WildSpoof - TTS Track

As per submission link

Team (TCS Speech)

Parth Khadse, Sunil Kumar Kopparapu

TCS Research - Mumbai

System specifications for training TTS

- CPU : Intel core i7-8700 @ 3.20GHz
- RAM : 62 GB
- GPU : Nvidia GeForce RTX 3090 (VRAM : 24GB)

Model description

We trained a **VITS-based Text-to-Speech** (TTS) architecture to synthesize speech from text. To represent speaker identity, we developed a separate speaker encoder that generates a 256-dimensional embedding for each unique speaker.

Additionally, we use embeddings extracted from a pretrained **YAMNet** model (an audio event detection pre-trained model) to encode noise (non-speech) characteristics in the audio sample used for training.

The model was trained on the TITW-Easy part of the dataset.

During inference, we provide

1. Text to be synthesised
2. The speaker embedding corresponding to the target speaker, and
3. The YAMNet embedding (we construct a 521-D synthetic representation)
4. values for non-speech labels are randomly sampled within the range [0, 0.5),
5. while the speech label is manually set to 1.
6. The entire embedding vector is then normalized so that its elements sum to 1.

Submission

The synthesized voice of TITW-KSKT and TITW-KSUT test set (16 kHz) attached

In zip file (Sample)

TITW-KSKT

```
45612  2025-12-02 11:36   TCSSPEECH/TITW-KSKT-200/TITW-KSKT_id10293_3.wav
73772  2025-12-02 11:36   TCSSPEECH/TITW-KSKT-200/TITW-KSKT_id10276_4.wav
26156  2025-12-02 11:36   TCSSPEECH/TITW-KSKT-200/TITW-KSKT_id10274_2.wav
49196  2025-12-02 11:36   TCSSPEECH/TITW-KSKT-200/TITW-KSKT_id10308_3.wav
```

TITW-KSUT

```
59436  2025-12-02 11:36   TCSSPEECH/TITW-KSUT-200/TITW-KSUT_id10302_1.wav
```

```
73260  2025-12-02 11:36   TCSSPEECH/TITW-KSUT-200/TITW-KSUT_id10307_2.wav
76332  2025-12-02 11:36   TCSSPEECH/TITW-KSUT-200/TITW-KSUT_id10299_1.wav
82988  2025-12-02 11:36   TCSSPEECH/TITW-KSUT-200/TITW-KSUT_id10298_2.wav
```

2025-12-02 (15:48) Last updated