INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Detection and analysis of attention errors in sequence-to-sequence text-to-speech

Cassia Valentini-Botinhao, Simon King

The Centre for Speech Technology Research, The University of Edinburgh, UK
cvbotinh@ed.ac.uk, simon.king@ed.ac.uk

Abstract

Sequence-to-sequence speech synthesis models are notorious for gross errors such as skipping and repetition, commonly associated with failures in the attention mechanism. While a lot has been done to improve attention and decrease errors, this paper focuses instead on automatic error detection and analysis. We evaluated three objective metrics against error detection scores collected by human listening. All metrics were derived from the synthesised attention matrix alone and do not require a reference signal, relying on the expectation that errors occur when attention is dispersed or insufficient. Using one of this metrics as an analysis tool, we observed that gross errors are more likely to occur in longer sentences and in sentences with punctuation marks that indicate pause or break. We also found that mechanisms such as forcibly incremented attention have the potential for decreasing gross errors but to the detriment of naturalness. The results of the error detection evaluation revealed that two of the evaluated metrics were able to detect errors with a relatively high success rate, obtaining F-scores of up to 0.89 and 0.96.

Index Terms: speech synthesis, attention, sequence-to-sequence modelling

1. Introduction

Many sequence-to-sequence models rely on an attention mechanism in order to learn to associate segments of the output sequence to segments of the input sequence. Sequence-to-sequence systems like DCTTS [1] significantly outperform frame-level systems like Merlin [2] in terms of naturalness and quality [3]. However, unlike Merlin, a frame-level system that has an explicit duration model, attention-based models are prone to so-called gross errors, commonly associated with failures in the attention mechanism. Such errors include skipping and repetition of phones or words, muffling and early stop (a special case of skipping). The same model might produce gross pronunciation errors for some sentences and yet synthesise other sentences with human-like naturalness.

Several types of attention mechanisms and workarounds have been proposed to mitigate such errors. TTS systems like DCTTS [1] employ the so-called dot-product attention mechanism [4] where the attention matrix is calculated as the multiplication between a key (derived from a text encoder) and a query vector (derived from an audio encoder). Dot-product mechanisms are computationally efficient but do not take into account location information: the product between the two vectors is the same independent of the location of the segment in the input sequence, making it harder for the system to learn to move along the input sequence. To stabilise attention and encourage a monotonic behaviour an additional loss component can be added into training. The so-called guided attention loss, the expected value of the attention matrix weighed by a diagonal guide, encourages the model to generate matrices whose non-zero components are closer to the diagonal. Another way to encourage monotonicity is to add positional information to the key and query vectors as done in Deep Voice 3 [5]. More complex types of attention mechanism like the additive model used in Tacotron [6, 7] and the GMM based [8] used in Char2wav [9] and VoiceLoop [10] are able to incorporate location information in different ways with mechanisms such as location sensitive attention [11] and the dynamic convolution and GMM-based methods proposed in [12]. Such mechanisms are however computationally expensive and have also been shown to benefit from an additional guided attention loss [13].

In this paper, rather than proposing new ways of stabilising attention, we focus on understanding when errors are more like to occur and how to automatically detect errors. Automatic error detection is a useful tool to reduce human listening effort when choosing which model to store while training, comparing systems and producing samples for a listening test. We focus our analysis on the DCTTS system as it can generate the large amount of sentences we required in a timely manner.

2. Attention errors and metrics

The most common errors associated with attention failures are skipping, repetition and the so-called "muffling". In Fig. 1 we present the attention matrix of three different sentences generated by the DCTTS text-to-speech model, each containing one of these errors. In the left most image we can see the attention matrix of a sentence with a skipped segment and an early stop (incomplete sentence). These are obvious from the clear discontinuity and the abrupt end of the "attention path". The image in the middle shows the attention matrix of a sentence with a repeated segment, where a different kind of discontinuity appears. Rather than attending to the current part of sentence the attention path "jumps" to a later encoder step, that is synthesised twice. This is marked by a stronger competing path. The right most image, that displays the attention matrix of a sentence with a "muffling" error, shows a similar pattern, but this time the competing path is not as pronounced. The "divided" attention between two segments of the text is what leads to the muffled quality.

These examples show that it is possible to identify that a synthesis error occurred by visually inspecting the attention matrix. The authors in [14] proposed a metrics for output translation confidence based on attention distributions. This metric is based on three separate metrics: the coverage deviation penalty (CDP) and the absentmindedness penalties (Ain and Aout).

