Adjustable Singing Synthesis Using Machine Learning

Using DDSP to learn and synthesize singing with adjustable latent space parameters

Harry Twigg (BEng Aeronautics and Astronautics) Supervisor Thomas Blumensath

Differentiable Digital Signal Processors are differentiable versions of traditional signal processing techniques, allowing them to be used in a machine learning network

A machine learning network based on DDSP was tasked with recreating singing samples, DDSP is a modular deep learning based architecture, enabling characteristics of the output signal to be explicitly adjusted (e.g loudness and pitch)

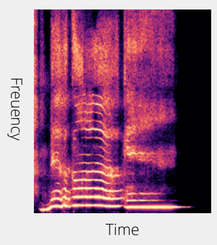
Timbral transfer was also demonstrated between musical instruments (e.g. making a flute sound like a violin). DDSP was originally designed with instruments in mind, however this research focuses on its application to singing

A targeted approach to synthesising singing using DDSP was devised, guiding a modified autoencoder based model to the use of explicitly defined fundamental frequency (F0) and loudness, as opposed to just obscure encoder outputs

Training dataset of only vocals

A 4-second sample has its spectrogram taken and fed into the encoder

A spectrogram is a 2D decomposition of an audio sample, brighter regions indicate loudness, spectrograms enable sounds to be used by image based deep learning technques



Each of the datasets were over 700Mb in size, and countained 2 albums worth of songs this was to help provide generality to overcome pronunciation difficulties

Autoencoder extracts key musical information

A pretrained model called CREPE was used to detect the fundamental frequency and confidence at each timestep

Obscure network outputs from the encoder are passed directly into the decoder (latent encoding)

Loudness at each timestep was statistically determined

Fundamental Frequency is passed directly to the harmonic oscillator, it is unnecessary to make the decoder learn it

Decoder is adapted to detect vocal features

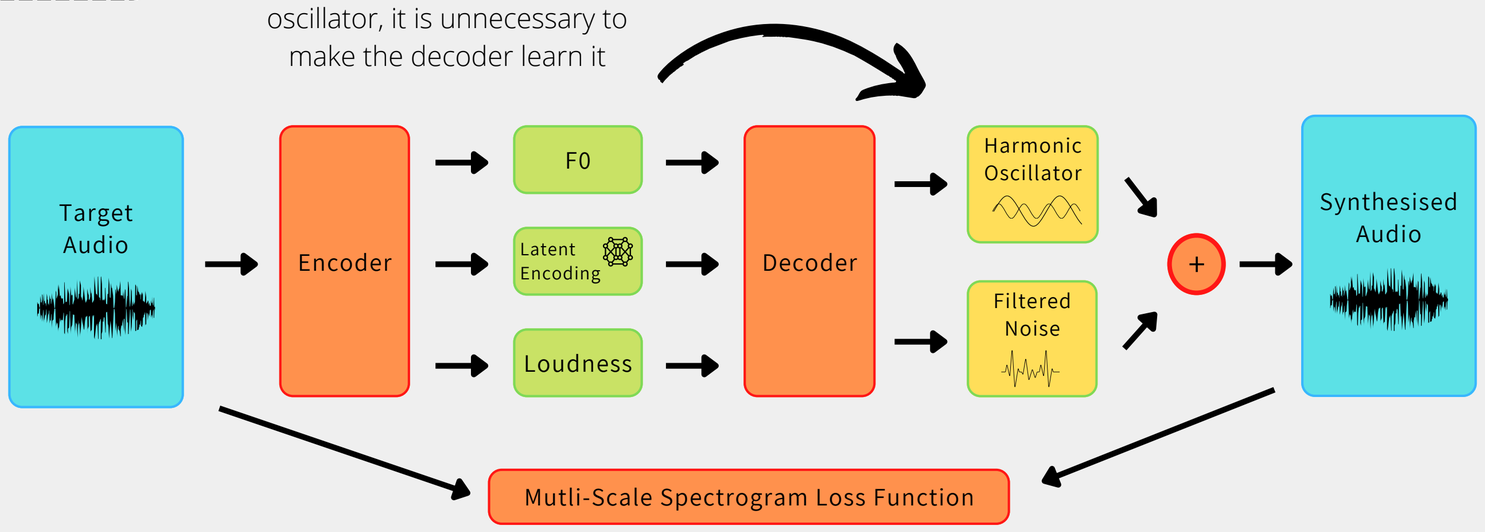
The decoder takes F0, loudness, and latent encoding, attempting to construct suitable inputs for the DDSP modules

The decoder is designed to learn Mel Frequency Cepstral Coefficients, (MFCCs), these capture the non-linearity of the human voice, helping the model to synthesize words and phonetics accurateley

Decoder outputs are inputs of the DDSP Modules

The Harmonic Oscillator module, outputs the sum of a bank of sinusoidal oscillators, each at an integer multiple of the fundamental frequency, it was hypothesised it would model harmonics of the human voice

The Filtered Noise Bank module, an LTI-FIR filter-bank that varies over time, takes white noise and outputs a noise signal, it was hypothesised it would model consonants of the human voice

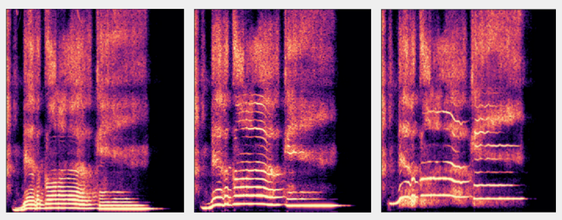


Outputs from both synthesizers are summed. Synthesised sound is outputted in the form of raw audio, this can be listened to or plotted on a spectrogram

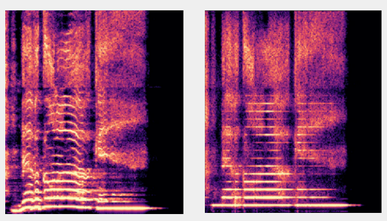
The model was tasked with minimising reconstruction loss. An L1 loss measure was used, this is the sum of the absolute differences between the magnitude spectrograms. This was necessary as 2 wave-forms could sound the same but have different point-wise characteristics

Results

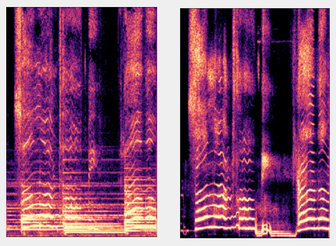
- All lyrics were mostly clear and understandable ✓
- Pitch transposition was successful ✓



- Constant F0 inference featured accurate pitch holding, however, outputs sounded slightly robotic ✓



- Separation of vocals from original track with instrumentals was possible showing specificity of training to singing ✓



- Generality outside of the training datasets was demonstrated ✓
- Timbral transfer (the uniqueness of an artist's voice) was not possible ⊗
- Control over loudness of the output signal did not yield differences in the output (loudness latent had no effect) ⊗

Conclusions

- DDSP has demonstrated itself as a powerful tool for learning and synthesising the vocal features of the human voice
- Discarded information such as phase should be incorporated into a future model to enable lossless synthesis
- Failed objectives such as timbral transfer and inability to modify loudness should be looked at in any future work
- Native phonetic conditioning could be implemented into DDSP, enabling modification of the underlying words and phonetics as latents similarly to F0 or loudness. This would open the way to text to speech using DDSP
- Similarly, the modular nature of DDSP should be used to integrate the technology into existing musical systems e.g. MIDI

Contact Email:
harrytwigg111@gmail.com

Listen to the results yourself!