Scalable Controllable Accented TTS

Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh,
Leibny Paola García-Perera, Sanjeev Khudanpur, Nicholas Andrews, Matthew Wiesner
Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, United States
xli257@jhu.edu

***Abstract*—We propose a method to scale accented TTS training to large, accent-diverse datasets that often lack consistent, high-quality accent labels. Our approach relies on a speech geolocation model to infer accent labels directly from audio. To improve speaker generalization and encourage disentangling speaker from accent we explore timbre augmentation through kNN voice conversion. We validate our approach on CommonVoice by fine-tuning XTTS-v2 with accent labels inferred or improved via geolocation. According to various automated metrics based on embeddings extracted from an accent identification model, the resulting accented TTS model produces speech with better accent fidelity compared to XTTS-v2 fine-tuned on self-reported accent labels in CommonVoice, or other existing accented TTS models. According to human evaluation, it was clear that the geolocation model based data discovery and enhancement improved the naturalness and accent fidelity of generated speech. However, the effect of different data augmentation strategies was less clear.**

***Index Terms*—TTS, accented TTS, label discovery**

I. Introduction

State-of-the-art (SOTA) text-to-speech (TTS) and Voice Conversion systems are increasingly designed to support zero-shot, speech-conditioned synthesis that enables them to mimic the timbre and speaking style of a reference utterance [1]–[7]. However, excplicit control over other speech characteristics, such as prosody, emotion, and especially accent, remains challenging.

A crucial bottleneck in training accented speech synthesis models is the scarcity of large, accent-annotated TTS datasets. Existing TTS datasets that include diverse accents with reliable labels, such as CMU Arctic [8], L2-Arctic [9], and VCTK [10], have total durations that only range between 10 and 50 hours.

TTS systems tend to produce American and British accents [11]–[13], the accents most well-represented in commonly used TTS datasets. This bias overlooks the needs of a broader global audience, many of whom speak English with diverse regional accents.

While several models for accented TTS have been trained on these datasets and demonstrate good accent similarity [13]–[16], they struggle to generalize to unseen speakers due to the limited number of distinct speakers per accent in these datasets. The high costs of collecting and annotating accent labels severely limit the scalability of these datasets and large, in-the-wild, accent-diverse, speech remains hard to leverage for training controllable accented TTS systems.

An alternative to manually curated accent-labeled datasets is the use of in-the-wild datasets with diverse accents, one notable example being CommonVoice [17], a crowd-sourced speech dataset with self-reported accent labels. These labels were used to construct the accent-annotated speech corpus CommonAccent [18], an accent-annotated corpus that has enabled accented TTS models such as AccentBox [12]. However, CommonAccent still has several limitations:

1) Low-quality accent labels, particularly due to L2 (non-native) speakers self-reporting as speakers of mainstream accents such as American or Southeastern English accents;
2) Limited applicability to datasets without self-reported accent labels, such as web-crawled corpora;
3) Certain accents are represented by very few speakers, resulting in difficulties disentangling speaker timbre from accent characteristics.

A. Towards Scalability in Controllable Accented TTS

In this work, we introduce Scalable Controllable Accented TTS. To address the challenge of limited scalable accent data, we leverage a speech geolocation model [19]—a system trained to predict the location on Earth where each utterance was spoken. By using such a model, we can automatically generate accent labels on speech data without accent labels, or enhance the quality of self-reported accent labels. Unlike classifier-based accent labeling models such as GenAID [12], this method can be extended to any accent in any language without existing labeled speech. We demonstrate that the precision of accent labels inferred from the geolocation model is similar to SOTA accent-ID systems.

Furthermore, in order to address the lack of speaker diversity for certain accents, we promote speaker-accent separation by using kNN-VC [20] to convert each utterance to a diverse array of speaker timbres. We show that kNN-VC preserves the accent of the original utterance, making it well-suited for data augmentation for accented TTS training. To validate our proposed method for scaling accent-labeled speech data, we fine-tune a pre-trained multilingual TTS model, XTTS-v2 [3], for accented TTS.

Our contributions are as follows:

1) We apply the speech geolocation model on accent label discovery and filtering, finding that label filtering using the geolocation model during training improves the quality of accent synthesis, while accent label discovery allows for accented TTS training with no pre-existing labels.
2) We find that applying kNN-VC as a data augmentation method for training accented TTS systems improves the quality of synthetic speech according to objective metrics.
3) We achieve comparable or stronger performance to SOTA systems on the accented TTS task on a wide range of accents through XTTS-v2 fine-tuning.

Sample utterances synthesized using our system are included in our demo[1].

II. Related Work

One way around the lack of labeled accented speech is to leverage the well-established research on accented phonetic transliteration. Since many of the TTS models developed during the years 2020 and 2023 contained an explicit phonemizer module, they could easily be cascaded with a dialectal phonemizer to produce accented

[1]https://hstehstehste.github.io/Projects/Demo/index.html

speech [21]–[25]. The idea has persisted in modern end-to-end TTS models that abandonned the phonemizer, but use large language models or other transliteration models to created for grapheme-level accent transliteration [11], [26].

