ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11462045

QUANTIFYING SPEAKER EMBEDDING PHONOLOGICAL RULE INTERACTIONS IN ACCENTED SPEECH SYNTHESIS

Thanathai Lertpetchpun[1*], *Yoonjeong Lee*[1*], *Thanapat Trachu*[2], *Jihwan Lee*[1], *Tiantian Feng*[1], *Dani Byrd*[3], *Shrikanth Narayanan*[1,2,3]

[1]Signal Analysis and Interpretation Lab, University of Southern California
[2]Thomas Lord Department of Computer Science, University of Southern California
[3]Department of Linguistics, University of Southern California

ABSTRACT

Many spoken languages, including English, exhibit wide variation in dialects and accents, making accent control an important capability for flexible text-to-speech (TTS) models. Current TTS systems typically generate accented speech by conditioning on speaker embeddings associated with specific accents. While effective, this approach offers limited interpretability and controllability, as embeddings also encode traits such as timbre and emotion. In this study, we analyze the interaction between speaker embeddings and linguistically motivated phonological rules in accented speech synthesis. Using American and British English as a case study, we implement rules for flapping, rhoticity, and vowel correspondences. We propose the phoneme shift rate (PSR), a novel metric quantifying how strongly embeddings preserve or override rule-based transformations. Experiments show that combining rules with embeddings yields more authentic accents, while embeddings can attenuate or overwrite rules, revealing entanglement between accent and speaker identity. Our findings highlight rules as a lever for accent control and a framework for evaluating disentanglement in speech generation.

***Index Terms*—** Text-to-Speech, Accent Control, Phonological Rules, Speaker Embeddings, Speech Generation

1. INTRODUCTION

Accents are a defining source of variation in spoken language, and English provides a striking case. Spoken by approximately one-fifth of the world's population, yet with only one-fourth native speakers [1, 2], the English language has developed a broad range of accent varieties from both native and non-native speaker backgrounds [3]. These include regional varieties such as American, British, and Australian English, as well as global variants such as Indian English. [4, 5]. Capturing such variation has become an important goal for speech generation technologies.

Accent in TTS is typically controlled by conditioning on speaker embeddings [6, 7]. However, embeddings also encode multiple traits unrelated to the accent, such as voice timbre [8, 9], speaker emotion [10, 11, 12], and background noise [13], making accent representation opaque and difficult to control. To address this, we use linguistically motivated phonological rules as targeted probes of accent control. Our goal is not to replace data-driven models with synthesis-by-rule but to test how explicit phoneme-level transformations interact with speaker embeddings in modern TTS systems.

We focus on three well-documented phonological processes that differentiate American and British English varieties: flapping, rhoticity, and vowel correspondences. In American English, intervocalic /t/ is often realized as a flap [ɾ] in unstressed contexts (e.g., *water* → [wɑɾɚ]), while British English typically retains [t]. Rhoticity also diverges: American English is rhotic, preserving post-vocalic ɹ (e.g., *car* → [kɑɹ]), whereas most British varieties are non-rhotic, deleting or vocalizing ɹ in coda position (e.g., *car* → [kɑː]). Finally, systematic vowel correspondences arise across lexical sets. For example, *bath* is pronounced as /bæθ/ in American English but /bɑθ/ in British English, and *goat* as [goʊt] versus [gəʊt].

These selected contrasts are not modeled to capture every nuance of accent variation. Instead, we deliberately operationalize them as big-stroke transformations—salient enough to shift perceived accent without overfitting dialectal micro-variation. This makes them a useful testbed for probing the balance between embedding-driven and linguistic control, focusing less on mimicking full dialects and more on revealing where embeddings and linguistic structure interact. Our substitutions target robust cross-accent pronunciation patterns that have been consistently documented in linguistic descriptions [3, 14] and behavioral studies of accent perception [15, 16], making them ideal probes for accent strength and a principled lens on disentanglement in speech generation.

