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

Jaynil Jaiswal*$^{[0000-0002-3739-9495]}$, Ashutosh Chaubey*$^{[0000-0002-8463-0012]}$, Bhimavarapu Sasi Kiran Reddy$^{[0000-0002-1090-4375]}$, Shashank Kashyap$^{[0000-0001-7500-9120]}$, Puneet Kumar$^{[0000-0002-4318-1353]}$, Balasubramanian Raman$^{[0000-0001-6277-6267]}$, and Partha Pratim Roy$^{[0000-0002-5735-5254]}$

Dept. of Computer Science and Engineering, Indian Institute of Technology Roorkee,
Roorkee - 247667, India
{jjaynil,achaubey,breddy,skashyap,pkumar99}@cs.iitr.ac.in,
{balarfcs,proy.fcs}@iitr.ac.in

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. The proposed method has been tested to evaluate the speech synthesized by state-of-the-art TTS models and it has reported the vMOS of 0.6675, 0.4945 and 0.4890 for Wavenet2, Tacotron and Deepvoice3 respectively while the vMOS for natural speech is 0.6682 on a scale from 0 to 1. These vMOS scores correspond to and are qualitatively explained by their manually calculated MOS scores.

Keywords: Automatic Speech Evaluation · Text-to-Speech · Conditional GAN · Binary Classifier · Virtual Mean Opinion Score

1 Introduction

The capability of generative models in generating realistic data examples has given birth to several text-to-speech (TTS) systems [1–3]. They find applications in areas such as - man and machine interaction, aid to visually impaired

* Denotes equal contribution

people, smart devices, personal digital assistants, security and authentication, vehicle control and automation, gaming and animation, etc. [4] 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. [5] is very challenging. It requires measuring the values of acoustic parameters such as - pitch, frequency, amplitude, etc. and getting their ranges corresponding to real natural speech so that the generated fake speech can be evaluated in comparison to these parameters. These methods suffer from the subjective variance caused by human intervention during the evaluation. Therefore, the need to build speech processing systems that are capable of automatically evaluating machine generated speech is increasing at a fast pace [6]. By automating the evaluation process of TTS models, model prototyping and testing can be accelerated significantly. The need for an objective metric builds upon this as it would be consistent, quick and cheaper than variants of MOS.

In this paper, a novel approach to evaluate the human-ness of a given speech sample has been proposed. The primary goal is to be able to automatically evaluate the speech generated by TTS models thereby excluding the need of having speech experts for calculating the MOS metric. The proposed model is an ensemble of a Generative Adversarial Network (GAN) and a binary classifier, resnet-v2-50 which outputs the scores about the quality of the generated speech. Then a 'Virtual MOS' (vMOS) is determined by ensembling the evaluation scores computed by the GAN and the binary classifier. The proposed system has been tested to evaluate the speech synthesized by state-of-the-art TTS models trained on benchmark speech datasets [7–10]. It has been shown that the scores outputted by the proposed model are in-line with the actual MOS scores already available for the TTS models. The correspondence of manual MOS and automatic vMOS scores proves the applicability of the proposed model to automatically evaluate the synthetic speech just as effectively as the manual approaches.

The rest of the paper is organized as follows. Existing techniques in context of speech evaluation have been reviewed in Section 2. Section 3 formulates the problem of developing an automatic speech evaluation method comparable to MOS approach in terms of its effectiveness. The proposed methodology has been detailed in Section 4. Section 5 presents the implementation details and results for various cases. The broader implications of the achieved results and the scope for future improvements have been concluded in Section 6.

2 Related Work

Speech evaluation is typically performed by human experts either manually by assigning opinion scores to speech samples or semi-automatically by formulating and optimizing relevant objective functions. The existing techniques in this field are briefly reviewed in this section.

2.1 Speech Evaluation Metrics

Evaluation of synthetic speech generated by TTS systems such as Deep Voice [1], Tacotron [2], Wavenet [4], etc. is done by natural speech experts. They are made to listen to the output speech and asked for their opinions on its humanness. The scores assigned by various experts are recorded and the evaluation metrics such as - Mean Opinion Score (MOS), paired-comparison score, etc. are calculated by taking their weighted average [5, 6].

