Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Pairwise Evaluation of Accent Similarity in Speech Synthesis

Jinzuomu Zhong[1], *Suyuan Liu*[2], *Dan Wells*[1], *Korin Richmond*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Department of Linguistics, University of British Columbia, Canada
{jinzuomu.zhong, dan.wells, korin.richmond}@ed.ac.uk, suyuan.liu@ubc.ca

Abstract

Despite growing interest in generating high-fidelity accents, evaluating accent similarity in speech synthesis has been underexplored. We aim to enhance both subjective and objective evaluation methods for accent similarity. Subjectively, we refine the XAB listening test by adding components that achieve higher statistical significance with fewer listeners and lower costs. Our method involves providing listeners with transcriptions, having them highlight perceived accent differences, and implementing meticulous screening for reliability. Objectively, we utilise pronunciation-related metrics, based on distances between vowel formants and phonetic posteriorgrams, to evaluate accent generation. Comparative experiments reveal that these metrics, alongside accent similarity, speaker similarity, and Mel Cepstral Distortion, can be used. Moreover, our findings underscore significant limitations of common metrics like Word Error Rate in assessing underrepresented accents.

Index Terms: speech synthesis, accent, preference evaluation, listening test, laboratory phonetics

1. Introduction

Motivated by the social and moral imperative for more inclusive speech technology, the community has witnessed a growing interest in systems capable of generating high-fidelity *accent*, across various speech generation tasks. In Zero-Shot Text-to-Speech (ZS-TTS), accent hallucination/mismatch, where the generated speech deviates from the reference speech in accent, is reported in [2, 3, 5] and addressed in [3, 4, 5]. In Accented TTS, numerous attempts have been made to generate high-fidelity accent based on pre-defined accent variety labels or intensity levels [6]–[13]. In Accent Conversion (AC), numerous attempts have been made to map speech from foreign (L2) to native (L1) accent, preserving content and speaker information while removing the foreign accent in the source speech [14]–[22]. However, how to evaluate accent similarity in speech is an under-researched topic that lacks consensus. Tab. 1 shows a non-exhaustive list of various (accent) generation systems/papers, along with the subjective and objective methods they employ for evaluation.

For *subjective* evaluation, there are mainly two categories: *reference-free* and *reference-based* listening tests. *Reference-free* methods ask listeners to rate speech utterances without any reference speech. Listeners are tasked to rate the degree of L2 accent, such as the accentedness test [26] commonly used in AC [14, 15, 16] or to rate the resemblance to a pre-defined accent label such [12, 13, 18]. These methods wrongly assume accents to be on a one-dimensional scale of intensity, or categorical (only comprising standard accent varieties) rather than varying on an individual basis [27]. *Reference-based* methods borrow the AB/MOS/CMOS listening tests from standard evaluation of naturalness or audio quality, and ask listeners to rate the accent similarity between reference speech X and target speech utterances, denoted as XAB/XMOS/XCMOS. However, no prior work has studied the validity of these tests specifically on evaluating accent similarity, with most relevant work focusing on test designs in evaluating naturalness [28, 29].

In the absence of reliable and cost-effective subjective evaluation methods, many systems/papers turn to *objective* evaluation, mostly using either classification results or cosine similarity of embeddings extracted from Accent Identification (AID) models. However, how well these proxy metrics represent accent similarity remains unstudied, with relevant work focusing on the correlation between objective and subjective evaluation of speaker similarity [30]. Meanwhile, these systems/papers report a broad range of objective metrics to proxy intelligibility, naturalness, and so on (see Tab. 1). However, some of these methods may be unfair for evaluating speech of underrepresented accents, namely the accent bias in WER [31].