The CDP is calculated per phone and averaged across phones. The metric increases when too much or too little attention is given to a particular phone. It should be able to detect skipping (and early stop) as well as repetition (too much attention). From the original definition in [14] we dropped the minus

http://dx.doi.org/10.21437/Interspeech.2021-286

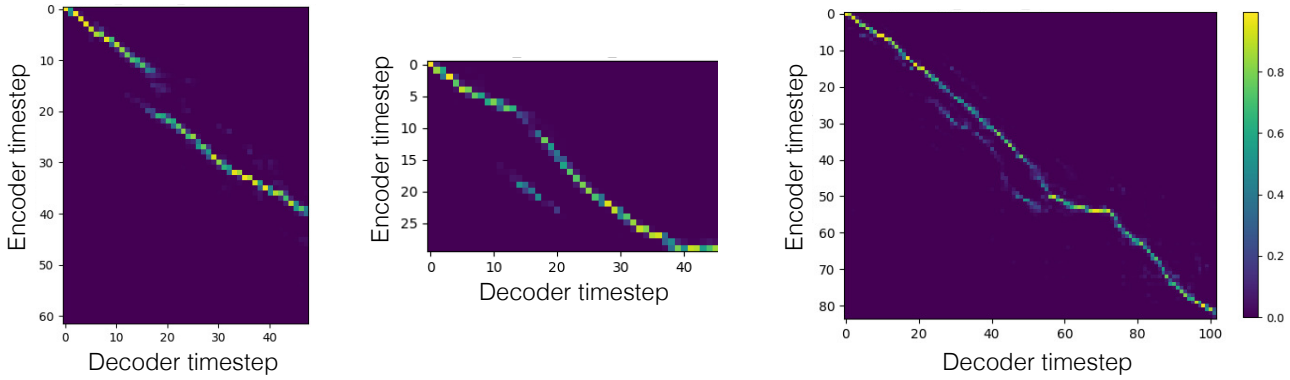

Figure 1: *Example attention matrix for skipping and early stop (left), repetition (middle) and "muffling" (right).*

sign so CDP is always a positive value:

$$CDP = \frac{1}{J}\sum_{j}\log(1+(1-\sum_{i}\alpha_{ij})^2) \quad (1)$$

where α is an element in the attention matrix, j indexes encoder steps (phones or letters) and i decoder steps (acoustic frames) and J the total number of encoded steps.

The other remaining metrics, the absentmindedness penalties, targets scattered attention either per input (Ain) or per output token (Aout). Dispersion is measured as the entropy of the predicted distribution. Ain and Aout increase when attention is more scattered and should be able to detect multiple modes (repetition and muffling).

Ain is then calculated as the entropy of the attention distribution per phone and averaged across phones:

$$Ain = -\frac{1}{J}\sum_{j}\sum_{i}\hat{\alpha}_{ij}\log\hat{\alpha}_{ij} \quad (2)$$

$$\hat{\alpha}_{ij} = \frac{\alpha_{ij}}{\sum_i \alpha_{ij}} \quad (3)$$

Aout is calculated as the entropy of the attention distribution taken along the encoded steps, calculated per frame and averaged across frames:

$$Aout = -\frac{1}{I}\sum_{i}\sum_{j}\hat{\alpha}_{ij}\log\hat{\alpha}_{ij} \quad (4)$$

$$\hat{\alpha}_{ij} = \frac{\alpha_{ij}}{\sum_j \alpha_{ij}} \quad (5)$$

where I is the total number of decoder steps (in our case, acoustic frames).

None of these measures require a reference natural speech signal and are very fast to calculate. They could be used during training as additional losses or as part of generation for monitoring purposes.

3. TTS system

DCTTS is a sequence-to-sequence model that synthesises frame-wise Mel spectrograms directly from text. DCTTS's acoustic model generates Mel spectogram features at a 50 ms frame shift. Then, a spectrogram super-resolution network, converts this coarse representation into a higher resolution linear scale spectrogram which is passed to a waveform generator for synthesis, in our case the phase reconstruction method Griffin-Lim [15]. DCTTS is fully convolutional and for that reason trains and generates faster than models that contain recurrent units.

