INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

RAD-MMM: Multilingual Multiaccented Multispeaker Text To Speech

Rohan Badlani, Rafael Valle, Kevin J. Shih, João Felipe Santos, Siddharth Gururani, Bryan Catanzaro

NVIDIA

rbadlani@nvidia.com,rafaelvalle@nvidia.com

Abstract

We create a multilingual speech synthesis system that can generate speech with a native accent in any seen language while retaining the characteristics of an individual's voice. It is expensive to obtain bilingual training data for a speaker and the lack of such data results in strong correlations that entangle speaker, language, and accent, resulting in poor transfer capabilities. To overcome this, we present RADMMM, a speech synthesis model based on RADTTS with explicit control over accent, language, speaker, and fine-grained F_0 and energy features. Our proposed model does not rely on bilingual training data. We demonstrate an ability to control synthesized accent for any speaker in an open-source dataset comprising of 7 languages, with one native speaker per language. Human subjective evaluation demonstrates that, when compared to controlled baselines, our model better retains a speaker's voice and target accent, while synthesizing fluent speech in all target languages and accents in our dataset.

1. Introduction

Recent progress in Text-To-Speech (TTS) has achieved human-like quality in speech synthesis through mel-spectrogram [1, 2, 3, 4] and direct waveform [5, 6] prediction. Most models support speaker selection during inference by learning a speaker embedding table [1, 2, 3] during training, while some support zero-shot selection by generating a speaker conditioning vector from a short audio sample [7]. However, most models support only a single language. This work focuses on factorizing out speaker and accent as controllable attributes, in order to synthesize speech for any desired combination of speaker, language and accent present in the training dataset.

It is very expensive to obtain bilingual datasets because most speakers are monolingual. Hence, speaker, language, and accent attributes are highly correlated in most TTS datasets. Training models with such entangled data can result in poor language, accent, and speaker transferability. Most TTS systems use different symbol sets for each language, sometimes even separate encoders [8], severely limiting representational sharing across languages. This aggravates speaker, language and text entanglement, especially in datasets with very few speakers per language. Approaches like [9] introduce an adversarial loss to curb this dependence of text representations on a speaker, but this can result in lower content quality. Other approaches use a union of linguistic feature sets of all languages [10] to simplify text processing for multi-language training. However, these solutions don't support code-switching situations where words from multiple languages appear in mixed order in the synthesis prompt.

Recently, there has been interest in factorizing out fine-grained speech attributes [11, 12, 13] like F_0 and energy. We extend this fine-grained control by factorizing out accent and speaker, with an ability to predict frame-level F_0 and energy, for a desired combination of accent, speaker and language. We analyze and show the benefits this explicit conditioning on fine-grained speech features brings to synthesized speech when transferring a voice to other languages.

Our goal is to synthesize speech for a target speaker in any language with a specified accent. Related methods include YourTTS [14] and VallE-X [15], but these methods focus on zero-shot multilingual voice conversion in the large data regime. Although promising results are presented for a few language combinations, the results in YourTTS indicate limited success on transferring from languages with limited speakers (like Portuguese as stated in [14]). Moreover, it uses a curriculum learning approach to extend the model to new languages, making the training process cumbersome. Closest to our work is [9], which describes a multilingual and multispeaker TTS model trained on a proprietary high-quality dataset with approximately 100 English professional voice actors and a handful of Spanish and Mandarin voice actors. We depart from [9], use open source datasets and focus on more challenging scenarios with one speaker, or a couple, per language.

In this work, we **(1)** demonstrate effective scaling of single language TTS to multiple languages using a shared alphabet set and alignment learning framework [4, 16]; **(2)** introduce explicit accent conditioning to control the synthesized accent; **(3)** propose and analyze several strategies to disentangle attributes (speaker, accent, language and text) without relying on parallel training data (multilingual speakers); and **(4)** explore fine-grained control of speech attributes such as F_0 and energy and its effects on speaker timbre retention and accent quality.

