Multi-MelGAN Voice Conversion for the creation of under-resourced child speech synthesis

Abstract: Voice conversion (VC) is an important technique for the development of text-to-speech voices in the use case of lacking speech resources. VC can convert an audio signal from a source speaker to a specific target speaker whilst maintaining the linguistic information. The benefit of VC is that you only require a small amount of target data which therefore makes it possible to build high quality text-to-speech voices using only a limited amount of speech data. In this work, we implement VC using a Melspectrogram Generatative Adversarial Network called MelGAN-VC. This technique does not require parallel data and has been proven successful on as little as 1 hour of target speech data. The aim of this work was to build child voices by modifying the original one-to-one MelGAN-VC model to a many-to-many model and determine if there is any gain in using such a model. We found that using a many-to-many model performs better than the baseline one-to-one model in terms of speaker similarity and the naturalness of the output speech when using only 24 minutes of speech data.

Keywords: voice conversion, speech technology, child speech

1. Introduction

When developing text-to-speech synthesis systems using neural architectures for resource-scarce languages, we are faced with the limitation of data. Deep Neural Networks (DNNs) are typically data hungry algorithms and therefore building voices that span across accents, speakers and speaking styles become expensive and impractical. Thus, finding methods whereby one can convert from a model that is rich in data to another model that lacks in richness is important. Such a technique is called Voice Conversion (VC). VC involves multiple speech processing techniques, such as speech analysis, spectral conversion, prosody conversion, speaker characterization, and vocoding. VC can convert an audio signal from a source speaker to a specific target speaker whilst maintaining the linguistic information. It is used in various real-world applications such as personalized text-to-speech (TTS) and video dubbing. With the recent advancements made in the field, it is now possible to produce human-like voice quality with high speaker similarity to the target voice. In this work, we are interested specifically in using voice conversion for the creation of child voices for under-resourced languages.

Early studies of VC techniques required parallel data, where speech of the same linguistic content was available from both the source and the target speaker. However, collecting parallel speech data is not always feasible and if such data is collected then one needs to place further efforts in performing time alignments. In general, the quality and conversion effect obtained with non-parallel methods are usually limited compared with methods using parallel data due to the disadvantage related to the training condition. Thus, developing non-parallel methods that produce high audio quality that are comparable with that of parallel methods can be very challenging especially when there is a large acoustic gap between the data collected from the source and target speakers. The recordings acquired from target speakers for customisation are usually of different acoustic conditions from the source speech data. For example, the adaptation data is usually recorded with diverse

speaking prosodies, styles, emotions, accents and recording environments. The mismatch in these acoustic conditions makes the source model difficult to generalize and leads to poor adaptation quality.

Recent studies on voice conversion are moving towards non-parallel training data [1, 2] and have opened up opportunities for new applications. The challenge now is geared towards establishing the mapping between non-parallel source and target utterances. VC has been previously applied using several methods such as Gaussian mixture models (GMMs) [3] and are gradually being replaced by DNNs, which include feed forward, recurrent and convolutional neural networks [4,5] and more recently GANS [6]. More recent VC research focuses on non-parallel VC that requires no parallel utterances for training as it is not easy to collect large amounts of parallel data from many speakers. The common non-parallel VC approach consists of two steps. Firstly, a VC model tries to disentangle speaker identity from the content information in the speech and generates a latent representation. Secondly, this latent representation is used to generate speech in the target speaker's voice by conditioning on some embedding of the target speaker.

VAE is a popular VC method which is based on encoder-decoder architecture and is fed a specific label as input and the network uses this input label to learn how to convert from the source to the target. However, a disadvantage of this method is that the decoder tends to be over smoothed and therefore results in a poor-quality output. GANs offer a general framework for training a generator network in such a way that it can deceive a real/fake discriminator network. GAN based models overcome this weakness and have shown promising results. GANs comprise a generator and a discriminator. The discriminator learns to distinguish between real and fake samples whilst the generator learns to generate fake samples that are indistinguishable from real samples. MelGAN-VC is a voice conversion method that relies on non-parallel speech data and is capable of converting audio signals of arbitrary length from a source to a target speaker. MelGAN-VC is derived from the GAN architecture which comprises a generator and discriminator. In addition, a Siamese network is used to preserve speech information during the conversion process whilst still maintaining the speaker information of the target speaker.

