MODELLING LOW-RESOURCE ACCENTS WITHOUT ACCENT-SPECIFIC TTS FRONTEND

Georgi Tinchev[*1], *Marta Czarnowska*[*1], *Kamil Deja*[2], *Kayoko Yanagisawa*[1], *Marius Cotescu*[1]

[1]Amazon Research, United Kingdom [2]Warsaw University of Technology, Poland

ABSTRACT

This work focuses on modelling a speaker's accent that does not have a dedicated text-to-speech (TTS) frontend, including a grapheme-to-phoneme (G2P) module. Prior work on modelling accents assumes a phonetic transcription is available for the target accent, which might not be the case for low-resource, regional accents. In our work, we propose an approach whereby we first augment the target accent data to sound like the donor voice via voice conversion, then train a multi-speaker multi-accent TTS model on the combination of recordings and synthetic data, to generate the donor's voice speaking in the target accent. Throughout the procedure, we use a TTS frontend developed for the same language but a different accent. We show qualitative and quantitative analysis where the proposed strategy achieves state-of-the-art results compared to other generative models. Our work demonstrates that low resource accents can be modelled with relatively little data and without developing an accent-specific TTS frontend. Audio samples of our model converting to multiple accents are available on our web page[3].

Index Terms— text-to-speech, accent conversion

1. INTRODUCTION

Recent advancements in text-to-speech (TTS) have been made to explore multilingual training of the acoustic models [1, 2]. They usually rely on huge amounts of multi-lingual and multi-speaker data and can convert between individual accents by conditioning specifically on the speaker and accent information. While attempts have been made to condition the model directly on character input [3], many of the approaches still rely on phoneme input since it (a) avoids pronunciation errors introduced by the model mislearning how to pronounce certain words and makes the task for the neural network easier [4]; and (b) allows for phonetic control of the synthesized speech [5]. Phoneme sequences can be generated by passing the text input through a TTS frontend. The latter consists of several modules including text normalisation and grapheme-to-phoneme (G2P) conversion, and is typically language- or accent-dependent. It is expensive to develop a dedicated frontend due to the need for expert language knowledge [6].

In this paper we propose a TTS system that is capable of modelling a target accent with a selected donor voice without the need for a dedicated frontend that can produce phonetic transcriptions matching the target accent. Our model produces state-of-the-art results in terms of naturalness and accent similarity. The model generalizes to learn differences in phonological contrasts between accents even when trained with phoneme sequences produced by a frontend not specifically developed for the accent. In summary, our contributions are as follows: 1) We propose the first approach, to our knowledge, for accent modelling without explicitly conditioning the model on phonemes generated with an accent-specific G2P for the target accent. Our method demonstrates state-of-the-art results compared with existing multi-lingual multi-accent baselines, modelled by normalizing flows, VAEs, or diffusion models, 2) we can reliably model the accent with augmented data regardless of the chosen TTS frontend and 3) we demonstrate that modelling the accent via synthetic data can be done with just 2000 utterances of the target accent.

2. RELATED WORKS

This section summarises relevant literature for the task of accent modelling. The review is split in two sections - research in grapheme-to-phoneme (G2P) and accent conversion.

TTS models usually use phoneme sequences as inputs to model speech, since directly modelling from character inputs adds an extra challenge for the network to learn the relationship between characters and pronunciation. Work on G2P traditionally involved rule-based methods [7, 8, 9], then evolved to machine learning based approaches using n-grams [10] or neural networks [11, 12]. Recently, there have been efforts to translate these methods to low-resource languages by using distance measures and linguistic expertise [13]. The main focus of our work is to develop a TTS system for an accent for which we do not have a dedicated frontend, as developing a new frontend requires linguistic expertise and is not scalable.