2. Methodology

We build upon RADTTS [4, 13] as deterministic decoders tend to produce oversmooth mels that require vocoder fine-tuning. Our model synthesizes mels($X \in \mathbb{R}^{C_{mel} \times F}$) using encoded text($\Phi \in \mathbb{R}^{C_{txt} \times T}$), accent($A \in \mathbb{R}^{D_{accent}}$) and speaker($S \in \mathbb{R}^{D_{speaker}}$) as conditioning variables, with optional conditioning on fundamental frequency($F_0 \in \mathbb{R}^{1 \times F}$) and energy($\xi \in \mathbb{R}^{1 \times F}$), where F is the number of mel frames, T is the text length, and energy is the per-frame mel energy average. We refer to our conditioning as accent instead of language because we consider language to be *implicit in the phoneme sequence*. The information captured by the accent embedding should explain the fine-grained differences between how phonemes are pronounced in different languages.

We propose the following modifications:

2.1. Shared text token set

Our goal is to train a single model with the ability to synthesize a target language with desired accent for any speaker in the

10.21437/Interspeech.2023-2330

dataset. We represent text tokens with the International Phonetic Alphabet (IPA) to enforce a phoneme-based shared textual representation. A shared alphabet across languages reduces the dependence of text on speaker identity, especially in low-resource settings (e.g. 1 speaker per language) and supports code-switching.

2.2. Scalable Alignment Learning

We utilize the alignment learning framework in [4, 16], to learn speech-text alignments $\Lambda \in \mathbb{R}^{T\times F}$ without external dependencies. A shared alphabet set simplifies this since alignments are learnt on a single token set instead of distinct sets. However, the same token can be spoken in different ways due to differences in speaker's accent, which can make alignments brittle. To curb this multi-modality, we learn alignments between (text, accent) and mel-spectograms using accent A as a conditioning variable.

2.3. Disentangling Factors

We focus on non-parallel data with a speaker speaking only one language, which typically has text Φ, accent A and speaker S entangled. We evaluate strategies to disentangle these attributes:

Speaker-adversarial loss In TTS datasets, speakers typically read different text and have different prosody. Hence, there can be entanglement between speaker S, text Φ and prosody. Following [9], we employ domain adversarial training to disentangle S and Φ by using a gradient reversal layer. We use a speaker classification loss, and backpropagate classifier's negative gradients through the text encoder and token embeddings.

$$L_{adv} = \sum_{i=1}^{N} P(s_i|\phi_i; \theta_{spkclassifier}) \tag{1}$$

Data Augmentation Disentangling accent and speaker is challenging, as a speaker typically has a specific way of pronouncing words, causing a strong association between speaker and accent. Hence, not addressing this issue in non-parallel data can lead to entangled representations because a speaker's language and accent can be trivially learned from the dataset. Since our goal is to synthesize speech for a speaker in a target language with a desired accent, disentangling speaker S and accent A is essential, otherwise either speaker identity is not preserved in the target language or the generated speech retains the speaker's accent from the source language. To overcome this problem and promote disentanglement between speaker and accent, we use data augmentations like formant, F_0, and duration scaling. For a given speech sample x_i with speaker identity s_i and accent a_i, we apply a fixed transformation $t \in \{1, 2, ...\tau\}$ to construct a transformed speech sample x_i^t and assign speaker identity as $s_i + t \cdot N_{speakers}$ and accent as original accent a_i, where τ is the number of augmentations. This creates samples with variations in speaker identity with fixed accent helping decorrelate them.

