EVALUATING SPEECH SYNTHESIS BY TRAINING RECOGNIZERS ON SYNTHETIC SPEECH

Dareen Alharthi[1], Roshan Sharma[1], Hira Dhamyal[1], Soumi Maiti[1], Bhiksha Raj[1,2], Rita Singh[1]

[1]Carnegie Mellon University, [2]Mohammed bin Zayed University of AI

ABSTRACT

Modern speech synthesis systems have improved significantly, with synthetic speech being indistinguishable from real speech. However, efficient and holistic evaluation of synthetic speech still remains a significant challenge. Human evaluation using Mean Opinion Score (MOS) is ideal, but inefficient due to high costs. Therefore, researchers have developed auxiliary automatic metrics like Word Error Rate (WER) to measure intelligibility. Prior works focus on evaluating synthetic speech based on pre-trained speech recognition models, however, this can be limiting since this approach primarily measures speech intelligibility. In this paper, we propose an evaluation technique involving the training of an ASR model on synthetic speech and assessing its performance on real speech. Our main assumption is that by training the ASR model on the synthetic speech, the WER on real speech reflects the similarity between distributions, a broader assessment of synthetic speech quality beyond intelligibility. Our proposed metric demonstrates a strong correlation with both MOS naturalness and MOS intelligibility when compared to SpeechLMScore and MOSNet on three recent Text-to-Speech (TTS) systems: MQTTS, StyleTTS, and YourTTS.

Index Terms— automatic speech quality assessment, speech synthesis, speech recognition, metric

1. INTRODUCTION

Speech synthesis systems, aka Text-to-Speech (TTS) systems are increasingly becoming better. TTS systems are generally judged using the following two criteria: intelligibility and naturalness of the synthesized speech to human listeners. These metrics are traditionally measured by calculating the Mean Opinion Score (MOS) (or intelligiblity score) of a panel of listeners, who annotate the synthesized speech with their subjective evaluation. However, as is generally the norm for any annotation process involving human evaluators, computing MOS is time and resource-expensive. As an alternative, there are other proposed efficient algorithmically computable metrics [1–5] which measure *proxies* of intelligibility, quality and naturalness of the synthesized speech for example, such as using an ASR model trained on real speech to evaluate Word Error Rate (WER) of the synthesized speech. However, we argue that these metrics fall short of capturing the real quality of the synthesized speech. In this paper, we propose a better approximation of these measures of synthetic speech, which we show to be highly correlated with MOS.

The top line of synthetic speech in intelligibility and naturalness is real speech, i.e. synthetic speech when closest to real speech would have the highest measure in these metrics. Therefore the question we pose in evaluating a TTS system is "How close is the quality and intelligibility of synthetic speech generated by the system to that of real speech and how can we evaluate this?". We hypothesize that quality and intelligibility differences between the synthetic and real speech are attributable to the distributional shift between the two, and any metric which attempts to quantify these differences must capture this shift. However explicit knowledge of the true distributions of the two is infeasible, and measurements must be made through mechanisms that invoke them implicitly.

Traditionally this is done by evaluating the synthetic speech on an ASR model trained on real speech. However, since speech synthesis is effectively a *maximum likelihood* generating process that attempts to produce the most likely speech signal for any text, this can result in unrealistically high recognition accuracies biased in favor of the synthetic speech and, consequently, anomalous measurements of the speech quality. We argue that on the other hand, an ASR model trained on the synthetic speech and evaluated on real speech better captures the statistical difference between the two, and would be a better approximation of the closeness of the real and synthetic speech. Since the ASR models the distribution of the synthetic speech, its ability to recognize the real speech exhibits how closely the distributions of the synthetic training data matches with that of real testing data.

This paper makes the following contributions:

1. We propose a new evaluation method for TTS that captures distributional similarity between real and synthetic speech as a proxy for perceptual speech quality tests.

2. We compare the proposed metric to multiple automatic metrics and Mean Opinion Score (MOS), and show that our metric correlates well with human-provided MOS.