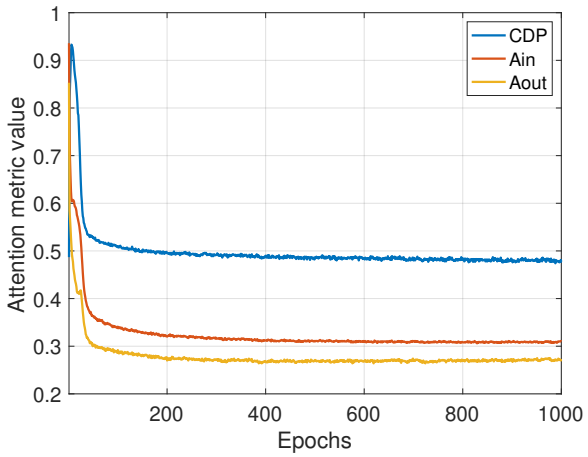

Figure 2: *CDP, Ain and Aout scores calculated using the attention matrices generated during training, averaged across all training sentences.*

DCTTS relies on a dot-product attention mechanism and an additional guided attention loss. At synthesis time it employs the forcibly incremented attention (FIA) mechanism to force attention to focus on neighbouring encoder inputs only. This is done by applying a mask to the generated attention matrix before the softmax operation. This mask is updated at every decoder step according to the current phone position.

The results reported in this paper were obtained from a DCTTS model trained with data from the Jane Eyre audiobook (14 hours) from the Blizzard 2013 dataset [16]. The acoustic model was trained for 1000 epochs using Ophelia [1]. Fig.2 shows the average CDP, Ain and Aout calculated from the training data for each trained epoch. Even though these metrics were not explicitly minimised during model training they do decrease over epochs, a possible side effect of minimising the guided attention loss.

4. Evaluation

In this section we evaluate how well these three metrics are able to detect errors. We do not evaluate whether these metrics can

[1] Code and samples: https://github.com/CSTR-Edinburgh/ophelia

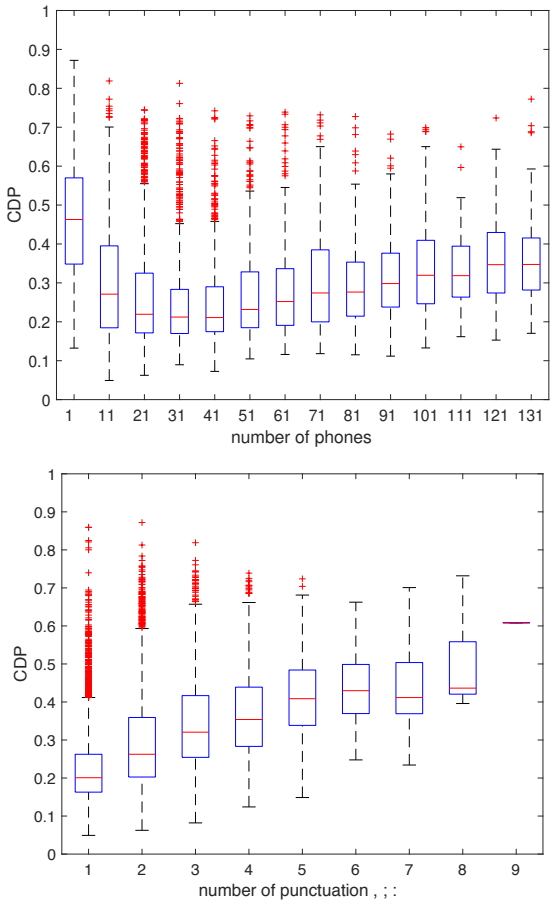

Figure 3: *Boxplot distribution of CDP scores over sentences grouped in terms of number of phones (top) and number of punctuation marks (bottom).*

tell where the error occurred or which kind of error occurred. We focus instead on whether they are able to detect if an error occurred in a sentence at all.

To evaluate this we annotated a large dataset of synthesised speech by asking human listeners to detect whether a synthesised sentence contained an error given the expected transcription. To create the stimuli for this we first performed an error analysis to identify when gross errors are more likely to occur.

4.1. Error analysis

Sequence-to-sequence models produce gross errors but they do not occur very frequently, particularly if a model has converged during training. In order to obtain a substantial amount of errors one would have to synthesise a large set material.

