Simulating Dysarthria using Parallel Neural Voice Conversion with Mel-spectrograms

B157108, 9222 words

THE UNIVERSITY
of EDINBURGH

Master of Science

Speech and Language Processing

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2021

Abstract

Large datasets of healthy and dysarthric speech are required to train voice repair models such as those developed at SpeakUnique. The rarity of dysarthria in the population and the expense of collecting audio data limit the amount of available training data. This dissertation presents an investigation into synthesising dysarthric data using neural parallel voice conversion techniques, with Mel-spectrograms as spectral representations. Fully connected and convolutional conversion models are first explored, then speaker embeddings are added and a multi-task training procedure using electromagnetic articulography readings is investigated. Listening tests conducted with an expert evaluator indicate that, while some speaker characteristics are successfully converted, converted samples do not sound clearly dysarthric, potentially due to high levels of noise and distortion.

Acknowledgements

First and foremost I'd like to thank my supervisors Cassia Valentini-Botinhao, Oliver Watts and Lovisa Wihlborg for their continual support and guidance. Without them I would have tried to train a GAN. Thanks to Patricia for being a great colleague and friend. Thanks to Lynda Tomarelli for lending her expertise to my evaluations.

Thanks to Simon King for his kindness and reassurance over the past two years. Thanks to Jacob for all the coffee. And thanks to my mum, for everything.

Table of Contents

Chapter 1

Introduction

Dysarthria is a speech disorder resulting from damage to neural motor mechanisms, found in patients with, for example, Motor Neurone Disease (MND) (also called Amyotrophic Lateral Sclerosis (ALS)) and Cerebral Palsy (CP). Dysarthria is characterised by perceptually recognisable abnormalities in speech articulation, described in section 2.1.1. Many people living with dysarthria use assistive technologies and communication aids, including text-to-speech software. Companies including SpeakUnique provide personalised text-to-speech voices for people living with MND. Creating these voices requires recordings of a client's speech, made either during the early stages of the disease while speech is relatively unaffected, or using voice repair techniques to restore a client's dysarthric speech to sound like themselves pre-MND. Voice repair techniques require large datasets of healthy and dysarthric speech data, which is expensive and difficult to collect due to the rarity of the disorder. The experiments I present in this dissertation attempt to simulate dysarthric speech characteristics, so that fake dysarthric speech data can be created from healthy data and used to train better voice repair models.

Many methods of dataset augmentation have previously been explored, including using signal processing methods to modify the speaking rate, pitch contours, fundamental frequency (F0) of samples [1]. While these are helpful in expanding the size and variability of a training dataset, dysarthric training datasets have the additional problem of few available dysarthric speakers. This problem, a consequence of the rarity of dysarthria and the expense of collecting speech data, results in datasets of dysarthric speakers which have limited variation in speaker characteristics such as age and regional accent. Since these speaker characteristics cannot be easily modified by signal processing methods, different approaches are required. While many avenues of research could be explored to address the problem of dysarthric speaker augmentation, including dysarthric text-to-speech, the project presented here explores deep neural parallel voice conversion models for transforming healthy to dysarthric speech. I took a voice conversion

approach because text-to-speech systems produce speech somewhat deterministically, whereas voice transformation allows for many varied input samples for a given utterance, and preserves some of this variation in the output. Under this approach, any high quality recordings of healthy speech could be transformed to dysarthric speech, providing essentially unlimited amounts of dysarthric training data. The main factors influencing the usefulness of this transformed speech as voice repair training data are the accuracy of synthetic dysarthric features, the naturalness of synthesised speech, and the audio quality.

In this dissertation I present five experiments on healthy-to-dysarthric speech conversion using neural networks to learn mapping functions between pairs of healthy and dysarthric Mel-spectrograms. I consider the questions:

- Can a framewise healthy-to-dysarthric Mel-spectrogram mapping function be learned?
- What effect does the presence of fundamental frequency (F0) have on the ability to learn such a mapping?
- Can speaker embeddings resolve F0 differences between target and source speakers?
- Can multi-task training using electromagnetic articulography (EMA) features improve dysarthric Mel-spectrogram predictions?

I begin by providing a technical background on dysarthria, Mel-spectrograms, and parallel voice conversion in Chapter 2, along with an overview of previous work on dysarthric voice conversion and Mel-spectrogram conversion. Chapter 3 describes the TORGO database used in experiments, and the data pre-processing steps I carried out. In Chapter 4 I present five voice conversion experiments. The first checks whether the dataset and healthy-to-dysarthric transformation task meet the assumptions of parallel voice conversion. The latter four describe and evaluate four parallel voice conversion models. I begin with fully-connected and fully-convolutional conversion models. I then add speaker embeddings to alleviate issues caused by differences in F0 between source and target speakers. Finally, I follow a multi-task training procedure using electromagentic articulograph readings with the aim of improving dysarthric prediction accuracy. In Chapter 6 I evaluate the two most successful models using subjective listening tests with an expert Speech and Language Therapist. Finally, I end with a discussion of the project's successes, limitations, and avenues for future research.

Chapter 2

Background

2.1 Technical background

2.1.1 Dysarthria

Dysarthria describes speech characteristics caused by decreased motor control resulting from neurological injury to the muscles or motor mechanisms which control speech production. Such injuries can be caused by degenerative disease (e.g. Motor Neurone Disease), abnormal brain development caused by injury (e.g. Cerebral Palsy), traumatic brain injury, or stroke.

Dysarthria ranges in severity, but speech features typically symptomatic of dysarthria include: hypernasality (greater nasal emission of air); imprecise consonant articulation; reduced vowel space; increased vowel duration; voice onset timing delays; general speech rate decrease and segment lengthening; decreased fundamental frequency (F0) variability; low volume resulting from respiratory weakness; breathiness; harshness or raspiness; stuttering; phoneme repetition, and unstable phonation. Speech sounds involving the tongue and larynx in articulation are especially affected by impaired motor control [2] [3].

2.1.2 Mel-spectrograms

Mel-spectrograms are two-dimensional time-frequency representations of sound, the result of multiplying the magnitude short time Fourier transform by a Mel-scaled nonlinear windowing function. The Mel-scale is inspired by human auditory frequency resolution, which is more discriminative of lower sound frequencies, which transmit the most content information in speech. Mel-spectrograms encode spectral frequency information representing the position of speech articulators over time.

In contrast to the more commonly used spectral representation Mel-frequency cepstral coefficients (MFCCs), Mel-spectrogram features are not decorrelated from fundamental frequency (F0), and so all features at a given time-step covary. My use of Mel-spectrograms as a feature representation was motivated by two things. Firstly, recent high-quality neural vocoders like HiFi-GAN [4] and MelGAN [5] use Mel-spectrograms as their intermediate representation. As described in section 2.2.1, performing voice transformation directly on these intermediate representations should allow for better quality resynthesis compared to older vocoding methods which use MFCCs as spectral representations, like STRAIGHT or Griffin-Lim. The naturalness and audio quality of synthesis is important if it will be used to train downstream models. Secondly I expect that, because they encode F0 information, training conversion models directly on Mel-spectrograms will model F0 effects, such as the reduced F0 variation which gives dysarthric speech its monopitch quality, jointly along with spectral features.

2.1.3 Parallel voice-conversion

Voice conversion techniques aim to convert speaker identity (voice characteristics) from a source speaker to a target speaker, while retaining the linguistic content (the words). A voice conversion module can then be understood as a mapping function[6]:

$$y = F(x)$$

The task I present in this dissertation is not voice conversion in the strictest sense, as it aims to modify some of those linguistic features which are affected by dysarthria, such as phoneme realisations, alongside spectral and speaker characteristics. Since this kind of conversion has been less-widely reported than traditional speaker identity conversion, the experiments presented here take inspiration from both research areas.

Voice conversion techniques can be broadly split into parallel and non-parallel. Parallel voice conversion techniques require parallel datasets, with pairs of source and target speakers saying the same utterances. These pairs must be time-aligned, removing durational differences so that the representation of each at a given time step corresponds to the same linguistic “thing”, e.g. the same point in a word. Non-parallel techniques do not require time-aligned data. In this project I employ parallel techniques as they are generally simpler and their efficacy will allow us to gauge the complexity of the conversion task. I employ deep learning conversion techniques using neural network models because they can easily approximate non-linear mappings between source and target speaker [6] for complex feature representations like Mel-spectrograms. For a detailed overview of other modelling approaches to parallel voice conversion, including both parametric and non-parametric statistical approaches, see [6].

2.2 Previous work

2.2.1 Healthy-to-dysarthric transformation

Healthy-to-dysarthric voice conversion for the purpose of training data augmentation has been previously investigated. Jiao et al. [7] take an approach incorporating both deep learning and signal processing methods. Waveforms of healthy speech were uniformly stretched to slow speaking rate. Dysarthric spectral features (Mel-cepstral coefficients (MCEPs) and band-aperiodicity parameters (BAPs)), were predicted from the extracted healthy ones using a Deep Convolutional Generative Adversarial Network (DC-GAN). The authors expected MCEPs would encode the precision of phoneme articulation, nasality, and breathiness; BAPs would encode harsh vocal quality. Fundamental frequency (F0) was linearly transformed to model reduced pitch variation. These features were then recombined using STRAIGHT vocoder. Subjective evaluations of the synthesised output highlighted the success of this approach: five speech and language therapists (SLTs) classified generated dysarthric speech as ALS 65% of the time, and 76% of samples which attained consensus from all SLTs were classified as ALS. The naturalness and audio quality of the generated samples was unclear: one SLT described some samples as having an "unnatural/robotic/tinny quality".

Waveform generation using a vocoder is a crucial part of voice conversion. We hope that using a recent neural vocoder (HiFi-GAN) for resynthesis will produce more natural results. While no direct comparative study of HiFi-GAN and STRAIGHT has been carried out, a subjective evaluation comparison of STRAIGHT with three neural vocoders (WaveNet, WaveRNN, and HiNet) showed the three neural vocoders produced synthesis with significantly better Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) scores than those vocoded by STRAIGHT [8]. In a comparative evaluation of HiFi-GAN with neural vocoders WaveNet, WaveRNN and MelGAN, all three HiFi-GAN versions attained higher average MOS scores (3.77, 3.69 and 3.61) on synthesised samples from unseen speakers than the other vocoders (3.52, 3.52 and 3.50, respectively) [4].

HiFi-GAN vocodes speech to Mel-spectrograms, enabling us to experiment with Mel-spectrogram voice conversion. Since Mel-spectrograms contain both spectral and F0 information, directly mapping from source to target Mel-spectrograms enables us to jointly modify these features. Rather than linearly transforming F0 as [7] did, I can forego the error-prone task of approximating this transformation function and instead assume a Mel-spectrogram based neural model will learn this transformation.

2.2.2 Mel-spectrogram voice conversion

Voice conversion using Mel-spectrograms as spectral representations has recently been studied. MelGAN-VC [5] is a voice conversion model which performs Mel-spectrogram source-to-target translation on non-parallel datasets using a DC-GAN. Chuks of a source spectrogram are fed into a generator, and consecutive pairs of generated target frames are concatenated and fed to a discriminator which distinguishes real from generated Mel-spectrograms. This approach can convert samples of arbitrary length, while the discriminator prevents discontinuities at concatenation points. A siamese network preserves linguistic content information by maintaining encoded vector distances between source and converted samples during conversion. Generated spectrograms are inverted to waveforms using the Griffin-Lim algorithm. While no formal assessments were undertaken, the resynthesised voices were deemed by the author (and by myself) as realistic and relatively natural, with the voice successfully converted and the linguistic content well preserved. Griffin-Lim has been shown to produce less-natural waveform reconstruction from Mel-spectrograms than neural vocoders WaveNet and WaveGlow, as measured by MUSHRA scores [9]. Using HiFi-GAN to resynthesise from predicted Mel-spectrograms should similarly produce more natural synthesis than using the Griffin-Lim algorithm.

2.2.3 Mel-spectrogram conversion for dysarthric speech

Mel-spectrogram transformation methods have been investigated in relation to dysarthria, but only for dysarthric-to-healthy voice repair. Korzekwa et al [10] attempted to learn a dysarthric latent space representation using a jointly-trained Variational Autoencoder (VAE) and dysarthria detector network. The VAE encoded input Mel-spectrograms and text transcriptions to a low-dimension latent representation, and a decoder resynthesised the VAE's output to a waveform. The authors hoped moving the latent encodings of dysarthric speech to approximate those of healthy latent encodings, then decoding to a waveform from these shifted encodings, would result in voice repair. The dysarthria detector was used as a training adversary to improve Mel-spectrogram conversion. The approach was unsuccessful; the latent space also inseparably encoded non-dysarthric speech features such as F0 and timbre, so modifying encodings also modified linguistic content and perceived pitch.

