REMAP, WARP AND ATTEND: NON-PARALLEL MANY-TO-MANY ACCENT CONVERSION WITH NORMALIZING FLOWS

Abdelhamid Ezzerg[1] *Thomas Merritt*[1] *Kayoko Yanagisawa*[1] *Piotr Bilinski*[1]
Magdalena Proszewska[2*] Kamil Pokora[1] Renard Korzeniowski[1] Roberto Barra-Chicote[1]
Daniel Korzekwa[1]

[1] Amazon Alexa
[2] Jagiellonian University, Poland
{*ezzerg, thommer, yakayoko, bilipiot, kamipoko, korenard, rchicote*}*@amazon.com*

ABSTRACT

Regional accents of the same language affect not only how words are pronounced (i.e., phonetic content), but also impact prosodic aspects of speech such as speaking rate and intonation. This paper investigates a novel flow-based approach to accent conversion using normalizing flows. The proposed approach revolves around three steps: remapping the phonetic conditioning, to better match the target accent, warping the duration of the converted speech, to better suit the target phonemes, and an attention mechanism that implicitly aligns source and target speech sequences. The proposed remap-warp-attend system enables adaptation of both phonetic and prosodic aspects of speech while allowing for source and converted speech signals to be of different lengths. Objective and subjective evaluations show that the proposed approach significantly outperforms a competitive CopyCat baseline model in terms of similarity to the target accent, naturalness and intelligibility.

Index Terms: Accent conversion, normalizing flows, Flow-TTS, CopyCat.

1. INTRODUCTION

Speech attributes' conversion aims to generate synthetic speech from a source one by changing only attributes of interest. Speech attributes' conversion includes two notable tasks: Voice conversion (VC) [1, 2, 3, 4], where the goal is to change the perceived speaker's identity while maintaining the linguistic and prosodical content of the source speech, and Accent Conversion (AC), whose goal is to generate new speech that sounds as if the source speaker was natively speaking the target accent. Accent conversion has a wide range of applications which include personalized voices for AI assistants and pronunciation feedback for language learners [5, 6].

There are two main families of models that convert speech attributes: parallel models and non-parallel models [7]. Parallel methods for AC require a dataset of pairs of the same utterance spoken in different accents. Such data is expensive to collect and difficult to obtain. Therefore, we focus in this paper on non-parallel approaches to accent conversion which only require transcribed speech.

Accent conversion approaches are varied in terms of the used tools and paradigms. Earlier AC methods relied on manipulating various features such as duration, intonation, pitch contours etc. using Hidden Markov Models (HMMs) [8, 9, 10, 11]. It is worth noting that these approaches were trained on parallel data and require access to a reference speech in the target accent at inference time. With the advent of more modern machine learning architectures, recent AC approaches started leveraging neural networks. Regardless of the used tools (DNNs or HMMs/GMMs), it is possible to distinguish multiple AC approaches in the literature. W. Li et al. [12] leverages a Text-to-Speech (TTS) system for accent conversion in a two-stage training pipeline in order to create synthetic parallel data which is later used to train a conversion model. S. Aryal et al. [13, 14] uses articulation data to improve accent conversion results; however, articulation data is difficult to collect and therefore limits the applications of this approach. Moreover, the approach is trained on a corpus of parallel data which further limits its applications. Other approaches have tackled the AC problem using phonetic posteriograms [10, 11]. While these approaches are non-parallel at inference time, they need a parallel training corpus. A common aspect of existing approaches is that they assume the availability of parallel data [15] or the availability of the source speech and a reference speech of the target accent at the time of inference [12]. A notable exception to the previous observation is the work conducted by Z. Wang [16] where they tackle both VC and AC simultaneously. In their work, the author use an ASR system, in order to extract content information only, followed by a Tacotron-like model, conditioned on both speaker and accent IDs, which generates mel-spectrogram from the ASR-extracted features. In our proposed approach, only one source spectrogram and its text transcription are needed during training and inference. Moreover, only the conversion model needs to be trained as opposed to the two-stage training approach of [16]. To the best of our knowledge, this is the first approach to satisfy both properties.