2.2 Subjective Methods of Speech Recognition and Evaluation

The evaluation of speech has been carried out in terms of social cues such as - ethnicity, social class, speaker-age, etc. through pair-wise comparision of a speech sample in contrast to another one and in a questionnaire-based manner [11, 12]. While these methods were able to provide a legitimate assessment of the output speech, they required human intervention to carry out the assessment process. In the context of speech recognition, Vivek et. al. [13] used selectively-biased linear discriminant analysis and Jan et. al. [14] used Attention-based model. Researchers have been successful to recognize user-specific [15] and multi-lingual speech [16]. Various efforts to improve speech recognition and evaluation have been made such as - modelling and evaluation of the speech duration [17], instrumental measure-based speech enhancement [18], speech evaluation by analysing the correlation between pitch frequencies of input spectrum and processed spectrum [19], etc. The speech processing methods thus developed performed decently to recognize the natural speech and evaluate it manually [6]. However, they were not adequate to automatically evaluate the machine synthesized speech.

2.3 Objective Function based Speech Evaluation

There are several issues with subjective speech evaluation methods, most prominent of them being the time and cost involved in the process. Another issue is that a separate analysis of speech needs to be done at each stage to check which properties or features are not properly modelled in the TTS system. This is not feasible to do with subjective evaluation methods. To overcome these problems, there have been some attempts at evaluating machine synthesized speech using objective methods. Objective evaluation of speech involves representing the speech evaluator as a function and using the output of this function as a score to rate a speech sample, thus eliminating the need of manual intervention. It helps in developing good quality TTS voice during its initial stages as well. Well-known objective measures, viz., Mel Cepstral Distortion and Dynamic Time Warping distance have been used as the objective measures during the optimization stage for TTS synthesis [20]. Their results concluded the need for further research on objective measures as human speech was too complicated to be evaluated on the basis of simple objective measures mentioned above.

Motivated by the successful use of neural networks as function approximators, we have attempted to model the evaluator as a function using deep neural network. Hereafter, this function is referred as the *scoring function* and its effectiveness for the task of automatic speech evaluation has been demonstrated as advocated by the results. In contrast to the currently used manual techniques of analysing the human-ness of synthetic speech, the proposed method uses an end-to-end neural network to automatically calculate a score qualitatively similar to manually obtained MOS.

3 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.

We need to find some mapping from the generated speech space to the vMOS space conditioned on the prior of input text. It has been shown that the vMOS outputted by our method for the speech generated by well-known TTS Models for input sentence $\hat{S}$, corresponds well to the MOS already available for them.

Definition 1: (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. MOS ranges from 0 to 5 with value 4.58 for natural speech while vMOS varies from 0 to 1 with value 0.6682 for the natural speech. vMOS value 1 for natural speech corresponds to 4.58 and not 5. 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.

$$A = \frac{vMOS\ for\ T}{vMOS\ of\ natural\ speech} \tag{1}$$

4 Proposed Methodology

The proposed model uses an ensemble of Generative Adversarial Network (GAN) Discriminator along with a binary classifier, resnet-v2-50 for calculating vMOS. The details of various steps involved in this process are described as follows.

4.1 Pre-processing

The transcripts to generate the corresponding speech are available along-with the training speech data. The speech generated by off-the-shelf TTS models and human voice audio are interpreted in terms of intensity of the signal and then plotted along the log mel scale to obtain images of the spectrogram of size $64 \times 64 \times 3$. These images are now used as the speech samples rather than the audio. The latent vector of dimension 100×1 has been used as the input to the generator (as random Gaussian noise) along with the pre-processed text sequence. The speech generated by the generator is also the image of the waveform with a dimension of $64 \times 64 \times 3$. We use speech (waveforms as images) as input and their corresponding ground truth. Real and fake nature of the speech is denoted by 1 and 0 respectively as the labels to train the classifier.

4.2 Training

A schematic of the proposed model during training is summarized in Fig. 1. The architecture has a discriminator along with a vanilla binary classifier to classify the input speech as real or fake. The reason behind adding a classifier on top of the network is that training the classifier encourages it to automatically learn features that are inherent to human voice. The discriminator network is trained alternatively with real speech and the enhanced generated speech.