At the time of writing, no research has been published exploring healthy-to-dysarthric voice conversion using Mel-spectrograms as spectral features.

Chapter 3

Data

3.1 TORGO database

TORGO [11] is a database of dysarthric speech collected by the University of Toronto, containing time-aligned audio and electromagnetic articulographt (EMA) features. It comprises data from 8 speakers (5 male, 3 female) with dysarthria caused by Motor Neurone Disease (MND) or Cerebral Palsy (CP), and 7 healthy control speakers (4 male, 3 female) gender-and-age-matched to the dysarthric speakers.

3.1.1 Audio

Audio recordings comprise speakers following 4 prompt sets: reading individual words; reading sentences; describing images in their own words; repeating phonemes. While there is some overlap between the utterances recorded by healthy-dysarthric speaker pairs, there is not a recording of every speaker completing every prompt. Audio files were recorded at 16kHz with a 16-bit depth by two microphones: one head-mounted and one array microphone placed in front of the speaker. I used the head-mounted microphone recordings as they were higher quality.

3.1.2 Electromagnetic articulography

The database includes time-aligned electromagnetic articulograph (EMA) readings made by a Carstens Medizinelektronik AG500 machine at 200kHz, which capture 3D position and orientation readings of the mouth and speech articulators using transmitter coils [12]. These EMA readings are used in a multi-task learning experiment in section 4.6.

3.1.3 Dataset limitations

One potential limitation of training voice conversion models on this dataset is that it includes speakers with both Motor Neurone Disease and Cerebral Palsy, both of which can result in different forms of dysarthria, such as ataxic, flaccid, spastic, with the most common forms being mixed flaccid-ataxic and mixed spastic-ataxic. However, the degree of feature overlap between speakers should be great enough that this difference should not significantly hinder models' ability to learn a general healthy-to-dysarthric voice mapping. If the experimental models are successful, the techniques could be applied to more specific dysarthric datasets in order to augment training data for specific purposes e.g. MND voice reconstruction.

Another limitation of this dataset is that speakers' severity of dysarthria ranges from highly intelligible to almost unintelligible speech. Training models on the entire speaker pool will likely cause learned healthy-to-dysarthric mappings to be most accurate to the dysarthria severity of the speakers with the most data available. While this is a risk, I did not want to balance speakers' training set sizes by excluding usable data, as the dataset was already relatively small.

3.2 Data pre-processing

Audio files were denoised by student B116359, my collaborator on this project, using RNNoise [13], a deep neural noise suppression model. Denoising was required to remove both background noise and the tone produced by the EMA machine. Audio was resampled to the sample rate expected by HiFi-GAN (22.05kHz).

I wrote a Python script to automatically reorganise and relabel the TORGO data. This script (alongside all other preprocessing, conversion model and analysis scripts) is publicly available in the GitHub repository I created for this dissertation [14].

This script restructures the database into two corpora: one containing all available audio files, and one containing only the audio files with associated EMA readings. Within these subcorpora, files are split by prompt type and renamed according to a consistent scheme so each utterance has a unique ID. 23 corrupted audio files were discarded.

I formed two training datasets from the reorganised TORGO database. The first set, used for training the fully-connected and fully-convolutional models in experiments 2 and 3, contained source and target utterances from healthy/dysarthric speaker pairs dictated by TORGOs age-and-gender-matched pairing scheme. It included only sentence and word files, since there were very few phoneme repetition files ($<$5 per speaker) and linguistic content of the freeform image description files differed across speakers. Despite not being an official TORGO match, I paired

the two unpaired female speakers (F04, FC02) as I deemed their speech characteristics similar enough, and wanted to maximise the size of the training set. The final dataset contained 813 recording pairs from 3 female and 4 male speaker pairs.

The second dataset, used for training speaker encoding and multitask learning models in experiments 4 and 5, contained source and target recordings from all gender-matched healthy/dysarthric speaker pairs. TORGO's pairing scheme was no longer required since the models' architectures were modified to alleviate speaker differences between source/target pairs. Unlike the first dataset, this one contained only audio files with corresponding EMA readings. The final dataset contained 8164 recording pairs from 3 female and 4 male speaker pairs.

From both datasets, 30 healthy/dysarthric sample pairs were held out as test data to test models' generalisability. Validation datasets were not systemically split, but rather randomly sampled from the shuffled training set before training each model.

Chapter 4

Experiments

4.1 General methodology

4.1.1 Code implementation

The models outlined in experiments, including data preprocessing, loading, and conversion modules, I built from scratch using the PyTorch deep learning library [15] and audio processing functions from TorchAudio [16] and Librosa [17]. I decided against using pre-existing voice conversion models, since building from scratch allowed me to better understand the complexity of the conversion problem and tailor all steps of the conversion pipeline specifically to healthy-to-dysarthric transformation. This decision meant rigorous hyperparameter tuning was not possible due to time constraints. Code for all models, along with preprocessing and analysis scripts, is publicly available on GitHub [14]. Models were trained for 100 epochs on a GPU using the ECDF Linux Compute Cluster.

4.1.2 Mel-spectrogram extraction

In this project, I take the approach of indiscriminately modelling all dysarthric speech characteristics captured by frame-wise differences between aligned healthy and dysarthric Mel-spectrograms. These could include any formant differences resulting from imprecise articulation and constrained articulator movement. Mel-spectrogram conversion cannot encode longer-term and sporadic features such as stuttering, phoneme repetition, or increased duration. I will modify duration separately (details in section 4.1.4), but stuttering and repetition could be better modelled using text-to-speech, which is out of the scope of this project. We generated Mel-spectrograms using the same hyperparameter settings used in HiFi-GAN, to ensure representations were compatible for resynthesis. Power Mel-spectrograms were computed using a

Hanning analysis window with a length of 1024 samples, a hop size of 256 (12ms), and 1024 points in the FFT computation. Mel-spectrogram features were normalised to the [0,1] range using min-max normalisation, with minimum and maximum values computed globally across the entire TORGO dataset. Pretrained HiFi-GAN vocoder model VCTK_V1 was used, since it attained the highest mean opinion score (MOS) of the three available models in listening tests evaluating the quality of vocoded samples from unseen speakers [4]. This model was trained by the authors of HiFi-GAN [4] on 100 speakers of both sexes from the VCTK corpus [18].

4.1.3 Duration modification

To simulate the slower speaking rate often present in dysarthric speech, I stretched healthy input samples by a factor of two, so the speaking rate was half the original rate. Jiao et al. found this speed reduction best approximated speaking rates of Motor Neurone Disease-induced dysarthric speech [7]. I modified duration by performing spline interpolation directly on their Mel-spectrograms, following the method described in [19] for the duration modification of Mel-spectrogram music representations. This method was preferable to using signal-processing methods which include time-to-frequency domain conversion, and vice versa, which risk degrading the signal quality and introducing artefacts more than is necessary.

4.2 Dynamic time-warping

Since parallel voice conversion methods aim to learn direct mappings between equivalent source and target speech samples, the success of such approaches depends on the accuracy of time alignment between sample pairs. Time alignment is often performed using dynamic time-warping (DTW), an algorithm which computes local distance costs between each allowable time-step pair of two temporal sequences, and finds the optimal alignment as the one with the lowest sum of local costs.

Because dysarthria is fundamentally characterised by speech production which diverges spectrally from healthy speech, accurately time-aligning spectral representations from healthy and dysarthric speakers may be difficult. I hypothesise that the imprecise articulation often present in dysarthric speech could result in a healthy and dysarthric speaker pair producing distinct speech sounds for a given word, and so the local cost metrics used to compute DTW alignments could result in some dysarthric speech sounds being misaligned to frames of healthy speakers' neighbouring phones. For example, impaired tongue motor control can cause difficulties producing the precise approximation required by /r/ sounds, often resulting in articulation

closer to a vowel [2] [3]. Dynamic time-warping could misalign this target sound to the healthy source speaker's neighbouring vowel instead of /r/. Given enough misalignments, the voice conversion system could fail to learn an accurate mapping for this articulation difference.

We evaluated the appropriateness of parallel approaches for healthy-to-dysarthric voice conversion by performing DTW on paired samples from the TORGO database, and inspecting their results. Each healthy source sample (S) was separately aligned to two same-utterance samples: its gender-and-age-matched dysarthric counterpart (D) and a second gender-matched healthy sample (H). The healthy source audio was warped, since dysarthric speakers often speak at a slower rate [2] [3] so TORGO's dysarthric samples are generally longer than the healthy ones. Warping a longer sequence to a shorter one would result in a loss of information which would later reduce the number of frame pairs my voice conversion systems use to learn spectral mappings. DTW was performed using dtw-python [20] with Euclidean distance as a similarity metric on Mel-frequency cepstral coefficients (MFCCs) extracted from the Mel-spectrograms intended for use in my voice conversion systems. This feature representation was used for alignment because, unlike Mel-spectrograms, it does not contain F0 information, which exhibits strong inter- and intra-speaker variation so could perturb local distance computations. 80 MFCCs were extracted from speech samples using Librosa [17], using an FFT window size of 1024 with a hop length (the number of samples the FFT window is shifted each time) of 256 samples, on power Mel-spectrograms. DTW alignment paths were governed by three constraints:

- **Fixed boundaries:** The warp path must begin at the first frames of the two sequences, and finish at their end-frames.
- **Monotonic:** Temporal order is preserved, so alignment paths cannot move backwards through time.
- **Continuous:** The warp path cannot skip frames of either signal. This constraint can only apply when the sequence being warped is shorter than the reference signal it is being warped to.

I tested multiple DTW step-patterns and windowing functions and found the best alignments were produced by a symmetric step with no windowing constraint (figure 4.1).

As expected, the average cost per time-step was higher for the warped-to-dysarthric alignment than the warped-to-healthy one (130.28 and 75.82, respectively). To compare the accuracy of healthy-dysarthric and healthy-healthy mappings, I warped healthy source Mel-spectrograms

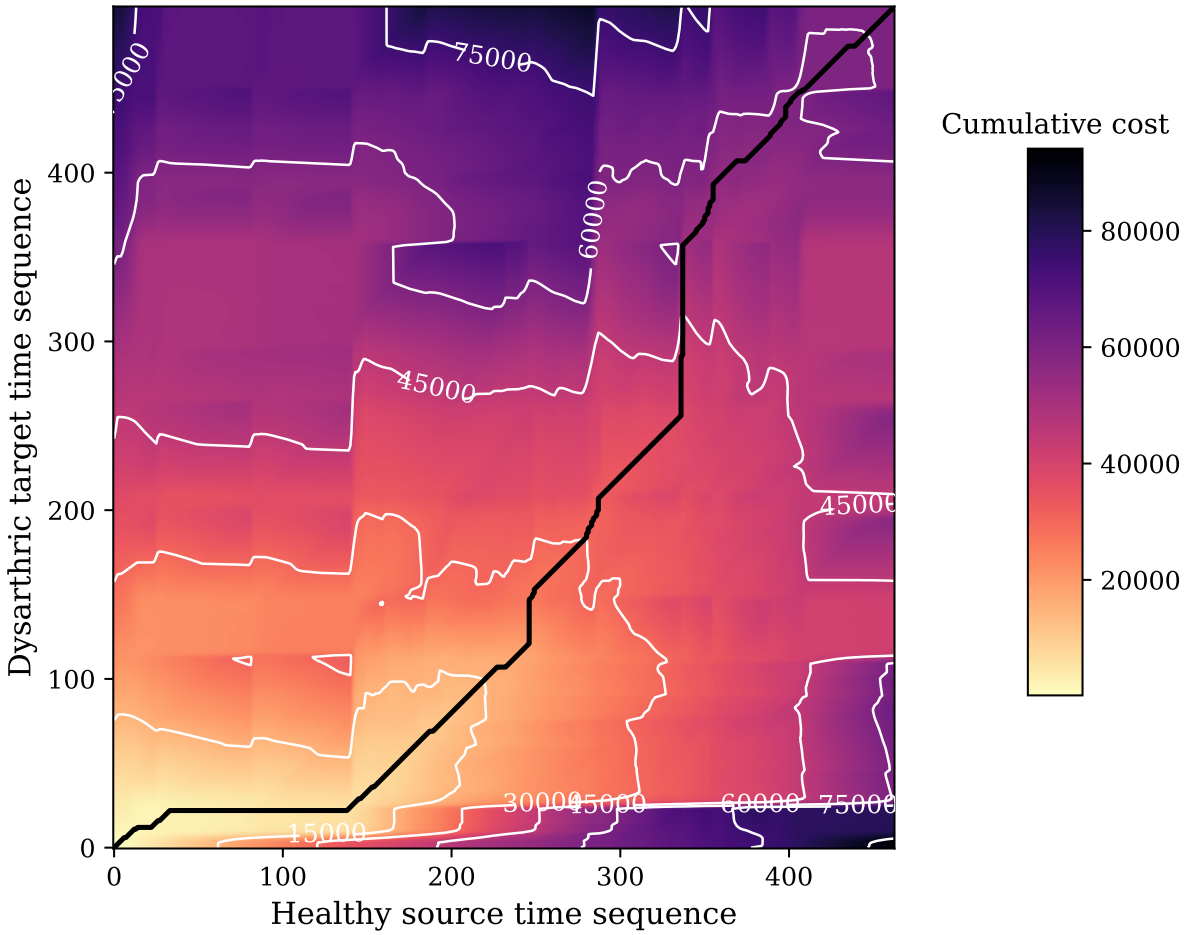

Figure 4.1: A symmetric, no-window warp path computed by performing dynamic time-warping from healthy source to target dysarthric speech. The path is monotonic, continuous, and has fixed start and end states.

according to the optimal DTW paths (figure 4.2), and resynthesized these to audio using HiFi-GAN. I inspected the resulting audio using auditory and visual (spectrogram) judgement in Praat [21].