Historically accent conversion (AC) has been achieved in two ways - combining spectral features via voice morphing [14, 15, 16] and voice conversion (VC) adaptation using spectral and phoneme frame matching [17, 18, 19]. There is also a line of work that relies on external features with frame-matching after the VC step [20, 21, 22, 23, 24, 25]. The main shortcomings of voice morphing methods is their requirement

[*]Indicates equal contribution. [2]Work done while at Amazon.
[3]https://bit.ly/3V52ZrF

for parallel data. For frame-matching it is hard to convert the duration and speaking rate of the target voice. In contrast to existing VC-based approaches for AC, we do not require an ASR model to disentangle the accent information. Instead we leverage VC as a first step to generate data of the *donor voice* in the *target accent*, since it preserves the accent [17].

Similarly to us, in the domain of speech recognition with foreign accent, other works observe that data augmentation through voice and accent conversion positively impacts the model's performance [26, 27]. Most similar to us, [28] also leverage data augmentation for accent transfer. Crucially, our model does not require text to phoneme transcriptions for the *target accent* and we are able to model the accent by using less data - just 27k utterances vs 425k utterances in [28].

3. METHODOLOGY

The goal of our method is to train a TTS model for a new low-resource accent (from here on referred to as the *target accent*) without an accent-specific frontend to generate phonetic transcriptions with. To achieve this we use a multi-speaker multi-accent dataset. We select a *donor* speaker from one of the non-target accent speakers. We want the synthesized speech to sound like the *donor* speaker but with the *target accent*. We require G2P model for one of the accents from the dataset, which we use to extract the phoneme sequence for all the utterances. Crucially, we do not require accent-specific G2P for any of the other accents, including the *target accent*. In this section we will explain the model and how we generate augmented data for the *donor* voice in the *target accent*.

3.1. Architecture

We divide our approach into two major steps - Voice Conversion (VC) and Text-To-Speech (TTS). We use VC to generate synthetic data of our *donor* voice speaking in the *target accent*. We then use this augmented data together with the original recordings to train a multi-speaker multi-accent TTS system. Below we briefly outline each of the two models.

3.1.1. Voice conversion

To augment the training data for the TTS model, we use a flow-based voice conversion model [29, 30]. During training the model encodes the input mel-spectrogram into a latent vector z, using phoneme sequence, speaker embeddings, f0, and binary voice/unvoiced flag as conditioning. The f0 is normalized per utterance to disentangle speaker information from the utterance prosody. We use pre-trained speaker embeddings based on [31]. The model is optimized with a negative log-likelihood loss. During inference we encode the mel-spectrogram of the *target accent* speaker into the latent z, using corresponding conditioning, and then decode it back to a mel-spectrogram, changing the speaker embedding to that of the *donor* speaker. We first train the VC model on the whole dataset. Using this model we convert the speaker from the *target accent* to sound like the *donor* voice. This way we are able to change the speaker identity but preserve the rest of the

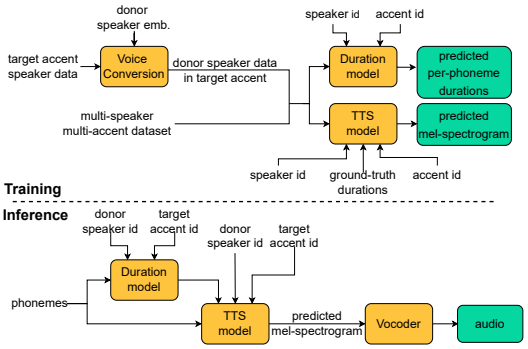


Fig. 1: TTS model training and inference procedure

speaking characteristics, namely accent, style, and duration, in line with [17]. We then train TTS acoustic and duration models on the original dataset and synthetic data generated by the VC model, as described in the following section.

3.1.2. Text-to-speech

Our text-to-speech (TTS) architecture is a sequence-to-sequence model with explicit duration [3, 32]. It has two separate models: the acoustic model and the duration model.