Another approach to address the lack of accented speech data in accents influenced by an L1 substrate language, is to employ zero-shot adaptation using a TTS model trained on the L1 substrate [27], [28]. The relative scarcity of accented speech data can also be addressed with data augmentation [29], [30]. Alternatively, training a TTS model from scratch without sufficient accented speech can be avoided with multi-stage training, where a TTS system is first pre-trained on clean read speech in English or US accent and then fine-tuned on accented speech [29].

Specific modeling strategies are employed to promote speaker-accent disentanglement when the training data is insufficient. The most common strategy is to employ a low-bandwidth accent bottleneck representation, followed by a VAE-based structure which attempts to reconstruct accented speech from the low-bandwidth bottleneck [14], [16], [29], [31], [32]. Another strategy to disentangle speaker and accent representations is domain adversarial training [12], [13], [33]. Data collection efforts have filled in some of the gaps within accented speech datasets. These include datasets that target specific regions of the world [34], or that aim to cover the widest possible a range of accents [35].

III. Method

A. Automatic Accent Label Discovery and Label Filtering through Speech Geolocation

[19] introduced a speech geolocation model trained to predict the broadcast location of speech clips from radio stations around the world. We noticed that this model, though originally developed for applications to language ID, appears, in the process, to learn something about dialect region. In our work, we extend this approach to uncover and refine English accent labels. Specifically, we define approximate geographic bounding boxes for each accent under study (see Figure 1). Next, we predict the location of each utterance in CommonVoice using the geolocation model (in a zero-shot fashion, **without any additional model training**). We **accept** utterances as belonging to a target accent if their predicted location falls within the corresponding bounding box.

We source training and testing data from version 20 of CommonVoice [17], which contains a large but potentially noisy collection of accented speech data. To construct accent-specific subsets of CommonVoice for TTS fine-tuning, we compare the following data selection strategies:

1) **Unfiltered**: using self-reported accented labels in CommonVoice without any filtering. This is the same approach that was used to construct CommonAccent [18].
2) **Filtered**: filtering self-reported accent labels using the geolocation model.
3) **Unlabeled**: discovery of accent labels using the geolocation model.

Table I shows the amount of self-reported accent data pre and post filtering, as well as the amount of available data when the geolocation model [19] is directly employed for accent label discovery. We additionally compute the precision of label discovery using the geolocation model on all the accents included in the CommonAccent dataset, as shown in table III. We observe that our method outperforms XLSR [36] fine-tuned on CommonAccent, and is comparable with GenAID [12] on many accents, in terms of label discovery precision, the most important metric for ensuring that the discovered accent data is of high quality.

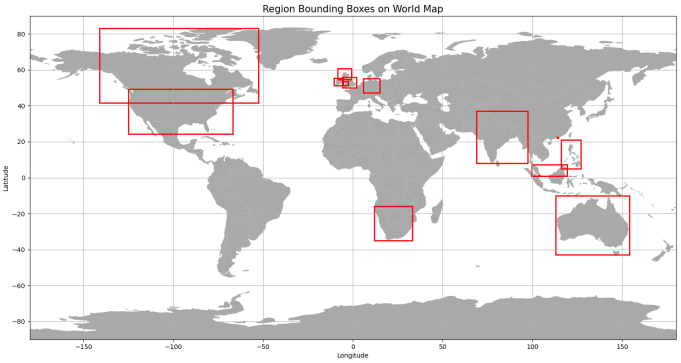

Fig. 1: Bounding boxes for each of the accents included in this study.

TABLE I: Accent Data Statistics in CommonVoice

	Amount of Data (hr)			
Accent	Labeled	Filtered	Found	↑ Found Prec.
US	432.5	222.7	511.9	72.8%
England	133.8	35.2	29.5	52.2%
India	130.7	113.1	196.4	88.8%
Canada	90.4	64.8	701.4	13.3%
Australia	67.5	34.8	68.9	80.3%
Africa	33.4	10.9	14.7	96.3%
Scotland	23.0	1.5	22.8	9.4%
Germany	95.9	0.9	16.7	15.9%
Philippines	7.7	2.7	9.6	60.6%
Ireland	23.8	0.8	7.2	20.4%
Malaysia	2.2	0.1	1.4	7.1%

The geolocation method for accent label discovery may theoretically be extended to any accents that are associated with a geographical range. In practice, since the model was trained on radio broadcasts, its performance depends on the availability and location of broadcasts spoken in the desired accent, as well as which broadcasts were chosen in training. The model works well for Indian and Australian accents, but somewhat worse for accents of the British Isles.

B. Fine-tuning XTTS-v2 for Accented TTS

XTTS-v2 [3] is a multilingual TTS model trained on CommonVoice. It consists of three main components: a discrete VAE [37] encoder, a discrete token sequence to sequence transformer [38], and a Hifi-GAN [39] vocoder which converts token sequence predictions into waveforms. The language of synthesized speech is controlled with a language token that is pre-pended to each sequence. Notably, XTTS-v2 exhibits some zero-shot accented TTS capability when the language token used during inference is different from the language of the underlying text.

