. On the one hand, Multi-Band MelGAN is the GAN-type vocoder model that shows best scoring in subjective tests [38]. On the other hand, the very recent VocGAN presents a better MOS score than Parallel WaveGAN [39]. Finally, we used VocGAN model\(^6\) which is already trained also with LJSpeech dataset.

3.2.2. Intelligibility measurement

Since any element that may cause speech to be hardly understood would directly increase the error rate in an ASR system, we quantify so by measuring the word error rate (WER) plus the token error rate (TER) per sample. Note that since our ASR system outputs graphemes as tokens, our TER score is just the grapheme error rate. If a word is mispronounced, the ASR is likely to yield a transcription different to the ground truth original text, which shall be reflected in a higher TER score. An ASR giving high TER values for a TTS checkpoint gives a clue that speech might be mispronounced or be hardly intelligible. Therefore, lower error rates denote higher speech intelligibility, whereas higher error rates may reflect less intelligible speech. We report results only with TER score, since both TER and WER scores are very correlated and TER contains more fine-grained information, acting at the token level, instead of the word level. TER measurement is computed by comparing the transcription from the ASR system, with the ground truth human transcription. Such score is based on the Levenshtein distance [40], which is the minimum number of edits that needs to be done in a text sequence to match another one. For example, TER score is computed by counting the number of token substitutions (S), deletions (D) and insertions (I) needed to convert the ASR transcription into the ground truth one. Then, these counts are summed up, dividing the result by the number of tokens in the ground truth transcription.

3.2.3. Naturalness prediction

As commented previously, the NISQA model for TTS assessment is more confident at system level. We highlight the limitation of this model, which is the length of the audio sample. Only clips between 5-12 seconds are accepted, so each checkpoint may have different number of accepted samples. Accordingly, we computed an estimated MOS for every trained checkpoint MOS\(_c\), like:

\[
\text{MOS}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} \text{MOS}_n
\]

being \(N_c\), the total number of suitable test samples of the \(c\) model checkpoint, and \(\text{MOS}_n\) the predicted MOS rating of the \(n\) suitable sample of \(c\) checkpoint.

3.2.4. Full Assessment Score

Once both measurements were obtained, we looked for a compact representation of the scores into a single one. However, TER and MOS values move through very different ranges—in % and integers from 1 to 5, respectively. So we normalized both ranges to \([0, 1]\), obtaining hence \(\text{TER}_c[0,1]\) and \(\text{MOS}_c[0,1]\). Then, assuming that an optimal TTS system performs with the highest naturalness and intelligibility, we looked to the checkpoint with the highest predicted MOS and the lowest TER. To fulfill both requirements, we proposed the Full Assessment Score (FAS):

\[
\text{FAS}_c = \text{MOS}_c[0,1] - \text{TER}_c[0,1]
\]

FAS is calculated for every checkpoint, giving to them values ranged between \([-1, 1]\). If the value tended to -1, that would mean the TTS system had a very low MOS and a very high TER rating, and the opposite if it was close to 1. If FAS went close to 0 that would indicate an inverse proportion of naturalness and intelligibility. Therefore, the checkpoint with a FAS closer to 1 would be the best of the training process.

4. Results

Here we check the FAS curve behavior, comment relevant observations and the results of the preference test that we carried on are also presented.

4.1. Observations

The direct comparison between the FAS and validation loss curves is printed in Figure 1. Note that we plotted both curves starting from the training step 10400, which was the point in which Tacotron2 alignment started to perform diagonally. The validation loss curve rapidly goes quite flat but presenting low peaks. However, FAS curve looks more uneven with more oscillation but with a slight upward trend. We marked with red dots the minimum validation (step 147200) and the maximum FAS (step 151200). The maximum FAS obtained was 0.71, while FAS value of checkpoint 147200 was 0.68. Thus, proposed FAS pointed to a checkpoint with few more thousands of training steps than the one with minimum validation loss.