In this paper, we propose a novel AC approach based on normalizing flows [17, 18, 19]. Our model shares similarities with other flow-based approaches applied to speech [20, 21, 22, 23, 24], however, to the best of our knowledge, this is the first application of normalizing flows to the AC task. Moreover, we extend the flow-based conversion approaches to allow for conversion between source and target sequences of different lengths by using an attention mechanism, similar to [25], and warping of speech phonemes' durations.

The main contributions of the paper are the following:

- We propose a flow-based accent conversion model that significantly outperforms the competitive CopyCat baseline [2]. The proposed model only requires a source utterance and its transcription and does not require the use of parallel data or a reference utterance at the inference stage.
- We introduce a novel scheme for the AC task, referred to as 'remap, warp and attend'. The remap stage consists of remapping the conditioning phonetic sequence to match that of the

*Work performed during an internship at Amazon.

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

target accent, the warp stage adjusts the durations to better suit the remapped phonemes while the attend stage learns to implicitly align the source and target speech signals whose phoneme sequences are of different lengths.

The remainder of the paper is organised as follows: Section 2 describes the proposed model, Section 3 presents the evaluation strategy used and discusses the results, while Section 4 presents the conclusions.

2. PROPOSED APPROACH

Our approach is inspired by [26] which used a Flow-TTS-like architecture [20] to achieve state-of-the-art results in the VC task. We extend this model in order to allow it to perform accent conversion. We achieve this goal by introducing a novel remap, warp and attend technique that significantly improves the quality of accent conversion. Figure 1 summarizes the architecture of the proposed model. Figures 2 demonstrates the accent conversion procedure.

Our model is based on Normalizing Flows, and is able to encode a mel-spectrogram x into a latent sequence z:

$$z = f^{-1}(x; ph, spk, acc), \tag{1}$$

and then decode the latent sequence z back into the original mel-spectrogram using the inverse transform f:

$$x = f(z; ph, spk, acc), \tag{2}$$

where ph is the phoneme sequence, spk is the speaker embedding, and acc is the accent embedding. The Flow-based generative model explicitly learns the data distribution $p(x)$, and therefore it is optimized by minimizing the negative log-likelihood. Similar to the VC model in [26], we use explicit durations, obtained using either a forced alignment [27] or predicted by a duration model, to upsample the phonetic representations, and we introduce utterance-level speaker and accent embeddings which are pre-trained as in [28]. However, we chose not to condition the model on f0 and voiced/unvoiced flags in order to assess separately the effects of our proposed approach and prosody modelling on the AC's performance. Once we establish the effectiveness of our model for AC, we can extend it to incorporate f0 and voiced/unvoiced flags as it was the case for [26].

2.1. Accent conversion via phoneme remapping 'remap'

The Flow VC [26] model is extended to perform accent conversion. We propose to encode a mel-spectrogram x_{source} into a latent sequence z using a phoneme sequence ph_{source} and an accent embedding acc_{source} corresponding to the source accent:

$$z = f^{-1}(x_{source}; ph_{source}, spk_{source}, acc_{source}), \tag{3}$$

and decode the latent sequence z into a mel-spectrogram x_{target} using a new phoneme sequence ph_{target} and a new accent embedding acc_{target} corresponding to the target accent:

$$x_{converted} = f(z; ph_{target}, spk_{source}, acc_{target}), \tag{4}$$

where $x_{converted}$ is the mel-spectrogram output of the conversion procedure and spk_{source} is the speaker embedding corresponding to the source speaker.

Accent embeddings help the model disambiguate how to pronounce phonemes that share the same symbol in different accents. The phoneme sequences ph_{source} and ph_{target} represent the same text but were extracted with two different grapheme-to-phoneme models, each dedicated to the source and target accents respectively. The 'remap' stage is altering the phoneme sequence between the encoding and decoding stages to match the target accent (Equations 3 and 4).