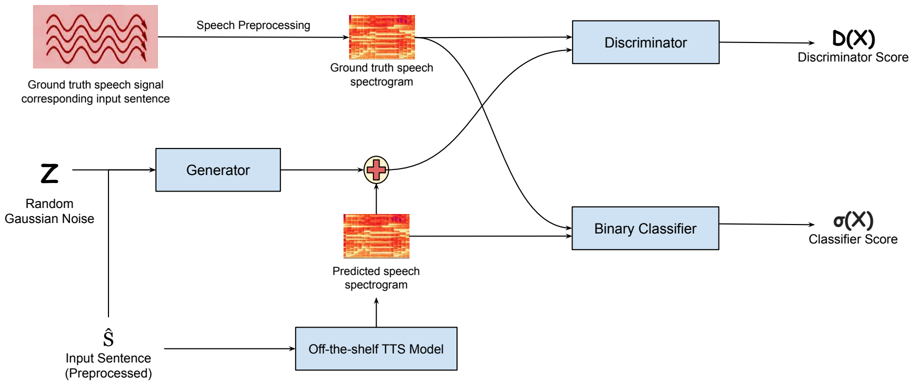

Fig. 1. Overview of the proposed model during training stage

Phase I: GAN Training Generative Adversarial Networks have been well known for their effectiveness in image generation tasks since their introduction in 2014 [21]. GANs are composed of two deep networks namely the Generator and the Discriminator. Training of GANs takes place as a minmax game between Generator and Discriminator where the Generator tries to minimize the value function $V_f(D, G)$ and the Discriminator tries to maximize it.

$$V_f(D, G) = E_{x \sim P(x)}[logD(x)] + E_{z \sim P(z)}[log(1 - D(G(z)))] \tag{2}$$

where $D(x)$ is the score outputted by the discriminator, which is the probability that the image input to the Discriminator is real and $G(z)$ is the image generated from the Generator given input z from a Gaussian prior.

The proposed method uses a *Conditional* GAN [22] for the training. It uses a Generator-Discriminator model for training the GAN as is used traditionally. During training, the random Gaussian noise $\mathbf{z}$ input to the Generator is conditioned with the input sentence $\hat{S}$. Then the Generator of the network is removed and only the Discriminator is used to output the discriminator score. Here GANs have been used despite them not being very good at waveform generation because the task is not to generate speech but to evaluate its goodness on a relative scale. To provide a way to calculate a metric very similar in quality to MOS, known generated speeches of previous models and datasets are passed from the proposed Discriminator.

Since the discriminator, in a speech synthesizing GAN discriminates between the generated speech and the real speech, it will itself learn the features which are specific to real speech. So, after the discriminator has been trained, we use it as the scoring function, and the score $D(X)$ produced by it for an input speech X is used as a measure of the human-ness of the input speech. However, the problem with this approach is that the speech output by a Conditional GAN [22] on the input $\hat{S}$, is very weak in quality compared to the natural speech, due to which our discriminator can fool the generator easily which will disable it from learning different features in the spectrogram. To tackle this issue, instead of the generator directly producing the speech spectrogram, our generator outputs the error to the speech spectrogram output by some off-the-shelf TTS model corresponding to input text sentence, so that the corrected spectrogram becomes close to natural speech spectrogram.

Phase II: Binary Classifier Training The second part of the proposed ensemble model is a binary classifier, resnet-v2-50 that is used to classify whether the input speech spectrogram is real or fake. Binary classifiers learn the features corresponding to the two different classes, i.e. real speech and generated speech in our case. Based on the learned features, it outputs the probability of the input speech belonging to one of the two classes. If we want to calculate the quality of some speech spectrogram, indirectly we want to know what is the likelihood of that speech being real. So, we can directly use this binary classifier as a model which outputs the quality score of some given input speech. To train the binary classifier, generated speech of some TTS models corresponding to some input sentence is labelled as fake (zero) and the natural speech corresponding to same sentence is considered as real (one). After training the model in such a way, it will output the quality score corresponding to some speech spectrogram input to it.