To quantify accent strength we use an accent classification model, where a higher predicted probability for a given accent indicate stronger accentedness. A caveat in this approach, however, is that predicted probabilities are inherently tied to the model's training task and the particular accent contrasts it encodes, making them task-specific rather than general measures. To complement this, we also compute accent embedding similarity. If two utterances yield embeddings that lie close together in the accent space, they are assumed to share similar accent characteristics.

While useful, neither of these measures directly captures the impact of phonological rules. Our rules are deliberately coarse, targeting the most salient cross-accent shifts, but the actual realization may be moderated by the speaker embedding. To evaluate this interaction, we introduce the phoneme shift rate (PSR), a novel metric that quantifies how much speaker embeddings preserve or overwrite the rule-driven phoneme mappings. For example, when we convert [lærəɪ] (*latter*) into a more British-like target [lɑtəə], the synthesized output may still realize [lætɚ], reflecting a partial pull back toward the American form. PSR measures such cases of gradient reinforcement or attenuation, offering a direct probe of where rule-based transformations hold versus where embeddings dominate. We

*Equal contribution

view PSR as a first step toward systematic evaluation of disentanglement in TTS, offering an interpretable way to steer accent strength that is linguistically grounded yet flexible enough for modern TTS.

The contributions of this study are: 1) We introduce a small set of linguistically guided phonological rules for highly salient dialect differences that transform accents between American and British English, offering a controlled and interpretable probe for accent in TTS. 2) We introduce the phoneme shift rate (PSR), a novel metric to evaluate how rule-based transformations are preserved or attenuated by speaker embeddings. 3) We present a systematic analysis of the interplay between phonological rules and speaker embeddings, showing how coarse, knowledge-driven constraints can reveal the degree of disentanglement in speech generation. Together, these results demonstrate that knowledge-driven rules, even at a coarse granularity, can be effective levers for accent control, providing interpretable guidance within modern data-driven TTS systems. Speech samples are available[1]. Code implementing the phonological rules is in https://github.com/linguistylee/KAtDial.

2. PHONOLOGICAL RULES

We define three sets of substitution rules to map American to British English phoneme sequences.

Flapping: Intervocalic /t/ realized as [ɾ] in American English is mapped to [t] in British English (e.g., *water*: [wɑɾɚ] → [wɑtə]).

Rhoticity: Post-vocalic /ɹ/ is retained in American English but deleted or vocalized in British English (e.g., *car*: [kɑɪ] → [kɑː])

Vowel correspondences: Systematic mappings are applied across lexical sets, such as TRAP, BATH, and GOAT (e.g., bath: /bæθ/ → /bɑθ/; goat: [goʊt] → [gəʊt]).

The rules are applied as one-to-one substitutions, with the phoneme character count strictly preserved across the American and British IPA sequences, ensuring that differences in synthesized accent strength arise only from segmental mappings and speaker embeddings. Although this work focuses only on the mapping between American to British, additional accents can be supported by defining corresponding rule sets and applying the same pipeline.

3. APPLICATION TO SPEECH GENERATION TASKS

To implement our phonological rules, we follow the pipeline illustrated in Fig 1: 1) obtain an American English phoneme sequence from normalized text using Misaki G2P [2]; 2) apply the proposed phonological rules to derive the corresponding British phoneme sequence; 3) synthesize speech outputs from both inputs.

Speech is generated using Kokoro TTS [3]. Inputs to the model include the phoneme sequence (American or British), a speaker embedding, and fixed phoneme durations. The phoneme character count and duration are strictly preserved across all conditions, ensuring that observed accent differences reflect the interaction of phonological substitution rules with speaker embeddings rather than confounds from timing or text normalization.

4. METRICS

We evaluate our approach along two dimensions: accent strength and naturalness. For indexing accent strength, we use two models: Vox-Profile [17] and Phoneme Recognition [18] (Wav2Vec2Phoneme).