We fine-tune our accented TTS model from XTTS-v2, replacing the language token with an accent token as we're limited to TTS in English for this study. During training, we up-sample lower-resource accents so that all the accents included in our training are equally represented in each batch.

C. Timbre and Acoustic Diversification using kNN-VC

In a low-resource accented TTS training, it is common for certain accents to have very few distinct speakers in the dataset, causing timbre generalization difficulties for the model on those accents. Table II shows that the CommonVoice dataset suffers from this very issue, with certain accents (German before filtering; German and

TABLE II: Number of distinct speakers in CommonVoice for various accents

Accent	England	US	India	Germany	Africa	Canada	Australia	Philippines	Scotland	Ireland	Malaysia	Wales
Filtered	583	4335	1171	**1**	71	599	493	55	73	17	**10**	**4**
Unfiltered	1507	5218	1254	**3**	178	635	540	83	109	123	60	49
Unlabeled	3580	21276	7124	2458	586	26459	5093	1315	3179	1392	391	449

TABLE III: Precision on CommonAccent test

Accent	Geolocation	XLSR	GenAID
US	78%	86%	**89%**
England	66%	67%	**87%**
India	**87%**	61%	**87%**
Canada	9%	11%	**21%**
Australia	**70%**	47%	**70%**
Africa	**95%**	16%	74%
Scotland	17%	23%	**77%**
Philippines	**77%**	5%	46%
Ireland	3%	11%	**62%**

Welsh after filtering) containing less than 10 distinct speakers in their respective training set. Previous works [7], [28], [30], [40] have employed various data augmentation techniques, including duration, pitch, phonetic feature perturbation, as well as voice conversion, in order to promote separation between different aspects of speech: timbre, accent and content.

In our work, we employ a voice conversion system known as kNN-VC [20] to augment our data. Due to the strict time alignment of kNN-VC's input and output speech, it is known to be able to faithfully produce speech with the timbre of the target speaker while preserving the input utterance's accent and speaking style [41]. We convert the timbre of each utterance to a randomly chosen speaker from LibriTTS [42], allowing the model to be exposed to a variety of speaker timbres in each accent.

IV. Experiments

A. XTTS-v2 Fine-tuning

We included all the accents in CommonVoice that has more than 2 hours of training data (as shown in table I). We do not measure the multilingual TTS capability of the model post fine-tuning. We fine-tune a separate model using each of the data label filtering/discovery strategy discussed in section III-A for comparison.

We perform ablation and comparison studies for kNN-VC augmentation on the XTTS-v2 model fine-tuned on the filtered (Approach 2) set, training one model without any data augmentation and another with pitchshift augmentation where the fractional step parameter is chosen at random between -4 and 4.

B. Baseline Systems

In addition to the original XTTS-v2 system, we compare our fine-tuned models with two external baselines: AccentBox[2] [12] and CosyVoice2[3] [1].

1) AccentBox: Accentbox was pretrained on LibriTTS-R [43] and subsequently fine-tuned on VCTK [10], which contains 11 different accents from the British isles. As accented TTS inference for AccentBox is performed using an accent embedding as conditioning, we generate accented speech using AccentBox using the following procedure: first, for each accent, we extract and average the embeddings of each utterance with that accent label in the CommonVoice development split; next, at inference time, we choose the average embedding of the desired accent label as the accent conditioning embedding.

2) CosyVoice2: While CosyVoice2 does not explicitly support accent-controlled TTS in English, it performs zero-shot style copying and demonstrates some degree of accent imitation capabilities. We approximate accented TTS using CosyVoice2 by randomly picking an utterance from CommonVoice development split with the desired accent label to serve as the style prompt.

3) XTTS-v2 baseline configurations: We experiment with two baseline configurations of XTTS-v2. In the first configuration, we run inference with English as the language label, which tends to generate US or English-accented speech. In the second configuration (labeled "XTTS approx" in result tables), we leverage XTTS-v2's limited zero-shot L2 accent synthesis capability, and pick an "approximate target" language label corresponding to the desired accent label whenever applicable (for example, we use language label "Hindi" when generating Indian-accented speech).

C. Model-based Evaluation of Accent Similarity

Following [44], we identified a number of candidates for model-based evaluation of accent similarity: vowel formants (VF), phonetic posteriorgrams (PPG)[4] [45], Mel ceptral distortion (MCD), WavLM [46] fine-tuned for speaker identity[5]), XLSR [36] fine-tuned on CommonAccent[6] [18], and GenAID[7].

We first observed the distribution of similarity scores compute using each of these metrics on content-normalized but accent-diverse speech data from the Speech Accent Archive [47]. We found no meaningful difference between the scores for different accents and those for the same accents in terms of VF, PPG, and MCD. The remaining three models all demonstrated the ability to distinguish speech accents in recorded speech; however, WavLM and XLSR were unable to meaningfully differentiate between synthesis systems, giving almost identical scores for each of our proposed and baseline systems. We therefore use GenAID as our primary model for accent evaluation. Figure 2, where different accents were grouped using the spectral clustering algorithm ($n = 7$) according to their pairwise average GenAID embedding cosine similarity, illustrates that the cosine similarities align to an extent with established typological groupings of English accents.

