DSA4199: Deep Learning for Evaluation of Synthetic Speech

Li Yang Hew

2024-02-06

Table of contents

Preface

⚠ Warning

This page is a work in progress

Project

The focus of this project is on automated evaluation of synthetic speech, ie: **Text-To-Speech (TTS)** models.

It aims to cover the following non-exhaustive list of topics in a gentle but technical manner:

1. Audio processing in general
2. How deep learning is applied to audio waveforms for different downstream tasks.
3. Existing literature on SoTA TTS models
    - Model architecture (spectrogram feature extractors [CNN], GPT, Transformers, Diffusers, Vocoders)
4. Existing literature on automated TTS evals
    - Types of metrics (MoS, prosody, naturalness)
    - Datasets (NISQA, MOSNet)
    - Model training & inference
    - Evaluation
5. Review of contributions & conclusion

Typesetting

This **Quarto** book serves as living documentation, which should later turn into a nicely formatted PDF.

Using Quarto over Overleaf

Quarto is able to export all this markup to **TeX** and then to a PDF document automatically thanks to Pandoc.
Not only do I get to write **Markdown**, I'm also able to version control everything using Git & automate publishing to both PDF & static HTML (this website) upon push to `main` with this GitHub Action

1 Introduction

1.1 Background

In comtemporary academia, the burgeoning field of synthetic speech synthesis has garnered substantial attention and interest owing to its multifarious applications spanning from human-computer interaction, such as voice powered ChatGPT, Google Home and Alexa to assistive technologies, like the Rabbit R1. Within this domain, the imperative for automated evaluations arises due to several paramount reasons.

Firstly, the proliferation of Text-To-Speech (TTS) systems across diverse sectors necessitates a systematic means of evaluating their performance. Different downstream use cases demand optimizing for different metrics. Long form speech generation such as audiobook readers require natural flow and consistent prosody, whereas voice cloning requires evaluating intonation, cadence and emotional nuances. In applications like virtual assistants and navigation systems, intelligibility, clarity and prompt delivery are paramount.

Secondly, these nuances are not directly present at the optimization step during training. Without jumping the gun on terminology, TTS models in general are trained on mel-spectrograms with the goal of minimizing their reconstruction loss, more formally known as *spectral accuracy*. This is typically computed by Mean Squared Error (MSE or $L2$ loss) or Mean Absolute Error (MAE or $L1$ loss). These metrics do not contain any information on the previously mentioned nuances in speech but merely how well it approximates the training data. This phenomenon gives rise to the need for manual assessments by human judgement. For example, the Mean Opinion Score (MoS) serves as a popular metric employed for this purpose. It operates on the premise of perceived quality of synthesized speech samples, typically on a Likert scale ranging from 1 (poor) to 5 (excellent).

It is no surpise then that the existing process of developing TTS systems is beset by the limitations inherent in manual evaluations. Primarily, they are labor-intensive, required significant time and resources to collect & analyze. This approach not only imposes constraints on the scalability of evaluation efforts but also introduces biases and variability stemming from inter-rater differences in subjective perception.

However, the recent advancements in deep learning such as the advent of **Autoregressive Transformers**, **Self Attention** and **Diffusers** has accelerated research on data-driven approaches to predict these subjective quality metrics with remarkable accuracy, including Lo et al. (2019) and Mittag et al. (2021). The significance of reliable evaluation systems can alleviate the reliance on manual human assessments but also offer a scalable and objective means of evaluating TTS systems.

1.2 Analog to Digital

2 Text To Speech Architectures

2.1 Tacotron 1 & 2

2.2 Speech T5

2.3 VITS

2.4 Tortoise TTS

3 Related Work

3.1 MOSNet

3.2 NISQA

4 Experimental Setup

4.1 Dataset Curation

We compile the commercially usable subset of the overarching NISQA training set as follows.

1. **Blizzard Challenge (2008 - 2023)**

The Blizzard Challenge is a TTS competition overseen by SynSIG, a special interest group of ISCA, the International Speech Communication Association.

Every year, research groups from numerous global institutions undertake the task of training a model to synthesise speech in various languages. The dataset provided by the challenge are often donated by external organisations, subject to an agreed level of licensing.

Generated speech submitted by participants are then evaluated via systematic subjective tests by a mixture of paid and volunteer human listeners. In particular, the subjects are asked to rate the quality of each synthesized sample according to a set of pre-defined criterion, including, speaker similarity, naturalness (MOS) and intelligibility (WER). This collection of these three metrics are largely consistent across the years.

A summary of the final dataset is described in the following table.

Source	Years	Comments	Metrics
Blizzard Challenge	2008 -> 2023	Only commercial allowed subset	`mos`, `wer`, `speaker_sim`

Additional, we note that there is a significant amount of available data not included in our dataset due to restrictions on commercial usage. These include Maniati et al. (2022) with over 300 thousand listener ratings.

5 Results

6 Conclusion

References

Lo, Chen-Chou, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion.” In *Interspeech 2019*. Interspeech_2019. ISCA. https://doi.org/10.21437/interspeech.2019-2003.

Maniati, Georgia, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. “SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis.” In *Proc. Interspeech 2022*, 2388–92. https://doi.org/10.21437/Interspeech.2022-10922.

Mittag, Gabriel, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets.” In *Interspeech 2021*. Interspeech_2021. ISCA. https://doi.org/10.21437/interspeech.2021-299.