[1] https://sav-eng.github.io/icassp_samples.html
[2] https://github.com/hexgrad/misaki
[3] https://github.com/hexgrad/kokoro

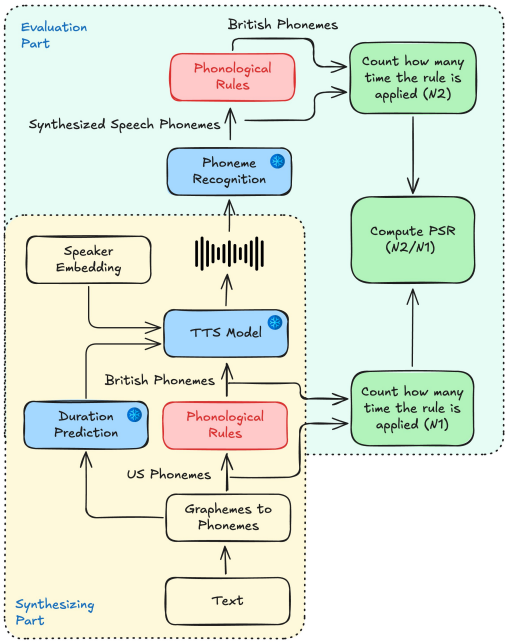

Fig. 1. Synthesis and evaluation pipeline. Inputs to Kokoro TTS are a phoneme sequence (American or British), a speaker embedding, and fixed phoneme durations. The yellow path shows rule-based phoneme substitutions prior to synthesis, and the green path shows evaluation of their effect on the synthesized accented speech.

Vox-Profile evaluates the synthesized speech with regard to the accent classifier output; while Phoneme Recognition measures how phoneme realizations shift under the influence of speaker embedding. The evaluation metrics are explicated further below.

4.1. Vox-Profile Accent Classifier

Vox-Profile [17] is a benchmark tool that predicts multiple speaker traits, including speaker accent. In this work, we use its Whisper-based broad accent classifier, trained on more than 100k unique utterances from 11 diverse data sources. Specifically, the classifier assigns each input utterance to one of three accent categories: North American English, British Isles English, and English spoken with other language backgrounds. To measure the accentedness in synthesized speech, we adopt two complementary measures, following prior work [19, 20, 21, 22]. The first is the **accent probabilities**: logits from the classifier's final layer are converted to class probabilities with the softmax function, and the average probability assigned to the target accent (North American or British Isles) is used as the index of accent strength. The second is **accent similarities:** cosine similarity is computed between the accent embedding of each synthesized speech and a group-level reference accent embedding. The reference group-level embedding is derived by averaging accent embeddings from real speech of each accent in the Vox-Profile test set.

4.2. Phoneme Shift Rate (PSR)

Speaker embeddings encode many dimensions of speaker traits, and therefore condition speech generation on them. They can influence not only timbre but also phoneme realization. To quantify this interaction with our phonological rules, we introduce a novel metric called phoneme shift rate (PSR).

Without loss of generosity, we use the example in generating the British target speech described earlier. In this case, PSR is defined as the ratio $\frac{N2}{N1}$, where $N1$ is the number of phoneme substitutions specified by our rules to convert a US phoneme sequence into a British target, and $N2$ is the number of substitutions that must still be applied when re-applying the same rules to the phoneme transcript of the synthesized British output. If the output perfectly respects the rule-based transformations, $N2 = 0$ and $PSR = 0$. If the speaker embedding completely overrides the rule inputs, $N2 = N1$ and $PSR = 1$, under the assumption that TTS and phoneme recognition have perfect accuracy.

Crucially, while our rules are modeled as categorical substitutions (e.g., mapping /t/ to [ɾ], deleting post-vocalic /ɹ/, or shifting /ae/ to ɑ), phonological categories are realized gradiently in natural speech. The same holds in synthesis: a rule may surface fully, partially, or not at all, depending on how strongly the speaker embedding shapes the acoustic outcome. For example, a British target like [lɑtɔː] (*latter*) may surface as [lætɚ], reflecting a partial drift back toward the American form.