Given this research gap, we explore the question: *How should accent similarity be evaluated both subjectively and objectively?* For subjective evaluation, due to the reliability of AB tests and the limitations found with MOS tests [28, 29, 32], we build upon the XAB listening test, with three additional components that achieve higher statistical significance with fewer listeners and lower costs. Our method involves providing listeners with transcriptions, having them highlight perceived accent differences, and implementing meticulous screening for reliability. For objective evaluation, inspired by [7, 33, 34], we propose to use pronunciation-related metrics based on distances between vowel formants and phonetic posteriorgrams (PPGs), both of which capture phonetic identities, to evaluate accent generation. We also comparatively experiment with a broad range of objective metrics and their correlation to the ranking of several systems (with hypothesised different qualities of accent generation). Our findings show these pronunciation-related metrics, alongside cosine similarity of Accent Identification (AID) or Speaker Verification (SV) embeddings, and Mel Cepstral Distortion (MCD), can be used for evaluating speech of underrepresented accents. Moreover, our findings underscore significant limitations of metrics like WER in assessing these underrepresented accents, calling for more research into fair and inclusive speech evaluation. To summarise, our contributions are:

- We propose several refinement strategies to make subjective evaluation of accent similarity more reliable.
- We propose to utilise pronunciation-related metrics, based on distances between vowel formants and PPGs, to objectively evaluate accent similarity in speech.
- Comparative experiments show that our refined XAB subjective evaluation, proposed pronunciation metrics, and several other objective metrics, can be used for evaluating underrepresented accents.

10.21437/Interspeech.2025-2283

Table 1: *Evaluation methods used in various (accent) generation systems/papers. Blue indicates objective evaluation methods.*

Category	System/Paper	Accent Similarity	Speaker Similarity	Intelligibility	Naturalness	Audio Quality	F0	Duration
Zero-Shot TTS	XTTS, 2024 [1]		XMOS, CosSim	CER	CMOS, UTMOS			
	VALL-E, 2025 [2]		CosSim	WER	CMOS			
	MaskGCT, 2025 [3]	XMOS, CosSim	CosSim	WER, CER	MOS, CMOS	MCD, FSD		
	Zhang et al., 2023 [4]	XMOS(4), AID-Acc	XMOS(4)		MOS		RMSE, PCC, DTW	
	AccentBox, 2025 [5]	XAB, CosSim	XAB, CosSim		AB			
Accented TTS	Zhou et al., 2024 [6]	XAB, XMOS, CosSim	XAB, XMOS, CosSim		AB, MOS	MCD	RMSE, PCC	FD
	Zhou et al., 2024 [7]	XAB, XBWS, PPG-KL			AB, MOS		RMSE, PCC	FD, RMSE
	CTA-TTS, 2024 [8]	XMOS, AID-Acc				MOS, BWS, MCD	Moments, DTW	MAE
	DART, 2024 [9]	XBWS	XBWS, CosSim	WER	MOS	MCD	FFE	
	RAD-MMM, 2023 [10]	XCMOS, CosSim	XCMOS	CER				
	AccentSpeech, 2022 [11]	XAB	XAB, CosSim		AB			MAE
	Deja et al., 2023 [12]	MUSHRA, AID-Acc	XMUSHRA		MUSHRA			
	Yang et al., 2023 [13]	MOS		CER	MOS	MCD		
Accent Conversion	Accentron, 2022 [14]	MOS(9)	XMOS(15)			MOS		
	Quamer et al., 2022 [15]	MOS(9)	XAB		MOS			
	Zhou et al., 2023 [16]	MOS(9)	XMOS(4)	WER, PER		MUSHRA		
	Bai et al., 2024 [17]	XBWS	XBWS		MUSHRA			
	Chen et al., 2024 [18]	MOS	XAB	WER	MOS			
	Jia et al., 2023 [19]	XMOS	XMOS		MOS			
	Jin et al., 2023 [20]	XAB	XMOS			MOS		
	Jia et al., 2024 [21]	AID-Acc, AID-Acc	CosSim		MOS, NISQA			
	MacST, 2025 [22]	MUSHRA, CosSim	CosSim	WER	MUSHRA			

X...: Reference speech is provided for similarity ratings. (`num`): A `num`-point rather than 5-point scale is used. AID-Acc: accent identification accuracy. CosSim: cosine similarity of accent/speaker embeddings. PPG-KL: KL divergence of phonetic posteriorgrams. WER/CER/PER: word/character/phone error rate. UTMOS / NISQA: predicted MOS by [23] / [24]. MCD: mel cepstral distortion. FSD: Fréchet speech distance. RMSE: root mean square error. PCC: Pearson correlation coefficient. DTW: DTW alignment cost. Moments: see [25]. FFE: F0 frame error. FD: frame difference. MAE: mean absolute error.