Salvador and Chan [22] computed DTW accuracy using the distance between the estimated and ground-trouth optimal paths, but this requires a ground-truth alignment which was unavailable in my case. Instead, I annotated samples' word-boundaries and compared the boundaries of the warped source samples to that of their targets. The differences in milliseconds between the warped source and target samples was approximately equal for the healthy-dysarthric pairs and the healthy-healthy pairs. Based on this judgement, I concluded dynamic time-warping performed on the TORGO healthy-dysarthric pairs was probably accurate enough that parallel voice conversion methods could be explored without misalignment being the main limiting fac-

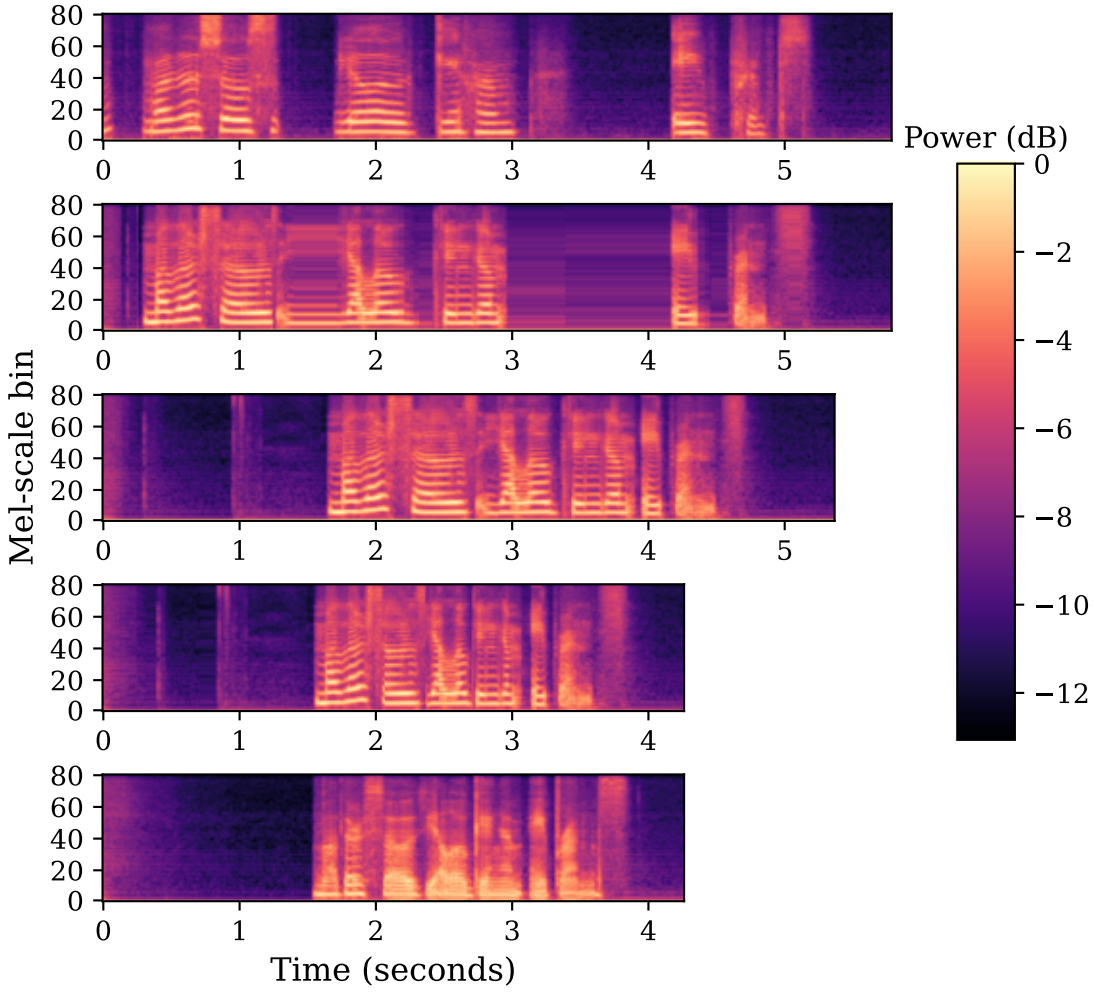

Figure 4.2: Mel-spectrograms produced by dynamic time-warping. From top to bottom: Unmodified dysarthric target; healthy source warped to dysarthric target; unmodified healthy source, healthy source warped to healthy target; unmodified healthy target (H).

tor in their success since neural networks' weight updates are averaged across all frames of all utterances.

4.3 Fully-connected neural network

To determine the difficulty of the healthy-to-dysarthric spectral transformation problem, I began by building the simplest possible neural model: a fully-connected artificial neural network. The quality and accuracy of Mel-spectrograms produced by this model will be used to inform the design of subsequent model architectures. The aim was to predict a dysarthric Mel-spectrogram frame-by-frame from its DTW-aligned healthy counterpart. This modelling

approach had successfully been used to convert source-to-target formants and Mel-cepstral features for healthy speakers in a Gaussian Mixture Model-based voice conversion system [23]. However, in both cases the spectral representations used for conversion were disentangled from F0. Target speaker F0 was estimated separately, and recombined with predicted spectral features using a vocoder. Since I am using Mel-spectrograms, whose features are F0-dependent and so vary both within and across speakers, it is unclear whether the network will be able to learn a universal mapping across speakers.

4.3.1 Model architecture and training

The feed-forward model I built had six dense layers each with 256 hidden units, with the rectified linear (ReLU) activation function to squash values into the [0,1] range. No activation function was applied to the final layer, so the output features were real values as in real Mel-spectrograms. Mean squared error (MSE) loss computed between predicted and target frames was the objective function, using stochastic gradient descent to compute weight updates and backpropagation to update model parameters. Through experimentation, I found satisfactory training was achieved with a learning rate of 10e-6 and an exponential learning rate decay of 0.02.

4.3.2 One input frame

Our first modelling approach was to input single frames of a DTW-aligned healthy source speaker Mel-spectrogram, and predict single frames of a target dysarthric Mel-spectrogram, in the hope that the model would learn a universal mapping function which could be used to transform any healthy Mel-spectrogram, including ones from new speakers, into a dysarthric one. The model was trained on 577 utterance pairs from TORGO's four male age-matched healthy-dysarthric speaker pairs. 10% of utterances were held-out as validation data, which I used to check the model's ability to convert unseen samples and to prevent overfitting to the training data. Only male samples were used initially, since I thought variation between source and target speakers' F0 could differ for male and female speakers, which could make finding a universal spectral mapping more difficult.

During inference, the dysarthric samples were time-warped to the healthy input samples, to prevent unnatural speech patterns in the output while allowing test loss to be calculated between the two parallel-frame samples.

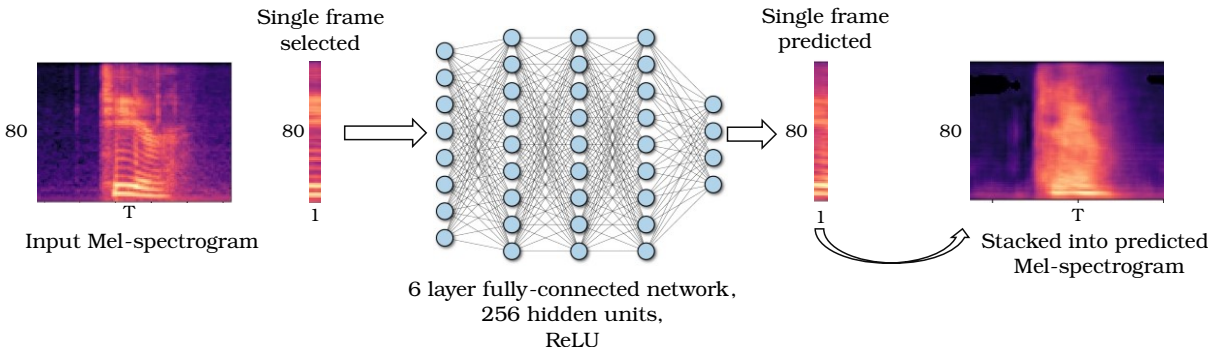

Figure 4.3: Frame-to-frame fully connected network architecture.

4.3.3 Results

This modelling approach was unsuccessful. The strong horizontal striation in the output translated to robotic tones after vocoding. The power present in each generated frame was highly distributed across the 80 features, which resulted in high noise levels and buzzing sounds in the resynthesised audio. This noise was too great to discern any successful synthesis of dysarthric features. The diffusion of power across the Mel-spectrogram features could suggest the model is not powerful enough to accurately predict the energy bands characteristic of human speech Mel-spectrograms. The MSE objective function is then optimised by spreading an average amount of energy across all frequencies.

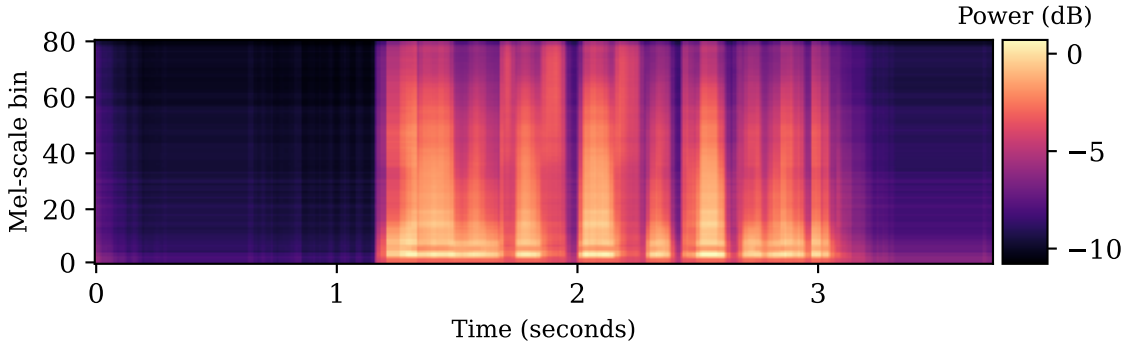

Figure 4.4: Dysarthric Mel-spectrogram predicted by frame-to-frame fully connected network for an unseen utterance.

Based on these results I concluded that obtaining accurate Mel-spectrograms from one-to-one frame mapping was unlikely, either because more context is required to learn the complex transformation, or because the alignments produced by DTW were so inaccurate that a significant portion of training frames were incorrectly paired, resulting in the network trying to map

between frames of different speech sounds.

4.3.4 Additional context with delta features

To increase the complexity of the learnable mapping and provide the model with additional speech context, I trained a second fully-connected model, with the same architecture and training procedure as the previous one. Instead of inputting a single frame of the spectrogram, I concatenated each frame with its static delta and dynamic delta features, computed with each contiguous frame to either side, to represent how the speech signal changes over time. The input to this model at each time-step was then a 400-by-one vector comprising the original 80 healthy Mel-spectrogram features, 160 static delta features, and 160 dynamic delta features.

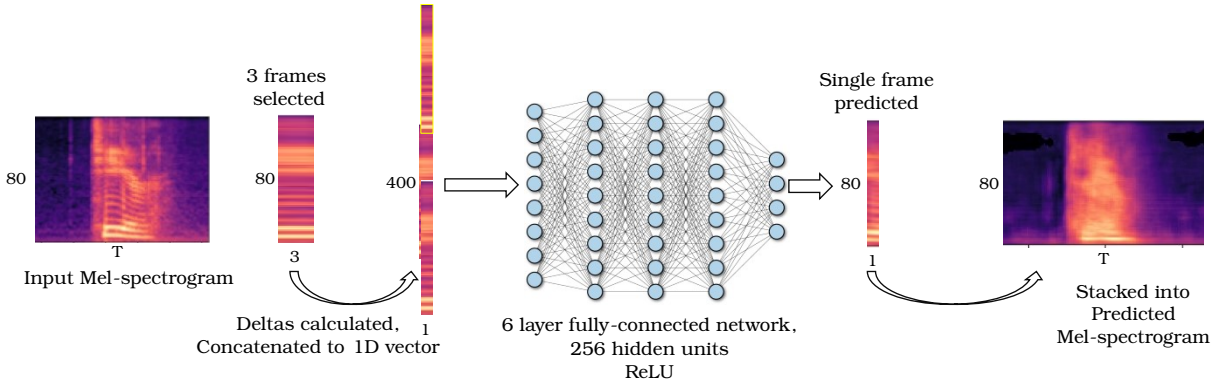

Figure 4.5: Frame-to-frame fully connected network architecture with additional delta context features.