PSR captures this continuum, quantifying how rule-based transformations interact with the gradient pressures of speaker embeddings. It thus goes beyond "rule errors" providing a diagnostic for accent control and a novel framework for analyzing how linguistic rules and embeddings compete or reinforce one another in synthesis.

4.3. Naturalness Evaluation with UTMOS

We also evaluate the naturalness of synthesized speech to ensure that applying phonological rules shifts accent strength without degrading the naturalness of the output speech. For this, we use UTMOS [23], a non-intrusive model trained to estimate human naturalness ratings (MOS) of speech without requiring subjective listening tests. UTMOS outputs a score ranging from 1 (least natural) to 5 (human-level naturalness). We compute UTMOS for each utterance and average across the set to obtain an overall naturalness measure.

5. DATASETS AND EXPERIMENTAL SETUPS

We use a pretrained TTS model, Kokoro-82M v0.19[4], a multilingual TTS model supporting eight languages including English, Japanese, and Mandarin. The model takes a speaker embedding and a phoneme sequence as input. For English, 28 preset speaker embeddings are available: 20 are derived from American English speakers, and 8 are from British speakers. Our experiments manipulate both the speaker embeddings and phonological rules to control accent strength. Synthesized utterances are based on transcripts from LibriTTS-R [24] on the train-clean-100 subset. For consistency, we use *af_heart* as the primary American speaker, *bm_fable* as the primary British speaker and assign the duration of each phoneme from *af_heart* uniformly to British speakers. In total, the synthesized speech spans 33k utterances and 55.4 hours of speech.

5.1. Rule Application Configurations

To assess the contribution of individual phonological rules, we conduct analyses under three configurations: (1) speaker embeddings alone, (2) embeddings plus a single transformation rule, and (3) embeddings plus the full set of rules. For the full set, we also perform ablation experiments in which one rule is removed at a time. This design allows us to isolate the effect of each rule on accent strength while also evaluating the combined benefit of data-driven and rule-driven approaches.

[4]The model is available at `https://huggingface.co/hexgrad/Kokoro-82M`.

6. RESULTS AND DISCUSSION

We present results from three sets of experiments. First, we fix the speaker embedding input to the TTS and vary the number of applied rules to assess how each rule contributes to accented speech. Second, we analyze synthesized phoneme sequences to see how speaker embeddings interact with or override rule-based transformations. Finally, we examine the role of individual speaker embeddings in shaping accented output.

6.1. Effects of Each Phonological Rule on Speech Synthesis

Table 1. Results on accented speech synthesis under different phonological transformation rules. All rules applied in this experiment are crafted for British English. We also present results conditioned on speaker embeddings from both accents to study how strongly these speaker embeddings adhere to or override these rule-driven transformations. "NA" denotes North American. "B" denotes British. "Spk Emb" denotes speaker embedding.

	UTMOS	**Accent Prob**		**Accent Sim**		**PSR↓**
		NA ↓	B ↑	NA ↓	B ↑	
NA Spk Emb	4.43	86.5	3.79	0.85	-0.05	0.856
+ All	4.42	**58.8**	**17.3**	**0.74**	**0.21**	**0.827**
B Spk Emb	3.74	17.6	67.8	0.33	0.67	0.775
+ Flapping	3.74	17.3	68.5	0.32	0.67	0.749
+ Rhoticity	3.73	9.5	68.8	0.14	0.78	0.739
+ Vowel	3.73	9.8	77.8	0.18	0.78	0.693
+ All	3.72	**5.3**	**78.4**	**0.03**	**0.85**	**0.628**
- Flapping	3.72	5.3	78.1	0.03	0.85	0.646
- Rhoticity	3.73	9.5	78.1	0.18	0.78	0.670
- Vowel	3.73	9.4	69.4	0.14	0.79	0.716

Table 1 summarizes naturalness (UTMOS), accent strength (probabilities and similarities for North American vs. British), and phoneme shift rate (PSR) under different rule configurations. The column 'Accent Prob' with '(NA/B)' indicates the average probabilities assigned by the Vox-Profile accent classifier to North American or British speech. The column 'Accent Sim' with '(NA/B)' denotes the average cosine similarity between synthesized utterances and group-level reference embeddings for each accent. PSR measures how often rule-driven phoneme substitutions persist in the output, with lower values reflecting stronger rule preservation.