2. Subjective Evaluation

2.1. XAB accent similarity listening test

In our listening tests, listeners are instructed as follows: *"Listen carefully to all speech recordings below in full. Then pick the candidate speech recording that is more similar in terms of accent to the reference speech recording. Please disregard the mismatch in voice, gender, and audio quality."* We ensure that the reference and candidate speech share identical content but are delivered by two different speakers of different genders. Having identical content allows listeners to focus on detailed pairwise comparisons of pronunciation and prosody, essential for assessing accent [33]. By rating the similarity of utterances from speakers of different genders, listeners are prevented from confusing speaker similarity with accent similarity.

2.2. Highlight of perceived accent difference

Inspired by Rapid Prosody Transcription [35], we propose to additionally provide the text transcript of test utterances to listeners, and ask them to highlight the perceived accent differences. This should further guide listeners to nuanced differences related to accents. The detailed instructions are as follows: *"i) Please carefully highlight only the specific parts of the sentence that helped you decide how similar one recording's accent is to another. ii) Avoid selecting the entire sentence. Instead, try to be as precise as possible. For example, if the 'tt' sound in 'bottle' influenced your decision, highlight just 'tt'. iii) Click and drag to select or deselect parts of the text. If you make a mistake, you can click again to undo your selection. iv) Use the 'Clear All Highlights' button below to remove all highlights if you wish to start over."*

2.3. Screening listeners

To screen out invalid submissions, we devise two components. First, we include several attention-check questions with speech samples, one (A) from the same accent as in the reference (X), the other (B) from a very different accent, expecting the listeners to select A. Second, we include an open-ended accent identification question at the end, requesting listeners to identify the accent in the reference speech as specifically as possible. We expect listeners to be able to identify at least the country if not the city where the accent is commonly spoken.

3. Objective Evaluation

3.1. Commonly used objective metrics

We first extract several representative and commonly used objective metrics in evaluating TTS/VC systems. To evaluate accent/speaker similarity, we calculate the cosine similarity of accent/speaker embeddings (`CosSim`) using two AID models: 1) CommonAccent (`ComAcc`) [36], 2) `GenAID` [5], and one SV model: 3) `WavLM-base-plus-sv` [37]. To evaluate intelligibility, we calculate the WER/CER using Whisper[1] with reference to ground truth transcription. To evaluate naturalness, we use `UTMOS` [23]. To evaluate audio quality and F0 related metrics, we use the Amphion [38] evaluation pipeline to extract MCD, F0 RMSE, F0 Periodicity RMSE, and F0 Pearson Correlation Coefficient (PCC).

3.2. Vowel Formant Distance

Vowel formants are extracted as a reasonably interpretable evaluation metric. The first and second formants (F1&F2) roughly correspond to vowel height and frontness, respectively [39]. We generate phone-level alignments using Montreal Forced Aligner [40] with pretrained `english_us_arpa` acoustic model/dictionary, and extract formants using Fast Track [41] estimation at vowel midpoints. Finally, we calculate pairwise vowel formants RMSE (`VF RMSE`) using the extracted F1&F2.

3.3. Pronunciation distance based on DTW-aligned PPGs

Mainly inspired by [34] which proposes using normalised Jensen-Shannon distance between PPGs as pronunciation distance, we propose to use PPGs to measure accent-related pronunciation distance. We first extract the PPGs following [34], then additionally align the PPGs using Dynamic Time Warping (DTW) alignment, to handle mismatch in audio length, serving both TTS and VC tasks. Different from [34], we drop the normalisation, since the normalisation in their work is biased by what General American phones are similar, based on the accent-imbalance data/model. We experiment with two alignment cost functions: cosine similarity distance and Jensen-Shannon distance, treating PPGs as features and probability distributions respectively. We calculate the mean cost over all alignment steps as the final pronunciation distance of the pairwise comparison.