4.3.5 Results

While the Mel-spectrogram predicted by this model exhibited slightly more tightly-banded power distribution than the previous model, the wide power distribution across Mel features again meant no speech features could be discerned from the vocoded audio. This model's training and test losses (table 4.1) were higher that the first's. This result indicates that even with additional context features, a fully-connected model is not capable of performing accurate healthy-to-dysarthric Mel-spectrogram transformation. The next chapter seeks to address the weaknesses of fully-connected spectral conversion discussed here, by replacing the model's dense layers with convolutional layers.

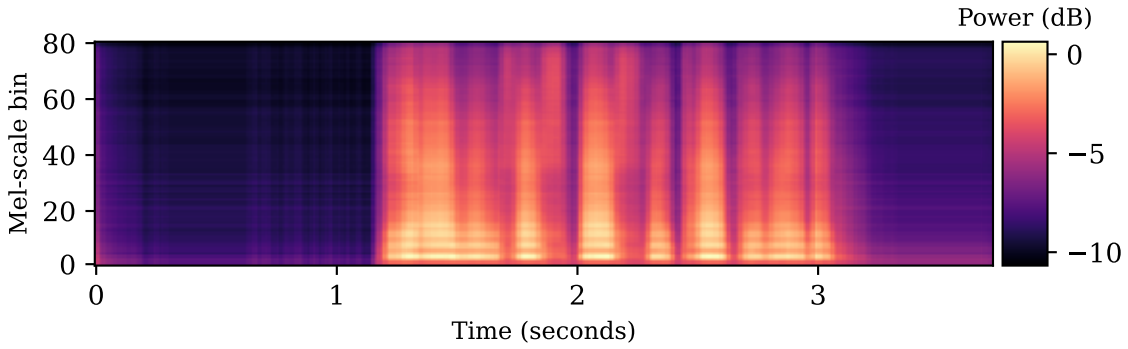

Figure 4.6: Dysarthric Mel-spectrogram predicted by fully connected network with delta context features for an unseen utterance.

Table 4.1: Final training, validation, and test losses for the fully-connected conversion models.

Model	Training loss	Validation loss	Test loss
1 input frame	0.00436	0.010475	0.007728
1 frame + deltas	0.006948	0.010373	0.007757

4.4 Convolutional neural network

4.4.1 Why convolution?

Since speakers are gender- and age-matched, their formants in a parallel Mel-spectrogram frame should appear in approximately the same region. This constraint means fully-connected layers should not be required to learn a frame conversion mapping. Instead, I turn to convolution in the hope of learning a more complex spectral transformation in smaller local Mel-spectrogram regions. Convolutional layers have successfully been used in neural voice conversion models, including in healthy-to-dysarthric voice conversion with MFCCs as spectral representations [7], and in Mel-spectrogram based healthy-to-healthy voice conversion [5]. Because 2D convolution with a kernel size >1 means multiple time-steps in a region are considered when predicting any given cell in a feature map, this approach could also alleviate the problem of inaccurate frame time-alignment.

4.4.2 Model architecture

I trained and evaluated four models with six and twelve convolutional layers, and with kernel sizes of three-by-three and five-by-five (table 4.2). These filter sizes are typical for image pro-

cessing convolutional networks (see, for example, [24]). It is conventionally understood that increasing the number of layers or the kernel size in a neural network can improve test accuracy [25], [26], but since increasing the kernel size quadratically increases the number of learnable parameters, model with larger kernels are more prone to overfitting.

Table 4.2: Convolutional network parameters for conversion models A-D.

Model	Conv. layers	Kernel size
A	3	6
B	3	12
C	5	6
D	5	12

4.4.3 Data & training

The fully convolutional models were trained on the same dataset as was used for the fully connected model. 10% of randomly selected pairs were held-out as validation data. Although a CNN could in theory work without time-warping given a large enough kernel, I still time-aligned sample pairs to avoid giving the model the additional task of duration mapping. Convolution was performed with a stride of one and no dilation. The input to each convolutional layer was re-scaled and re-centered by batch normalisation, which improves the speed and stability of network training by smoothing the optimisation landscape of the network's objective function (see [27] for an explanation). The ReLU activation function was applied after each convolutional layer to rescale cell values to the [0, 1] range.

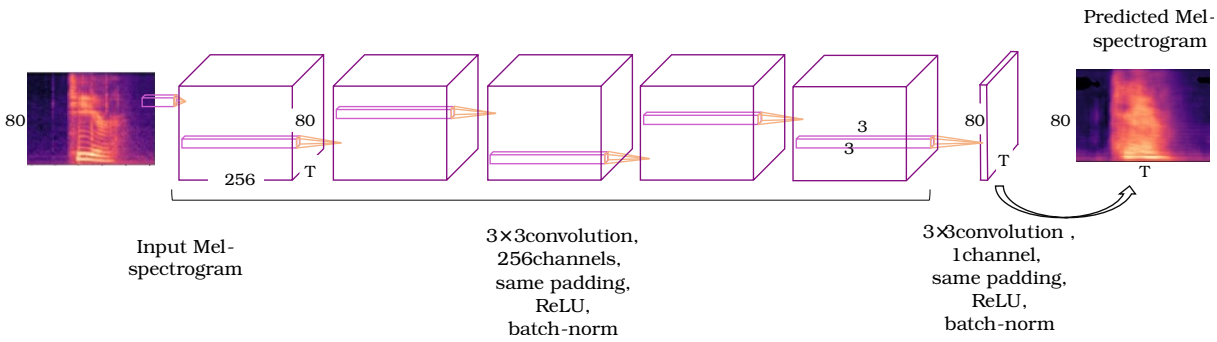

Figure 4.7: Fully convolutional voice conversion model architecture.

Models were trained by optimising mean squared loss error computed between the predicted and real dysarthric Mel-spectrograms. Through initial testing I found a learning rate of 10e-5 and an exponential learning rate decay of 0.03 were the largest hyperparameter values which successfully approximated an optimum. Because convolutional models have a huge number of parameters, which increases quadratically with kernel size, training was slow. The largest model (5x5 kernel, 12 layers) took 922105 milliseconds per epoch- more than twice the time for the 3x3 kernel model with the same number of layers (402430 milliseconds). In hindsight, increasing the kernel dilation to at least two, instead of (or as well as) increasing the kernel size and number of layers, would have allowed the model to learn mappings over a wider receptive field without increasing the number of model parameters. For this model and all subsequent models, training batch size was one so the models could accept variable-sized input without zero-padding. This decision may have contributed to the long run-time: since the matrix operations involved in network training can be parallelised when run on a GPU, the process could be quicker with a larger batch size. Reducing time per epoch would have allowed us to train models more, potentially resulting in better conversion systems.

4.4.4 Results

The model with the most parameters (D) exhibited the lowest training, validation and test losses after 100 epochs (table 4.3). These loss values corresponded to my judgments of resynthesised speech quality. The largest model produced the most natural-sounding speech, but the quality was still very poor. The Mel-spectrograms predicted by the fully-convolutional models (figure 4.8) looked closer to natural speech Mel-spectrograms than the ones produced by the fully-connected models presented in section 4.3. Power was clustered more tightly into formant bands, and these bands had fluid two-dimensional trajectories instead of the straight striations present in the fully-connected output, translating as fewer robotic tones after resynthesis. However, power was still diffused across Mel features, resulting in loud noise which prevented us from discerning whether any dysarthric spectral features had been successfully synthesised. This diffusion could be caused by the presence of F0 in all Mel-spectrogram features, and F0 discrepancies between source and target speakers. Since the difference in F0 between a source and target speaker varies between speaker pairs, it is likely the convolutional model could not accurately approximate every source-to-target feature mapping simultaneously. Based on these results, it seems unlikely there is a universal healthy-to-dysarthric Mel-spectrogram mapping which is learnable across many speakers. Section 4.5 investigates adding speaker embeddings to the convolutional model to resolve differences in F0 between speakers.

Table 4.3: Final training, validation, and test losses for fully convolutional voice conversion models A-D.

Model	Training loss	Validation loss	Test loss
A	0.008837	0.109707	0.009401
B	0.006091	0.009645	0.007416
C	0.008648	0.010515	0.008913
D	0.005838	0.009723	0.006842

Loss values (table 4.3) indicate that increasing the number of layers is more effective in improving image prediction accuracy than increasing the kernel size: the model with a 3x3 kernel and 12 layers attained lower train, val and test loss values than the model with a 5x5 kernel and 6 layers.

4.5 Speaker embedding model

4.5.1 Speaker embeddings

As described in section 4.4, the presence of F0 in all Mel-spectrogram features complicates the task of learning a function which approximates healthy-to-dysarthric spectral differences across many speakers. I considered two approaches to addressing the issue of F0 mismatch. Firstly, the F0 of the target speaker could be normalised to match the F0 of the source speaker, using a signal processing algorithm like Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA). This approach was considered inappropriate considering the intended downstream use of the synthesised dysarthric data- training future models- requires high-quality, natural speech, and overlap-and-add methods are known to introduce signal distortion and other artefacts [28] [29], which the voice conversion model could learn and reproduce. The approach I took instead was to explicitly represent target speaker identity in the model's input.

Korzekwa et al. [10] modelled a two-dimensional latent space directly from healthy and dysarthric Mel-spectrograms, and found it encoded interpretable differences between healthy and dysarthric speakers, which were used along with text transcriptions to resynthesise speech. I instead learn a latent speaker space, as in Deng et al. [30], allowing us to directly control the identity of the target speaker during voice conversion. Speaker codes are learned fixed-dimension representations of target speaker identity. They are learned by modelling a latent space of speaker characteristics using backpropagation of gradients, so the vectors learned for

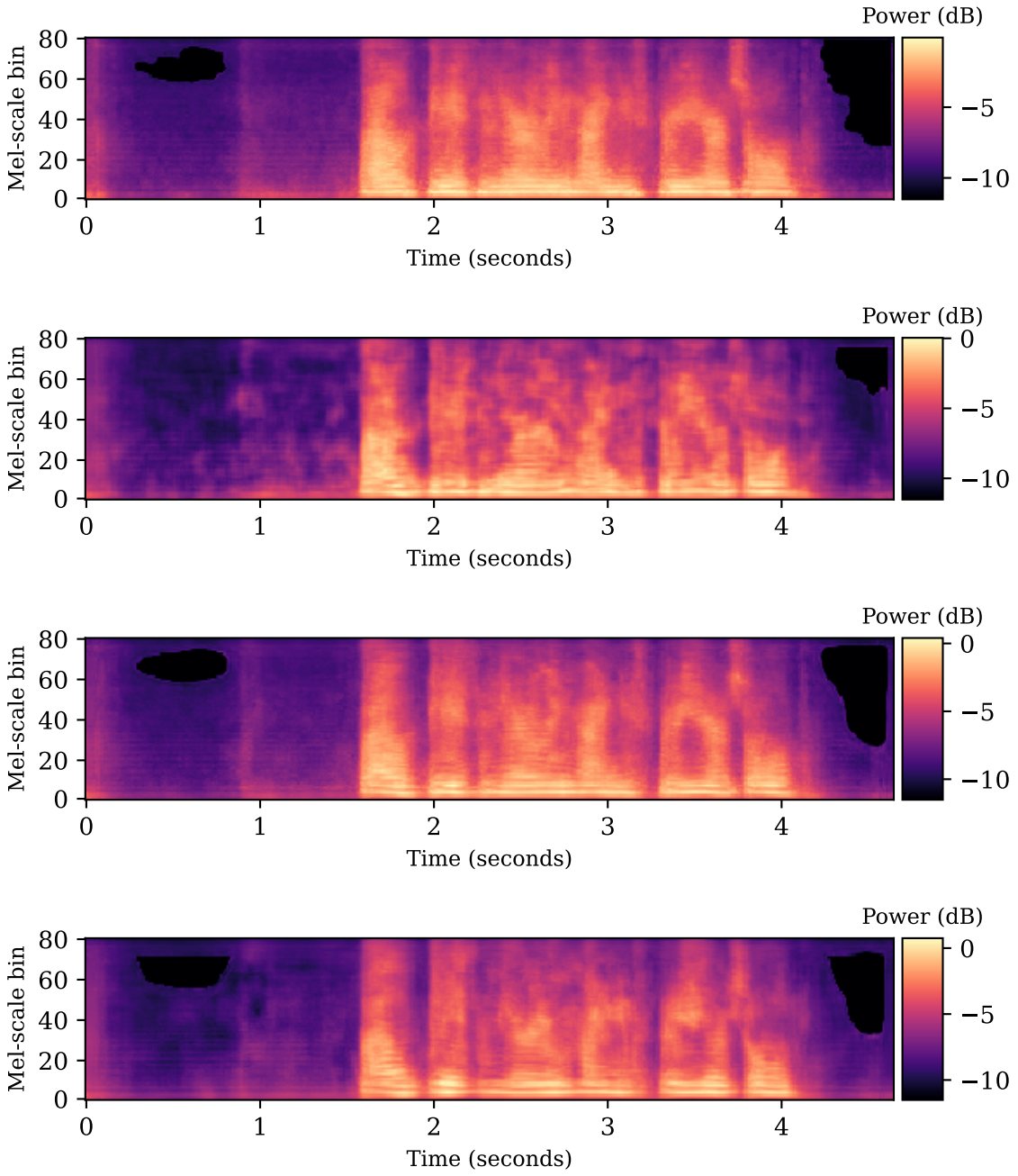

Figure 4.8: Dysarthric Mel-spectrograms produced by fully convolutional models for an unseen utterance. From top to bottom: A, B, C, D.

each speaker are maximally useful in improving the model's predictions.