We compute the following metrics using GenAID as our embedding extractor:

1) Table Va: Accent embedding cosine similarity to the ground truth utterance ;
2) Table Vb: Motivated by the LRE22 challenge [48], we similarly construct an accent evaluation use GenAID as a "frontend" model. We use the evaluation data as enrollment. We train a Gaussian backend model on the ground-truth data, treating it as enrollment. We then use the synthetic speech as the evaluation data, where

[2]https://github.com/jzmzhong/coqui-TTS/tree/accentbox/
[3]https://github.com/FunAudioLLM/CosyVoice2
[4]https://github.com/liusongxiang/ppg-vc/tree/main
[5]https://huggingface.co/microsoft/wavlm-base-plus-sv
[6]https://huggingface.co/Jzuluaga/accent-id-commonaccent_xlsr-en-english
[7]https://github.com/jzmzhong/GenAID/tree/GenAID

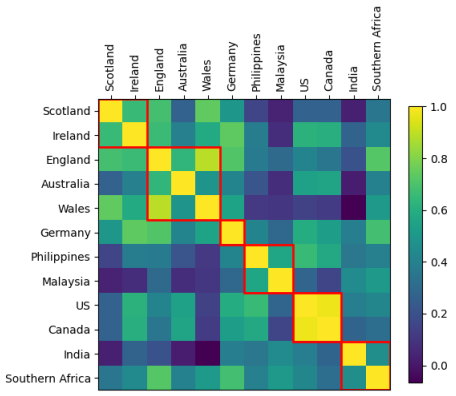

Fig. 2: Cosine Similarity Matrix of average GenAID embeddings for various accents in CommonVoice.

the goal is to produce accented enough speech that it can be easily detected. We measure the detection cost function (DCF) at two operating points, as done in [48] to compute how close the generated speech was to the target accent. The goal is effectively to fool an accentID model. Due to the high dimensionality of GenAID embeddings and the relatively small sample size of our test set, we used a very reduced dimension of 18 for PCA when training the Gaussian backend.

Metric 1 is the most commonly used objective metric for accented TTS evaluation. However, it assumes that accent embeddings are disentangled from speaker or content information, a condition that is often not met in practice. Metric 2 addresses these limitations by aggregating the reference accent embeddings. In particular, metric 2 accounts for different degrees of intra-accent embedding variance for each target accent. Furthermore, as metric 2 does not require synthesized utterances to have corresponding ground-truth utterances, we are able to greatly increase the test set (especially for low-resource accents) and control for factors such as speaker timbre and speech content when comparing across accents.

D. Human Subjective Evaluation

For each target accent, we enlist 15 human evaluators through prolific[8] who are based in the country or region where the target accent would be prevalent or dominant. For each utterance, evaluators are tasked with producing two scores, both on a 1–5 scale: naturalness and accent plausibility. Following [49], we narrowly define naturalness as "sounding like it was produced by a real, human speaker rather than by a computer or artificial system". Accent plausibility is defined as "how closely the accent in the audio matches the way natives of [country/region] would naturally speak English." Each human evaluator is presented with 30 real utterances from the CommonVoice test split with the target accent label, as well as 30 utterances from each system included in the human evaluation with the same content as the real utterances. An exception is made for Filipino-accented English, where, due to lack of data availability, only 20 utterances from each system are included.

In this work, as we trained and evaluated on a much greater number of accents than is commonly reported, human evaluation of every system output for all the accents seen in training is prohibitively expensive. Furthermore, many accents share many similarities and appear to be quite confusable, i.e., US and Canadian English (see Figure 2). We therefore focused human evaluation on a subset of accents with wide geographic coverage, unique linguistic features or data particularities, and which facilitate comparison with prior work.

[8]https://app.prolific.com/

V. Results

A. Speaker Similarity and Content Preservation

We analyze the speaker timbre copying capability of the various models using Resemblyzer.[9] As shown in Table IV, our systems trained with data processed with numerous strategies (filtered, unfiltered and unlabeled) achieve very similar speaker similarity scores that are comparable with that of CosyVoice2. On the content preservation side, we measured the word error rate (WER) of synthesized speech using Whisper[10] [50]. CosyVoice2 outperforms all other systems, although it is worth noting that synthesized accented speech presents a challenge to ASR systems.

TABLE IV: Cosine similarity of Speaker Embedding and WER. Cells are colored based on their relative value in their respective column: darker indicates better performance.

System	Speaker Sim. ↑	WER (%) ↓
Ground Truth		7.1
Filtered	0.843	13.1
Unfiltered	0.847	14.0
Unlabeled	0.858	18.2
Accentbox	0.664	14.2
CosyVoice2	0.855	7.1
XTTS-v2	0.812	8.6
XTTS approx	0.813	23.2

B. Our system vs. Baselines

1) Objective Evaluations: As shown in table Va, CosyVoice2 outperforms all of our systems on metric 1, while Accentbox lags substantially behind. However, different patterns emerge when the influence of speaker similarity is eliminated, either by aggregating the accent embedding similarity target or using the entire pool of target accent embeddings as enrollment data (metric 2). Shwon in table Vb, in most categories, the filtered system stands out for producing the most accent-appropriate utterances according to GenAID embeddings. Accentbox, whose training data included only accents from the British Isles, nevertheless stands out in a number of L1 accents including English, Australian, Scottish and Irish. CosyVoice2 remains competitive for nearly every accent, including L2 accents such as Indian, German and Malaysian. Finally, the unaccented XTTS-v2 baselines are not competitive with other baselines and models, except in rare cases like zero-shot generation of German-accented speech.