Embedding Regularization Ideally, the information captured by the speaker and accent embeddings should be uncorrelated. To promote disentanglement between accent and speaker embeddings, we aim to decorrelate the following variables: (1) random variables in accent embeddings; (2) random variables in speaker embeddings; (3) random variables in speaker and accent embeddings from *each other*. While truly decorrelating the information is difficult, we can promote something close by using the constraints from VICReg [17]. We denote $E^A \in \mathbb{R}^{D_a\times N_a}$, $E^S \in \mathbb{R}^{D_s\times N_s}$ as the accent and speaker embedding tables respectively. Column vector $e^j \in E$ denotes the j'th embedding in either table. Let μ_E and $Cov(E)$ be the means and covariance matrices. By using VICReg, we constrain standard deviations to be at least γ and suppress the off-diagonal elements of the covariance matrix ($\gamma = 1, \epsilon = 1e{-}4$):

$$L_{var} = \frac{1}{D}\sum_{i=j} \max\left(0, \gamma - \sqrt{Cov(E)_{i,j} + \epsilon}\right) \tag{2}$$

$$L_{covar} = \sum_{i\neq j} Cov(E)^2_{i,j} \tag{3}$$

Next, we attempt to decorrelate accent and speaker variables from *each other* by minimizing the cross-correlation matrix from batch statistics. Let $\tilde{E^A}$ and $\tilde{E^S}$ be the sampled column matrices of accent and speaker embedding vectors sampled within a batch of size B. We compute the batch cross -correlation matrix R^{AS} as follows (μ_{E^A} and μ_{E^S} computed from embedding table):

$$R^{AS} = \frac{1}{B-1}(\tilde{E}^A - \mu_{E^A})(\tilde{E}^S - \mu_{E^S})^T \tag{4}$$

$$L_{xcorr} = \frac{1}{D_a D_s}\sum_{i,j}(R^{AS}_{i,j})^2 \tag{5}$$

2.4. Accent conditioned speech synthesis

We introduce an extra conditioning variable for accent A to RADTTS [4] to allow for accent controllable speech synthesis. We call this model RADTTS-ML, a multilingual version of RADTTS. The following equation describes the model:

$$P_{radtts}(X, \Lambda) = P_{mel}(X|\Phi, \Lambda, A, S)P_{dur}(\Lambda|\Phi, A, S) \tag{6}$$

2.5. Fine-grained frame-level control of speech attributes

Fine-grained control of speech attributes like F_0 and energy $\mathcal{E}$ can provide high-quality controllable speech synthesis [13]. We believe conditioning on such attributes can help improve accent and language transfer. During training, we condition our mel decoder on ground truth frame-level F_0 and energy. Although we believe attribute predictors can be generative models, however due to the preference of deterministic models over generative models[13] by human evaluators, we train deterministic attribute predictors to predict phoneme durations Λ, F_0, and energy $\mathcal{E}$ conditioned on speaker S, encoded text Φ, and accent A. We standardize F_0 using the speaker's F_0 mean and standard deviation to remove speaker-dependent information. This allows us to predict speech attributes for any speaker, accent, and language and control mel synthesis with such features. We refer to this model as RADMMM:

$$P_{radmmm}(X) = P_{mel}(X|\Phi, \Lambda^h, A, S, F_0^h, \mathcal{E}^h) \tag{7}$$

3. Experiments

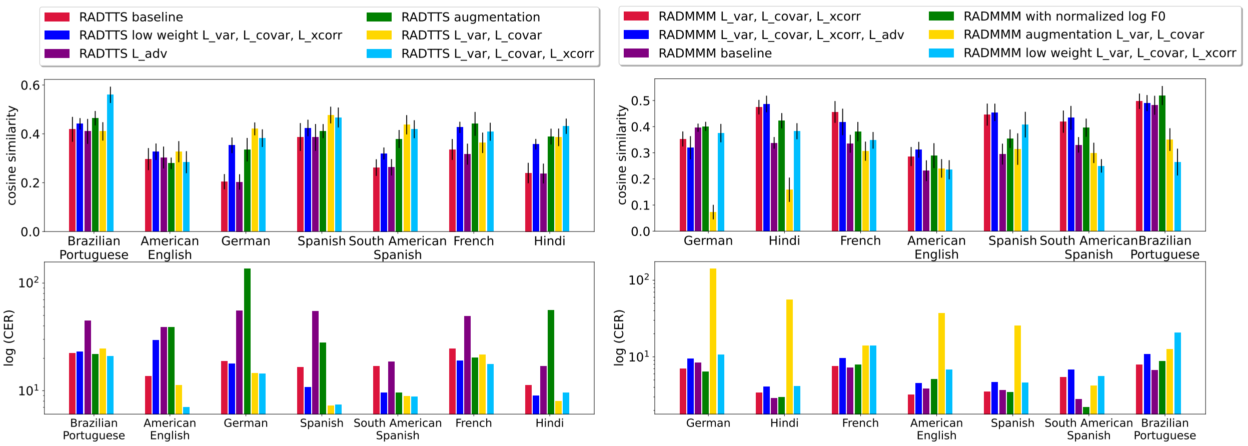
We conduct our experiments[1] on an open source dataset[2] with a sampling rate of 16kHz. It contains 7 different languages (American English, Spanish, German, French, Hindi, Brazilian Portuguese, and South American Spanish). This dataset emulates low-resource scenarios with only 1 speaker per accent with strong correlation between speaker, accent, and language. We use HiFiGAN vocoders trained individually on selected speakers in the evaluation set. We focus on the task of transferring the voice of 7 speakers in the dataset to the *other* 6 language and accent settings. Herein we refer to RADTTS-ML as RT and RADMMM as RM for brevity.