Speaker codes have successfully been used to adapt and control acoustic estimates produced by fully-connected acoustic model components of multi-speaker text-to-speech systems [31]. Luong et al. showed that augmenting the input of a standard DNN-based acoustic model with speaker codes in a hidden Markov model text-to-speech system expanded the model's capabili-

ties from speaker-dependent- that is, only capable of producing acoustic estimates for the single speaker it is trained on- to a multi-speaker model, trained on 135 speakers and able to produce acoustic estimates for each, given the corresponding speaker code during synthesis [31].

4.5.2 Implications for dataset augmentation

While the previous experiments explored dataset augmentation as a healthy-to-dysarthric voice transformation problem, seeking to find a general mapping which preserves similar voice characteristics between source and target speaker pairs, this experiment instead takes a true voice conversion approach. This approach incurs a tradeoff between conversion quality and generalisability; while unseen target speakers can be simulated by inputting unseen speaker embeddings, the resulting speech quality will be lower since the model was not trained on any relevant data. I deemed it was worth pursuing this model at the cost of flexibility: determining whether dysarthria simulation by Mel-spectrogram transformation is possible in a constrained speaker-specific use case would inform whether more generalisable transformation systems are viable.

Despite being unable to convert healthy speech to dysarthric speech without altering speaker identity, a conversion model with speaker codes could still be useful for training data augmentation provided the system is trained on enough target speakers. In use cases where a speaker has a lot of healthy data and a small amount of dysarthric data available, a speaker-dependent mapping model could be trained to generate parallel dysarthric training examples for the healthy ones. Further, it could be modified to estimate encodings for unseen target speakers from small amounts of data, using the methodology detailed in [31]. Finally, inputs could be further augmented with aditional codes representing sex and age, as detailed in, or even regional accent, to allow further controllability over target voices by varying individual speaker characteristics. A larger database of dysarthric speech would likely be required for this, in order for the embedding layers to learn meaningful representations of specific characteristics.

4.5.3 Data

The systems presented here are intended to convert the identity of a speech sample from that of a source speaker to a target speaker, with speaker identity represented explicitly as an embedding. As such, sex-and-age-matched speaker pairs are not required, as explicitly modelling speaker identity should resolve differences in F0 across speaker pairs. Athough not required, same-sex source-and-target pairs were used again, to reduce the amount of training data due to time limitations, and to allow for fairer comparison with the earlier single-sex models. 4166 utterance pairs were formed from the subset of the TORGO database which had associated EMA

readings, so this model could be more fairly compared to the multitask model presented in section 4.6 which only uses data with associated EMA readings available. Healthy source speakers were paired with all available healthy and dysarthric target speakers, so models were trained to perform both healthy-to-dysarthric and healthy-to-healthy conversion. Before training, this set was randomly split into 95% training data and 5% validation data.

I used model B, presented in section 4.4, as a baseline system and modified for this experiment. The project's limited time-frame meant using the best model (D), which took over twice as long to train per epoch, was not feasible. Each speaker was assigned a unique integer ID. These IDs were converted to two-dimensional speaker code vectors by an embedding layer which was jointly trained along with the rest of the model. A two-dimensional embedding space was chosen because of its interpretability; it could easily be plotted and visually inspected.

4.5.4 Model architecture

Integrating speaker codes into a fully-connected model is trivial, since all units in a layer have learnable weights, so by concatenating a code to an input it is made available to every element of the first dense layer. However, since each unit of a convolutional feature map is determined only by the input features within its receptive field, incorporating the speaker embeddings so they are usable and useful to learning voice conversion mappings is less straightforward. It is unclear whether concatenating a speaker embedding to the Mel-spectrogram will have any effect on converting parts of the Mel-spectrogram out with its receptive field.

I investigated two methods of incorporating the speaker embeddings: in the first (model se-1-fc-last), for each source-target Mel-spectrogram pair, the target speaker's embedding was concatenated frame-wise to the input source speaker's Mel-spectrogram. I deemed this the most useful way to utilise embeddings, as it would allow the system to convert speech from a new, unseen speaker to that of a target speaker selected from the training set by specifying a target ID. This augmented Mel-spectrogram constituted the input for the convolutional network. A final fully-connected layer was added to reduce the output's vertical dimensionality back to 80-features.

The embedding layer was trained along with the rest of the network so the embeddings learned were maximally useful for performing conversion from source to target speaker. The network was trained using stochastic gradient descent with MSE as the objective function. Through initial testing, the optimal hyperparameters were found to be a learning rate of 10e-5 with an exponential decay of 0.03.

The second model (se-1-fc-first) differed only in that the fully-connected layer was moved

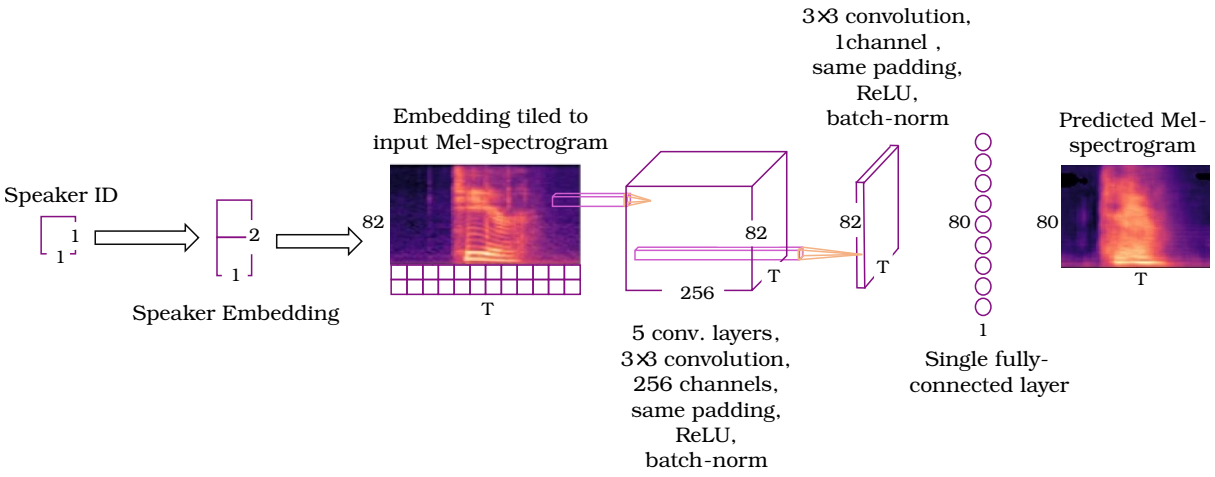

Figure 4.9: Conversion model (se-1-fc-last) architecture with target speaker embeddings.

from after the convolutional layers to immediately before. This modification makes the speaker embedding available to every unit of the first convolutional layer. Training followed the same procedure as the previous model.

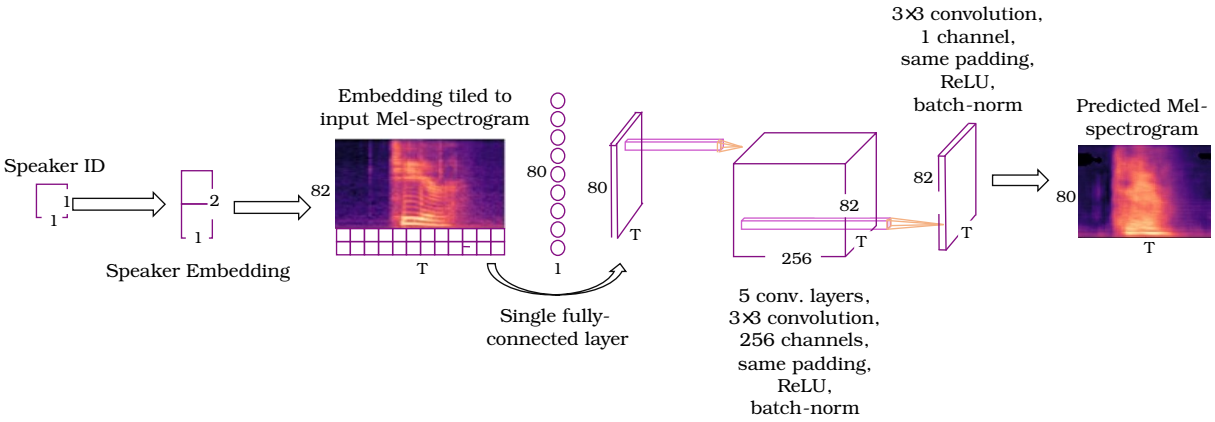

Figure 4.10: Conversion model (se-1-fc-first) architecture with target speaker embeddings.

A third model was trained, identical in architecture to the second, but with three consecutive fully-connected layers instead of one. This design choice was motivated by my suspicion that the additional complexity introduced by multiple dense layers would allow the network to better approximate the complex dysarthric mapping, reducing robotic artefacts in the resynthesised output.

4.5.5 Results

Model se-1-fc-last failed to resolve F0 differences between speakers, likely because the speaker embedding was not available to most cells in most convolutional layers. Its training and validation loss were the highest of the three speaker embedding models (table 5.4). The single fully-connected layer at the end was not powerful enough to accurately combine the final feature map with the speaker embeddings in a way which improved conversion. F0 sounded approximately the same in all output samples, both male and female, and inputting different target speaker codes did not result in audibly different output. This result was corroborated by the model's low Mel-cepstral distortion (MCD) score (table 5.4). This score indicates the model failed to accurately convert spectral features. I attempted to compute F0 root mean-squared error scores between real and predicted dysarthric sample pairs to summarise F0 conversion accuracy using pyreaper [32], but the high noise levels impacted the accruacy of F0 estimation. Output Mel-spectrograms (figure 4.11) exhibited the same horizontal striation produced by the fully-connected models presented in section 4.3, which translated to noisy and robotic speech after synthesis.

Table 4.4: Final training, validation, and test losses for the three conversion models with target speaker embeddings.

Model	Training loss	Validation loss	Test loss
se-1-fc-last	0.002504	0.010275	0.010275
se-1-fc-first	0.004815	0.007176	0.007435
se-3-fc-first	0.004912	0.00707404	0.007333

Table 4.5: Mel-cepstral distortion for speaker embedding models.

Model	MCD
se-1-fc-last	11.319
se-1-fc-first	9.997
se-3-fc-first	9.592

Models se-1-fc-first and se-3-fc-first performed conversion more successfully. Model se-3-fc-first exhibited the lowest validation and test losses after 100 epochs. Auditory judgement indicated the perceived pitch was converted from source to target speaker. Inputting different target speaker IDs resulted in different-sounding output, including different pitch (figure 4.12).