2. BACKGROUND: SPEECH SYNTHESIS AND EVALUATION

Recent advancements in speech synthesis systems have reached a point where they are often indistinguishable from human speech [6]. However, evaluating these systems has become increasingly complex. The most dependable method for evaluating speech synthesis systems from various perspectives is the Mean Opinion Score (MOS), in which human raters listen to synthesized speech and assess its naturalness, quality, and intelligibility using a 5-point Likert scale. However, this process is time-consuming, expensive, and subject to subjective judgments. To address these challenges, researchers have developed automatic metrics aimed at reducing evaluation costs. However, each metric is typically limited to evaluating a specific aspect of speech synthesis system performance, necessitating the use of multiple metrics to comprehensively assess these systems. Recent studies have tackled this challenge through the training of regression models using pairs of speech MOS scores [7] or by utilizing semi-supervised learning methods to acquire MOS scores. An important constraint associated with this method is the need for labeled datasets in the same domain, making it less generalizable [8] to any text-to-speech (TTS) system.

Unsupervised metrics have also been employed to assess various aspects of speech synthesis, such as the Equal Error Rate for measuring speaker similarity in synthesized speech and metrics like Frechet DeepSpeech Distance [5] (FDSD) and Frechet Speech Distance (FSD) [4] to measure the quality and diversity of synthetic speech. However, it's important to note that each of these metrics focuses on a single factor and cannot serve as standalone measures. Recently, the utilization of speech-language models to assess speech quality has revealed a correlation with MOS scores. The SpeechLM-Score [2] calculates the perplexity of synthetic speech by employing pretrained autoregressive unit speech language models (uLM) [9]. Another avenue of exploration involves Automatic Speech Recognition (ASR)-based metrics. One approach involves measuring the distance between synthetic and real speech [10] by computing various distance metrics to assess speaker, prosody, and environmental similarity within real distributions. A commonly used ASR evaluation method is the computation of Word Error Rate (WER) [1] for synthetic speech using pre-trained ASR models to measure intelligibility. Our proposed ASR evaluation approach seeks to evaluate both the naturalness and intelligibility of synthetic speech by quantifying the distribution shift between synthetic and real distributions.

3. PROPOSED METHOD

3.1. Divergence metric for Distributional Shift

Given a text T, let X_r be a random variable that represents real speech signals produced by humans to convey text T. Let
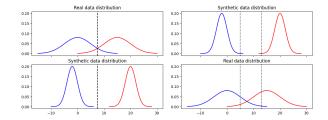

Fig. 1. Left: Model trained on real data and tested on synthetic data.
Right: Model trained on synthetic data and tested on real data.

X_s be a random variable that represents synthesized speech from a TTS model for text T. $P(X_r, T)$ and $P(X_s, T)$ are the joint distributions of the speech and text.

To evaluate the TTS Model, we want to compare these joint distributions - if the distributions are similar, synthetic speech has relatively high quality. Therefore, we want to compute a divergence $\text{div}(P(X_r, T), P(X_s, T))$ between the probability distributions, that measures the distance between the two distributions, i.e. distributional shift.

Distributional shifts are typically computed using divergence metrics such as the Kulback-Leibler (KL) divergence [11], Jensen-Shannon divergence [12], the Earth-mover distance [13] etc. However, these metrics require explicit knowledge of the distributions, or at least the ability to compute the probability of a given instance, if sampling-based approaches are to be used, which is infeasible, since it requires explicit modelling of $P(X_s, T)$ (or $P(X_r, T)$) whereas neural models only approximate the conditional probability of T. Furthermore, even with explicit sampling, given the high dimensionality and time-series nature of the data, it would require sampling an infeasibly large number of pairs of (X, t), where X is either X_r or X_s, for reliable estimates.

To address the limitations of existing distributional similarity metrics, we propose an alternate metric that uses classification performances as a proxy to get distributional shifts. This classification-based pseudo-divergence uses probability distributions to get accuracy metrics. Given the two data distributions, input and labels, below we present the general case of the divergence metric.