2.2. Duration warping 'warp'

Remapping the phoneme sequences between the encoding and decoding stages improves accent control. However, since we remap the source phoneme sequence into a target sequence without further manipulations, the durations of the target phonemes are still the same as those of the source sequence's. We argue that it is possible to generate even more natural-sounding speech in the target accent by forcing both the phonemes and their durations to better match the target accent.

To achieve this, we propose duration warping: the goal of duration warping is to make the encoded sequence of latent representations (whose length is T_{source}) have the same length as speech in the target accent (the target speech length is denoted by T_{target}). We first train a duration model. The duration model is similar to the phonetic encoder as it uses the same inputs (phonemes and accent embedding) and has the same architecture. The only difference being that the duration model is trained separately using L2 loss. Using the predicted durations, we create a warping matrix W that will serve to create a warped version of the latent sequence z_{warped}:

$$z_{warped} = W \times z_{source}, \tag{5}$$

where z_{source} is the sequence, of length T_{source}, of latent representations encoded from the source mel-spectrogram, z_{warped} is the warped sequence of representations that will be used at the decoding stage (of length T_{target}), and W is the warping matrix of size $T_{target} \times T_{source}$. Multiplying z_{source} with W has the same effect as applying interpolation to z_{source} (similar to using *interpolate* in common deep learning frameworks) but with a scaling factor that varies across phonemes instead of being fixed for the whole utterance.

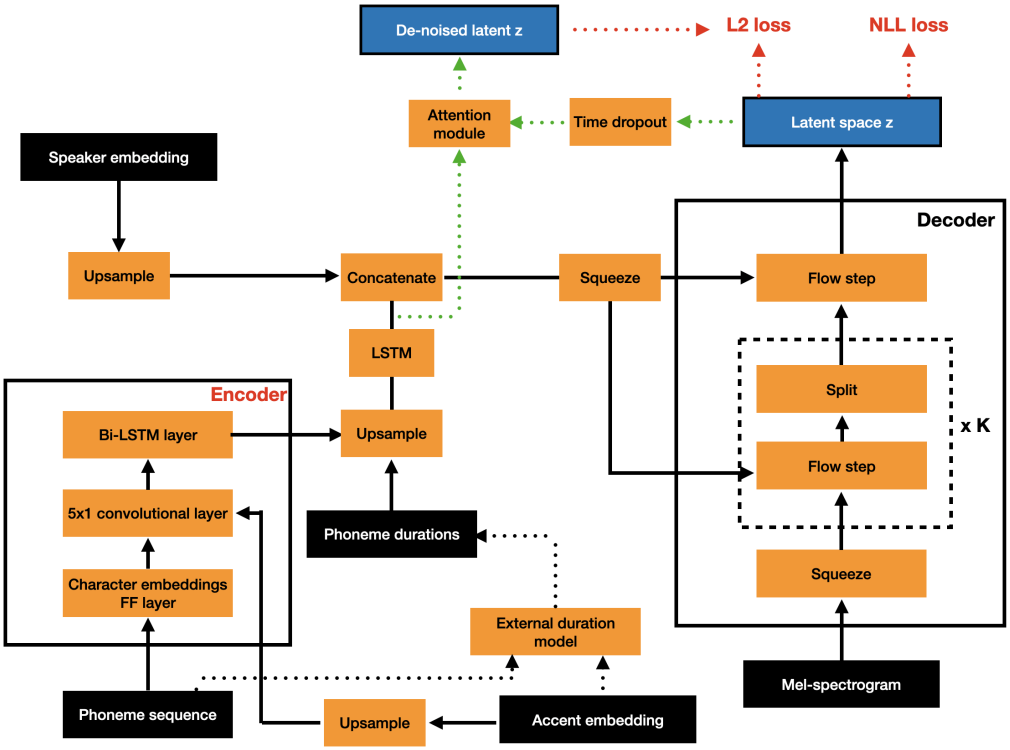
The accent conversion with duration warping can be summarized as running the same encoding stage as in Equation 3, but then replacing z by z_{warped} in Equation 4 during the decoding stage.

