VoiceSplit

Targeted Voice Separation by Speaker Conditioned Spectrogram

Partial Report

SCC0251/5830 Image Processing

Edresson Casanova - 11572715
Pedro Regattieri Rocha - 8531702

1. Introduction

VoiceSplit's goal is the development of a system that, given an audio input, is able to separate overlapping voices through the use of Mel Spectrograms, based on the characteristics of each speaker's speech patterns. With this the system will be able to separate people having a conversation into different entries to help populate data sets.

VoiceSplit's neural architecture is based on VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking, a paper proposed by Google researchers. This paper uses the Librispeech data set, which will also be used for this project, along the Speech2Phone and Voxceleb 1 & 2 data sets, also used in the Google paper.

The initial step is the use of the normalise-resample.sh script to convert .flac lossless sound files into .wav, single channel mono sound files that will be used by VoiceSplit.

The process for data processing is the same as outlined in VoiceFilter, using the preprocess_by_csv.py script that reads .csv files and creates a data set from it using the pandas library.

After that, we remove the initial and final silence from the files, and, as is the case for VoiceFilter, only three seconds of each audio is considered. Furthermore, the audio in these three seconds contains the overlapping voices of the speakers in the file.

The expected audio, the predicted audio and the embedding reference are then saved. VoiceSplit supports three different embedding systems: GE2E CorentinJ, a model trained with the Librispeech and both Voxceleb data sets; GE2E Seungwonpark, used during our initial training, which was trained using only Voxceleb 2 and Speech2Phone, which contains one hour of audio featuring 40 different speakers.

We used MSE as our loss function so we could start training our models as soon as possible, as training time is the bottleneck of our project.

We had access to a V100 machine for training purposes, allocated to us for one week. However, due to issues with data pre-processing, the long time needed to train each model, and the limited time frame we had access to the V100, ultimately only two days worth of training were performed using the V100 before we had to downgrade to a Nvidia GTX 1080, slowing down the training process.

Once the implementation of Powerlaw Compression Loss function is complete, it will replace MSE as the loss function in further training.

Using the V100, we were able to perform 820 thousand steps with a batch size of 24, for a total of 19.680.000 iterations of our training.

2. Sample Images of the Spectrogram Processing

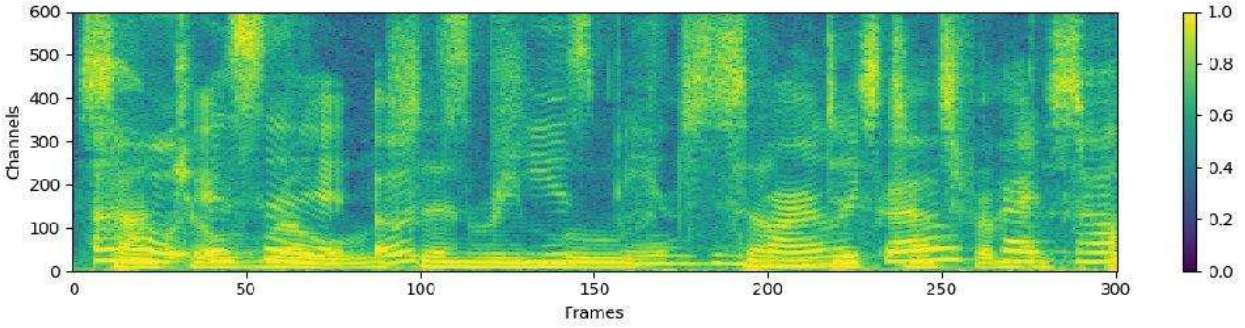

Figure 1: Processed .wav file spectrogram, with overlapping speakers

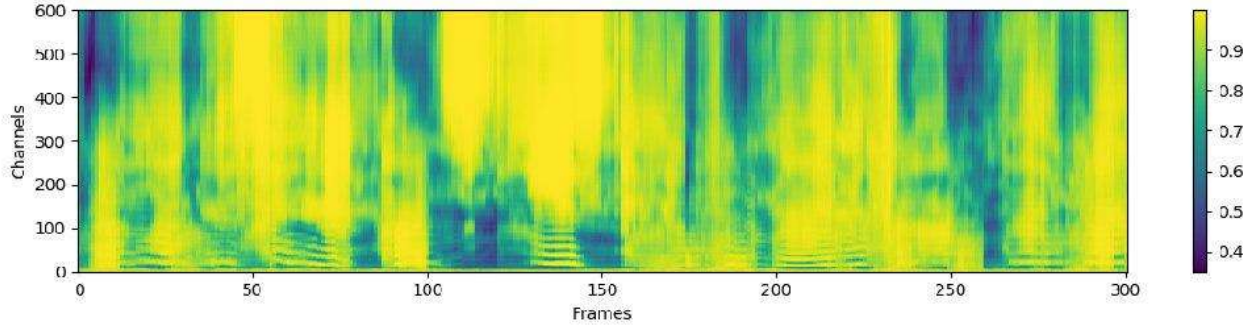

Figure 2: Predicted Mask spectrogram, used to separate the speakers.