3.1. Ablation of Disentanglement Strategies

We evaluate the effects of disentanglement strategies on the transfer task by measuring speaker timbre retention using the cosine similarity (Cosine Sim) of synthesized samples to source speaker's reference speaker embeddings obtained from the speaker recognition model Titanet [18]. We measure character error rate (CER) with transcripts obtained from Conformer [19] models trained for each language. We use the following definition of Character Error Rate(CER), defined by the following

[1] Samples available at https://bit.ly/radmmm-samples

[2] More details available at https://research.nvidia.com/labs/adlr/projects/radmmm/.



(a) *RADTTS-ML (RT)*

(b) *RADMMM (RM)*

Figure 1: *Comparing speaker cosine similarity and CER of considered disentanglement strategies for every accent.*

equation. CER can be larger than 100 when the number of insertions is larger than the number of correct characters.

$$CER = \frac{substitutions + deletions + insertions}{substitutions + deletions + correct} \times 100 \qquad (8)$$

Table 1 and Figure 1 demonstrate overall and accent grouped effects of various disentanglement strategies. The RT baseline uses the shared text token set, accent-conditioned alignment learning, and no additional constraints to disentangle speaker, text, and accent. The RM baseline uses this setup with F_0 and energy conditioning.

Table 1: *Ablation results comparing disentanglement strategies using Cosine Sim and CER defined in 3.1. Fields are N/A for cases where experiment is not applicable or non informative.*

	RADTTS-ML		**RADMMM**	
Disentanglement Strategy	**Cosine Sim(↑)**	**CER(↓)**	**Cosine Sim(↑)**	**CER(↓)**
Baseline (B)	0.306 ± 0.017	17.7	0.343 ± 0.013	5.1
(B) + normalized F_0 pred	N/A	N/A	0.394 ± 0.014	5.3
(B) + L_{adv}	0.302 ± 0.017	39.9	N/A	N/A
(B) + augmentation	0.385 ± 0.014	44.3	0.217 ± 0.013	41.7
(B) + L_{var}, L_{covar}	0.403 ± 0.014	13.7	N/A	N/A
(B) + L_{var}, L_{covar}, L_{xcorr}	0.422 ± 0.015	12.2	0.419 ± 0.015	5.5
(B) + L_{var}, L_{covar}, L_{xcorr}, L_{adv}	N/A	N/A	0.416 ± 0.016	7.2

Speaker Adversarial Loss (L_{adv}) We observe that the addition of L_{adv} loss to RT and RM does not affect speaker retention when synthesizing the speaker for a target language. However, we observe a drop in character error rate. We believe the gradients from the speaker classifier tend to remove speaker and accent information from encoded text Φ, which affects the encoded text representation leading to worse pronunciation.