Across conditions, **naturalness** is stable. UTMOS remains around 4.4 for North American and 3.7 for British settings, whether or not rules are applied. This indicates that integrating phonological rules alongside speaker embeddings does not degrade the naturalness of the synthesized speech. Consistently lower UTMOS scores across the board for the British-accented speech likely reflects bias in the predictor, which was trained on data with heavier North American representation [20] rather than any actual degradation.

Second, **accent probabilities** show clear rule effects. With a North American speaker embedding, the baseline system yields an 86.5% probability of the North American accent. Adding all three phonological transformation rules drops this to 58.8% and raises the British probability to 17.3%, showing that rules shift the classifier's decision away from North American toward British. Conversely, with a British speaker embedding, the baseline system al-

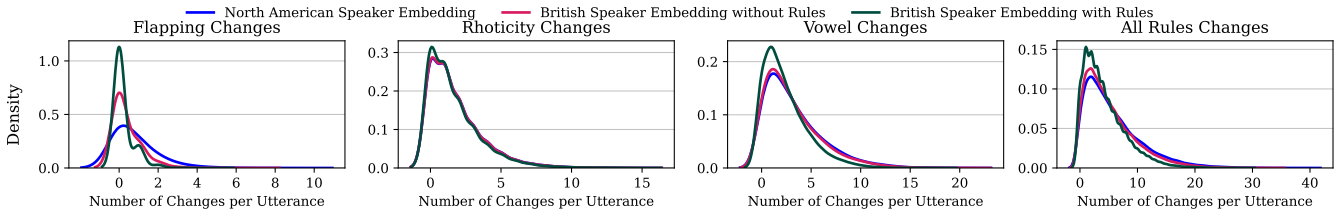

Fig. 2. Kernel Density Estimation (KDE) of histograms on the number of changes of the rules.

ready achieves 67.8% British probability, but this increases to 78.4% when all British English substitution rules are applied.

Third, **accent similarities** follow the same trend. With North American embeddings, British similarity increases from -0.05 to 0.21 when rules are applied, pulling synthesized speech closer to the British reference. With British embeddings, similarity to the British reference rises from 0.67 to 0.85.

The **role of individual rules** emerges when applied selectively with British embeddings. Vowel correspondences drive the largest gains, raising British probability to 77.8% and reducing PSR to 0.693. Rhoticity increases similarity to the British embedding space (0.78) even when classifier probabilities remain modest. Flapping alone has minimal impact but contributes additively when combined with other rules. Ablations confirm vowel correspondences as the single most influential factor.

Finally, PSR captures how often rule effects survive in output phonemes. Lower values indicate stronger preservation of rule-driven changes. With British embeddings, PSR falls from 0.775 without rules to 0.628 with all rules, showing that rules measurably shape outputs even when embeddings exert a strong pull in the opposite direction.

6.2. How Do Rules Affect Phonemes of the Synthesized Speech?

Table 2. Total number of times we apply the rule (in thousands). $N1$ and $N2$ are the notation from Fig 1 in the green highlight. NA and B denote North American and British accents.

	Flapping	Rhoticity	Vowel	All Rules
NA ($N1$)	12.8	83.5	125.1	221.4
NA ($N2$)	25.3	57.9	106.3	189.5
B w/o rules ($N2$)	12.3	57.4	101.7	171.5
B w/ rules ($N2$)	**6.7**	**53.7**	**78.5**	**139.0**

To analyze rule retention at the phoneme level, we compare the number of substitutions ($N1$) expected with the number observed in recognized outputs ($N2$), as summarized in Table 2. Ideally, if all substitutions survived, $N2$ would be zero after applying rules. It is often observed that speaker embeddings overwrite some substitutions, pushing outputs back toward their accent biases.