The acoustic model consists of three components: an encoder, an upsampling layer, and a decoder. The encoder architecture is based on Tacotron2 [3] and the decoder is composed of residual gated convolution layers and an LSTM layer. We use a Flow-VAE reference encoder [2]. We train this architecture with 2000 utterances of the *donor speaker* in the *target accent*, generated with VC. However, we empirically found out that training on a multi-speaker multi-accent dataset improves the naturalness (Section 4.4). We therefore modify the architecture to include speaker and accent conditioning. We condition both the encoder and decoder on speaker and accent embeddings, trained together with the model. During training, ground-truth phoneme durations are used to upsample the encoded phonemes sequence, but during inference we use the durations predicted by the duration model. The acoustic model is optimized with an L2 loss between ground truth and predicted mel-spectrogram and KL divergence for the VAE.

The duration model uses the same encoder architecture as the acoustic model, followed by a projection layer. The model is also conditioned on speaker and accent information. It is optimized with an L2 loss between ground-truth and predicted phoneme durations. During inference we predict the duration of the phonemes conditioned on the *target accent* and the *donor* speaker. We then run inference with the acoustic model, providing the same speaker and accent conditioning, predicted durations, and VAE centroid computed from the synthetic data generated by the VC model. We vocode predicted mel-spectrograms using a universal vocoder [33].

3.2. Dataset

In our experiments we chose to model Irish English (en-IE) accent using a British English (en-GB) *donor speaker*. We use 25k utterances for the *donor speaker* and between 0.5k

Table 1: MUSHRA comparison to the state-of-the-art. For accent similarity we used 25 test cases from 4 en-IE speakers as reference. We used two upper anchors: i) different recording of the reference speaker ii) recording of a different en-IE speaker.

System	Naturalness	Accent similarity all speakers	Accent similarity speaker 1	Accent similarity speaker 2	Accent similarity speaker 3	Accent similarity speaker 4
Upper anchor	74.42	82.24	72.56	89.33	84.79	79.24
Upper anchor different speaker	N/A	52.55	43.40	49.42	57.10	54.89
Proposed method	**64.61**	**50.73**	**61.40**	**40.81**	**45.16**	**50.44**
Grad-TTS	63.23	**50.68**	59.29	**40.77**	**46.32**	**50.71**
Flow-TTS	56.46	**50.32**	59.09	**40.28**	**45.19**	**51.21**

Table 2: MUSHRA comparison to the state-of-the-art.

System	Naturalness	Accent similarity
Upper anchor different speaker	90.46	65.85
Proposed method	**72.33**	**74.14**
Polyglot	47.24	49.29
Lower anchor (en-GB)	N/A	12.17

Table 3: Ablation study for different data configurations.

System	Naturalness	Accent similarity
VC multi-speaker multi-accent	78.58	60.85
TTS multi-speaker multi-accent	**65.07**	**58.12**
TTS single speaker IE finetuned	59.07	**57.32**
Lower anchor (en-GB)	N/A	11.50

- 4.5k utterances for the rest of the speakers from 6 supporting accents - British (en-GB), American (en-US), Australian (en-AU), Indian (en-IN), Welsh (en-GB-WLS), and Canadian English (en-CA) [34]. For en-IE, which is our *target accent*, we have the same 2000 utterances recorded by 12 speakers. Having this type of parallel data is not a requirement for either the VC or TTS models. In Section 4.4 we demonstrate that training on just 2000 utterances by a single speaker is enough for our approach to model the accent. We extracted phonetic transcriptions using the en-GB frontend, which is the accent of the *donor speaker*. We use a unified representation for the phonemes - we map the same phonemes across different accents to the same input tokens [2].

4. EXPERIMENTS

In our work we have chosen to model en-IE accent with an en-GB *donor speaker*. In this section we demonstrate how our method achieves state-of-the-art results compared to other TTS models. We also present qualitative and quantitative evaluations of the proposed model using two different G2P models to generate phonetic transcriptions: en-GB and en-US. Finally, through ablation studies, we will demonstrate that modelling accents can be done with low-resource data.