Data Augmentation We use Pratt [20] to apply six augmentations with a scaling factor randomly chosen between a specific range: formant scaling down ($\times[0.875 - 1.0]$) and up ($\times[1.0 - 1.25]$), scaling F_0 down ($\times[0.9 - 1.0]$) and up ($\times[1.0 - 1.1]$), and scaling durations to make samples faster($\times[0.9 - 1.0]$)) or slower($\times[1.0 - 1.1]$). We augment the dataset with transformed audio defining a new speaker identifier, but retaining the original accent. In RT, this leads to a significant boost in speaker retention. We believe that creating more speakers per accent enhances disentanglement of accent and speaker. However, in RM, where F_0 is predicted and the model is explicitly conditioned on augmented F_0, we observe a significant drop content quality (higher CER) with augmentations, likely due to conditioning on noisy augmented features.

Embedding Regularization We conduct three ablations with regularization: one that adds variance (L_{var}) and covariance (L_{covar}) constraints to the baseline, and two more involving all three constraints (L_{var}, L_{covar}, L_{xcorr}). We observe an improvement in speaker similarity with the best speaker retention with all three constraints in both RT and RM. Moreover, we observe similar CER to the baselines suggesting similar pronunciation quality.

Our final models include regularization constraints, but we don't use augmentation and L_{adv} due to worse pronunciation quality and limited success on speaker timbre retention.

3.2. Comparing proposed models with existing methods

We compare our final RT and RM with the Tacotron 2-based model described in [9], call it T2, on the transfer task. We reproduced the model to the best of our ability, noting that training on our data was unstable, possibly due to data quality, and that results may not be representative of the original implementation. We tune denoising parameters [21] to reduce audio artifacts from T2 over-smooth generated mels [3, 22]. We attempted to implement YourTTS [14] but ran into issues reproducing the results on our dataset and hence we don't directly compare to it.

Speaker timbre retention Table 2 shows the speaker cosine similarity of our proposed models and T2. We observe that both RT and RM perform similarly in terms of speaker retention and achieve better speaker timbre retention than T2. However, our subjective human evaluation below shows that RM samples are overall better than RT in timbre preservation and pronunciation.

Table 2: *Speaker timbre retention using Cosine Sim (Sec 3.1)*

Model	**Cosine Similarity**
RADTTS-ML (RT)	0.4186 ± 0.0154
RADMMM (RM)	0.4197 ± 0.0149
Tacotron2 (T2)	0.145 ± 0.0119

3.3. Subjective human evaluation

We conducted an internal study with native speakers to evaluate accent quality and speaker timbre retention. Raters were pre-screened with a hearing test based on sinusoid counting. Since MOS is not suited for finer differences, we use comparative mean opinion scores (CMOS) with a 5 point scale (-2 to 2) as the evaluation metric. Given a reference sample and pairs of synthesized samples from different models, the raters use the 5 point scale to indicate which sample, if any, they believe is more similar, in terms of accent or speaker timbre, to the target language pronunciation or speaker timbre in reference audio.

Accent evaluation: We conduct accent evaluation with native speakers of every language. Fig 2 shows the preference scores of native speakers with 95% confidence intervals in each language for model pairs under consideration. Positive mean

Table 3: *Role assignment ablation comparing 1 speaker per accent against 3 speakers per accent.*