2.3. Improving duration warping with attention 'attend'

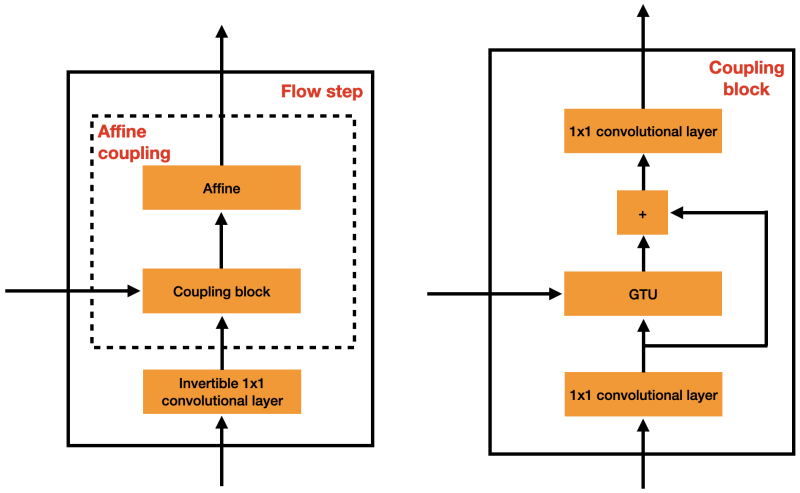
The combination of phoneme sequence remapping and duration warping leads to a better control of accent. However, such approach has one main limitation: constructing the warping matrix requires access to an explicit alignment between the source and target phoneme sequences which is then used to align the target phonemes ph_{target} with the encoded latent representations z_{source}.

In order to circumvent the above limitation, we introduce an attention block which takes a phoneme sequence as queries and the sequence of latent representations as both keys and values. The idea behind the attention block is to enable the model to learn an implicit alignment between the target sequence of warped latent representations and the source sequence of representations, encoded from the source speech signal, without the need for an explicit definition of a warping matrix. The attention can also be seen as a more powerful generalization of interpolating using a warping matrix W.

During the training phase, the attention block trained using L2 loss to denoise a noisy version of the sequence of the source latent representations. We use a noisy version of the source latent representations as input to avoid the attention collapsing into a point where



(a) model architecture



(b) Flow step (c) Flow coupling block

Fig. 1. Overview of our accent conversion approach using normalizing flows (a) Flow step (b) and Coupling block (c). Arrows show the data flow during training. Black dotted arrows correspond to elements used with a duration model only, see Section. 2.2. Green dotted arrows correspond to elements used with the attention block only, see Section. 2.3. Red dotted arrows correspond to loss computation.

it learns an identity transformation between its input and output latent representations. We generate the noisy version of the source latent representations by applying time-dropout where a percentage of frames are zeroed-out at training time.

3. EXPERIMENTS AND RESULTS

We evaluate the performance of the proposed approach in terms of intelligibility, accent similarity, naturalness and speaker similarity.

We chose CopyCat as a baseline model. CopyCat showed state-of-the-art performance in the Voice Conversion task. We adapt CopyCat to the AC task by conditioning the model on speaker and accent embeddings where both embeddings are concatenated to the