Let P_1 and P_2 be two data distributions of random variables X and y, where X is the input signal and y is the label. The predicted label for the given input x can be written as:

$$\hat{y_1}(x) = \text{argmax}_y P_1(y|x)$$

When the classification boundaries are learned from P_1, and used to classify the data coming from the same distribution, the accuracy of this classification can be written as.

$$\mathbb{E}_{P_1(x)}[P_1(\hat{y_1}(x)|x)]$$

Using the classification boundaries learned on the distribution P_1 and used to classify the data coming from the dis-

tribution P_2, the accuracy can be written as:

$$\mathbb{E}_{P_2(x)}[P_2(\hat{y}_1(x)|x)]$$

The difference between the two classification accuracies captures the distributional shift between the P_1 and P_2. This can be written as:

$$d(P_1, P_2) = |\mathbb{E}_{P_1(x)}[P_1(\hat{y}_1(x)|x)] - \mathbb{E}_{P_2(x)}[P_2(\hat{y}_1(x)|x)]|$$

The absolute value is needed since this difference could be negative. Note that $d(P_1, P_2)$ is a pseudo-divergence that goes to zero when $P_1 = P_2$ and is non-negative. It is also asymmetric, so $d(P_1, P_2) \neq d(P_2, P_1)$.

We can use the above formulation to calculate the pseudo-divergence of the real and synthetic speech. The distributions P_1 and P_2 can be estimated using an ASR Model trained on the data samples taken from the real and synthetic speech respectively. Since this is asymmetric, it is important to note which divergence to calculate $d(P_1, P_2)$ or $d(P_2, P_1)$. Either the ASR Model trained on real speech and tested on synthetic speech or vice versa. Empirically we show that the model trained on synthetic and tested on real data is a more accurate metric for the distributional shift of the two distributions than doing it vice versa. We explain the intuition behind this in the following section.

3.2. Real vs Synthetic data distribution

Figure 1 shows the joint distributions X and y where red curve shows when class $y = 1$ and blue curve shows class $y = 0$. Note that the real data has more variance than the synthetic data (which is true for the real and synthetic speech). When the classification boundary for the two classes is learned on the real data, there is some natural Bayes error associated with the class overlap present in the real data. When this classification boundary is used to do classification in the synthetic data, the error is zero, since the data distributions are far apart and there is no overlap in the two. In fact, there are multiple decision boundaries associated with different errors on the real data that would ensure zero error in the synthetic data. This zero error does not say anything about how different the real and synthetic data joint distributions are. The synthetic data distribution could be far off the chart, be highly unlikely compared to the real data, and still have zero error.

On the other hand, let's consider the case where the lower variance data, i.e. the synthetic data, is used for learning the classification boundary. The right part of Figure 1 shows this scenario. The dotted line shows the range where the decision boundary can lie such that the error rate on the synthetic data would be zero. However, this range of boundary would always be associated with greater than zero error on the real data. The higher the difference in the joint distributions in the real and synthetic distributions, the greater the range of errors in the real data.

Therefore, the second scenario is better representative of the distributional differences in the real and synthetic data distributions. We believe that this would hold for the real and synthetic speech distributions. An ASR model trained on synthetic speech and evaluated on real speech would be a better metric of the quality of the synthetic speech than doing it the other way around.

4. EXPERIMENTAL SETUP

4.1. Text-to-Speech-Synthesis

We evaluate the proposed method using three state-of-the-art open-source TTS systems: StyleTTS [14], MQTTS [15], and YourTTS [16]. These models utilize different techniques for synthesis, but all use a reference encoder to extract both speaker and style information from the input speech. For our assessments, we made use of the publicly released pre-trained models. The StyleTTS model, MQTTS, and YourTTS models we used were trained on LibriTTS [17], Gigaspeech [18] without audiobooks, and VCTK [19] respectively. A Deep-Phonemizer [20] was used to extract phonemes from the text for synthesis.

