





esnya / **japanese_speecht5_tts**   like 18

 Text-to-Speech

 Transformers

 Safetensors

 Japanese


speecht5


text-to-audio


jvs


pyopenjtalk


speech-to-text





 Train ▾


 Deploy ▾

 Use this model ▾

 **Model card**

 Files

 xet

 Community **5**

Downloads last month
270



 **Safetensors** 

Model size 0.1B params Tensor type F32  Files info

Inference Providers NEW

 Text-to-Speech

This model isn't deployed by any Inference Provider.

 Ask for provider support

Model tree for esnya/japanese_speecht5_tts

Quantizations

1 model

 Space using esnya/japanese_speecht5_tts 1

SpeechT5 (TTS task) for Japanese

SpeechT5 model fine-tuned for Japanese speech synthesis (text-to-speech) on JVS. This model utilizes the JVS dataset which encompasses 100 speakers. From this dataset, speaker embeddings were crafted, segregating them based on male and female voice types, and producing a unique speaker embedding vector. This 16-dimensional speaker embedding vector is designed with an aim to provide a voice quality that is independent of any specific speaker.

Trained from microsoft/speecht5_tts. Modified tokenizer powered by Open Jtalk.

🔗 Model description

See [original model card](#) My modified codes licensed under MIT Licence.

🔗 Usage

Install requirements

```
pip install transformers sentencepiece pyopenjtalk # or pyopenjtalk-prebuilt
```

Download a modified code.

```
curl -O https://huggingface.co/esnya/japanese_speecht5_tts/resolve/main/speecht5_c
```

(SpeechToTextPipeline is not released yet.)

```
import numpy as np
from transformers import (
    SpeechT5ForTextToSpeech,
    SpeechT5HifiGan,
    SpeechT5FeatureExtractor,
    SpeechT5Processor,
)
from speecht5_openjtalk_tokenizer import SpeechT5OpenjtalkTokenizer
import soundfile
import torch

model_name = "esnya/japanese_speecht5_tts"
with torch.no_grad():

    model = SpeechT5ForTextToSpeech.from_pretrained(
        model_name, device_map="cuda", torch_dtype=torch.bfloat16
    )

    tokenizer = SpeechT5OpenjtalkTokenizer.from_pretrained(model_name)
    feature_extractor = SpeechT5FeatureExtractor.from_pretrained(model_name)
```

```

processor = SpeechT5Processor(feature_extractor, tokenizer)
vocoder = SpeechT5HifiGan.from_pretrained(
    "microsoft/speecht5_hifigan", device_map="cuda", torch_dtype=torch.bfloat16
)

input = "吾輩は猫である。名前はまだ無い。どこで生れたかとんと見当がつかぬ。"
input_ids = processor(text=input, return_tensors="pt").input_ids.to(model.device)

speaker_embeddings = np.random.uniform(
    -1, 1, (1, 16)
) # (batch_size, speaker_embedding_dim = 16), first dimension means male (-1.0)
speaker_embeddings = torch.FloatTensor(speaker_embeddings).to(
    device=model.device, dtype=model.dtype
)

waveform = model.generate_speech(
    input_ids,
    speaker_embeddings,
    vocoder=vocoder,
)

waveform = waveform / waveform.abs().max() # normalize
waveform = waveform.reshape(-1).cpu().float().numpy()

soundfile.write(
    "output.wav",
    waveform,
    vocoder.config.sampling_rate,
)

```

[🔗](#) Background

The motivation behind developing this model stems from the noticeable lack of Japanese generation models in SpeechT5 TTS, or their scarcity at best. Additionally, the g2p functionality of Open Jtalk (pyopenjtalk) enabled us to achieve a vocabulary closely resembling English models. It's important to note that the special modifications and enhancements were primarily applied to the tokenizer. Unlike the default setup, our modified tokenizer separately extracts and retains characters other than phonation to ensure more accurate text-to-speech conversion.

[🔗](#) Limitations

One known issue with this model is that when multiple sentences are fed into it, the latter parts may result in extended silences. As a temporary solution, until this is rectified, it is recommended to split and generate each sentence individually.

[🔗](#) License

Model inherits [JVS Corpus](#).

[🔗](#) See also

- Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint, 1908.06248, Aug. 2019.