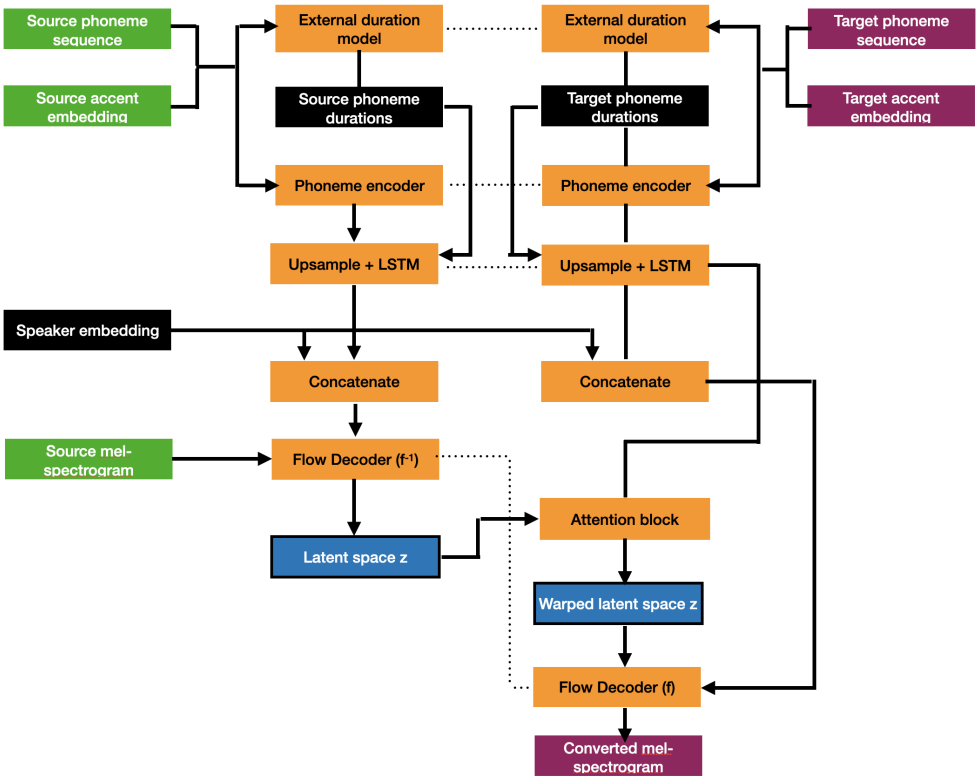


Fig. 2. Accent conversion procedure

upsampled phoneme sequence before being passed to the phoneme encoder, which is in line with [29]. We also applied phoneme remapping and duration warping to the CopyCat baseline.

3.1. Dataset

To evaluate accent conversion, we train our flow model and the CopyCat baseline on a multi-speaker, multi-accent proprietary dataset. The dataset comprises 3173 speakers distributed across 6 English accents: American, Australian, British, Canadian, Indian and Welsh. Each accent has a different number of utterances, ranging from around 18k to 300k. The number of utterances also differs across speakers, ranging from 100 to 20k. The recording conditions vary across the dataset, with some speakers recorded in studio quality conditions whilst other speakers were recorded with lower quality microphones in more ambient surroundings.

3.2. Perceptual evaluations setup

All perceptual evaluations were crowdsourced using the same settings unless explicitly stated otherwise. The evaluations were conducted on the conversion from a male American English speaker to the 5 remaining accents: Australian, British, Canadian, Indian and Welsh. A test set of 25 utterances was selected, resulting in 125 effective utterances for evaluation (25 utterances * target accents). The utterance lengths ranged from 4 words to 25 words. All audio samples were vocoded using a universal vocoder [30] except for recordings. The evaluations were conducted with 300 listeners, each rating 25 utterances. To test for statistically significant differences between systems we perform paired t-tests. Holm-Bonferroni correction is applied due to the number of systems being compared. In the remainder of this paper, we use the term *statistically significant* to refer to $p \leq 0.05$.

It is also worth noting that the proposed AC approach of remapping the phoneme sequence to the target accent, relies on the assumption that the target phoneme sequence can still be aligned with the source mel-spectrogram to be converted. This alignment is implicitly learned by our final model: *remap-warp-attend*. However, the *remap-warp* model and the *CopyCat* baseline need an explicit alignment which requires aligning the source and target phoneme sequences. We simplify the evaluation by restricting the test set to utterances for which ph_{source} and ph_{target} have the same number of phonemes. This evaluation restriction allows us to perform an ablation study of the components without the need to externally align source and target phoneme sequences. We note that *remap-warp-attend* can handle any general case and doesn't need such restriction to be evaluated.