Besides, observing TER separately, we perceived very high peaks appearing throughout its evolution (see Figure 2), even at checkpoint regions where the validation loss was relatively stable in comparison. Such peaks were present for all the 5 ASR systems, giving a clue that a fair amount of speech synthesized by the corresponding checkpoints might be incorrect, unintelligible or corrupted in some manner. We observed the synthesis of the associated checkpoints and compared them against neighbor checkpoints with much lower errors. For instance, mean TER across the 5 ASRs at checkpoint 170400 was 5.4 ± 0.3%, whereas the following checkpoint 171200, the adjacent one, scored a 33.1 ± 0.2% mean TER. Validation loss was only


\[\text{https://github.com/rishikksh20/VocGAN}\]
0.02 points larger in the latter checkpoint, a smaller peak given the numeric ranges of such curve, which can be seen at Figure 2. Many samples observed from the latter checkpoint revealed severe suppression of sentence parts and mispronunciation of words. We have published some of these samples for perceptual evaluation.\footnote{https://bit.ly/3mcTBBa} Just to name an example, sample LJ031-0041 uttered by checkpoint 171200 contains many skipped words. Ground truth is “he noted that the president was bluewhite or ashen in color had slow spasmodic agonal respiration without any coordination”, checkpoint 170400 hypothesis is “he noted that the president was blue terrific” and checkpoint 171200 hypothesis is “he noted that the president was blue white or ashen in car had slow spasmodic aah aah gonal respiration without any coordination” and checkpoint 171200 hypothesis is “he noted that the president was blue terrific”.

Figure 2: High TER peak location compared with validation loss curve behavior.

4.2. Subjective test

To prove whether suggested checkpoint by FAS had a better performance, we carried out a subjective test to observe preferences of the listeners. To do that, we collected 9 out-of-domain sentences from BBC news: 3 short (6-8 words), 3 medium (11-15 words) and 3 large (15-26). Each of these were synthesized by the checkpoint with the lowest validation loss (VAL-C) and the one with the highest FAS score (FAS-C). We asked listeners to choose their preferred sample version based on intelligibility on the one hand, and naturalness on the other hand. We enabled a ‘both’ option in case any of these aspects sounded ambiguous for the listener. In total, we collected answers from 28 subjects. The average results are shown in Table 2, where we can see a higher preference in both aspects to the samples generated by FAS-C. In terms of intelligibility, the majority of the subjects decided between FAS-C or ‘both’, so a low percentage were clearly decided for the VAL-C. There is a more significant preference to FAS-C in terms of naturalness, more than 62% rather than VAL-C or ‘both’. We assume, thus, FAS-C provided more adequate prosodic attributes. Therefore, we also observed perceptually that highest FAS checkpoint model performs better than the one obtained from validation loss.

Table 2: Preference test averaged results.

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>FAS-C</th>
<th>VAL-C</th>
<th>Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligibility</td>
<td>47.6</td>
<td>15.1</td>
<td>37.3</td>
</tr>
<tr>
<td>Naturalness</td>
<td>62.3</td>
<td>23.4</td>
<td>14.3</td>
</tr>
</tbody>
</table>

5. Conclusions and future work

In this work, we have exposed the need of a full speech assessment for the training of any TTS. Minimizing common objective functions does not necessarily reflect an improvement of artificial speech. To show this, we have used equivalent artificial metrics to monitor the two main aspects of speech that any TTS aims to maximize: intelligibility and naturalness. For the former, we robustly measured WER and TER from 5 different Transformer-based ASR models. And for the latter, we used the recently trained NISQA model to predict MOS ratings. Then, we proposed a metric termed Full Assessment Score (FAS), which combines averaged TER with predicted MOS and made the measurements over the test set at every training checkpoint. After following up the Tacotron2 training, we observed that minimum validation loss and maximum FAS were located in different checkpoints. So we performed a subjective test which resulted in a clear tendency to the FAS checkpoint in both aspects. Moreover, we observed high TER peaks between consecutive checkpoints where loss validation presented an irrelevant difference. This work motivates us to continue in our research on improving TTS speech assessment for a better model performance. We want to extend this work to the loss function of TTS models, towards more dedicated objective functions or the use of ASRs to help TTS systems to train optimally.

6. Acknowledgements

This work is a part of the INGENIOUS project, funded by the European Union’s Horizon 2020 Research and Innovation Programme and the Korean Government under Grant Agreement No 833435. The third author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE). This work has been carried out using an NVIDIA GPU Titan Xp generously provided by NVIDIA Company.

7. References