4.2. Automatic Speech Recognition

In order to make evaluations robust and meaningful, we need to select strong End-to-End models. In this paper, we therefore elect to fine-tune Whisper rather than train from scratch using 10h of real/synthetic speech. We use the `Whisper-medium multilingual model` [21] as the initialization. We then fine-tune it within ESPNet [22, 23] using CTC loss [24]. ASR Inference was performed using beam search with a beam size of 5.

4.3. Datasets

To generate synthetic speech for our evaluation, we utilized the LibriTTS [17] dataset, which is based on Librispeech [25]. From this dataset, we sample one subset of 10 hours containing speech data from all available speakers. All three TTS models used a speaker encoder to clone the identity of a given speech reference. It's worth noting that we excluded speech samples that were less than 4 seconds in duration and those exceeding 30 seconds in length. This exclusion was necessary as MQTTS and StyleTTS do not support short samples as references.

4.4. Evaluation Metrics

MOS-Naturalness (MOS-N) : We conducted a crowd-sourced Mean Opinion Score (MOS) evaluation to assess the naturalness of synthetic speech generated by each system, in comparison to real speech. We obtained 50 sentences from

Table 1. This table shows the scores for real and synthetic speech on multiple metrics for LibriTTS test-clean. MOSNet and SpeechLMScore scores are on the same 50 samples of MOS-N. Relative ranking among synthetic speech systems are shown in red inside the brackets.

Model / Metric	Ground Truth	StyleTTS	MQTTS	YourTTS
WER ↓ [1]	20.57	**18.7**(1)	29.35(3)	22.1(2)
SpeechLMScore ↑ [2]	3.98	3.62 (3)	**4.13**(1)	3.96 (2)
MOSNet ↑ [7]	4.30	**4.49**(1)	3.57(3)	4.01(2)
MOS-N ↑	**3.69**	3.68 (1)	3.66 (2)	3.59 (3)
MOS-I ↑	-	**0.698**(1)	0.618(2)	0.566 (3)
Ours 10h ↓	**3.1**	3.3 (1)	3.9(2)	4.5 (3)

the LibriTTS test-clean dataset and another 50 from the LibriTTS test-other dataset, resulting in a total of 100 samples each for real speech, MQTTS, YourTTS and StyleTTS. Each sample was evaluated by 10 raters, who were instructed to rate the naturalness of the speech on a scale of 1 to 5, with 1 indicating poor and 5 indicating excellent quality.

MOS-Intelligibility(MOS-I): We assessed intelligibility of spoken words by using nonsense sentences [26], effectively eliminating sentence structure and grammar from the evaluation. This absence of structure allowed listeners to only focus on the quality of the synthesized speech and not be distracted by the grammar. Participants were presented with a choice between the original sentence and a transcription generated by the `Whisper-medium`. We specifically selected 60 sentences with relatively high Word Error Rate (WER) from a pool of 200 random sentences generated by ChatGPT [27]. Among these, 30 sentences were short (less than 10 words), while the other 30 were long. This allowed us to evaluate the impact of sentence length variation on intelligibility. We generated synthetic speech using the three TTS systems for the 60 sentences using a test-clean set as a reference for the model's speaker and style encoder. We used WebMushar [28] to create a test form along with Prolific for crowd-sourcing.

Intelligibility of Synthetic Speech using WER from Pre-trained ASR: We computed the WER for synthetic speech generated by three different systems using the `Whisper medium multilingual`. This model is pre-trained on real speech and evaluated on synthetic speech. This setting of training / testing demonstrates the traditional way that speech synthesis evaluation is performed. This evaluation was performed on both the test-clean and test-other datasets from LibriTTS.