3.3. Intelligibility

We measure the intelligibility of the models using Word Error Rate (WER). WER was computed by comparing the original text of an utterance with the transcription, obtained using AWS Transcribe ASR system, of the converted speech.

WER results summarized in Table 1 show that speech generated by the proposed approaches (remap-warp and remap-warp-attend) are significantly more intelligible than other models, with up to 56% reduction in WER compared to the CopyCat baseline. As such it can be seen that remap-warp provides a more stable conversion compared to the baseline and that the attention block in remap-warp-

attend allows the flow model to handle the general accent conversion case without loss to the intelligibility of the converted speech. Remap had a sub-par performance compared to the other flow models in terms of intelligibility. We hypothesize that this performance is explained by the poor match between the source durations and the target phoneme sequence to be upsampled. The remap model naively assumes that all changes between source and target phoneme sequences are the result of phoneme substitutions if both sequences are of the same length. However, it is likely that insertions and deletions of phonemes did occur which is not handled by said model.

Table 1. WER, relative WER and WERR (WER Reduction) per accent conversion system baselined against CopyCat.

System	WER	Relative WER (WERR)
CopyCat	26.86% $\pm$2.74	1
Remap	20.40% $\pm$2.49	0.76 (24%)
Remap-warp	13.43% $\pm$2.11	0.50 (50%)
Remap-warp-attend	**11.84% $\pm$2.00**	**0.44 (56%)**

3.4. Accent similarity: duration warping effect

We assess the impact of duration warping on the accent conversion using a MUSHRA test. For each MUSHRA screen, listeners listen to a reference recording from a different speaker with the same gender speaking in the target accent. The testers were then asked to "*rate how similar the accents in each system sounds compared to the reference speaker*".

Results, summarized in Table 2, confirm significant changes from the source accent towards the target accent for all proposed models. All the differences between the systems in this evaluation were significant except for the difference between remap-warp and remap-warp-attend. The results from this evaluation suggest that both the remap and warp steps played an important role towards achieving better accent conversion by making the phonemes and their duration better match the target accent. The attend step allows the model to handle the general conversion case, where source and target phoneme sequences are not explicitly aligned, without degrading the accent similarity score.

3.5. Accent similarity: baselining

We compare our flow approaches against CopyCat using a MUSHRA test where we use the same setup as in section 3.4. Results of this evaluation are summarized in Table 3.

All differences between systems in the above evaluation are significant except for the difference between remap-warp and remap-warp-attend. These results show that the proposed flow-based approach significantly improves AC accuracy, with both remap-warp and remap-warp-attend significantly outperforming CopyCat. A potential explanation to the previous observation, is that the lossless nature of flow-models, where they are able to exactly reconstruct their input mel-spectrogram if the conditionings stay the same, is more beneficial to the AC task than the lossy representation of speech that CopyCat uses, where speech is decoded after using a time bottleneck. The results also confirm the findings from the previous accent similarity test indicating that the attention block adds flexibility to flow models without degrading the accent similarity scores.

Table 2. Average scores for the accent similarity evaluation which assesses the effect of duration warping.

Accent similarity - duration warping effect	
Source speaker recordings - lower anchor	57.53
Target accent recordings - upper anchor	68.57
Remap	60.41
Remap-warp	**62.13**
Remap-warp-attend	61.51

Table 3. Average scores for the accent similarity evaluation which baselines our flow model against CopyCat.

Accent similarity - baseline	
Source speaker recordings - lower anchor	62.14
Target accent recordings - upper anchor	71.44
CopyCat	63.05
Remap-warp	65.26
Remap-warp-attend	**65.31**

Table 4. Average scores for naturalness evaluations comparing accent conversion approaches.