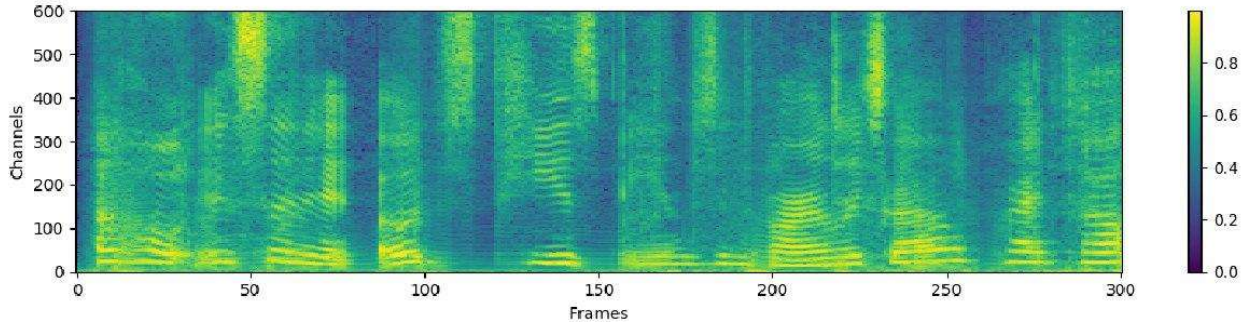

Figure 3: Predicted Audio spectrogram, a single speaker with minor interference from the other. Obtained by multiplying the overlapping speaker spectrogram by the predicted mask spectrogram.

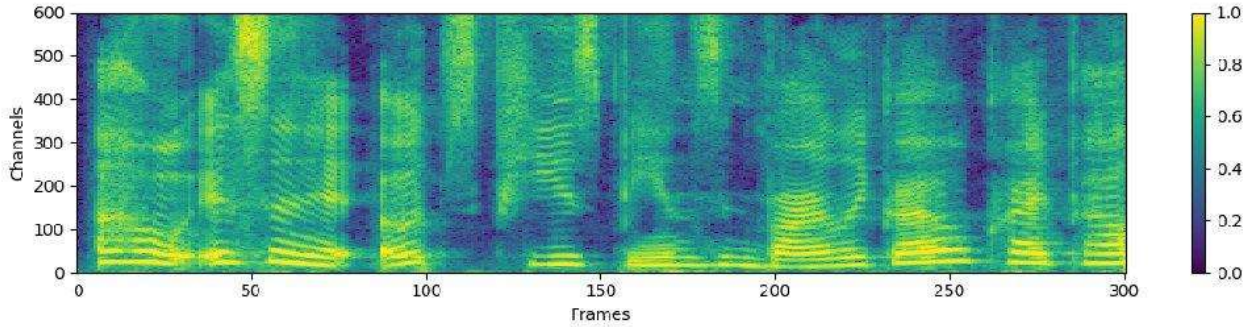

Figure 4: Expected result spectrogram, the spectrogram of a sound file that only has a single speaker, with no interference from another speaker, used as a baseline.

3. Sample Code

The following sample code, the mix_wavfiles function, is a part of generic_utils.py. Its purpose is to extract the spectrogram from an audio sample. It checks if the sound file is three seconds long and also removes silences at the start and end of the sound file, as the presence of silence could cause the model to overfit.

```
def mix_wavfiles(output_dir, sample_rate, audio_len, ap, form, num,
embedding_utterance_path, interference_utterance_path, clean_utterance_path):
    data_out_dir = output_dir
    emb_audio, _ = librosa.load(embedding_utterance_path, sr=sample_rate)
    clean_audio, _ = librosa.load(clean_utterance_path, sr=sample_rate)
    interference, _ = librosa.load(interference_utterance_path, sr=sample_rate)
    assert len(emb_audio.shape) == len(clean_audio.shape) == len(interference.shape) ==
1, \
        'wav files must be mono, not stereo'

    # trim initial and end  wave file silence using librosa
    emb_audio, _ = librosa.effects.trim(emb_audio, top_db=20)
    clean_audio, _ = librosa.effects.trim(clean_audio, top_db=20)
    interference, _ = librosa.effects.trim(interference, top_db=20)

    # calculate frames using audio necessary for config.audio['audio_len'] seconds
    audio_len_seconds = int(sample_rate * audio_len)

    # if merged audio is shorter than audio_len_seconds, discard it
    if clean_audio.shape[0] < audio_len_seconds or interference.shape[0] <
audio_len_seconds:
        return

    clean_audio = clean_audio[:audio_len_seconds]
    interference = interference[:audio_len_seconds]
    # merge audio
    mixed_audio = clean_audio + interference

    # normlise audio
    norm_factor = np.max(np.abs(mixed_audio)) * 1.1
    clean_audio = clean_audio/norm_factor
    interference = interference/norm_factor
    mixed_audio = mixed_audio/norm_factor

    # save normalized wave files and wav emb ref
    target_wav_path = glob_re_to_filename(data_out_dir, form['target_wav'], num)
    mixed_wav_path = glob_re_to_filename(data_out_dir, form['mixed_wav'], num)
    emb_wav_path = glob_re_to_filename(data_out_dir, form['emb_wav'], num)
    librosa.output.write_wav(emb_wav_path, emb_audio, sample_rate)
    librosa.output.write_wav(target_wav_path, clean_audio, sample_rate)
    librosa.output.write_wav(mixed_wav_path, mixed_audio, sample_rate)

    # extract and save spectrograms
    clean_spec = ap.get_spec_from_audio_path(target_wav_path) # we need to load the wav
to maintain compatibility with all audio backend
    mixed_spec = ap.get_spec_from_audio_path(mixed_wav_path)
    clean_spec_path = glob_re_to_filename(data_out_dir, form['target'], num)
    mixed_spec_path = glob_re_to_filename(data_out_dir, form['mixed'], num)
    torch.save(torch.from_numpy(clean_spec), clean_spec_path)
    torch.save(torch.from_numpy(mixed_spec), mixed_spec_path)
```