2) Human Evaluations: CosyVoice2 was included in 4 sets of tests: Irish, Indian, Australian, and US accents. Annotators consistently placed CosyVoice2 above every other system included in this study, even preferring it over real utterances in the US accent test (more discussions in section V-E). CosyVoice2 also scored well on accent plausibility in every test except in the case of Irish.

Accentbox was included in 2 sets of tests: Irish and US. Despite only being trained on accents from the British Isles, the original authors reported their capability to generate the unseen US accent. This was confirmed in our study, which found that annotators consistently preferred Accentbox over every other system in both tests where it was included. In the case of Irish accent, this came at the cost of slightly reduced naturalness.

[9]https://github.com/resemble-ai/Resemblyzer
[10]https://github.com/openai/whisper

TABLE V: Automated evaluation based on GenAID embeddings. † AccentBox was conditioned on the GenAID embeddings during training for accent generation, so this may not be a fair comparison. * The training data may not contain all accents, which could have contributed to poor performance on certain accents. XTTS-approx is the XTTS-v2 system where the English language token is replaced with a non-English L1 language that might more closely align with the accent, i.e., Hindi, for Indian accented English.

(a) Cosine similarity of ground-truth utterances and generated accented speech samples prompted with different utterances from the same ground-truth speakers. Darker colors indicate better similarity.

System	**Accent** (Cosine Similarity ↑)												
	All	England	US	India	Germany	Africa	Canada	Australia	Philippines	Scotland	Ireland	Malaysia	Wales
XTTS-v2	0.356	0.358	0.437	0.270	0.382	0.302	0.412	0.348	0.304	0.348	0.371	0.245	0.401
XTTS approx	0.411	0.364	0.440	0.459	0.470	0.374	0.427	0.341	0.475	0.361	0.370	0.496	0.411
Unlabeled	0.485	0.396	0.489	0.606	0.368	0.392	0.524	0.469	0.510	0.480	0.391	0.407	0.487
Unfiltered	0.499	0.436	0.488	0.602	0.444	0.371	0.549	0.471	0.534	0.485	0.447	0.522	0.518
Filtered	0.520	0.451	0.507	0.629	0.409	0.396	0.579	0.489	0.568	0.546	0.451	0.593	0.527
AccentBox†*	0.376	0.384	0.386	0.334	0.376	0.307	0.415	0.439	0.338	0.426	0.380	0.202	0.430
CosyVoice2*	0.542	0.510	0.560	0.597	0.428	0.466	0.603	0.509	0.566	0.439	0.432	0.521	0.561

(b) The detection cost function (DCF) averaged over two operation points ($p_{tgt} = 0.1$ and $p_{tgt} = 0.5$). The DCF is computed using a Gaussian backend with shared covariance. The real, ground-truth evaluation data is used for enrollment, and the generated synthetic data are the test set.

System	**Accent** (Detection Cost Function ↓)												
	All	England	US	India	Germany	Africa	Canada	Australia	Philippines	Scotland	Ireland	Malaysia	Wales
XTTS-v2	1.191	1.454	1.285	1.001	1.434	1.133	1.184	1.137	1.044	1.146	1.074	1.013	1.385
XTTS approx	1.136	1.199	1.142	0.967	1.439	1.133	1.044	1.098	1.042	1.094	1.053	0.911	1.157
Unlabeled	0.936	1.092	0.869	0.553	1.269	1.088	0.799	0.861	0.752	0.998	0.996	1.046	0.906
Unfiltered	0.735	0.730	0.795	0.651	0.756	0.933	0.748	0.580	0.750	0.764	0.725	0.788	0.600
Filtered	0.571	0.652	0.674	0.494	0.638	0.869	0.672	0.382	0.489	0.442	0.701	0.329	0.511
AccentBox†*	0.862	0.926	0.771	0.945	1.227	0.920	0.784	0.855	0.902	0.449	0.683	1.001	0.882
CosyVoice2*	0.748	0.963	0.937	0.539	0.570	0.861	0.847	0.604	0.526	0.683	0.891	0.600	0.956

TABLE VI: Performance against SOTA systems

System	**NMOS** ±95% **CI**				**AMOS** ±95% **CI**			
	Ireland	**India**	**Australia**	**US**	**Ireland**	**India**	**Australia**	**US**
Ground Truth	3.66 ±0.11	3.44 ±0.11	3.04 ±0.12	3.59 ±0.12	3.49 ±0.13	3.33 ±0.11	3.04 ±0.12	3.43 ±0.12
CosyVoice2	2.92 ±0.11	3.30 ±0.11	3.18 ±0.12	3.70 ±0.11	2.09 ±0.12	3.01 ±0.12	2.95 ±0.13	3.70 ±0.11
Accentbox	2.77 ±0.11			3.69 ±0.11	3.03 ±0.12			3.63 ±0.11
Filtered	2.96 ±0.12	3.23 ±0.12	2.95 ±0.11	3.34 ±0.12	2.63 ±0.12	3.04 ±0.12	3.13 ±0.12	3.42 ±0.12

The original XTTS-v2 was included in evaluations for Australian, Filipino, and US accents. As XTTS-v2 produces mostly acoustically clean, US or English-accented speech, it unsurprisingly scored very well on naturalness on all tests as well as accent plausibility on the US accent test. It scored poorly on accent plausibility in the Australian and Filipino accent evaluations.