2. Objectives

The objectives of this work were as follows:

1. To convert speech of a source adult speaker to a target child speaker
2. To convert speech using a limited amount of target speech data
3. To convert speech using non-parallel data
4. To convert speech of several speakers using a single model

3. Methodology

In this work, the MelGAN-VC method was adopted. Benefits of MelGAN-VC include: (1) Requires no parallel utterances, transcriptions or time alignments, (2) Conversion happens in real time, (3) Requires only an approximation of 1 hour of training examples for source and target speaker (has not yet been tested with less data) (4) produces mel-spectrograms which can be used in a MelGAN vocoder to generate high quality speech output. These advantages help us achieve the objectives set out in Section 2.

The training procedure of MelGAN-VC is illustrated in Figure 1. Source speaker A is converted into target speaker B. The source and target speakers' data can be non-parallel. The model works by first converting the speech waveforms of both the source and target speakers into mel-spectrograms. To deal with the arbitrary length, a specific length is chosen, and all mel-spectrograms are split into equal chunks of the specific length. During training, mel-spectrogram ***a*** is further split into *n* equal parts. In our case, it is split into 2 parts ***a1*** and ***a2***. Similarly mel-spectrogram ***b*** (from target speaker) is split into 2 equal parts ***b1*** and ***b2***. Each part is fed to the Generator. ***a*** is fed to the generator for mapping of distribution ***a*** to distribution **b**. The Generator predicts a mel-spectrogram ***G(a1)*** and ***G(a2)*** which is then concatenated. This is referred to as "fake" ***b***. This concatenated output together with the target mel ***b*** before being split) is passed to the Discriminator. The Discriminator works as a classifier to determine whether the input is from the real distribution or the fake generated distribution. The Generator and Discriminator work together. The loss of the Generator is minimized so that it gets better at producing mel-spectrograms that resemble the target and the Discriminator loss is maximized so that it gets better at differentiating between the real and fake which in turn improves the generator. The Siamese network takes the 2 splits ***a1*** and ***a2*** as input and the output of those splits ***G(a1)*** and ***G(a2).*** A travel loss is included which aims at making the generator preserve vector arithmetic in the latent space produced by the Siamese network. Furthermore, an identity loss is included due to loss of linguistic information during the conversion process.

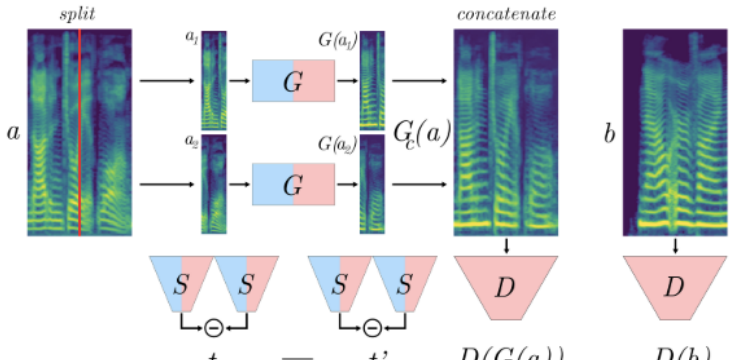

Figure 1: Overview of MelGAN-VC architecture consisting of Generator (G), Discriminator (D) and Siamese network (S) taken from [6]

4. Experimental Design

4.1 Dataset

The data used in this work is a subset of an in-house Afrikaans adult and child male and female TTS corpora . The adult male and female corpora were recorded in a studio with a professional voice artist at a 44.1 kHz sampling rate with 16 bits precision. The male child corpus was recorded in studio at a 44.1kHz sampling frequency with 16 bits precision. The female child was not studio recorded and was recorded at a 44.1kHz sampling frequency with 16 bits precision. Approximately, 2170 (+-3 hours) utterances of adult speech per speaker, 1890 (+-1 hour) utterances of the male child and 400 (+-24minutes) of speech data was used for training.

4.2 Pre-processing

The speech data for all speakers were either up sampled or down sampled to 24kHz sampling frequency which can later be compatible with plugging in our trained multiband-MelGAN vocoder. No text annotations were needed for the training of the VC models.

To prepare the data for training, some digital signal processing steps were first carried out. These included loading the speech files into waveform arrays, generating spectrograms from these arrays and then transforming them into mel-spectrograms. The mel-spectrograms needed to be split into equal size chunks and then cropped, shuffled and prepared into training batches.