To simplify this task and understand when errors are more like to occur we used the CDP as an automatic analysis tool. We synthesised close to 8,000 sentences derived from a book ("Far from the Madding Crowd") and calculated the CDP scores of each sentence from the generated attention matrix.

4.1.1. Sentence structure

The boxplot in Fig. 3 shows the distribution of the CDP values grouped in terms of number of phones and number of pause related punctuation marks. We identify that errors occur more often (higher CDP values) with sentences that contain between 20 to 60 phones and one or two of pause related punctuation marks (, ; :).

The fact that we observed higher CDP values with increasing number of pause related punctuation marks could indicate a failure of the trained model to learn to insert pauses, a notorious issue of sequence-to-sequence models.

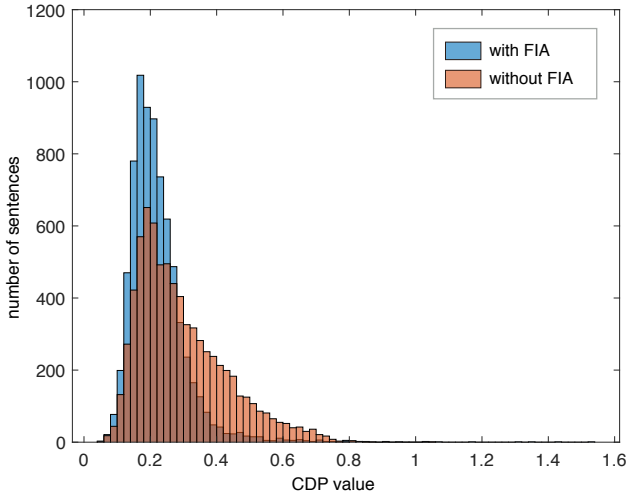

Figure 4: *Distribution of CDP scores over sentences. The two distributions refer to sentences synthesised with FIA (blue) and without FIA (orange)*

4.1.2. Forcibly incremented attention

We also noticed that higher CDP values are more likely to occur when forcibly incremented attention is turned off, as shown by the boxplot in Fig. 4.

Forced incremented attention has a substantial effect on the CDP values as it forces the attention matrix to focus on neighbouring phones, limiting coverage and dispersion. FIA has however a negative effect on naturalness. To test this and find whether FIA is necessary when no gross errors occur we performed an online listening test.

Listeners were asked to choose which sample was more natural (a sample synthesised with FIA and one synthesised without, both synthesised from the same model). Each test covered 100 sentence pairs: 10 focus checks pairs (TTS x vocoded) and 90 test pairs. The order of samples in the pair and the order of sentences was randomised. The focus checks were used for participant removal. The sentences were selected from the Harvard set [17] . These contained between 21 to 52 phones and no pause related punctuation marks (CDP$<$ 0.42, Ain$<$ 0.22). None of the synthesised sentences presented gross errors (with our without FIA).

We analysed the results of 49 participants (11 participants were excluded out of the 60 that took the test). A total of 4,398 preference scores were analysed. The preference score was calculated as the percentage of times a system was preferred (pooled across all sentences and participants). Results showed that participants significantly preferred samples generated without FIA (63.2%) over samples generated with FIA (36.8%).