C. Filtering vs. No Filtering vs. Label Discovery

1) Objective Evaluations: The model trained with data filtered using the geolocation model outperformed the model trained with self-reported accent labels in CommonVoice on nearly every accent. Some notable exceptions include German and Irish. Interestingly, contrary to our expectations, the filtering method resulted in smaller improvements in accents where we expect the labels to be noisier, namely English and US accents.

According to both metrics, the model trained using labels discovered with the geolocation model underperformed the other data settings, unfiltered and filtered. However, it substantially outperformed the XTTS-v2 baselines on most accents.

2) Human Evaluations: Comparisons of human evaluation of the different filtering systems are shown in table VII. The unfiltered system, trained using self-reported accent labels from CommonVoice, was included in evaluations for Australian, Filipino, and US accents. The filtered system outperformed the unfiltered system substantially in both the Australian and the Filipino accents, demonstrating that data filtering using the geolocation model is effective at improving accented speech synthesis.

When it comes to the system trained with geolocation-discovered labels, results demonstrate decent accent generation capability of the resulting model, although annotators consistently prefer other systems over the unlabeled one in terms of accent plausibility.

D. Effects of kNN-VC Data Augmentation

1) Objective Evaluations: To study the effect of kNN-VC data augmentation, we perform two ablation studies: one with no data augmentation during training, and another using a signal-processing-based pitch augmentation method known as pitchshift as timbre augmentation during training. Table VIII shows that on most accents, training with kNN-VC augmentation improves the accent similarity of the resulting model according to objective metrics. It is worth mentioning that kNN-VC introduces a significant training overhead, and that if training is time-constrained, then training with simple augmentation methods like pitchshift can also effectively improve the accent similarity of the model.

2) Human Evaluations: We first test the accent preservation effects of kNN-VC conversion. Real utterances voice converted using kNN-

TABLE VII: Human Evaluation: Comparison between different training data filtering/discovery strategies

System	NMOS ±95% CI			AMOS ±95% CI		
	Australia	Philippines	US	Australia	Philippines	US
XTTS-v2	2.85 ±0.11	2.85 ±0.19	3.45 ±0.12	2.22 ±0.11	2.66 ±0.19	3.61 ±0.12
Unlabeled	2.58 ±0.11	2.56 ±0.19	3.29 ±0.12	2.81 ±0.12	2.52 ±0.18	3.37 ±0.12
Unfiltered	3.04 ±0.10	2.84 ±0.19	3.33 ±0.12	2.92 ±0.11	2.77 ±0.18	3.41 ±0.11
Filtered	2.95 ±0.11	3.17 ±0.19	3.34 ±0.12	3.13 ±0.12	2.98 ±0.18	3.42 ±0.12
Ground Truth	3.04 ±0.12	3.68 ±0.18	3.59 ±0.12	3.04 ±0.12	3.57 ±0.17	3.43 ±0.12

TABLE VIII: Automated evaluation using the detection cost function (DCF) as described in Table Vb on models trained using different data augmentation methods.

Method	Accent (Detection Cost Function ↓)												
	All	England	US	India	Germany	Africa	Canada	Australia	Philippines	Scotland	Ireland	Malaysia	Wales
None	0.744	0.709	0.690	0.587	0.864	0.989	0.690	0.489	0.678	0.612	0.905	0.884	0.827
Pitchshift	0.666	0.715	0.699	0.576	0.726	1.025	0.704	0.445	0.575	0.625	0.845	0.319	0.744
kNN-VC	0.571	0.652	0.674	0.494	0.638	0.869	0.672	0.382	0.489	0.442	0.701	0.329	0.511

TABLE IX: Human Evaluation: Effect of data augmentation during training

System	NMOS ±95% CI			AMOS ±95% CI		
	Australia	Scotland	US	Australia	Scotland	US
Filtered - No Aug.	2.81 ±0.11	3.33 ±0.11	3.40 ±0.11	2.90 ±0.11	3.37 ±0.11	3.55 ±0.11
Filtered - Pitchshift Aug.	3.04 ±0.11	3.09 ±0.12	3.43 ±0.11	3.13 ±0.11	3.16 ±0.12	3.55 ±0.11
Filtered - kNN-VC aug.	2.95 ±0.11	3.19 ±0.11	3.34 ±0.12	3.13 ±0.12	3.17 ±0.11	3.42 ±0.12
Ground Truth	3.04 ±0.12	3.63 ±0.11	3.59 ±0.12	3.04 ±0.12	3.54 ±0.12	3.43 ±0.12

