A Generative Adversarial Network based Ensemble Technique for Automatic Evaluation of Machine Synthesized Speech

Jaynil Jaiswal*, Ashutosh Chaubey*, Bhimavarapu Sasi Kiran Reddy, Shashank Kashyap, Puneet Kumar, Balasubramanian Raman, Partha Pratim Roy

(*Equal Contribution) Indian Institute of Technology, Roorkee

Abstract

In this paper, we propose a method to automatically compute a speech evaluation metric, Virtual Mean Opinion Score (vMOS) for the speech generated by Text-to-Speech (TTS) models to analyse its human-ness. In contrast to the currently used manual speech evaluation techniques, the proposed method uses an end-to-end neural network to calculate vMOS which is qualitatively similar to manually obtained Mean Opinion Score (MOS). The Generative Adversarial Network (GAN) and a binary classifier have been trained on real natural speech with known MOS. Further, the vMOS has been calculated by averaging the scores obtained by the two networks. In this work, the input to GAN's discriminator is conditioned with the speech generated by off-the-shelf TTS models so as to get closer to the natural speech. It has been shown that the proposed model can be trained with a minimum amount of data as its objective is to generate only the evaluation score and not speech.

1. Motivation

The capability of generative models in generating realistic data examples has given birth to several text-to-speech (TTS) systems. However, the task of measuring the quality of the speech produced by TTS systems using traditional speech evaluation methods such as - Mean Opinion Score (MOS), paired-comparison test, etc. is very challenging. due to the following reasons:

- These methods suffer from the subjective variance caused by human intervention during the evaluation.
- The cost and time involved for carrying out such manually done analysis.

2. Problem Formulation

Consider the input text $\hat{S}$ and the speech generated by a TTS system T conditioned on $\hat{S}$ denoted as $X = \hat{G}(\hat{S})$. An automatic speech evaluation system is supposed to produce the score $V(\hat{G}(\hat{S}))$ in a way to fulfil following constraints:

- The generated score $V(\hat{G}(\hat{S}))$ should correspond to the quality of the generated speech X in terms of human-ness and emotion.
- The score $V(\hat{G}(\hat{S}))$ should be comparable to till date used metric for generated speech evaluation, i.e. MOS.

3. Anthropomorphic Score

Anthropomorphic Score is defined to gauge the goodness of a TTS model in terms of synthesizing human-like speech. Normalization of vMOS in order to directly compare it with MOS is not valid as they vary in their absolute scales. In this case, Anthropomorphic Score provides a better comparative framework to evaluate the capability of a TTS model by comparing the values of same scale. Anthropomorphic Score A of a TTS model T is determined as per the following formula.

$$A = \frac{vMOS\ for\ T}{vMOS\ of\ natural\ speech} \tag{2}$$

7. Conclusion and Future Work

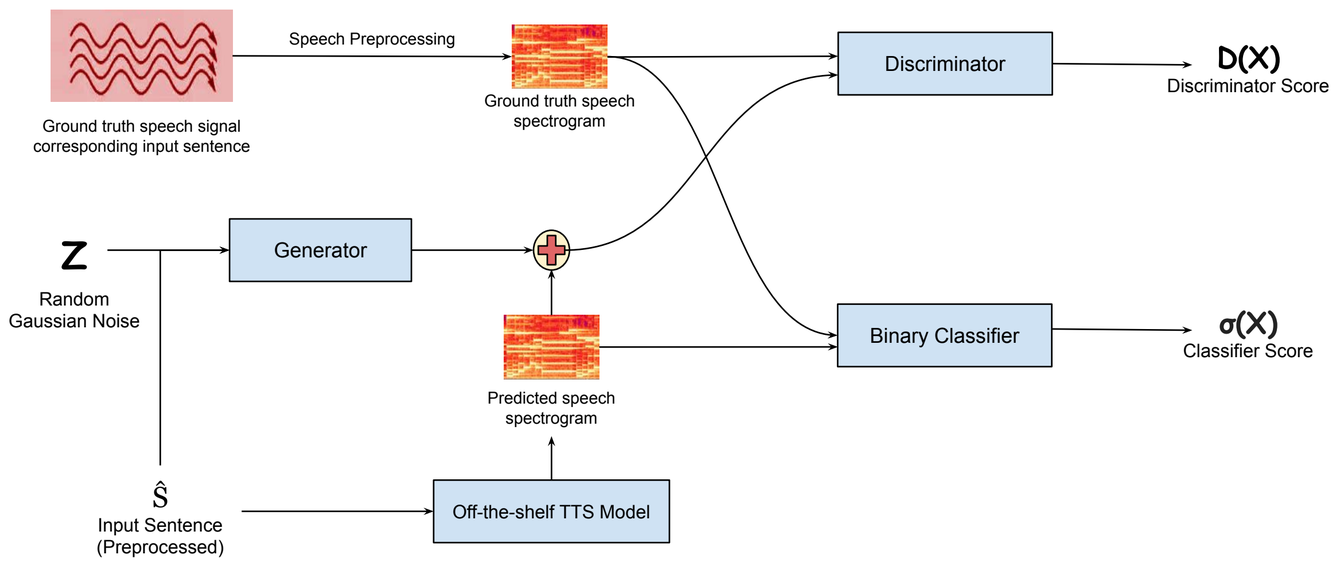
We have shown that, in spite of being trained on single-speaker speech dataset, the proposed model performed well in evaluating the speech including multi-speaker utterances and emotional variations. In future, we will focus to improve the proposed model by explicitly training it with multi-speaker and emotional datasets. Another research dimension could be to use the proposed methodology, in evaluating the correctness of emotion in speech generated by TTS models.