4.3 Inference through Ensembling

Fig. 2 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

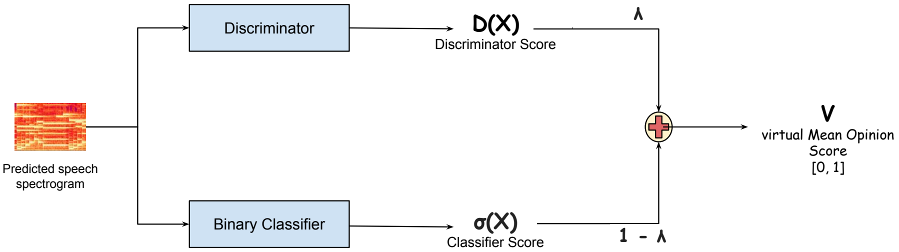

Fig. 2. Overview of the proposed model during inference stage

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. (3). The final vMOS score is a weighted combination of the scores outputted by the Discriminator and the binary classifier. The computed vMOS scores for various evaluation-cases have been presented in Section 5.6 and the intermediate calculations have been included in the Supplementary Material.

$$V(\hat{G}(\hat{S})) = \lambda * D(\hat{G}(\hat{S})) + (1 - \lambda) * \sigma(\hat{G}(\hat{S})) \tag{3}$$

5 Implementation and Results

This section discusses the experimental implementation and analyses the results. The proposed model has been trained with LJSpeech dataset using PyTorch [1]. Four test-cases have been formulated to generate the vMOS evaluation scores for synthetic speech samples containing single-speaker, multi-speaker, emotional and gender-specific speech utterances.

5.1 Experimental Set-up

The model training has been carried out on NVidia Tesla K80 GPU machine with 24GB RAM and 4992 CUDA cores. Model evaluation has been performed on Intel(R) Core(TM) i7-7700U, 3.60 GHz, 16GB RAM CPU machine. The implementation results for all the use-cases are presented in the following sections.

5.2 TTS Models evaluated

The following off-the-shelf TTS models have been considered for evaluation.

- **Tacotron** is a sequence-to-sequence architecture. It takes raw text as input and converts it into spectrogram. Then Griffin-Lim algorithm is used to synthesize speech by approximating the spectrogram into waveforms [2].

[1] https://pytorch.org/

- **DeepVoice3** is a fully convolutional end-to-end architecture for speech generation from text. It uses alternate vocoders instead of Griffin Lim algorithms such as – WORLD, WaveNet, etc. It is capable of scaling up to 2000 different voices with good quality speech output [3].
- **WaveNet2** is a generative model for raw-audio. It takes acoustic and linguistic features as input and generates human-like voices. It models the waveforms directly using a network trained with real speech recordings [4].

5.3 Datasets

Following datasets have been considered for evaluation using proposed method.

- **LJSpeech dataset** [7] consists of 13,100 samples of sentences spoken by a single-speaker from 7 non-fiction books. It contains short audio clips of single speaker whose length vary from 1 to 10 seconds.
- **VCTK Dataset** [8] is a multi-speaker data released by CSTR (Centre for Speech Technology Voice Cloning Toolkit) containing speech utterances by native English speakers in various accents.
- **RAVDESS** [9] The Ryerson Audio-Visual Database of Emotional Speech and Song contains audio samples of 24 professional speakers of North American accent. It contains spoken and sung speech utterances including angry, calm, fear, happy, neutral and sad emotions.
- **IEMOCAP** [10] The Interactive Emotional Dyadic Motion Capture database is a multi-speaker database containing scripted and improvised speech samples annotated into emotion categories anger, excited, frustration, happiness, neutral and sadness.

The proposed model has been trained with LJSpeech data and evaluated with the samples of aforementioned datasets. The intention of considering these datasets is to capture variations in terms of spoken or sung utterances, scripted or improvised samples and the samples labelled with gender and emotion information. For LJSpeech and VCTK data, the evaluation is done for the synthetic speech. However, for the emotional datasets RAVDESS and IEMOCAP, the evaluation has been carried out for natural speech only. The aim is to check whether human-ness for the scripted emotional speech is maintained or not.