[1] `https://huggingface.co/openai/whisper-large-v3`

Table 2: *Results of different objective evaluation metrics and their correlation with hypothesised ranking (Hyp. Rank), †: $p > 0.05$, i.e. not statistically significant. VF: vowel formants. JS: Jensen-Shannon distance. SRCC: Spearman rank correlation coefficient.*

System	Hyp. Rank	Pronunciation			Accent Similarity		Speaker Similarity	Intelligibility		Naturalness	Audio Quality	F0		
		VF RMSE↓	PPG CosSim↑	PPG JS↓	GenAID CosSim↑	ComAcc CosSim↑	WavLM CosSim↑	Whisper WER↓	Whisper CER↓	UTMOS↑	MCD↓	F0 RMSE↓	F0 Per. RMSE↓	F0 PCC↑
`copysyn`	1	62.12	0.9744	0.0684	0.9927	0.9558	0.9443	1.737	0.724	3.732	2.800	299.1	0.01076	0.8138
`xtts`	2	177.98	0.8793	0.1936	0.9447	0.7792	0.9248	1.262	0.499	3.912	5.571	482.7	0.01296	0.6164
`corrupt30k`	3	217.96	0.8717	0.1981	0.8399	0.7215	0.8778	0.364	0.151	4.108	6.067	431.2	0.01135	0.6427
`corrupt60k`	4	212.47	0.8706	0.1992	0.8403	0.7108	0.8758	1.209	0.839	4.115	6.093	456.9	0.01045	0.6027
`corrupt90k`	5	225.88	0.8594	0.2072	0.8302	0.7089	0.8726	1.896	1.578	4.121	6.232	441.6	0.01132	0.6316
`corrupt120k`	6	224.04	0.8388	0.2241	0.8394	0.7104	0.8722	3.909	3.845	4.085	6.658	425.6	0.01077	0.6695
`corrupt150k`	7	230.67	0.8404	0.2225	0.8369	0.7095	0.8676	2.331	2.091	4.104	6.576	441.6	0.01008	0.6272
SRCC w/ Hyp. Rank		0.9286	0.9643	0.9643	0.8571	0.8929	1.0000	0.6429	0.8214	0.4643	0.9643	0.1071	0.4643	0.1786
p-value		0.0025	0.0005	0.0005	0.0137	0.0068	0.0000	0.1194†	0.0234	0.2939†	0.0005	0.8192†	0.2939†	0.7017†

4. Experiments

4.1. Subjective evaluation

Test stimuli: We use utterances 001–023 from VCTK[2], i.e. the rainbow passage and elicitation paragraph read by all speakers, as our test stimuli. Utterance 024 is arbitrarily chosen as speech prompt, for inferencing ZS-TTS systems.

Systems: In all XAB listening tests, the ground truth audio is used as the reference (X). The `copysyn` system (A) involves passing the mel-spectrogram from the ground truth to the pretrained HiFi-GAN model [42]. The `xtts` [1] system (B) generates audio from text and a speech prompt. As `xtts` lacks explicit accent control and produces noticeable accent errors, while `copysyn` merely reconstructs the audio with some vocoder artefacts, we hypothesise that listeners should prefer `copysyn` over `xtts` with reference to ground truth.

Test accent & speaker: We focus on the Edinburgh accent, typically underrepresented in English TTS/VC systems. From VCTK's five self-reported Edinburgh accent speakers (1 female, 4 male), we select the female speaker p262 as reference (X). A pilot study identified the male speaker p252 as most closely resembling the accent of the reference speaker, so we select p252 to be the test speaker in evaluated systems (A & B)

Listening tests: We conduct three primary listening tests: baseline `XAB` as detailed in Sec. 2.1, `XAB+trans` (with text transcription shown), and `XAB+trans+highlight` (highlighting perceived accent differences, as per Sec. 2.2). Each of these initial tests involves 15 listeners. We further screen out invalid submissions using criteria from Sec. 2.3 until we reach 15 *valid* submissions, leading to additional sets of test results: `XAB+screen`, `XAB+trans+screen`, and `XAB+trans+highlight+screen`. Listeners in all listening tests are recruited on `Prolific` (paid £12/hour) with self-reported normal hearing, and birthplace & current residence in the test accent region (i.e. Scotland). To avoid learning effects, there is no overlap in listeners between different listening tests.