4.1. Evaluation method

We used a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test for comparing our proposed method to state-of-the-art baselines [35]. Each evaluation consisted of 100 unique testcases, not seen during training of the models. The testcases were evaluated by 24 native Irish speakers who rated each system on a scale between 0 and 100. We evaluated naturalness and accent similarity of the samples. For naturalness testers were asked to judge if the samples sound like a human speaker and for accent similarity how similar are the samples to a provided reference sample. Both the reference and the hidden upper anchor were the recordings of en-IE speakers. To ensure results are statistically significant, a paired t-test with Holm-Bonferroni correction was performed ($p \leq 0.05$). Bold results indicate the best performing systems in each column, up to statistically significant differences.

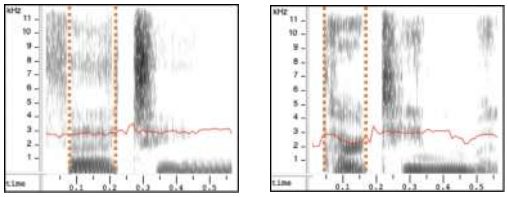
4.2. Comparison to the state-of-the-art

We compared our method to Polyglot - a multi-speaker multi-accent sequence-to-sequence attention-based model [34]. It achieves state-of-the-art results using phonetic transcriptions generated by accent specific G2P. We trained both our approach and the baseline using phonetic features extracted with an en-GB G2P model. Table 2 shows our method achieving significantly better results in naturalness and accent similarity, indicating that accent and speaker conditioning alone are not sufficient for accent transfer without accent specific phonetic transcriptions. In the accent similarity evaluation we used a different en-IE speaker for the reference and the upper anchor, both had the same variety of en-IE accent. The evaluators were good at distinguishing this subtle difference and thus rated our model better than the upper anchor. Therefore, in following experiments we introduced a second upper anchor with different recordings of the reference speaker.

We also compared our method to two additional baseline models, trained on the same dataset with phonemes for each accent extracted with en-GB G2P: 1) Grad-TTS [36] with explicit speaker and accent conditioning and 2) Flow-TTS-based accent conversion model [37, 29]. The results are presented in Table 1. For naturalness our method achieves significantly better results than both baselines. As Grad-TTS and Flow-TTS baselines model a generic en-IE accent, instead of speaker-specific en-IE accent, we decided to use 4 different en-IE speakers as the reference speaker. Our method is significantly better when the reference is the same en-IE speaker (we used speaker 1's accent for data augmentation), but there is no significant difference between the models for different reference speakers (speaker 2, 3, 4). This confirms that by using VC samples our proposed approach is better tailored towards a particular speaker's accent.

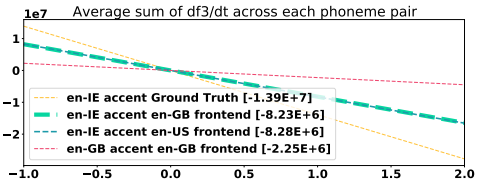
4.3. Analysis of Rhoticity

Next, we show how we can reproduce features of the *target accent* when we use a frontend for an accent that is missing that feature. In all our experiments above, we presented the



(a) en-GB synthesis (b) en-IE ground truth

(c) en-IE with en-GB G2P (d) en-IE with en-US G2P



(e) Average sum of F3 derivatives

Fig. 2: (a-d) F3 contour (in red) showing the lowering of the formant in the /ɝ/ or /ɜː/ vowel in "thirteen", indicated between orange dashed lines. (e) Illustration of the average slope of the postvocalic /r/ contexts.

results for the model trained with en-GB G2P ("en-IE with en-GB G2P"), the accent of the *donor voice*. This model is capable of generating samples with an accent closer to en-IE than en-GB (lower anchor), as shown in Table 2. en-IE differs from en-GB in one major aspect, rhoticity, which means that the /r/ is pronounced in postvocalic contexts when not followed by another vowel (e.g. /kɑːr.pɑːrk/ for "car park"). en-GB, on the other hand, is non-rhotic (/kɑː.pɑːk/). Despite this difference, our model trained with en-GB G2P was able to reproduce the rhoticity in the synthesised en-IE samples.