5. EXPERIMENTAL RESULTS

Table 1 reports the results of our experiments on Libri-TTS with the proposed evaluation method. We consider multiple metrics and report raw scores of the metric in the rows and rel-

ative ranking scores in brackets next to the raw score. The first row, named WER shows the case when the model is trained on real data and evaluated on synthetic data. The last row shows our setting, where the model is trained on synthetic data and evaluated on real data. Based on the absolute raw numbers of the metric, we rank the TTS systems from 1 to 3 based on which one performs the best to worst. For example in the row MOS-N, Style-TTS has the highest score and therefore has rank 1, followed by MQ-TTS and then YourTTS. In order to assess whether our metric is a good representation of the quality of synthetic speech, we compare the relative ranking of our metric with the other metrics. Two metrics with a matched relative ranking means that the metrics evaluate the quality of speech similarly and agree with each other.

First, we see that the Mean Opinion Score tests on naturalness (MOS-N) and intelligibility (MOS-I) agree on relative rankings between the synthetic speech models. Further, we observe that the traditionally used WER metric shown in the first row does not actually correlate completely with the MOS results. We observe similar issues with other popular metrics including SpeechLMScore and MOSNet.

From the last row, we observe that our metric evaluation of synthetic speech has a similar trend as the reported MOS scores, matching both MOS-N and MOS-I. Compared with the inconsistent result from the first row and the consistent result from our metric, we demonstrate the importance of the proposed evaluation method.

6. CONCLUSION

In this paper, we address the challenge of automatic evaluation for synthetic speech by modeling the similarity/dissimilarity between the distributions of synthetic and real speech. Existing divergence metrics require a large number of samples to capture the joint distribution and hence it is infeasible to employ them to calculate the distributional shift. In this paper, we introduce a new divergence measure that can be computed without knowledge of the joint distribution. The metric uses an ASR model as an approximation for the data distributions and the WER as a proxy for the quality of the synthesized speech. The metric is asymmetric, and it matters what the speech recognition models are trained and tested on. We show that in practice it is more accurate to train the model on synthetic speech and assess the resulting model's performance on real speech than doing it vice versa. Experiments using 3 public open-source speech synthesis systems show that our model correlates positively with subjective human Mean Opinion Scores for naturalness and intelligibility, while previously used ways for evaluating ASR performance trained on real and evaluated on synthetic does not correlate. Further, we show that it only takes small amounts of synthetic speech to train the ASR model to be able to make reliable judgments on the quality of the synthesized speech.

7. REFERENCES

[1] M. Cerňak, M. Rusko, and M. Trnka, "Diagnostic evaluation of synthetic speech using speech recognition," in *Procs. of the 16th International Congress on Sound and Vibration (ICSV16), Kraków, Poland*, 2009, pp. 5–9.

[2] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," in *Proc. ICASSP*, 2023, pp. 1–5.

[3] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, "Squid: Measuring speech naturalness in many languages," in *Proc. ICASSP*, 2023, pp. 1–5.

[4] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *arXiv preprint arXiv:2306.15687*, 2023.

[5] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *Proc. ICLR*, 2019.

[6] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[7] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[8] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[9] K. Lakhotia *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[10] C. Minixhofer, O. Klejch, and P. Bell, "Evaluating and reducing the distance between synthetic and real speech distributions," *arXiv preprint arXiv:2211.16049*, 2022.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[12] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[13] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 59–66.

[14] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[15] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," *arXiv preprint arXiv:2302.04215*, 2023.

[16] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, pp. 2709–2720.

[17] 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," *Interspeech 2019*, 2019.

[18] G. Chen *et al.*, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech*, 2021, pp. 3670–3674.

[19] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[20] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Transformer based grapheme-to-phoneme conversion," *Proc. Interspeech 2019*, pp. 2095–2099, 2019.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[22] S. Watanabe *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[23] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[26] O. Kang, R. I. Thomson, and M. Moran, "Empirical approaches to measuring the intelligibility of different varieties of english in predicting listener comprehension," *Language Learning*, vol. 68, no. 1, pp. 115–146, 2018.

[27] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[28] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st Web Audio Conference*, 2015, pp. 1–6.