Statistical testing: We employ one-sided t-tests with the null hypothesis that `copysyn` is not preferred over `xtts`. Due to the correlation between ratings across multiple utterances by a single listener, we calculate 95% confidence intervals (CI) across multiple listeners/submissions rather than multiple utterances to avoid underestimates [28]. We also randomly sample a subset from all valid submissions, and calculate the expected statistical significance (p-value) of these subsets with respect to the number of valid submissions.

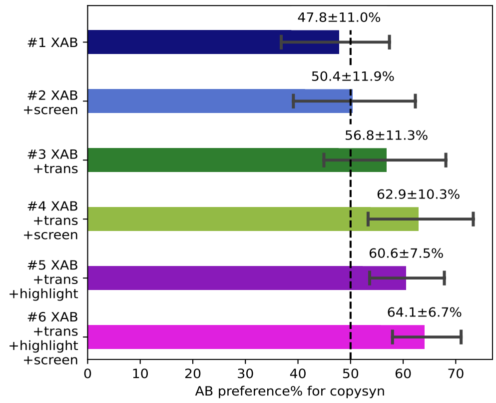
4.2. Objective evaluation

Additional systems: Objective evaluation allows for more systems to be evaluated, as it avoids costly listening tests. In addition to `copysyn` and `xtts`, we introduce a series of corrupted systems which are finetuned from original `xtts` on a single-speaker General American corpus, LJ Speech[3], with a batch size of 252 and learning rate of 5e-6. We hypothesise that longer fine-tuning leads to catastrophic forgetting, worsening non-General American accents. These corrupted systems are denoted as `corrupt30k`, `corrupt60k`, `corrupt90k`, `corrupt120k`, and `corrupt150k`, with finetuning conducted over steps ranging from 30k to 150k.

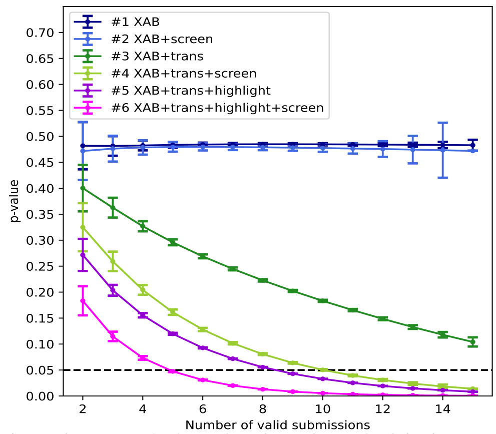
Evaluation methods: We use the same test stimuli as in subjective evaluation, with all speakers in Edinburgh accent. We calculate the mean of each objective metric listed in Section 3.

5. Results & Analysis

5.1. Subjective Evaluation



(a) AB preference% (mean±95%CI) of 15 valid submissions.



(b) p-value (mean±95%CI) of random subsets of valid submissions.

Figure 1: *Comparison of different accent similarity XAB listening tests (`copysyn` vs `xtts`), with final results shown in (a) and statistical significance investigated in (b).*