To quantify this, we analysed the third formant (F3) for the segments where the rhoticity contrast is found, as lowering of F3 is acoustically correlated to the phoneme /r/ [38]. For comparison, we trained an additional model where the phonemes were extracted with the en-US G2P, which is also a rhotic accent ("en-IE with en-US G2P"). To identify the regions where the contrast occurs, we extracted phonemes for our test set with both en-GB and en-US frontends and aligned the phoneme sequences using Dynamic Time Warping (DTW) [39] with a cost function based on phoneme similarity. We use the alignment to identify contexts for which the /r/ is present in rhotic accents but not in non-rhotic accents. For example, in the phrase "car park", en-GB /ɑː/ is aligned with en-US /ɑːr/. We use Kaldi external aligner [40] to find each phoneme position in the audio file and LPC analysis to extract the F3 for those contexts and compute its slope.

Fig. 2 shows how monolingual en-GB (non-rhotic) model sample does not show a lowering of F3 (2a) whereas the

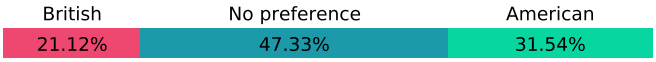


Fig. 3: Preference test results for samples from model trained with en-GB and en-US frontends.

ground truth en-IE (rhotic) recording does (2b). en-IE samples generated from models trained with our approach show the lowering of F3, regardless of whether the model was trained with en-GB G2P (2c) or en-US G2P (2d). This is confirmed in Fig. 2e which shows the average gradient of the F3 slope across 134 contexts with the rhotic contrast. There is no statistically significant difference between the gradient for "en-IE with en-GB G2P" and "en-IE with en-US G2P".

We subsequently compared "en-IE with en-GB G2P" and "en-IE with en-US G2P" in a preference test with 24 listeners each listening to 100 pairs of test cases. Fig. 3 shows that there is a significant preference for the model trained with en-US G2P, although the majority had no preference. These results show that we can reproduce features in the *target accent* even if we use the frontend for an accent that is missing that feature, but since accents can vary in many different aspects, we can further improve our model performance by carefully selecting the frontend used for annotation of the data.

4.4. Ablation Studies

As a final experiment we present ablation studies of our two models with different data configurations. We trained our TTS model with only the *donor voice* and fine-tuned with the *target accent* synthetic data. Table 3 demonstrates that there is no significant difference between the accent similarity of a multi-speaker, multi-accent model and the model trained on the *donor speaker* and fine-tuned with the *target accent* synthetic data. In line with [34], training in a multi-speaker, multi-accent scenario improves the naturalness. Results from this experiment show that using data from a single target accent speaker is sufficient for the task of modelling the accent.

5. CONCLUSIONS

In this work we presented an approach for accent modelling based on data augmentation with voice conversion and text-to-speech model trained on the combination of recording and synthetic data. Crucially, our approach does not require transcriptions generated with accent-specific G2P. We show that our model can reliably model features in the *target accent* and achieves state-of-the-art performance compared to existing strategies for accent modelling. We show that our approach works in a low resource scenario where we have little data in the *target accent*. We also found that the performance can be improved further by selecting a more suitable accent for the G2P. The strategy for choosing the best accent for G2P is left for future work. We also plan to extend our methodology to accents of language other than English.

6. ACKNOWLEDGEMENTS

We thank Andre Canelas, Tom Merritt, Piotr Bilinski, Maria Corkery, and Dan McCarthy for their feedback and insights.

7. REFERENCES

[1] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[2] A. Sanchez, A. Falai, Z. Zhang, O. Angelini, and K. Yanagisawa, "Unify and Conquer: How Phonetic Feature Representation Affects Polyglot Text-To-Speech (TTS)," in *Interspeech*, 2022.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018.

[4] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 4693–4702, PMLR.

[5] A. Perquin, E. Cooper, and J. Yamagishi, "An investigation of the relation between grapheme embeddings and pronunciation for tacotron-based systems," 2021.

[6] M. Gakuru, "Development of a kenyan english text to speech system: A method of developing a TTS for a previously undefined english dialect," in *Speech Comm.*, 2009.