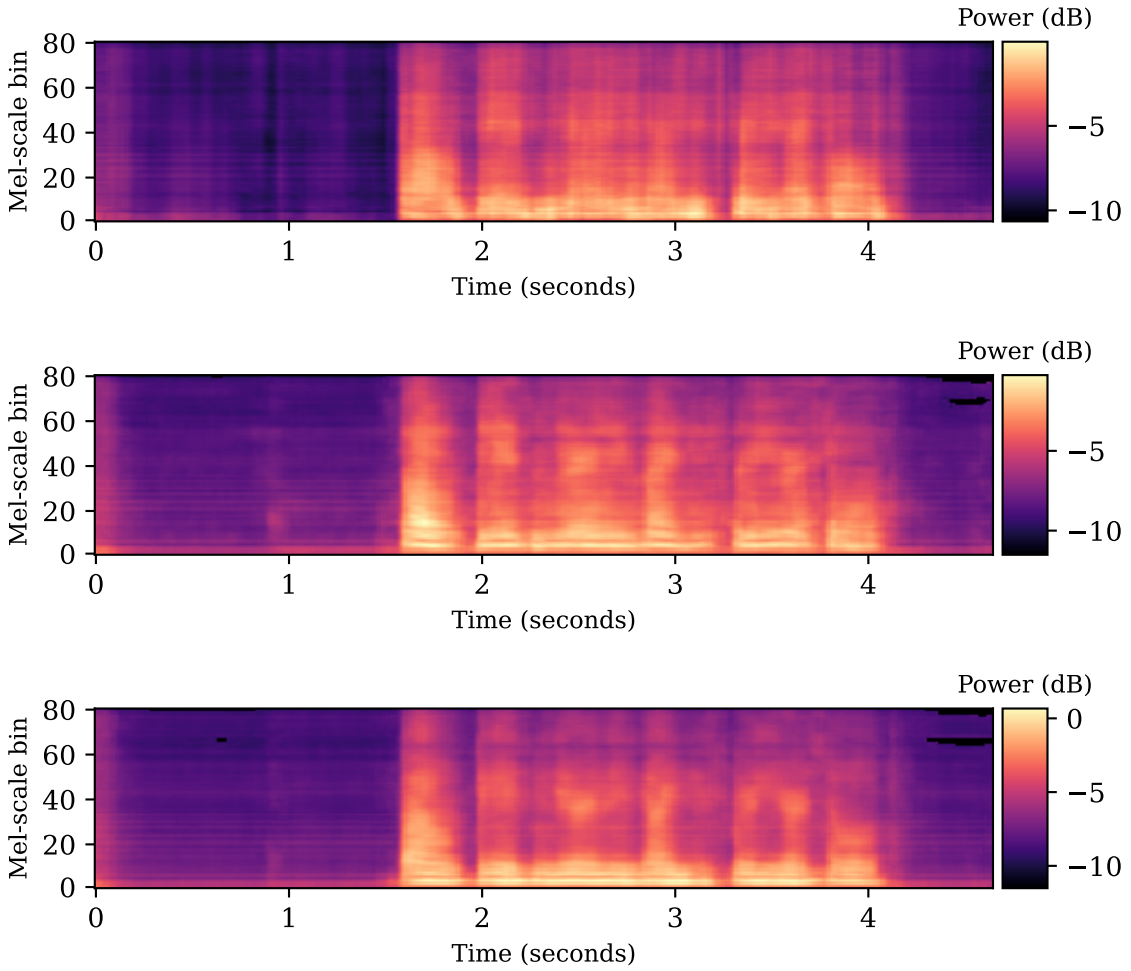

Figure 4.11: Dysarthric Mel-spectrogram predicted by conversion models with target speaker embeddings for an unseen utterance. From top to bottom: se-1-fc-last, se-1-fc-first, se-3-fc-first.

These perceived improvements were supported by lower Mel-cepstral distortion (MCD) values (table 4.5). The intelligibility and audio quality of resynthesised samples were higher for samples resynthesised from the output predicted by the model with three dense layers than the model with one dense layer. The resynthesised converted samples produced by model se-3-fc-firs will be evaluated in section 6. However, even the best model's output was noisy and poor quality, possibly due to the model overfitting to noise in input Mel-spectrograms. Section 4.6 takes a multitask learning approach to reducing overfitting and improving Mel-spectrogram predictions.

The plotted embedding spaces show no clear clustering of interpretable speaker characteristics I expected would be useful for performing conversion, such as dysarthria or sex. Future research could investigate what these models' latent spaces encode.

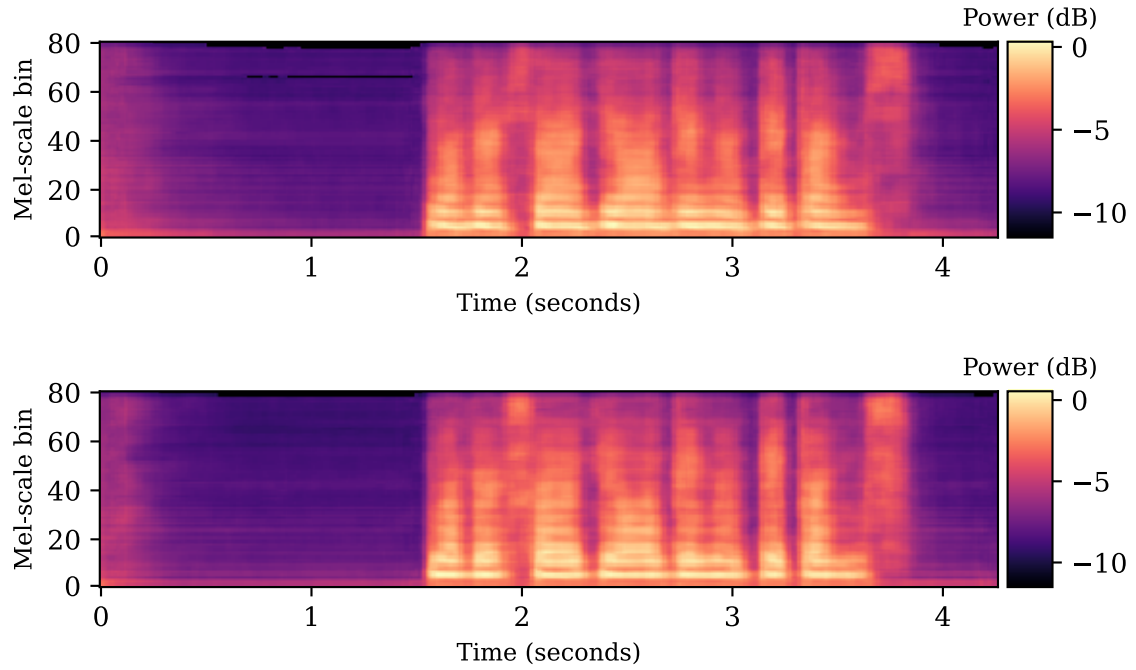

Figure 4.12: Mel-spectrograms predicted by model se-3-fc-first, produced using different target speaker IDs (top dysarthric, bottom healthy) for the same unseen source sample.

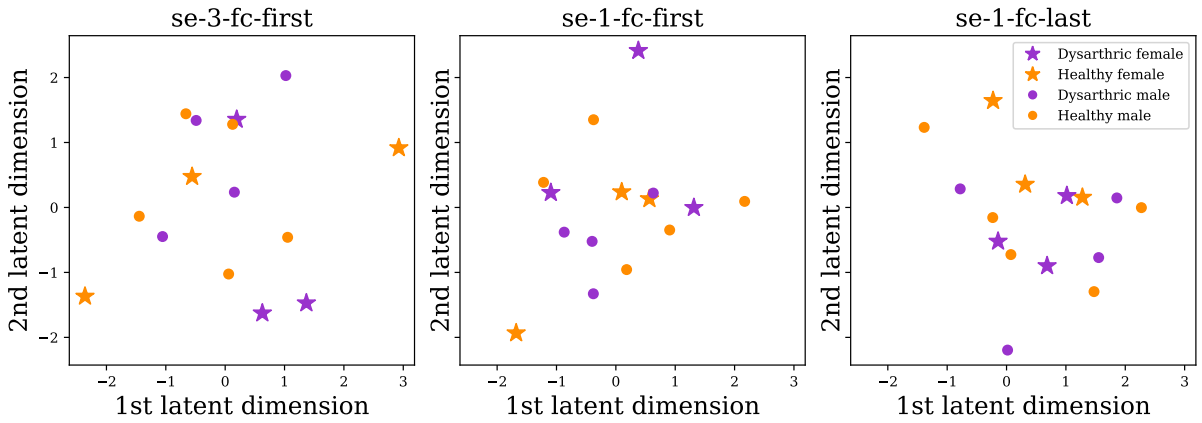

Figure 4.13: Latent embedding spaces learned by speaker embedding models.

4.6 Multi-task learning model

4.6.1 Multi-task learning

In this final experiment, I adopted a multi-task learning (MTL) approach to improving healthy-to-dysarthric conversion, using TORGO's electromagnetic articulograph (EMA) features. MTL is a training paradigm which updates model parameters in a way that improves performance on multiple separate but related parallel tasks, so all tasks are simultaneously solved by the

same model [33]. In speech technology research, MTL has been successfully used to improve model performance in automatic speech recognition [34] [35], and speech synthesis [36] [37], but limited research exists on the use of MTL for voice conversion.

The MTL approach taken here - known as 'hard parameter sharing'- involves training a model with shared layers for multiple related tasks which learns a common underlying representation for all output signals. Doing so exploits similarities between the two signals by leveraging the additional information contained in signal A to learn an underlying representation for solving the task required for signal B which is both more useful for making accurate predictions and more generalisable [33]. In other words, the information the neural network learns to solve task A can help it in solving task B.

4.6.2 Electromagnetic articulograph features

The use of EMA features was motivated by their ready availability in the TORGO dataset, as well as their direct correlation with Mel-spectrogram speech representations; both are time representations of speech; the Mel-spectrogram representing the acoustic signal produced by the articulatory mechanisms the EMA features monitor. I hoped predicting dysarthric EMA features as a secondary output would improve predicted Mel-spectrograms in two ways: firstly, since the two feature representations are highly correlated, the extra information provided by EMA features could improve predicted Mel-spectrogram feature accuracy. Secondly, it could reduce noise in predicted Mel-spectrograms since forcing the model to output two separate features from a shared underlying representation should reduce the model's capacity to overfit to noise in the training data.

The three tongue readings (tip, mid and back), along with lower lip and lower incisor readings were used, since the tongue, lips and jaw are the three tracked articulators most severely affected by dysarthria [2] [3], and most involved in speech production. I hypothesised these readings would contain the most salient dysarthria information and could be the most strongly correlated with the differences between healthy and dysarthric Mel-spectrograms, compared with the other articulator readings. Only the Y and Z coordinates were used, since I assumed the X coordinate (side-to-side movement) would not exhibit systematic or patterned behaviour associated with dysarthria, despite tongue deviation to one side while at rest being a known affliction of dysarthria [2] [3]. Plotted EMA features (figure 4.14) show clear differences in articulator trajectories between healthy and dysarthric speakers, with the dysarthric speaker's features exhibiting greater range in both Z and Y dimensions. This difference could be representative of the imprecise articulation symptomatic of dysarthria, but could also be attributed to

the dysarthric speaker's articulatory apparatus being larger.

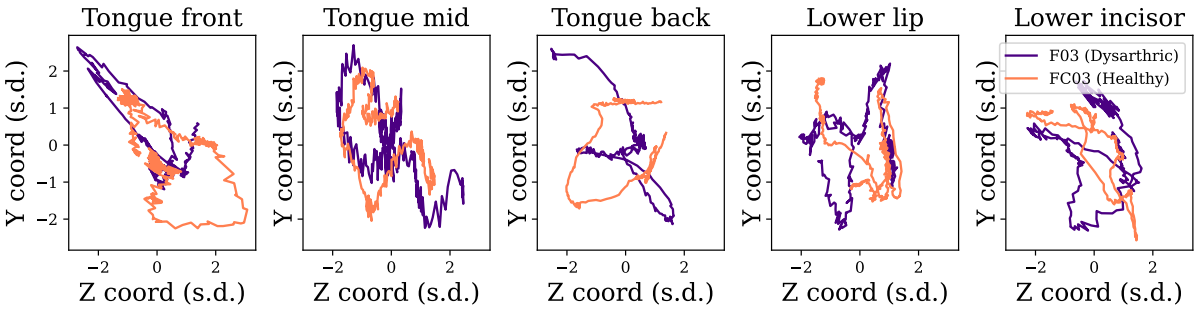

Figure 4.14: The five electromagnetic articulography features used in the multi-task learning model, for one utterance spoken by a TORGO female speaker pair.

The EMA readings provided in the TORGO database came head-normalised- that is, each reading was normalised by the participant's orientation-normalised head position at that time point. Since the TORGO EMA readings were taken at a 200Hz sample rate, I linearly interpolated them then resampled this interpolated function at a rate of 22050Hz, offset by half the STFTwindow length, so they aligned with the center of each FFT window used to produce the Mel-spectrograms. As previously described for normalising Mel-spectrogram features, EMA features were min-max normalised using the minimum and maximum values of all 10 articulator features, computed globally across all speakers, to squash values into the [0,1] range. These features then formed a ten-dimensional feature vector for every time-step, which was concatenated to its corresponding Mel-spectrogram frame.