[2] `https://datashare.ed.ac.uk/handle/10283/3443`

[3] `https://keithito.com/LJ-Speech-Dataset/`

AB Preference% and statistical significance: The effects of +trans, +highlight, and +screen are shown in Fig. 1. Notably, test designs #1 and #2 do not reveal a preference for copysyn, evidenced by below/near 50% preference in Fig. 1(a), and constantly high p-values in Fig. 1(b). This suggests that naive XAB listening tests, even with listeners screening (+screen), cannot meaningfully distinguish two systems in accent similarity. Providing listeners with transcription (+trans) in tests #3 and #4 leads to a preference for copysyn, with statistically significant preference achieved in test #4, i.e. p-value dropping below 0.05. This suggests that listeners are more attentive to the accent-related differences between utterances when provided with reference transcriptions. Statistical significance can be achieved by either careful screening within 10 valid submissions or potentially by more than 15 valid submissions (see the p-values of system #3 that seem on track to drop below 0.05). Requesting listeners to complete an auxiliary task of highlighting accent-related differences (+highlight) in tests #5 and #6 leads to both higher preference for copysyn and higher statistical significance. The most effective listening test design #6 requires as few as 5 valid submissions to reach statistical significance, and reaches the highest 64.1±6.7% preference among all listening tests.

Duration, rejection rate, and cost analysis: On average, participants take 16.9±3.4 / 18.5±2.6 / 30.6±7.7 minutes to complete listening tests #2 / #4 / #6, respectively. For the duration analysis only, we exclude data from 1 participant who has spent more than two standard deviations above the mean time. Screening listeners (+screen) minimally impacts duration, therefore duration statistics of #1 / #3 / #5 listening tests are not reported. Providing listeners with transcription (+trans) does not increase the duration much; however, the auxiliary task of highlighting accent-related differences (+highlight) nearly doubles the time required to complete the test. This indicates that for accents with enough participants, the statistical significance gain of +highlight is offset by longer completion time, resulting in a similar budget cost; for accents with a limited pool of available participants, +highlight still remains an effective method. Screening listeners (+screen) leads to a rejection rate of 16.7% / 11.8% / 25.0% for listening tests #2 / #4 / #6, respectively, incurring slightly higher costs. The rejection rate, probably, has less to do with test design, but more to do with the luck of a random sample from the Prolific participants. All "rejected" listeners pass the attention-check questions, but fail the AID question: they either cannot tell what the accent of the reference speech is (answering e.g. "not sure") or misidentifying it to be a non-Scottish accent (answering e.g. "Southern England").

5.2. Objective Evaluation

The results of all objective metrics and their correlation with the hypothesised ranking of various synthesis systems are shown in Tab. 2. The three proposed pronunciation metrics, alongside accent similarity, speaker similarity, and MCD metrics, exhibit strong correlations with the hypothesised ranking (all SRCCs > 0.85 with statistical significance). However, the high correlation of speaker similarity may be influenced by factors beyond purely accent characteristics, given the catastrophic forgetting introduced by our corruption scheme, which impacts both accent and unseen speaker modelling. This interaction warrants further investigation. The unreliability of WER/CER (where copysyn is worse than xtts, and xtts worse than corrupted30k, despite $p < 0.05$ for CER) likely stems from inherent accent bias in these models. Similarly, UTMOS is possibly inconsistent due to accent biases in the training data used for MOS prediction. Therefore, we recommend against using WER/CER/UTMOS metrics in evaluating accent generation. F0 metrics correlate poorly with the hypothesized ranking across different systems, likely because the corruption affects acoustic modelling aspects other than F0. Finally, we visualise the vowel space for p252, the test speaker, in Fig. 2, with near-perfect overlapping of vowel space/distribution between ground truth and copysyn and clear mismatching between ground truth and xtts. Unlike the ground truth and copysyn, the xtts vowel space appears to avoid overlapping between vowels distributions, especially for mid/central vowels. This clearly confirms the validity of our proposed VF RMSE metric and reveals the limitations of sota ZS-TTS in modelling accents.