Naturalness - baseline	
Source speaker recordings - upper anchor	75.30
CopyCat	68.15
Remap-warp	69.11
Remap-warp-attend	**69.62**

3.6. Naturalness

After establishing the proposed model's capabilities in terms of accent control, we evaluate how natural-sounding is its synthesized speech. To this end, we measure the naturalness of the proposed approach using a MUSHRA test. The testers were asked to "*rate the audio samples in terms of their naturalness*". Results are summarized in Table 4.

All the differences were significant except the differences between remap-warp and remap-warp-attend. The results suggest that both remap-warp and remap-warp-attend are the most natural-sounding AC systems surpassing the CopyCat baseline. The results also show that the attention block doesn't degrade the naturalness while still allowing more flexibility to the AC flow model. This is in-line with similar observations made previously in the accent similarity evaluations.

3.7. Speaker similarity

The focus of the model is to perform accent conversion while leaving the perceived speaker identity unchanged. We measure potential alteration to the speaker identity using a MUSHRA test. Testers are presented with a recording from the source speaker and are then asked to "*rate how similar the speakers in each system sound compared to the reference speaker*". The reference audio was always a different text content from the text of the samples evaluated.

Table 5. Average scores for speaker similarity evaluation comparing accent conversion approaches.

Speaker similarity - baseline	
Target accent recordings - lower anchor	49.11
Source speaker recordings - upper anchor	75.67
CopyCat	**63.53**
Remap-warp	62.53
Remap-warp-attend	61.06

Results from this evaluation are summarized in Table 5. All differences in the evaluation are significant. All AC systems show some perceived changes to the speaker identity. The significant difference between remap-warp and remap-warp-attend suggests that the attention mechanism does alter the perceived speaker identity. Following informal listening, we hypothesise that this may be due to a tendency of the attention to flatten the prosody of converted speech. However, more investigation is required to better understand how the prosody changes may have impacted the speaker similarity test's results without impacting the accent similarity. CopyCat seems to alter the speaker identity the least. This observation may be explained by the fact that CopyCat makes smaller changes to the source speech, as suggested by its lower accent similarity scores, and is thus more likely to restore the source speaker's identity in the converted speech.

3.8. Many-to-many accent conversion

To confirm the proposed approach's AC performance is stable across all combinations of source and target accents, we perform further evaluations. The MUSHRA test has the same setup as in section 3.4 except for the selection of the test set: one male and one female speakers were selected per source accent, we then select 27 utterances per speaker and run accent conversion towards all remaining target accents. The only exception is the Canadian English accent for which we selected only one female speaker. This setup yields an effective test set of 1485 evaluated utterances (27 utterances * 11 speakers * 5 target accents). Each evaluation screen showed the following systems: reference recordings in the target accent, our remap-warp-attend samples, source speaker recordings (lower anchor), and recordings from a speaker speaking the target accent who is different from both the source and reference speakers (upper anchor).

We report in Table 6 the ratio between the means of the MUSHRA scores of our remap-warp-attend model and the target accent recordings for each source-target accent pair. We chose to report the ratio between our system's and the upper anchor's MUSHRA scores instead of the absolute numbers since the upper anchors scored differently in each accent. We omit results on conversion to and from Indian English as we noticed later in the development process some labelling issues with the Indian English data. We confirmed after fixing the issue that the other accents were not affected.

Overall, we observe consistently high accent similarity scores ratios. It is worth noting that in multiple conversion combinations, the average accent similarity rating for the proposed approach achieved a ratio close to, or higher than, 1. This can be interpreted as the listeners not being able to distinguish the accent of converted speech from that of recordings of a native speaker of that accent. The observed ratios are lower for conversion to Canadian English compared to other locales. This is likely because the target accent speaker used for the reference was the same speaker as the target accent speaker used as upper anchor for this accent. This resulted in Canadian English upper anchor being scored much higher in comparison to the other accents'.