TABLE X: Human Evaluation: Effects of kNN-VC augmentation

Input	NMOS ±95% CI		AMOS ±95% CI	
	Australia	Philippines	Australia	Philippines
kNN-VC	2.4 ±0.11	3.17 ±0.19	2.47 ±0.12	3.14 ±0.11
Ground Truth	3.04 ±0.12	3.68 ±0.18	3.59 ±0.12	3.57 ±0.12

VC (into randomly-selected target speakers from LibriTTS) were included in two human evaluations: Filipino and Australian accents, as shown in table X. In the Filipino accent human evaluation, kNN-VC achieved by far the second best accent plausibility score, second only to real utterances. In the Australian accent human evaluation, kNN-VC received very poor scores on both naturalness (last) and accent plausibility (second from last, only above original XTTS-v2). We will further discuss this result in section V-E.

Next, we compare the systems trained using different augmentation strategies. Ablation models - ones trained with no speaker augmentation and with pitchshift augmentation - were included in the studies for Australian, Scottish, and US accents. Human evaluators assessed these systems to be very similar in all three cases.

E. Limitations of Human Evaluation

We observe that human annotators never give significantly higher accent plausibility scores than naturalness scores. Moreover, naturalness scores by human evaluators are known to be sensitive to acoustic conditions and content. These appear to be in effect in the US-accented human evaluation, where the relative noisiness of real utterance in CommonVoice resulted in many annotators preferring systems that generate acoustically clean utterances over real utterances. Similar effects likely contributed to the results in the Australian-accented evaluation, where both real utterances and kNN-VC voice-converted real utterances received relatively poor naturalness and accent plausibility scores.

VI. Conclusions

In this work, we apply a geolocation model for accent label discovery and cleaning, a technique that can be extended to any quantity of unlabeled accented speech data as well as to previously unseen accent labels. We achieve precision comparable to SOTA baselines for accent ID. We subsequently fine-tune XTTS-v2 on CommonVoice data that had been accent-labeled or filtered using the geolocation method. To promote speaker-accent disentanglement, we apply kNN-VC augmentation to diversify the timbre and acoustics of the training data. Using these methods, we achieve competitive accented TTS results to existing approaches.

Our future work will explore additional label discovery methods, such as pseudo-labeling using accent ID models, and will aim to bridge the gap between model-based and human evaluation of accented speech generation.

Acknowledgment

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the ARTS Program under contract D2023-2308110001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

We would also like to mention generous help from Jan Melechovsky and Jinzuomu Zhong with setting up their models as baselines, Niyati Bafna on setting up accent ID baselines, and Lin Zhang for help with drafting the paper.

References

[1] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "CosyVoice: A Scalable Multilingual Zero-Shot

Text-to-Speech Synthesizer Based on Supervised Semantic Tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[2] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer," in *International Conference on Learning Representations*, 2025.

[3] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, 2024, pp. 4978–4982.

[4] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, M. Wiesner, and N. Andrews, "Genvc: Self-supervised zero-shot voice conversion," 2025. [Online]. Available: https://arxiv.org/abs/2502.04519

[5] T. Dang, D. Aponte, D. Tran, and K. Koishida, "Livespeech: Low-latency zero-shot text-to-speech via autoregressive modeling of audio discrete codes," in *Interspeech 2024*, 2024, pp. 3395–3399.

[6] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[7] H.-S. Choi, J. Yang, J. Lee, and H. Kim, "NANSY++: Unified voice synthesis with neural analysis and synthesis," in *The Eleventh International Conference on Learning Representations*, 2023.

[8] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[9] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," in *Interspeech 2018*, 2018, pp. 2783–2787.

[10] C. Veaux, J. Yamagishi, and K. MacDonald, "Superseded - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," https://datashare.ed.ac.uk/handle/10283/2651, 2016, university of Edinburgh. Centre for Speech Technology Research (CSTR).

[11] S. Inoue, S. Wang, W. Wang, P. Zhu, M. Bi, and H. Li, "Macst: Multi-accent speech synthesis via text transliteration for accent conversion," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[12] J. Zhong, K. Richmond, Z. Su, and S. Sun, "Accentbox: Towards high-fidelity zero-shot accent generation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[13] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Accent conversion in text-to-speech using multi-level vae and adversarial training," in *Proc. of IEEE Tencon, Singapore*, 2024.

[14] ——, "Accented text-to-speech synthesis with a conditional variational autoencoder," in *Proc. of IEEE Tencon, Singapore*, 2024. [Online]. Available: https://arxiv.org/abs/2211.03316

[15] J. Melechovsky, A. Mehrish, D. Herremans, and B. Sisman, "Learning accent representation with multi-level vae towards controllable speech synthesis," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 928–935.

[16] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Dart: Disentanglement of accent and speaker representation in multispeaker text-to-speech," in *Audio Imagination: NeurIPS 2024 Workshop*, 2024.

[17] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[18] J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan, "Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice," in *Interspeech 2023*, 2023, pp. 5291–5295.