	German Speaker (Berndt Ungerer)			English Speaker (LJ Speech)			Hindi Speaker (Indic TTS)			French Speaker (Nadine Eckert)		
Evaluation Setup	**Cosine Sim(↑)**	**CER(↓)**	**PER(↓)**	**Cosine Sim(↑)**	**CER(↓)**	**PER(↓)**	**Cosine Sim(↑)**	**CER(↓)**	**PER(↓)**	**Cosine Sim(↑)**	**CER(↓)**	**PER(↓)**
RM Resynthesis (1 spk per accent)	0.8309 ± 0.0362	0.047 ± 0.033	–	0.8218 ± 0.0584	0.036 ± 0.042	–	0.7809 ± 0.0583	0.153 ± 0.027	–	0.7118 ± 0.0866	0.068 ± 0.028	–
RM Resynthesis (3 spk per accent)	0.8294 ± 0.0321	0.024 ± 0.010	–	0.8183 ± 0.0485	0.030 ± 0.044	–	0.7658 ± 0.0526	0.124 ± 0.035	–	0.8183 ± 0.0485	0.062 ± 0.040	–
RM Cross-Lingual (1 spk per accent)	0.4312 ± 0.0375	0.133 ± 0.033	0.19 ± 0.04	0.5022 ± 0.0404	0.132 ± 0.028	0.17 ± 0.04	0.4012 ± 0.0254	0.071 ± 0.035	0.09 ± 0.02	0.4644 ± 0.0412	0.1532 ± 0.0413	0.14 ± 0.04
RM Cross-Lingual (3 spk per accent)	0.6611 ± 0.0263	0.109 ± 0.027	0.18 ± 0.04	0.5546 ± 0.0274	0.103 ± 0.026	0.14 ± 0.03	0.4772 ± 0.0342	0.043 ± 0.021	0.07 ± 0.02	0.4663 ± 0.0424	0.0970 ± 0.0263	0.12 ± 0.04

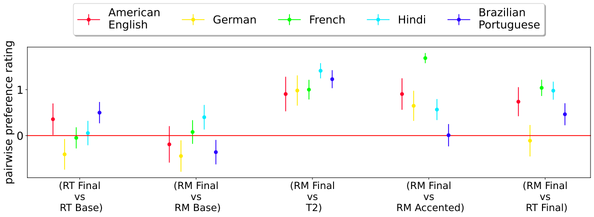


Figure 2: *Accent Similarity CMOS for model pairs.*

scores imply that the top model was preferred over the bottom model within the pair. Given limited access to native speakers, we show results for only 5 languages. We observe that there is no strong preference between RT final and its baseline in terms of accent quality. We find similar results for RM final and its baseline, suggesting that accent and pronunciation are not compromised by our suggested disentanglement strategies. To evaluate controllable accent, we synthesize samples from our best model (RM final) for every speaker in languages other than the speaker's native language. Samples using non-native language and accent are referred to as RM final, and samples with the new language but native accent are referred to as RM accented. Raters preferred samples using the target accent (RM final) over the source speaker's accent (RM accented), indicating the effectiveness of accent transfer. Finally, RM final is preferred over T2 in terms of accent pronunciation.

Speaker timbre evaluation: Table 4 shows CMOS scores with 95% confidence intervals. First, we observe that in both RT and RM, the final models with disentanglement strategies applied are preferred over baseline models in terms of speaker timbre retention. RM accented synthesis (RM accented) is rated as having similar speaker timbre as native accent synthesis with RM (RM final), indicating that changing accent doesn't change speaker timbre in RM, thus showcasing the disentangled nature of accent and speaker. Finally, RM final is preferred over T2 in terms of speaker timbre retention on transferring speaker's voice to target language.

Effects of control with F_0 and $\mathcal{E}$: Comparing RM final with RT final, we see that RM is preferred for most languages except German, indicating that explicit conditioning on F_0 and energy results in better pronunciation and accent. Moreover, as illustrated in Table 1, RM final achieves a better CER than RT final. Table 4 demonstrates that explicit conditioning on F_0 and energy in RM results in much better speaker timbre retention compared to RT. RM results in the best speaker retention, accent quality and pronunciation among our models.

Table 4: *CMOS for speaker timbre similarity.*

Model Pair	**CMOS**
RT Final vs RT Base	0.300 ± 0.200
RM Final vs RM Base	0.750 ± 0.189
RM Final vs RT Final	0.733 ± 0.184
RM Final vs RM Accented	0.025 ± 0.199
RM Final vs T2	1.283 ± 0.144

3.4. Speaker and Accent Role Assignment

In the one speaker per accent data setup described in our experiments, the role assignment of speaker and accent to respective embeddings can be ambiguous. The embedding regularization and cross-correlation minimization constraints attempt to decorrelate the variables but doesn't guarantee independence and role assignment. Whereas speaker timbre is a factor that globally modulates speech, accent determines how each phoneme (text tokens) will be pronounced and is a factor that locally modulates speech. In our formulation, speaker embeddings are concatenated to the input of the decoder and, hence, are only able to globally bias the output. Unlike speaker embeddings, accent embeddings are concatenated to text inputs and then passed to a bi-directional LSTM encoder, hence being able to locally bias the output. These design choices help with the role assignment of the speaker and accent to their respective embeddings. This role assignment ambiguity does not arise in more than one speakers per accent (or augmentation setup with a few additional augmented speakers per accent) setup because the number of speakers and accents are different.