5.4 Parameter Settings

The generator-discriminator pair has been trained using Adam Optimizer [23] ($\beta_1 = 0.5$) with learning rate of LR= 0.0002. The number of iterations for the training was set to be 3000. This number was decided after observing the convergence of loss values during the training. As depicted in Fig. 3 the losses started converging after just 1000 iterations. Both the generator and discriminator were trained in a 1:1 ratio of optimization step. A single iteration consists of one discriminator optimization step followed by one generator optimization step. Due to the unavailability of sufficient GPU vRAM, the batch size was kept to 1.

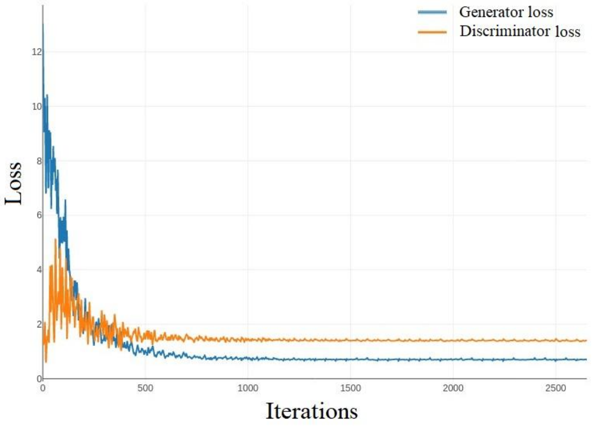

Fig. 3. GAN loss curve during training

5.5 Model Architecture

GAN The complete architecture of the generator and the discriminator has been provided in Table 1. The Generator contains 5 transpose convolution layers which upsample the latent z dimension conditioned on the input speech $\hat{\mathrm{S}}$, to the speech spectrogram dimension. The Discriminator contains 5 convolution layers which classify the speech spectrogram input to it as real or fake.

Binary Classifier The binary classifier is a neural network with ResNetv2-50 [24] architecture and has been trained completely separately from the rest of the model. In a separate pipeline to train the binary classifier, we take a pre-trained ResNet model [25] with its weights frozen for first 7 layers. The weights of the rest of the layers are initialized from scratch and the output dimension of the last fully connected layer is kept to 128. The detailed architecture of the Binary classifier has been depicted in the Supplementary Material.

5.6 Results and Discussion

This section presents the results for the four evaluation-cases that are formulated as per the datasets mentioned in Section 5.3.

Case-1 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 Table 2. As defined in Eq. (1), the Anthropomorphic scores, denoting the good-ness of TTS models in terms of producing human-like speech from text are presented in Fig. 4.

The vMOS scores are in correspondence with the MOS scores. Tacotron and DeepVoice3 have a MOS of 3.82 and 3.62 suggesting Tacotron to be a slightly

Table 1. Architecture details of the Generator and Discriminator

Generator

Layer (kernel-size)	**Stride**	**No. of filters**
ConvTranspose2D (4x4)	1	512
BatchNorm	-	-
ReLU	-	-
ConvTranspose2D (4x4)	2, pad=1	256
BatchNorm	-	-
ReLU	-	-
ConvTranspose2D (4x4)	2, pad=1	128
BatchNorm	-	-
ReLU	-	-
ConvTranspose2D (4x4)	2, pad=1	64
BatchNorm	-	-
ReLU	-	-
ConvTranspose2D (4x4)	2, pad=1	128
Tanh	-	-

Discriminator

Layer (kernel-size)	**Stride**	**No. of filters**
Conv2D (4x4)	2, pad=1	64
LeakyReLU (α=0.2)	-	-
Conv2D (4x4)	2, pad=1	128
BatchNorm	-	-
LeakyReLU (α=0.2)	-	-
Conv2D (4x4)	2, pad=1	256
BatchNorm	-	-
LeakyReLU (α=0.2)	-	-
Conv2D (4x4)	2, pad=1	512
BatchNorm	-	-
LeakyReLU (α=0.2)	-	-
Conv2D (4x4)	2	1
Sigmoid	-	-

Where α the negative slope of LeakyReLU function

better model than DeepVoice3 which is also depicted by the vMOS scores. The vMOS scores for *Natural Speech* reported in Table 2 correspond to the training samples from LJSpeech and VCTK datasets respectively. The difference between them is because of the variations in speech quality while recording human voice.