[19] P. Foley, M. Wiesner, B. Odoom, L. P. Garcia Perera, K. Murray, and P. Koehn, "Where are you from? geolocating speech and applications to language identification," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., 2024, pp. 5114–5126.

[20] M. Baas, B. van Niekerk, and H. Kamper, "Voice Conversion With Just Nearest Neighbors," in *Interspeech 2023*, pp. 2053–2057.

[21] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Accented text-to-speech synthesis with limited data," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1699–1711, 2024. [Online]. Available: https://doi.org/10.1109/TASLP.2024.3363414

[22] G. Karakasidis, N. Robinson, Y. Getman, A. Ogayo, R. Al-Ghezi, A. Ayasi, S. Watanabe, D. R. Mortensen, and M. Kurimo, "Multilingual TTS Accent Impressions for Accented ASR," in *Text, Speech, and Dialogue: 26th International Conference*, 2023, p. 317–327.

[23] B. Kolluru, V. Wan, J. Latorre, K. Yanagisawa, and M. J. F. Gales, "Generating multiple-accent pronunciations for tts using joint sequence model interpolation," in *Interspeech 2014*, 2014, pp. 1273–1277.

[24] R. Liu, H. Zuo, D. Hu, G. Gao, and H. Li, "Explicit intensity control for accented text-to-speech," in *Interspeech 2023*, 2023, pp. 22–26.

[25] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2188–2201, 2024.

[26] K. Yamauchi, Y. Saito, and H. Saruwatari, "Cross-dialect text-to-speech in pitch-accent language incorporating multi-dialect phoneme-level bert," 2024. [Online]. Available: https://arxiv.org/abs/2409.07265

[27] Y. Zhou, Z. Wu, M. Zhang, X. Tian, and H. Li, "Tts-guided training for accent conversion without parallel data," *IEEE Signal Processing Letters*, vol. 30, pp. 533–537, 2023.

[28] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "Rad-mmm: Multilingual multiaccented multispeaker text to speech," in *INTERSPEECH*, 2023, pp. 626–630.

[29] Y. Zhang, Z. Wang, P. Yang, H. Sun, Z. Wang, and L. Xie, "Accentspeech: Learning accent from crowd-sourced data for target speaker tts with accents," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 76–80.

[30] G. Tinchev, M. Czarnowska, K. Deja, K. Yanagisawa, and M. Cotescu, "Modelling low-resource accents without accent-specific tts frontend," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[31] L. Ma, Y. Zhang, X. Zhu, Y. Lei, Z. Ning, P. Zhu, and L. Xie, "Accent-vits:accent transfer for end-to-end tts," 2023. [Online]. Available: https://arxiv.org/abs/2312.16850

[32] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 5180–5189.

[33] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Multi-scale accent modeling and disentangling for multi-speaker multi-accent text-to-speech synthesis," 2025. [Online]. Available: https://arxiv.org/abs/2406.10844

[34] S. Ogun, A. T. Owodunni, T. Olatunji, E. Alese, B. Oladimeji, T. Afonja, K. Olaleye, N. A. Etori, and T. Adewumi, "1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis," in *Interspeech 2024*, 2024, pp. 1855–1859.

[35] W. Wang, Y. Song, and S. Jha, "Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech," in *Interspeech 2024*, 2024, pp. 1365–1369.

[36] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Interspeech 2021*, 2021, pp. 2426–2430.

[37] J. T. Rolfe, "Discrete variational autoencoders," in *International Conference on Learning Representations*, 2017.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[39] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.

[40] E. Cooper, C.-I. Lai, Y. Yasuda, and J. Yamagishi, "Can speaker augmentation improve multi-speaker end-to-end tts?" in *Interspeech 2020*, 2020, pp. 3979–3983.

[41] Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola Garc'ia-Perera, S. Khudanpur, Nicholas Andrews, and Matthew Wiesner, "Hltcoe jhu submission to the voice privacy challenge 2024," in *4th Symposium on Security and Privacy in Speech Communication*, 9 2024.

[42] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*, 2019, pp. 1526–1530.

[43] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," in *Interspeech 2023*, 2023, pp. 5496–5500.

[44] J. Zhong, S. Liu, D. Wells, and K. Richmond, "Pairwise evaluation of accent similarity in speech synthesis," 2025. [Online]. Available: https://arxiv.org/abs/2505.14410

[45] C. Churchwell, M. Morrison, and B. Pardo, "High-fidelity neural phonetic posteriorgrams," in *ICASSP 2024 Workshop on Explainable Machine Learning for Speech and Audio*, April 2024.

[46] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[47] S. Weinberger, "Speech accent archive," http://accent.gmu.edu.

[48] Y. Lee, C. Greenberg, L. Mason, and E. Singer, "Nist 2022 language recognition evaluation plan," Language Recognition Evaluation, Tech. Rep., 2022.

[49] E. Cooper, S. L. Maguer, E. Klabbers, and J. Yamagishi, "Good practices for evaluation of synthesized speech," 2025. [Online]. Available: https://arxiv.org/abs/2503.03250

[50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html