In order to showcase the role assignment in RADMMM, we conduct experiments[3] comparing 1 speaker per accent (where role assignment can be ambiguous) against 3 speakers per accent (without any role assignment ambiguity) in 4-language setting. We evaluate both monolingual (speaker speaking their own language) and cross-lingual (speaker speaking 3 languages other than their own language). Table 3 compares the results based on the content, accent and speaker identity using CER, Phoneme Error Rate(PER) and cosine similarity of Titanet speaker embeddings respectively. We observe that speaker identity for monolingual synthesis is about the same in both settings. Although the speaker identity retention in cross-lingual synthesis is better with 3 speakers than 1 speaker, both are quite close suggesting that speaker is retained even in 1 speaker setting which indicates speaker and accent roles are well assigned. We observe a better content quality with 3 speakers (indicated by lower CER) in both monolingual and cross-lingual synthesis which could be due to larger quantity of data available to model per language. Since accent relates to the pronunciation of text, spoken phonemes are a proxy to measure accent quality. We use an ASR model trained to predict phonemes from speech and compute the phoneme error rate(PER). We observe a lower PER with 3 speakers but with overlapping confidence intervals. This indicates accent quality doesn't change much indicating the adequate role assignment in 1 speaker per accent setting.

4. Conclusion

We present a multilingual, multiaccented and multispeaker TTS model based on RADTTS with novel modifications. We propose and explore several disentanglement strategies resulting in a model that improves speaker, accent and text disentanglement, allowing for synthesis of a speaker with closer to native fluency in a desired language without multilingual speakers. Internal ablation studies indicate that explicitly conditioning on fine-grained features (F_0 and $\mathcal{E}$) results in better speaker retention and pronunciation according to human evaluators. Our model provides an ability to predict such fine-grained features for any desired combination of speaker, accent and language and user studies show that under limited data constraints, it improves pronunciation in novel languages. Future work involves scaling the model to large-resource conditions with more speakers.

[3] Samples available at https://bit.ly/radmmm-role-assignment

5. References

[1] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[3] Rafael Valle, Kevin J Shih, Ryan Prenger, and Bryan Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," in *International Conference on Learning Representations*, 2020.

[4] Kevin J. Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro, "RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis," in *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[5] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "Naturalspeech: End-to-end text to speech synthesis with human-level quality," 2022.

[6] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021.

[7] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *CoRR*, vol. abs/1806.04558, 2018.

[8] Eliya Nachmani and Lior Wolf, "Unsupervised polyglot text-to-speech," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[9] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *CoRR*, vol. abs/1907.04448, 2019.

[10] Bo Li and Heiga Zen, "Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis," in *Interspeech*, 2016.

[11] Adrian Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[12] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[13] Kevin J. Shih, Rafael Valle, Rohan Badlani, João Felipe Santos, and Bryan Catanzaro, "Generative modeling for low dimensional speech attributes with neural spline flows," 2022.

[14] Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir Antonelli Ponti, "Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," *CoRR*, vol. abs/2112.02418, 2021.

[15] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," 2023.

[16] Rohan Badlani, Adrian Łańcucki, Kevin J. Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[17] Adrien Bardes, Jean Ponce, and Yann LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *International Conference on Learning Representations*, 2022.

[18] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8102–8106.

[19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[20] Paul Boersma and David Weenink, "Praat: doing phonetics by computer," 2003.

[21] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[22] Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu, "Revisiting over-smoothness in text to speech," *arXiv preprint arXiv:2202.13066*, 2022.