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.

Table 2. Evaluation of the speech synthesized by TTS models

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

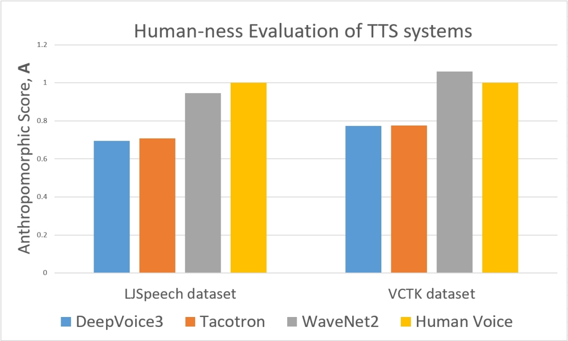

Fig. 4. Anthropomorphic Scores denoting the human-ness of TTS models

Case-2 Emotional speech data of spoken and sung samples Table 3 depicts the vMOS scores for the evaluation of emotion-labelled spoken and sung speech samples from RAVDESS dataset.

Table 3. Evaluation of emotional speech from **RAVDESS** dataset

Emotion Class	**vMOS (speaking)**	**vMOS (singing)**
Angry	0.6380	0.6372
Calm	0.6497	0.6227
Fear	0.6350	0.6390
Happy	0.6183	0.6352
Neutral	0.6457	0.6210
Sad	0.6290	0.6191

Case-3 IEMOCAP data with scripted and improvised samples The MOS scores for scripted and improvised speech samples from IEMOCAP speech dataset are shown in Table 4.

Case-4 Gender specific speech samples This evaluation-case evaluates the speech samples for gender classification. The resulting vMOS scores are shown

Table 4. Evaluation of emotional speech from **IEMOCAP** dataset

Emotion Class	**vMOS (scripted)**	**vMOS (improvised)**
Anger	0.6716	0.6669
Excited	0.6685	0.6654
Frustration	0.6729	0.6839
Happiness	0.6771	0.6688
Neutral	0.6724	0.6629
Sadness	0.6733	0.6600

in Table 5. Though the proposed model was trained on single female speaker dataset, it performed equivalently while evaluating the male speaker's voice.

Table 5. Evaluation of gender-specific speech from RAVDESS & IEMOCAP datasets

Gender	**vMOS (RAVDESS)**	**vMOS (IEMOCAP)**
Male	0.6232	0.6607
Female	0.6352	0.6715

Discussion The evaluator assigns average vMOS of 0.6682 to an actual natural speech rather than a score of 1. This behaviour has been implicitly learned to be the same as how even MOS of real speech is not 5. We observe that the model producing more human-like speech for each word performs better on our evaluation standards. WaveNet2 performs significantly better than its competition and comes very close to how our evaluator judges natural speech.

The first evaluation-case demonstrates the effectiveness of the proposed model to evaluate the human-ness of the synthetic speech. TTS models capable of generating speech output corresponding to various emotional classes are not yet completely developed [26]. For that reason, we have evaluated natural emotional speech samples in case 2, 3 and 4. The vMOS scores computed for various emotion and gender classes ranged from 0.6183 to 0.6839 which is very close to the average vMOS score of natural speech i.e. 0.6682. Although these score showed negligible differences among themselves, the usability of the proposed model to detect the human-ness of the speech samples involving variations in terms of emotional and gender-specific information has been demonstrated.