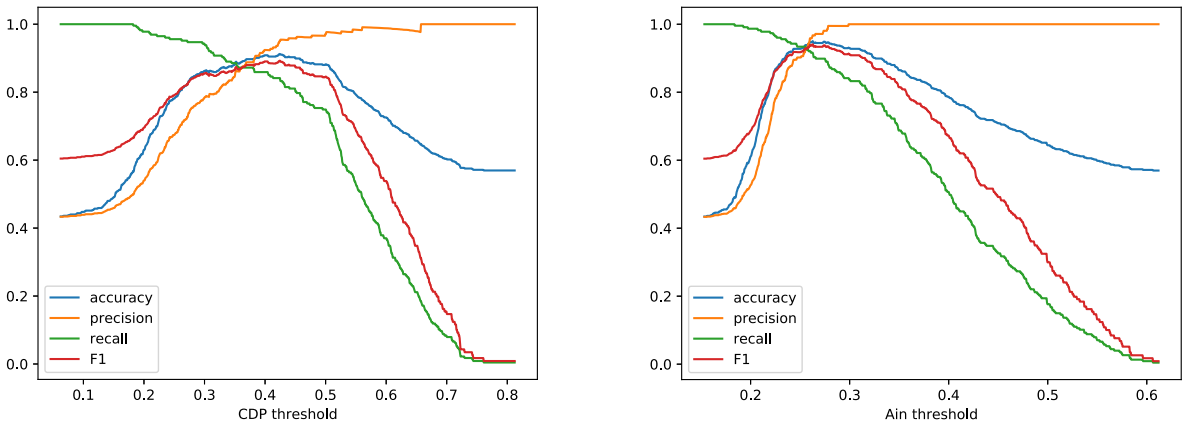

Figure 5: *Classification results for CDP (left) and Ain (right).*

4.2. Error detection

To evaluate if the metrics are able to detect gross errors in synthesised speech, we performed a listening test to gather human annotation of when gross errors occur. The listening test was carefully designed to assure sensible coverage of the material, best use of participant's effort and that the task was manageable. For this to happen we designed the test such that at least 50% of the material each listener was asked to annotate could potentially have errors (assuming errors occur when CDP > 0.5).

Listeners were presented with examples of what kind of errors they should detect, alongside speech samples and their transcriptions: early stop, skipping (missing phones and missing words) and repetition. Each participant listened to 100 sentences synthesised from a DCTTS system: 25 with errors we intentionally created ("focus" checks), 25 with suspected errors (CDP $> .5$) (selected from a set of 199 sentences) and 50 without suspected errors (CDP $< .5$) (selected from a set of 1490 sentences). The focus check sentences were used to check if the participant was doing the task correctly and were used to exclude participants from the final analysis.

4.2.1. Results

We analysed the results of 52 participants (6 participants were excluded from the total of 58 that took the test). We calculated the detection agreement for every sentence as the percentage of participants that marked that an error occurred.

We found that participants agreed on 84% of sentences, out of these, 40% of sentences were marked with errors, which indicates the task was reasonably easy and balanced.

As an initial evaluation metric we calculated the correlation coefficient between the human detection agreement and the scores each metric produced for a particular sample. We found a strong correlation (.73 and .78) for CDP and Ain and a week correlation for Aout (-.08). This result could be related to the fact that the most common error occurred was skipping, an error that Aout would not be able to detect.

To further evaluate the metrics we converted the detection agreement into zero (no error) and one (error) values and calculated the classification performance of each metric for this two class problem. Fig. 5 presents the accuracy, precision, recall and F scores for different detection thresholds (an error occurs when the measure is above the detection threshold). The maximum F-score=0.89 was obtained for a CDP threshold of 0.42 while a maximum F-score=0.94 was obtained for Ain threshold of 0.26, as presented in Fig. 5.

5. Conclusions

In this paper we evaluated automatic metrics for detecting gross errors in sequence-to-sequence speech synthesis. We focused on gross errors associated with attention failures and on metrics that rely on the attention metric alone without the need of a natural speech reference. We observed that a metric based on attention coverage and one based on dispersion are able to detect when a gross error occur with an F-score of .89 and .94 respectively. Using one of the measures to analyse a large number of synthesised samples we noticed that gross errors are more likely to occur in sentences with pause related punctuation marks and when the forcibly incremented attention mechanism is turned off. We showed, however, that imposing focus on attention by using a fixed mask degrades naturalness. These results motivate alternative mechanisms that takes into consideration coverage and dispersion.

Acknowledgements: The work presented in this paper was supported by Samsung Electronics Co., Ltd.

6. References

[1] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. ICASSP*, 2018, pp. 4784–4788.

[2] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[3] O. Watts, G. Henter, J. Fong, and C. Valentini-Botinhao, ""Where do the improvements come from in sequence-to-sequence neural TTS?"," in *Proc. SSW*. International Speech Communication Association, Sep. 2019, pp. 217–222.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf

[5] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *ICLR*, 2018.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, and Y. Zhang *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4799–4783.

[8] A. Graves, "Generating sequences with recurrent neural networks," 2014.

[9] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *ICLR*, 2017.

[10] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *ICLR*, 2018.

[11] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[12] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020.

[13] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65 955–65 964, 2019.

[14] M. Rikters and M. Fishel, "Confidence through attention," *Machine Translation Summit XVI, Nagoya, Japan, September 2017*, vol. abs/1710.03743, 2017. [Online]. Available: http://arxiv.org/abs/1710.03743

[15] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, 1984.

[16] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Challenge Workshop*, 2013.

[17] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.