From the many-to-many conversion results, we conclude that the proposed approach is capable of converting accent across all seen combinations of source/target accents.

Table 6. Matrix summarizing ratios of accent similarity scores between remap-warp-attend and target recordings for the conversion from source accents (rows) to target accents (columns).

	en_au	en_ca	en_gb	en_us	en_wls
en_au	-	0.66	0.94	0.99	0.98
en_ca	0.86	-	0.96	0.98	0.87
en_gb	0.97	0.69	-	0.97	0.99
en_us	0.95	0.66	1.00	-	0.91
en_wls	1.04	0.64	0.98	0.98	-

4. CONCLUSION

We propose an accent conversion method based on an innovative extension to normalizing flows called 'remap, warp and attend'. The remap step converts the sequence of phoneme conditioning between source and target speech to match the target accent. The warp step adjusts the phoneme durations, enabling conversion between source and target speech signals of different durations. Finally, the attend step allows the model to handle the general AC case, by removing constraints on its input sequences, while maintaining the same naturalness and accent similarity scores. Through objective and perceptual evaluations, we show that the approach significantly improves conversion compared to the CopyCat baseline in terms of accent similarity, naturalness and intelligibility. While the proposed approach is tailored to accent conversion, the "remap, warp and attend" technique paves the way for conversion of multiple speech attributes, where any speech property can be independently adjusted.

5. REFERENCES

[1] S. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," *Interspeech*, Oct 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1251

[3] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 5210–5219.

[4] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," 2019. [Online]. Available: https://arxiv.org/abs/1904.04169

[5] K. Probst, Y. Ke, and M. Eskénazi, "Enhancing foreign language tutors - In search of the golden speaker," *Speech Communication*, vol. 37, pp. 161–173, 2002.

[6] S. Ding, C. Liberatore, S. Sonsat, I. Lucic, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden Speaker Builder -An Interactive Tool for Pronunciation Training." in *Speech Communication 115*, 2019, pp. 51–66.

[7] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," in *The Speaker and Language Recognition Workshop*. ISCA, 2018.

[8] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7879–7883.

[9] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Interspeech*, 2013.

[10] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams," in *Proc. Interspeech*, 2019, pp. 2843–2847.

[11] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5314–5318.

[12] W. Li, B. Tang, X. Yin, Y. Zhao, W. Li, K. Wang, H. Huang, Y. Wang, and Z. Ma, "Improving accent conversion with reference encoder and end-to-end text-to-speech," *ArXiv*, vol. abs/2005.09271, 2020.

[13] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with deep neural networks," in *Interspeech*, 2015, pp. 3385–3389.

[14] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7694–7698.

[15] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021.

[16] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, and X. Li, "Accent and speaker disentanglement in many-to-many voice conversion," 2020. [Online]. Available: https://arxiv.org/abs/2011.08609

[17] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2020.2992934

[18] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving Flow-Based Generative Models with Variational Dequantization and Architecture Design," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 2722–2730.

[19] D. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[20] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209–7213.

[21] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Advances in Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 8067–8077. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/5c3b99e8f92532e5ad1556e53ceea00c-Paper.pdf

[22] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. de Oliveira, A. C. Junior, A. d. S. Soares, S. M. Aluisio, and M. A. Ponti, "SC-GlowTTS: an Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," in *Interspeech*, 2021.

[23] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Ig53hpHxS4

[24] J. Serrà, S. Pascual, and C. Segura, "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[25] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, p. 631–644, Mar 2019. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2019.2892235

[26] T. Merritt, A. Ezzerg, P. Bilinski, M. Proszewska, K. Pokora, R. Barra-chicote, and D. Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalising flows," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[28] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[29] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-Resource Expressive Text-To-Speech Using Data Augmentation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6593–6597, 2021.

[30] Y. Jiao, A. Gabrys, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal Neural Vocoding with Parallel WaveNet," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.