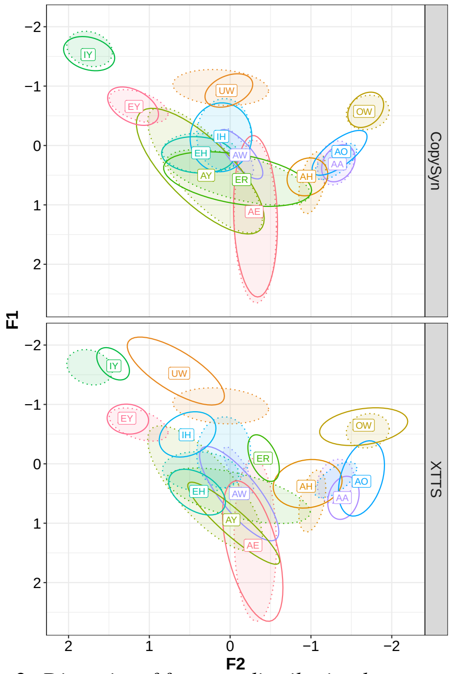


Figure 2: *Disparity of formant distribution between ground truth (shaded ellipses, dotted lines) and copysyn/xtts (hollow ellipses, solid lines) for speaker p252. Vowel symbols are ARPABET. F1/F2 axes are normalised for each speaker.*

6. Conclusions

In this work, we enhance subjective accent evaluation by adding transcription information, auxiliary highlight tasks, and meticulous screening. Additionally, we demonstrate that pronunciation-related metrics like vowel formant distances and phonetic posteriorgrams distances, serve as reliable indicators for objective accent evaluation. Based on our findings, we recommend assessing underrepresented accents using pronunciation, accent similarity, speaker similarity, and audio quality metrics while avoiding WER, CER, and UTMOS, which may not accurately capture accent characteristics. Future work will explore alternative test designs (e.g. XMOS, XCMOS), a broader range of L1/L2 accents, and additional synthesis systems. We also aim to examine the limitations of these metrics, to ensure a fair and inclusive framework for accent evaluation.

7. Acknowledgements

This work was supported in part by the UKRI AI CDT in Responsible and Trustworthy in-the-world NLP (Grant EP/Y030656/1), UKRI CDT in NLP (Grant EP/S022481/1), School of Informatics, and School of Philosophy, Psychology & Language Sciences, the University of Edinburgh.

8. References

[1] 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 *Proc. Interspeech*, 2024, pp. 4978–4982.

[2] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *IEEE/ACM TASLP*, vol. 33, pp. 705–718, 2025.

[3] 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 *ICLR*, 2025.

[4] M. Zhang, X. Zhou, Z. Wu, and H. Li, "Towards Zero-Shot Multi-Speaker Multi-Accent Text-to-Speech Synthesis," *IEEE Signal Processing Letters*, vol. 30, pp. 947–951, 2023.

[5] J. Zhong, K. Richmond, Z. Su, and S. Sun, "AccentBox: Towards High-Fidelity Zero-Shot Accent Generation," in *IEEE ICASSP*, 2025, pp. 1–5.

[6] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Multi-Scale Accent Modeling with Disentangling for Multi-Speaker Multi-Accent TTS Synthesis," *arXiv preprint arXiv:2406.10844*, 2024.

[7] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Accented Text-to-Speech Synthesis With Limited Data," *IEEE/ACM TASLP*, vol. 32, pp. 1699–1711, 2024.

[8] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable Accented Text-to-Speech Synthesis With Fine and Coarse-Grained Intensity Rendering," *IEEE/ACM TASLP*, vol. 32, pp. 2188–2201, 2024.

[9] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "DART: Disentanglement of accent and speaker representation in multispeaker text-to-speech," in *NeurIPS Workshops*, 2024.

[10] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "RAD-MMM: Multilingual Multiaccented Multispeaker Text To Speech," in *Proc. Interspeech*, 2023, pp. 626–630.

[11] 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 *IEEE ISCSLP*, 2022, pp. 76–80.

[12] K. Deja, G. Tinchev, M. Czarnowska, M. Cotescu, and J. Droppo, "Diffusion-based Accent Modelling in Speech Synthesis," in *Proc. Interspeech*, 2023, pp. 5516–5520.

[13] L.-J. Yang, C.-H. H. Yang, and J.-T. Chien, "Parameter-Efficient Learning for Text-to-Speech Accent Adaptation," in *Proc. Interspeech*, 2023, pp. 4354–4358.

[14] 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, p. 101302, 2022.

[15] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech*, 2022, pp. 4920–4924.