4.3 Implementation

A baseline voice was trained using the open-source code-base by the original author (https://github.com/marcoppasini/MelGAN-VC) with minor changes such that the code was able to train successfully on our internal speech corpora. The baseline voices were one-to-one voices that was trained using speech from the adult male as source and male child as target. Additionally, we trained the female adult to the female child.

Initially, a baseline voice was trained with only 1 hour of source data, but it was discovered that the output voice sounded more intelligible when using all the source data for the relevant speakers which approximate to 3 hours of speech data.

The proposed model is a multi-MelGAN-VC model. Unlike the original codebase which only converts from a single source speaker to a single target speaker, the multi-MelGAN-VC model converts from multiple source speakers to multiple target speakers. This is a more practical approach as it does not require separately trained models every time a particular voice needs to be trained. The multi-MelGAN-VC can train many speakers simultaneously which makes it less costly and more practical. Furthermore, the hypothesis at the start of this work is that the output speech quality of the multi-MelGAN-VC model will be higher than that of the single model architecture due to more data being used to update the model parameters. In order to implement this new model architecture, GAN conditioning was required. The model needed to be conditioned using speaker information of both the source and target speakers. The speaker information was captured in a 1-hot vector.

5. Results

To evaluate the quality of the proposed model a listening test was set-up through a web-browser. The evaluation consisted of a MUSHRA test to evaluate naturalness and speaker similarity. 10 audio samples for each model were generated for each of the relevant tests. A separate listening test was conducted for the male target speaker (Eval A) and female target speaker (Eval B). The voices used in the evaluation are presented in Table 1. 30 native Afrikaans speakers, aged 24 to 61 years, were recruited for Eval A and 12 native speakers, aged 25 to 57 years were recruited for Eval B.

Table 1: Summary of systems built for evaluation for Eval A and Eval B

	Source Speaker(s)	Target Speaker(s)	Type
Eval A: System A	Antowan	Johan	One-to-One (2spkr)
Eval A: System B	Antowan, Nadia	Johan	Many-to-One (3spkr)
Eval A: System C	Antowan, Nadia	Johan, Elsje	Many-to-Many(4spkr)
Eval B: System A	Nadia	Elsje	One-to-One (2spkr)
Eval B: System B	Nadia, Antowan	Johan, Elsje	Many-to-Many(4spkr)

For Eval A, the speech listened to by the listeners were always the conversion of Antowan to Johan for each of the models . The results of the MUSHRA test for Eval A are presented in Figure 1.

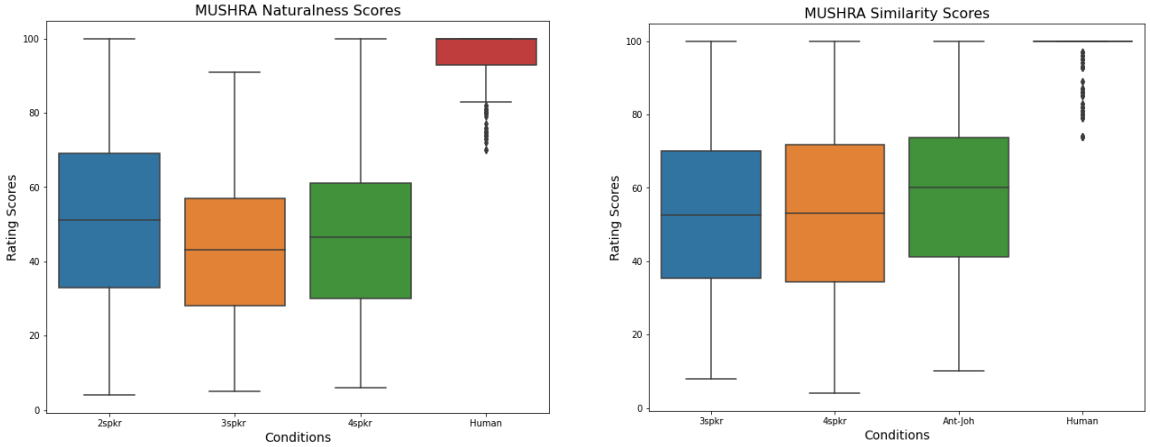

Figure 2: (left) Eval A Naturalness scores, (right) Eval A Similarity scores

The results show that the baseline one-to-one model was rated slightly more natural than the proposed models and slightly more like the human reference in terms of speaker similarity. However, using a one-way ANOVA statistical test, the F statistic was 2.72 with a *p_value* of 0.07 for naturalness and F statistic was 0.55 with a *p_value* of 0.58 for similarity. Alpha is 0.05. Since the *p_value* > alpha, we confirm the null hypothesis. Therefore, the systems compared in Eval A were not found to be statistically different from one another. We believe that no gain was found as 1 hour of target speech was sufficient to model a good enough output voice quality that is comparable to the many-to-many model.