4. Proposed Methodology

The proposed methodology uses an ensemble of a GAN and a vanilla binary classifier trained as follows:

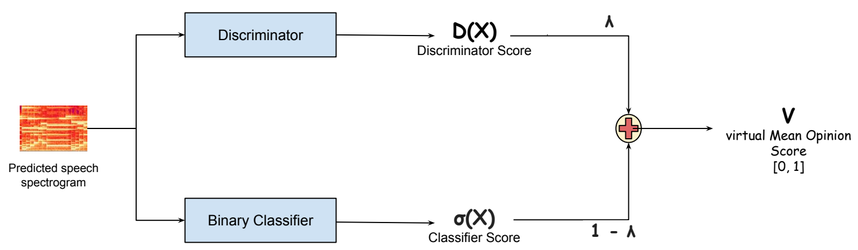
Phase I: GAN Training Training the GAN as shown in the Fig. , by conditioning it on the input sentence $\hat{S}$ end to end. Generator acts as an error corrector to the spectrogram generated by some state-of-the-art TTS model

Phase II: Binary Classifier Training Training a simple binary classifier to classify fake or real input speech spectorgram



5. Inference

Fig. illustrates the inferencing process to compute the vMOS by ensembling the evaluation scores produced by GAN and the binary classifier. Here $D(\hat{G}(\hat{S}))$ is the score output by the Discriminator for a speech spectrogram $\hat{G}(\hat{S})$ generated by some TTS model corresponding to a sentence $\hat{S}$, and $\sigma(\hat{G}(\hat{S}))$ is the classifier score outputted by the binary classifier. These two scores are combined as per the hyper-parameter λ to compute the speech quality score as per Eq. (1). The final vMOS score is a weighted combination of the scores outputted by the Discriminator and the binary classifier.



$$V(\hat{G}(\hat{S})) = \lambda * D(\hat{G}(\hat{S})) + (1 - \lambda) * \sigma(\hat{G}(\hat{S})) \tag{1}$$

6. Results

Case :- Non-emotional Single and Multi-speaker speech data The speech synthesized by various off-the-shelf TTS models for the transcripts of single-speaker LJSpeech data and multi-speaker VCTK data have been evaluated in this evaluation-case. The vMOS scores are presented in the following Table.

TTS Model	LJSpeech dataset		VCTK dataset	
	MOS	vMOS	MOS	vMOS
DeepVoice3	3.62	0.4863	3.62	0.4917
Tacotron	3.82	0.4966	3.82	0.4923
WaveNet2	4.53	0.6624	4.53	0.6725
Natural Speech	4.58	0.7009	4.58	0.6354

The Anthropomorphic Score,A for VCTK dataset with WaveNet2 came out to be 1.06.A >1 suggests the quality of synthesized speech to be better than that of natural speech. We analysed the speech outputs of WaveNet2 and found them to be actually clearer and louder than the corresponding data samples.