[7] R. Weide et al., "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[8] C. R. Rosenberg and T. Sejnowski, "Nettalk: A parallel network that learns to read aloud," Tech. Rep., Johns Hopkins University Press, 1986.

[9] P. Kingsbury, S. Strassel, C. McLemore, and R. MacIntyre, "Call-home american english lexicon (pronlex)," *Linguistic Data Consortium*, 1997.

[10] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[11] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "S2s neural net models for grapheme-to-phoneme conversion," in *Interspeech*, 2015.

[12] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4225–4229.

[13] A. Deri and K. Knight, "Grapheme-to-phoneme models for (almost) any language," in *ACL)*, 2016.

[14] M. Huckvale and K. Yanagisawa, "Spoken language conversion with accent morphing," 2007.

[15] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Comm.*, 2009.

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

[17] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?," in *ICASSP*. IEEE, 2014.

[18] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, et al., "End-to-end accent conversion without using native utterances," in *ICASSP*. IEEE, 2020.

[19] 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," in *ISCSLP*. IEEE, 2021, pp. 1–5.

[20] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *ICASSP*. IEEE, 2018.

[21] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Interspeech*, 2019.

[22] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[23] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, pp. 101302, 2022.

[24] S. Ding, C. Liberatore, S. Sonsaat, I. Lučić, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden speaker builder–an interactive tool for pronunciation training," *Speech Comm.*, vol. 115, pp. 51–66, 2019.

[25] C. Liberatore and R. Gutierrez-Osuna, "An exemplar selection algorithm for native-nonnative voice conversion.," in *Interspeech*, 2021, pp. 841–845.

[26] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data augmentation improves recognition of foreign accented speech," in *Interspeech*, 2018, number September, pp. 2409–2413.

[27] H. Huang, X. Xiang, Y. Yang, R. Ma, and Y. Qian, "Aispeech-sjtu accent identification system for the accented english speech recognition challenge," in *ICASSP*. IEEE, 2021.

[28] L. Finkelstein, H. Zen, N. Casagrande, C.-a. Chan, Y. Jia, T. Kenter, A. Petelin, J. Shen, V. Wan, Y. Zhang, et al., "Training text-to-speech systems from synthetic data: A practical approach for accent transfer tasks," in *Interspeech*, 2022, number September.

[29] 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 *ICASSP*. IEEE, 2020.

[30] P. Bilinski, T. Merritt, A. Ezzerg, K. Pokora, S. Cygert, K. Yanagisawa, R. Barra-Chicote, and D. Korzekwa, "Creating New Voices using Normalizing Flows," in *Interspeech*, 2022, pp. 2958–2962.

[31] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*. IEEE, 2018.

[32] R. Shah, K. Pokora, A. Ezzerg, V. Klimkov, G. Huybrechts, B. Putrycz, D. Korzekwa, and T. Merritt, "Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech," in *Interspeech Workshop (SSW11)*, 2021.

[33] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel wavenet," in *ICASSP*. IEEE, 2021.

[34] Z. Zhang, A. Falai, A. Sanchez, O. Angelini, and K. Yanagisawa, "Mix and Match: An Empirical Study on Training Corpus Composition for Polyglot Text-To-Speech (TTS)," in *Interspeech*, 2022.

[35] ITU-R Recommendation, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *ITU, BS*, 2001.

[36] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for TTS," in *ICML*, 2021.

[37] A. Ezzerg, T. Merritt, K. Yanagisawa, P. Bilinski, M. Proszewska, K. Pokora, R. Korzeniowski, R. Barra-Chicote, and D. Korzekwa, "Remap, warp and attend: Non-parallel many-to-many accent conversion with normalizing flows," in *SLTW*. IEEE, 2023.

[38] C. Espy-Wilson, S. Boyce, M. Jackson, S. Narayanan, and A. Alwan, "Acoustic modeling of american english /r/," *The Journal of the Acoustical Society of America*, vol. 108, pp. 343–56, 08 2000.

[39] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in *ASRU Workshop*. IEEE, 2011.