For Eval B, the listeners listened to the audio produced by the Nadia to Elsje conversion. The results of the MUSHRA test for Eval B are presented in Figure 2. The results show that the many-to-many model performs better than the baseline one-to-one model in both naturalness and speaker similarity.

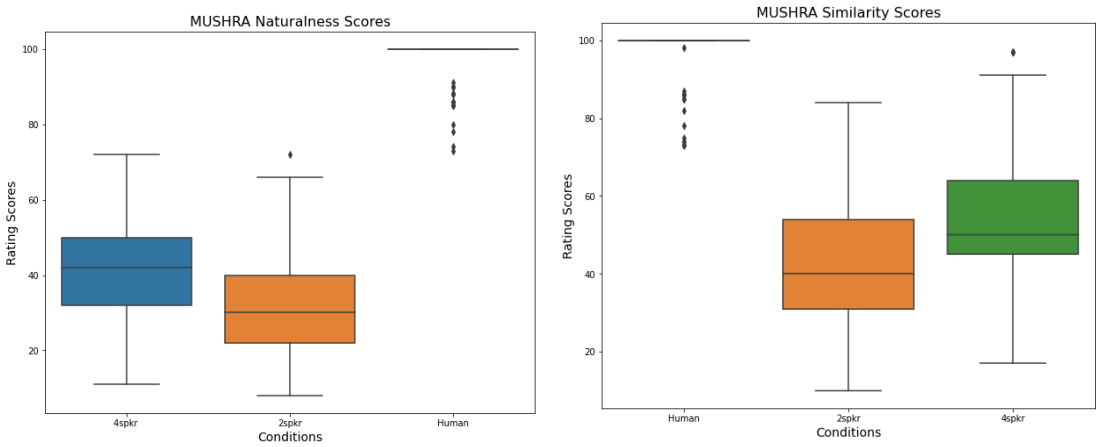

Figure 3: (left) Eval B Naturalness scores, (right) Eval B Similarity scores

Using a one-way ANOVA statistical test, the F statistic was 6.0 with a *p_value* of 0.002 for naturalness and F statistic was 3.9 with a *p_value* of 0.05 for similarity. Alpha is 0.05. Since the *p_value* < alpha, we reject the null hypothesis. Therefore, the systems compared in Eval B were found to be statistically different from one another. Therefore, the many-to-many model is shown to perform significantly better than the one-to-one model in the scenario of using only 24 minutes of target speech data.

6. Business Benefits

The business benefits of this work are that building unique text-to-speech voices is both time consuming and expensive. Building voices that span across speakers, accents and languages would require training separate model on the relevant data suited to match the target speaker desired. However, by using a many-to-many voice conversion model, only a single model is trained, and several target speakers voices can be generated, therefore reducing both time and cost of the voice building process.

7. Conclusions

By using the proposed many-to-many MelGAN-VC method we were able to successfully convert speech from a source adult speaker to a target child speaker using only a limited amount of parallel speech data. In addition, the model can generate more than one target speaker from training only a single model. The results showed that when the target data is limited, in this case only 24 minutes of data, the multi model can compensate for the lack of data and produce a voice of better quality than using a single speaker model. In the male child voice both the baseline and multi-model were of the same level of quality because there was sufficient data for the single speaker model to produce speech of high quality.

References

[1] Zhang, Mingyang, et al. "Transfer learning from speech synthesis to voice conversion with non-parallel training data." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 1290-1302.

[2] Kaneko, Takuhiro, et al. "Cyclegan-vc2: Improved Cyclegan-based non-parallel voice conversion." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[3] Ohtani, Yamato, et al. "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation." (2006).

[4] Zhang, Jing-Xuan, et al. "Sequence-to-sequence acoustic modelling for voice conversion." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.3 (2019): 631-644.

[5] Mohammadi, Seyed Hamidreza, and Alexander Kain. "Voice conversion using deep neural networks with speaker-independent pre-training." *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014.

[6] Pasini, Marco. "MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms." *arXiv preprint arXiv:1910.03713* (2019).