4.6.3 Model architecture

The most successful speaker encoding model (se-3-fc-first) was used as a baseline for adaptation. A second parallel final convolutional layer was added, followed by a fully-connected layer to reduce the dimensionality from 80 to 10 (figure 4.15). These two layers predicted dysarthric EMA features, while the other final convolutional layer predicted dysarthric Mel-spectrograms. The model's training procedure and hyperparameters followed the one outlined in the previous section (section 4.5). Model parameters were optimised by backpropagating updates calculated using stochastic gradient descent. The objective function was the sum of two MSE terms computed between expected and predicted features (one for EMA features, one for Mel-spectrograms). During inference, the EMA time series was warped according to the alignment used to warp dysarthric to healthy audio samples.

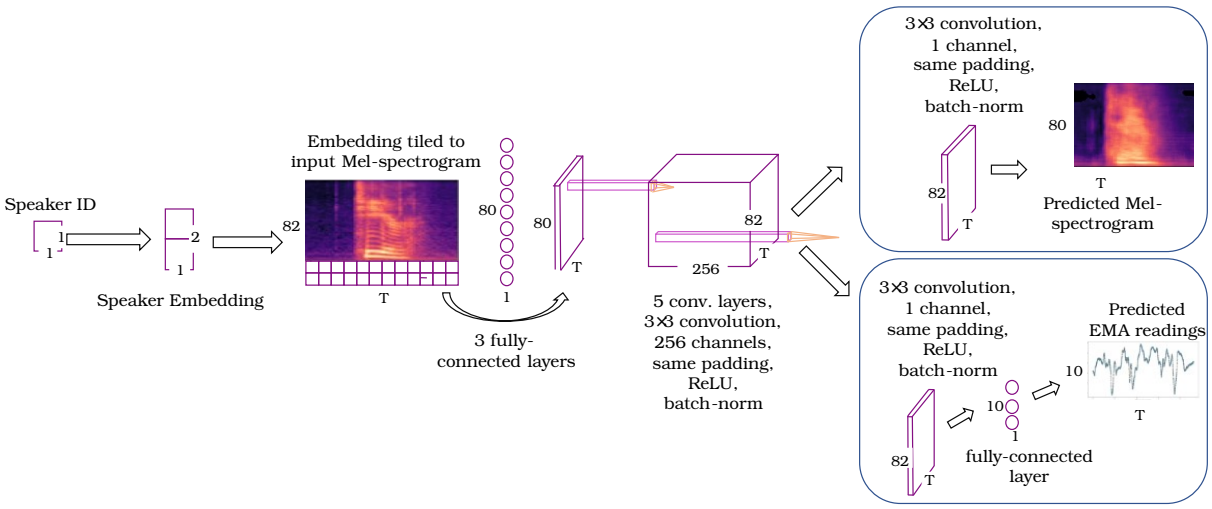

Figure 4.15: Multi-task conversion model (mtl-3-fc-first) architecture with target speaker embeddings.

4.6.4 Results

The multi-task model attained lower validation and test losses than any other model with speaker embeddings (table 5.6). However, its converted Mel-spectrograms exhibited significantly more noise than those produced by the best speaker encoding model (se-3-fc-first) described in section 4.5. Resythesised samples had higher Mel-cepstral distortion than se-3-fc-first (table 5.6). Samples resynthesised from these two models' predicted Mel-spectrograms will be evaluated by an expert speech and language therapist in section 5.

Table 4.6: Final training, validation, and test losses and Mel-cepstral distortion for the multi-task learning model

Model	Training loss	Validation loss	Test loss	MCD
mtl-3-fc-first	0.005065	0.006351	0.007242	9.9801

The lack of significant improvements in the MTL model's Mel-spectrogram predictions could be due to the difficulty of accurately predicting EMA features. Mean absolute differences between ground truth and predicted EMA features computed on the test set (figure 4.17) exhibit greater loss for dysarthric samples than healthy samples. This observation could be the result of greater variation within dysarthric EMA features than healthy features, caused by dysarthric articulatory constraints and lack of articulatory control. However, since spectral and EMA features are highly correlated, if the model was able to learn an accurate healthy-to-dysarthric

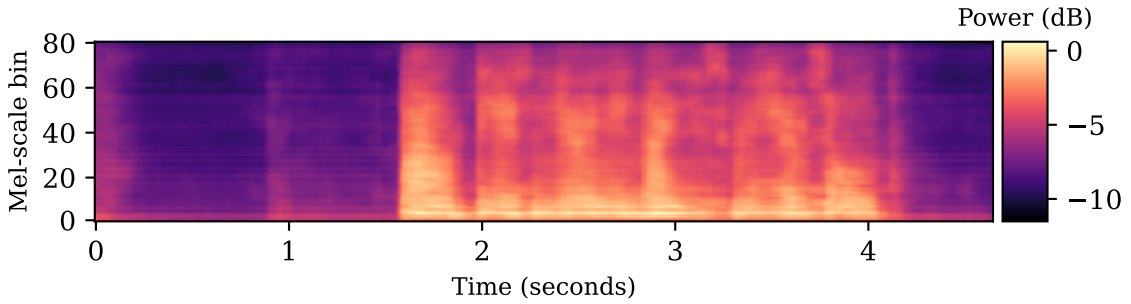

Figure 4.16: Healthy-to-dysarthric Mel-spectrogram produced by the multi-task model (mtl-3-fc-first) for an unseen utterance.

Mel-spectrogram conversion mapping, the internal feature maps could be used to predict EMA features accurately. The model's greater inaccuracy in predicting dysarthric EMA features supports my judgement in section 4.5 that models performed healthy-to-healthy conversion more successfully than healthy-to-dysarthric conversion.

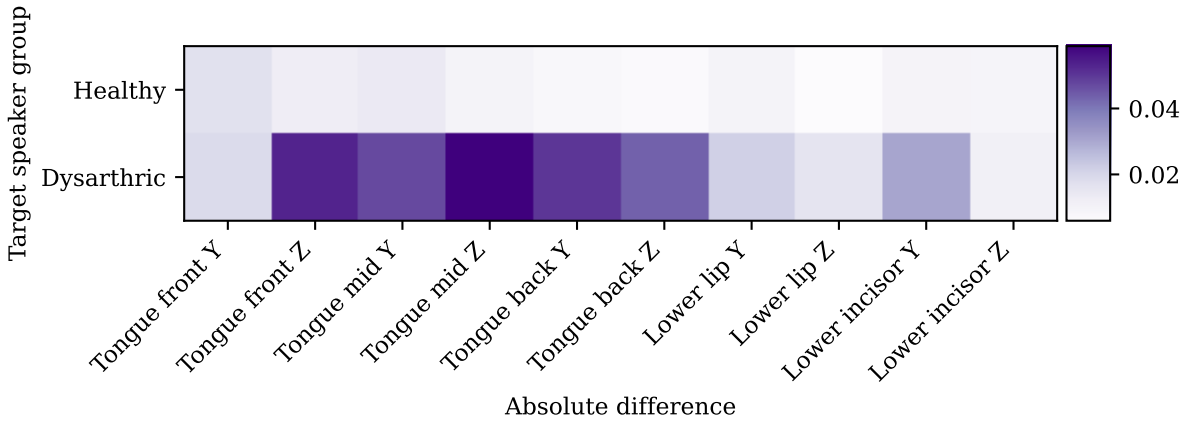

Figure 4.17: Mean absolute differences between ground truth electromagnetic articulograph features and those predicted by the multi-task learning model, computed on the test set for healthy and dysarthric target speakers.

The multi-task model's latent space appears to cluster dysarthric and healthy speakers more distinctly than previous models', with dysarthric speakers centered lower on the y-axis than healthy speakers. The dysarthric speaker outside of this cluster (F04) exhibited the least severe dysarthria of the TORGO dataset. This result could suggest that that the extra information provided by EMA features allows the model to learn a latent space more relevant to dysarthria, perhaps because EMA healthy and dysarthric readings are more distinctly different than healthy

and dysarthric Mel-spectrograms.

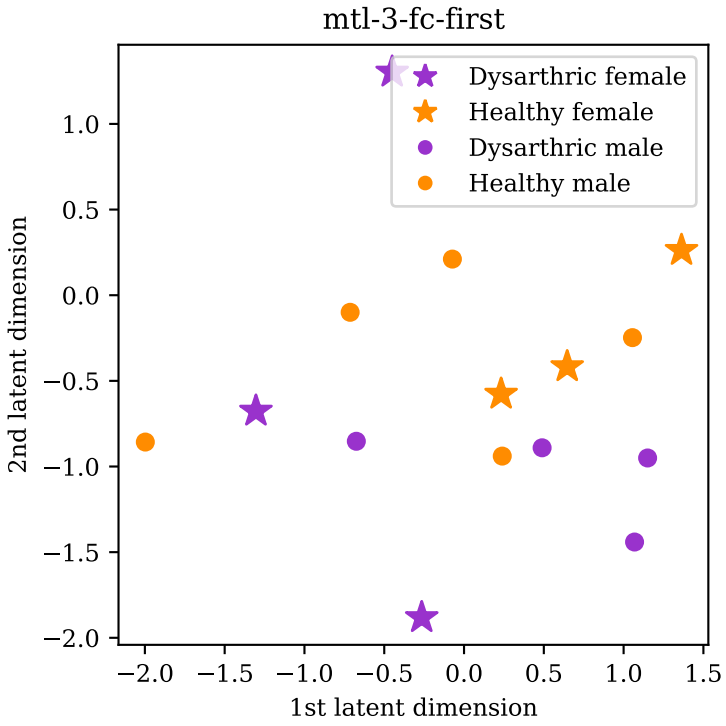

Figure 4.18: Latent embedding space learned by the multi-task model.

Chapter 5

Evaluation

5.1 Experimental setup

To evaluate the two most promising healthy-to-dysarthric conversion systems, I conducted two expert evaluation listening tests. I deemed expert evaluation the most appropriate method since recruiting enough participants with strong knowledge of dysarthria to run a multi-listener test would be difficult. '**se-3-fc-first**' refers to the conversion model with speaker embeddings and 3 fully-connected layers before the convolutional layers presented in section 4.5. '**mtl-3-fc-first**' refers to the multitask learning model presented in section 4.6.

The same set of 10 held-out healthy samples from five TORGO control speakers was converted by both systems, to enable direct sample-to-sample comparison between systems. I originally planned to test systems' ability to generalise to unseen source speakers but this was not feasible due to time limitations. Duration of healthy input Mel-spectrograms was modified following the method outlined in section 4.1. Dysarthric Mel-spectrograms were time-aligned to the stretched healthy ones to allow for mean squared error (MSE) to be calculated on the test set. Conversion was additionally performed between two healthy source and target speakers, using the same utterance set, allowing us to determine whether systems could better convert to healthy or dysarthric speakers, and whether conversion artefacts present in healthy-to-healthy samples were misdiagnosed as dysarthria. Duration of healthy-to-healthy samples was not modified. As a control, the 10 held-out natural healthy samples were time-stretched and vocoded using HiFi-GAN, to determine whether vocoding artefacts were misdiagnosed as dysarthria, and whether my test systems performed more dysarthric-sounding conversion than simply modifying the speaking rate of natural healthy samples. 10 held-out natural dysarthric speech samples vocoded using HiFi-GAN were included, to validate my expert evaluator's classification accuracy.

5.1.1 Speaker similarity

The first listening test sought to determine the accuracy of source-to-target speaker conversion by measuring speaker similarity. It was loosely based on that used to evaluate models submitted to the 2016 Voice Conversion Challenge [38]. Each of the 40 converted samples was presented alongside two natural samples from its source and target speaker. The evaluator was asked to choose which natural speaker the converted sample sounded most like. The natural samples were different utterances than those in the test set, to ensure classifications were made based on voice characteristics and not common linguistic content or utterance-specific features. The order of systems and test utterances was randomised. The order of source and target samples were randomised.

5.1.2 Dysarthria classification

The second test investigated whether spectral features produced during conversion sounded like dysarthria. All 40 converted samples and 20 control samples were presented, and the evaluator was asked whether each sample sounded ‘dysarthric‘ or ‘not dysarthric‘. ‘Not dysarthric‘ was used as a class instead of ‘healthy‘, since properties of converted speech could sound unnatural or impeded in ways uncharacteristic of dysarthria. In both tests, text boxes were presented under each question, allowing the evaluator to comment on samples. N1 Chi-squared tests (α=0.05, df=9) were used to determine the statistical significance of results between models and conditions.

5.2 Results

5.2.1 Speaker similarity

Our expert evaluator classified at least 50% of all converted samples as sounding more like their target speaker than their source speaker (figure 5.1). This result suggests speaker embeddings were somewhat successful in converting speaker characteristics.