6 Conclusion

In this paper, we present an automatic speech evaluation method capable of assessing the speech generated by TTS models as effectively as manual approaches such as - MOS, paired-comparison tests, etc. On evaluating the speech generated by benchmark TTS models, i.e. Wavenet2, Tacotron and Deepvoice3, the proposed method computed the vMOS as 0.6675, 0.4945 and 0.4890 respectively. It was also able to analyse the goodness of a TTS model in terms of average Anthropomorphic Scores of 0.7338, 0.7417 and 1.0017 for DeepVoice3, Tacotron and WaveNet2 respectively where a score of 1 corresponds to completely naturalistic speech synthesis. 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. It can be improved further by training with multi-speaker and emotional speech datasets. 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 focus on conditioning the existing TTS systems for emotional speech synthesis by incorporating a new evaluation score, along-with the already proposed score of human-ness, that could measure the correctness of desired emotion in the output speech.

References

1. Sercan Ö Arik, Mike Chrzanowski, Adam and Diamos, Gregory and Gibiansky, Andrew and Kang, Yongguo and Li, Xian and Miller, et al. Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning* (ICML). **70**, pp. 195–204. (2017).
2. Yuxuan Wang, RJ Skerry-Ryan, Daisy and Wu, Yonghui and Weiss, Ron J and Jaitly, Navdeep and Yang, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).
3. Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*. (2017).
4. Oord, Aaron and Dieleman, Sander and Zen, Heiga and Simonyan, Karen and Vinyals, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
5. Pier Luigi Salza, Enzo Foti, Luciano Nebbia, and Mario Oreglia. MOS and pair-comparison combined methods for quality evaluation of text-to-speech systems. *Acta Acustica united with Acustica*. **82**(4), pp. 650–656. (1996).
6. Pravin Ghate and S D Shirbahadurkar. A survey on methods of tts and various test for evaluating the quality of synthesized speech. *International Journal of Development Research*. **07**, pp. 15236–15239. (2017).
7. Keith Ito. The LJ Speech dataset. www.keithito.com/LJ-Speech-Dataset. (2017).
8. Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research* (CSTR). (2017).
9. Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *Public Library of Science* (PLOS). **13**(5), (2018).
10. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*. **42**(4), p. 335. (2008).

11. Raghav C Dwivedi, Suzanne St. Rose, Edward J Chisholm, Peter M Clarke, et al. Acoustic parameters of speech: Lack of correlation with perceptual and questionnaire-based speech evaluation in patients with oral and oropharyngeal cancer treated with primary surgery. *Head & neck.* **38**(5), pp.670–676. (2016).
12. Richard J Sebastian and Ellen Bouchard Ryan. Speech cues and social evaluation: Markers of ethnicity, social class, and age. In *Recent advances in language, communication, and social psychology.* pp. 112–143. (2018).
13. Vivek Tyagi, Aravind Ganapathiraju, and Felix Immanuel Wyss. Method and system for selectively biased linear discriminant analysis in automatic speech recognition systems. US Patent 9679556. (June 2017)
14. Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems.* pp. 577–585. (2015).
15. Bojana Gajic, Shrikanth Narayanan, Sarangarajan Parthasarathy, Richard Rose, and Aaron Rosenberg. System and method of performing user-specific automatic speech recognition. US Patent 9058810. (June 2015)
16. Steve Renals and Automatic Speech Recognition-ASR Lecture. Multilingual speech recognition. (2017).
17. M Russell and A Cook. Experimental evaluation of duration modelling techniques for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* **12**, pp. 2376–2379. (1987).
18. Alastair H Moore, P Peso Parada, and Patrick A Naylor. Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures. *Computer Speech & Language.* **46** pp. 574–584. (2017).
19. Takeshi Otani, Taro Togawa, and Sayuri Nakayama. Speech evaluation apparatus and speech evaluation method. US Patent App. 15/703,249. (March 2018).
20. Hardik B Sailor and Hemant A Patil. Fusion of magnitude and phase-based features for objective evaluation of TTS voice. In *The 9th IEEE International Symposium on Chinese Spoken Language Processing.* pp. 521–525. (2014).
21. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems.* pp. 2672–2680. (2014).
22. Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
23. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations* (ICLR). (2015).
24. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computing Research Repository* (CoRR). (2015).
25. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR). pp. 770–778. (2016).
26. Shumin An, Zhenhua Ling, and Lirong Dai. Emotional statistical parametric speech synthesis using LSTM-RNNs. In *2017 IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA-ASC). pp. 1613–1616. (2017).