[16] 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.

[17] Q. Bai, S. Wang, Z. Liu, M. Zhang, W. Rao, Y. Wang, and H. Li, "Diffusion-Based Method with TTS Guidance for Foreign Accent Conversion," in *IEEE ISCSLP*, 2024, pp. 284–288.

[18] X. Chen, J. Pei, L. Xue, and M. Zhang, "Transfer the Linguistic Representations from TTS to Accent Conversion with Non-Parallel Data," in *IEEE ICASSP*, 2024, pp. 12 501–12 505.

[19] D. Jia, Q. Tian, K. Peng, J. Li, Y. Chen, M. Ma, Y. Wang, and Y. Wang, "Zero-Shot Accent Conversion using Pseudo Siamese Disentanglement Network," in *Proc. Interspeech*, 2023, pp. 5476–5480.

[20] M. Jin, P. Serai, J. Wu, A. Tjandra, V. Manohar, and Q. He, "Voice-Preserving Zero-Shot Multiple Accent Conversion," in *IEEE ICASSP*, 2023, pp. 1–5.

[21] Z. Jia, H. Xue, X. Peng, and Y. Lu, "Convert and Speak: Zero-shot Accent Conversion with Minimum Supervision," in *Proc. ACM MM*, 2024, p. 4446–4454.

[22] 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 *IEEE ICASSP*, 2025, pp. 1–5.

[23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[24] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech*, 2020, pp. 1748–1752.

[25] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2021.

[26] M. J. Munro and T. M. Derwing, "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language :earners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[27] N. Markl and C. Lai, "Everyone has an Accent," in *Proc. Interspeech*, 2023, pp. 4424–4427.

[28] J. Camp, T. Kenter, L. Finkelstein, and R. Clark, "MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors," in *Proc. Interspeech*, 2023, pp. 1090–1094.

[29] D. Wells, A. L. Aldana Blanco, C. Valentini, E. Cooper, A. Pine, J. Yamagishi, and K. Richmond, "Experimental Evaluation of MOS, AB and BWS Listening Test Designs," in *Proc. Interspeech*, 2024, pp. 2695–2699.

[30] S. Liu, M. Babel, and J. Zhu, "A Comparison of Voice Similarity through Acoustics, Human Perception and Deep Neural Network (DNN) Speaker Verification Systems," in *Proc. Interspeech*, 2024, pp. 3674–3678.

[31] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR," in *IEEE ICASSP*, 2023, pp. 1–5.

[32] S. Le Maguer, S. King, and N. Harte, "The limits of the Mean Opinion Score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[33] M. Huckvale, "ACCDIST: a Metric for Comparing Speakers' Accents," in *Proc. Interspeech*, 2004, pp. 29–32.

[34] C. Churchwell, M. Morrison, and B. Pardo, "High-Fidelity Neural Phonetic Posteriorgrams," in *IEEE ICASSP Workshops*, 2024, pp. 823–827.

[35] E. Gutierrez, P. Oplustil-Gallegos, and C. Lai, "Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm," in *ISCA SSW*, 2021, pp. 25–30.

[36] J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan, "CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice," in *Proc. Interspeech*, 2023, pp. 5291–5295.

[37] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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, Oct. 2022.

[38] X. Zhang, L. Xue, Y. Gu, Y. Wang, J. Li, H. He, C. Wang, S. Liu, X. Chen, J. Zhang, Z. Fang, H. Chen, T. Y. Tang, L. Zou, M. Wang, J. Han, K. Chen, H. Li, and Z. Wu, "Amphion: an Open-Source Audio, Music, and Speech Generation Toolkit," in *IEEE SLT*, 2024, pp. 879–884.

[39] J. C. Wells, *Accents of English: Volume 1*. Cambridge University Press, 1982, vol. 1.

[40] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[41] S. Barreda, "Fast Track: Fast (nearly) Automatic Formant-tracking Using Praat," *Linguistics Vanguard*, vol. 7, no. 1, p. 20200051, 2021.

[42] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, 2020, pp. 17 022–17 033.