The speaker embedding model appears to perform healthy-to-healthy conversion more successfully than healthy-to-dysarthric conversion. An N1 Chi-squared test indicated this difference was on the borderline of the statistical significance at the α=0.05 level (p=0.0501). This difference could be due to greater variation in dysarthric speakers’ data, making it more difficult to learn dysarthric speaker embeddings which are useful for accurately outputting dysarthric Mel-spectrograms.

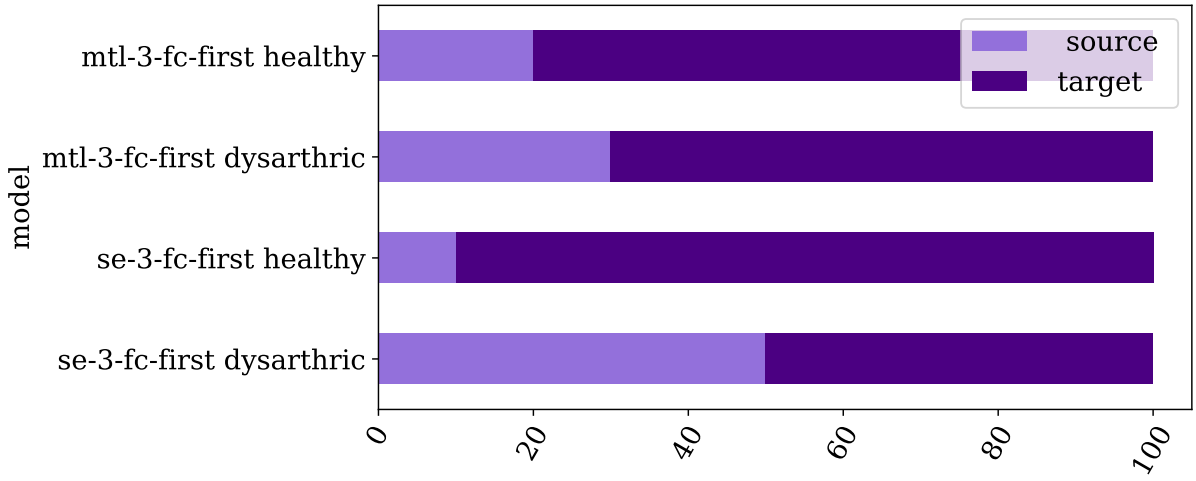

Figure 5.1: Proportion of converted samples classified as sounding more like source/target speakers (N=10).

The multi-task model exhibited no statistically significant difference in classification of healthy-to-healthy and healthy-to-dysarthric converted samples (p=0.6056). The multi-task model's dysarthric converted samples were classified as sounding more like the target speaker more often than the speaker embedding model's, but this difference does not appear to be statistically significant (p=0.792).

This mode of evaluation was clearly limited by the small sample size. Running evaluations with more samples and more participants would better allow differences between models to be analysed.

5.2.2 Dysarthria classification

Our expert classified the two conversion models' healthy-to-dysarthric samples as mostly 'not dysarthric' (figure 5.2). While the multitask model's were classified more often as 'Dysarthric' than the speaker embedding model's, the N1 Chi-squared at (α=0.05) determined this difference was not statistically significant (p=0.0671). The multitask model's healthy-to-dysarthric samples were classed as 'Dysarthric' at the same rate as the duration-modified vocoded natural healthy speech. N1 Chi-squared tests performed between each conversion model's healthy-to-healthy and healthy-to-dysarthric classification results indicated there were no statistical differences between the groups (mtl-3-fc-first p=0.0671, se-3-fc-first p=0.3173). These results suggest that the spectral mappings the conversion models learned did not produce speech features which sound dysarthric.

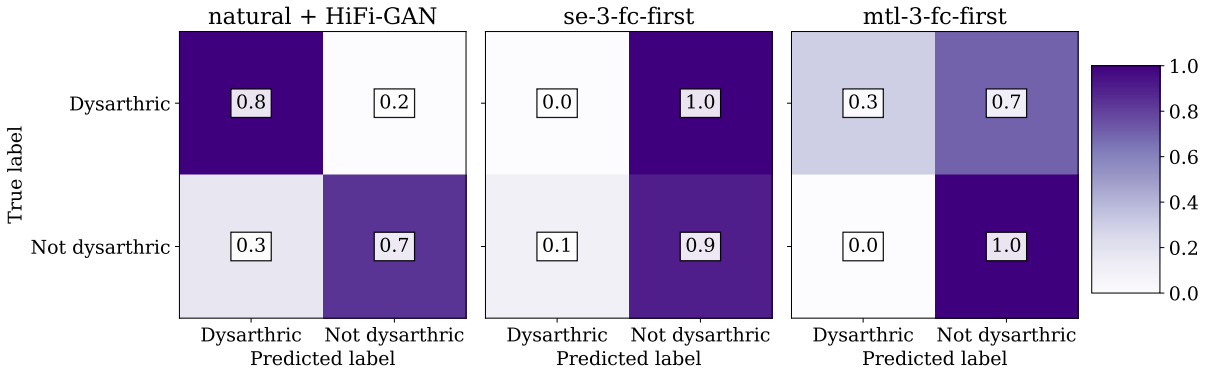

Figure 5.2: Proportion of converted and natural vocoded samples classified as 'Dysarthric' or 'Not dysarthric' (N=10).

Chapter 6

Discussion

The results presented in section 5 appear to show that both conversion models successfully transform some speaker characteristics, so that the majority of samples sounded more like the target speaker than the source speaker, but that healthy-to-dysarthric conversion does not produce output which sounds convincingly dysarthric. In terms of dysarthria simulation, this result is worse than Jiao et al.'s, whose healthy-to-dysarthric samples were classified as 'Dysarthric' 65% of the time, and samples with consensus from all SLTs were classified as 'Dysarthric' 76% of the time. My samples' poor intelligibility and audio quality likely affected these results. Lynda commented that many converted samples were difficult to classify, and that she sometimes had to guess, because of the amount of noise and non-dysarthric distortions generated by the conversion model. This limitation means the results presented here are not enough to conclude that my approach to modelling a healthy-to-dysarthric speech transformation was unsuccessful. Effort should be made to reduce noise in predicted Mel-spectrograms, then samples should be re-evaluated to more accurately determine conversion success.

Even with speaker encodings, the predicted Mel-spectrograms contained widespread diffusion of power across many frequencies instead of being clustered in bands, which translated as noise artefacts in the resyntheised samples. Given the high noise levels, evaluation results were difficult to obtain and analyse. Noise obscured speech features so much that detecting dysarthric features in converted samples was near impossible. However, objective evaluations of the speaker embedding model se-3-fc-first and the multitask model mtl-3-fc-first both showed potential. Model se-3-fc-first's samples had the lowest Mel-cepstral distortion (9.592). Model mtl-3-fc-first learned a latent speaker embedding space which appeared to model some difference between dysarthric and healthy speakers, but its MCD score was higher than se-3-fc-first's, probably due to increased noise levels. Future work could investigate whether moving speakers' y-axis position in this latent space varies dysarthric severity in the output.

Further research should focus on modifying model parameters to reduce noise in the output. The networks' objective function- mean squared error loss between expected and predicted Mel-spectrograms- aims to produce spectrograms which are similar to the natural dysarthric ones in absolute pixel values, without any consideration of larger patterns. Learned mapping functions may be improved with the addition of a secondary loss term which calculates contrast in a small area of each spectrogram frame and penalises high energy spread. This combined loss could result in more distinct clustering of power in formants. An alternative way of encouraging the network to produce more realistic dysarthric Mel-spectrograms would be to use a loss function which considers similarity to both the dysarthric target Mel-spectrogram and the healthy source one, e.g. by introducing a penalty for remaining too close to the source. Another method could be to implement a Generative Adversarial Network (GAN), as used in [7] and [5]. A binary classifier distinguishing real from generated Mel-spectrograms could be jointly trained along with the generator, and the network set to optimise the adversarial objective functions of both generating dysarthric Mel-spectrograms and distinguish real from generated Mel-spectrograms. If trained successfully, this addition should cause the network to generate more 'real'-looking spectrograms, with less noise and more tightly clustered power. Patch-GAN [39] would be a suitable discriminator for the CNN models, since it computes class likelihoods from many small patches of a given image, so can take variable sized inputs without using batching strategies which affect the random sampling of training batches. GANs were not explored in this project due to time constraints and their difficulty to train.

It is also possible that inaccurate time-alignment between source and target samples contribute to noisy output. To address this problem, non-parallel methods which remove the problem of time-alignment entirely, are worth exploring.

Chapter 7

Conclusion

In this work I investigated the appropriateness and success of parallel voice conversion techniques using Mel-spectrograms as spectral features for healthy-to-dysarthric voice transformation. I first evaluated the accuracy of dynamic time-warping for time-aligning healthy and dysarthric spectrograms, and determined that accuracy was likely high enough to justify experimenting with parallel conversion methods. I presented four neural voice conversion models, which varied in their ability to generate dysarthric Mel-spectrograms. The fully-connected and fully convolutional models were unsuccessful, likely due to their inability to resolve differences in source and target speakers' fundamental frequencies across all training samples. I used speaker embeddings to alleviate this problem, and explored three ways of incorporating the embeddings into the model. Simply concatenating embeddings to input spectrograms was unsuccessful, since it was outside most feature map cells' receptive fields. Adding fully-connected layers before convolutional layers made speaker embeddings available to every cell in the first convolutional layer. This method successfully modified F0 with the input of different target speaker embeddings. I finally adopted a multi-task training procedure using electromagnetic articulograph features, which enabled the model to learn a latent speaker embedding space with more interpretable division between source and target speakers, but increased noise and non-dysarthric distortion in output Mel-spectrograms. Listening test evaluations conducted with a speech and language therapist indicated that while conversion of to target speaker was somewhat successful, simulated dysarthric features were not recognisable, possibly due to high noise levels.

Bibliography

[1] Y. Hwang, H. Cho, H. Yang, D.-O. Won, I. Oh, and S.-W. Lee, "Mel-spectrogram augmentation for sequence to sequence voice conversion," *arXiv preprint arXiv:2001.01401*, 2020.

[2] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[3] E. K. Hanson, "Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment, and augmentative and," *Journal of Medical Speech-Language Pathology*, vol. 19, no. 3, pp. 12–30, 2011.

[4] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[5] M. Pasini, "Melgan-vc: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms," *arXiv preprint arXiv:1910.03713*, 2019.

[6] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[7] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6009–6013, IEEE, 2018.

[8] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 7–12, 2019.

[9] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep griffin–lim iteration: Trainable iterative phase reconstruction using neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 37–50, 2020.

[10] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," *arXiv preprint arXiv:1907.04743*, 2019.

[11] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[12] M. Stella, A. Stella, F. Sigona, P. Bernardini, M. Grimaldi, and B. G. Fivela, "Electromagnetic articulography with ag500 and ag501.," in *Interspeech*, pp. 1316–1320, 2013.

[13] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pp. 1–5, IEEE, 2018.

[14] E. Williams, "msc-project-voice-conversion." `https://github.com/evelyndjwilliams/msc-project-voice-conversion`, 2021.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[16] PyTorch, "Torchaudio 0.9.0," 2021.

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, Citeseer, 2015.

[18] J. Yamagishi, C. Veaux, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[19] H. Schreiber and M. Müller, "A single-step approach to musical tempo estimation using a convolutional neural network.," in *Ismir*, pp. 98–105, 2018.

[20] dtw python, "dtw-python 1.1.10," 2021.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009.

[22] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[23] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech communication*, vol. 16, no. 2, pp. 207–216, 1995.

[24] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020.

[25] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Conference on learning theory*, pp. 907–940, PMLR, 2016.

[26] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[27] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," in *Proceedings of the 32nd international conference on neural information processing systems*, pp. 2488–2498, 2018.

[28] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[29] J. A. Longster, *Concatenative speech synthesis: a Framework for Reducing Perceived Distortion when using the TD-PSOLA Algorithm*. PhD thesis, Bournemouth University, 2003.

[30] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[31] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling dnn-based speech synthesis using input codes," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4905–4909, IEEE, 2017.

[32] R. Yamamoto, "pyreaper: A python wrapper for reaper (robust epoch and pitch estimator)." `https://github.com/r9y9/pyreaper`, 2020.

[33] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[34] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6209–6213, IEEE, 2021.

[35] P. Wang, T. N. Sainath, and R. J. Weiss, "Multitask training with text data for end-to-end speech recognition," *arXiv preprint arXiv:2010.14318*, 2020.

[36] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4460–4464, IEEE, 2015.

[37] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7774–7778, IEEE, 2020.

[38] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results.," in *Interspeech*, pp. 1637–1641, 2016.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.