The results show that vowel correspondences account for the largest number of changes, consistent with their strong effect in Table1. For flapping, $N2$ increases under North American embeddings, suggesting that the G2P baseline phonemes were not fully American and the embedding reinforced flapping. For rhoticity and vowels, $N2$ is significantly lower than $N1$ under North American embeddings, demonstrating that one speaker embedding can simultaneously represent different accent tendencies depending on the phoneme context.

Fig 2 visualized these effects using KDE histograms; rule application increases the skewness toward smaller numbers of changes, indicating the effectiveness of the rules. However, distributions remain broad, highlighting how speaker embedding can override or dilute specific rule effects.

6.3. On the Role of Individual Speaker Embeddings in Accented Speech Synthesis.

Table 3. Results on accented speech synthesis under different speaker embeddings. -R suffix denotes the addition of rules. NA and B denote North American and British accents.

	Accent Prob		Accent Sim		PSR ↓
	NA ↓	B ↑	NA ↓	B ↑	
Isabella	2.7	85.3	0.03	0.88	0.633
Isabella-R	**1.1**	**90.1**	**-0.07**	**0.92**	**0.481**$_{-15.2\%}$
Lily	2.0	89.3	-0.01	0.91	0.660
Lily-R	**0.9**	**91.6**	**-0.12**	**0.93**	**0.494**$_{-16.6\%}$
Fable	17.6	67.8	0.33	0.67	0.775
Fable-R	**5.7**	**78.4**	**0.03**	**0.86**	**0.628**$_{-14.7\%}$
Daniel	4.7	89.8	0.07	0.86	0.706
Daniel-R	**1.5**	**93.2**	**-0.07**	**0.93**	**0.543**$_{-16.3\%}$

Table 3 examines how individual speaker embeddings interact with rule application. Results are shown for Kokoro TTS voices[5]. Across voices, rules consistently strengthen accent probabilities. For example, the Fable embedding alone yields 67.8% British probability; with rules, this rises to 78.4%. Similarity to the British reference also improves, and PSR decreases from 0.775 to 0.628. For Isabella and Lily, rule application produces 15-17% absolute reductions in PSR, confirming stronger preservation of rule-driven transformations.

The degree of interaction varies across embeddings. Some embeddings like Daniel, already produce high British probabilities ($\sim$ 89.8%) but still show improvements when rules are applied (to 93.2%). Others, like Fable, rely more heavily on rule guidance. This variation suggests that embeddings encode accent characteristics with differing levels of entanglement, which rules can either reinforce or correct.

7. CONCLUSION AND FUTURE WORK

Taken together, our results show that phonological rules are a coarse but effective lever for accent control. They preserve naturalness, strengthen accent attributes, and reveal how speaker embeddings can reinforce or override explicit transformations. By offering interpretable, linguistically grounded modifications, rules complement speaker embeddings and highlight future directions for designing embeddings that disentangle accent from other traits while remaining responsive to rule-based inputs. However, as this study relies on automated metrics and is subject to the noise of the specific phoneme recognition model used, future work will incorporate a broader array of recognition architectures and human evaluations.

[5]https://huggingface.co/hexgrad/Kokoro-82M/blob/main/VOICES.md

8. 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.

9. REFERENCES

[1] David Crystal, *English as a Global Language*, Cambridge University Press, 2 edition, 2003.

[2] Ethnologue, "English language profile," `https://www.ethnologue.com/language/eng`, 2024, Accessed: 2025-09-15.

[3] J. C. Wells, *Accents of English*, Cambridge University Press, 1982.

[4] Braj B. Kachru, *The Alchemy of English: The Spread, Functions, and Models of Non-Native Englishes*, University of Illinois Press, 1990.

[5] Edgar W. Schneider, *Postcolonial English: Varieties around the World*, Cambridge University Press, 2007.

[6] Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li, "Multi-scale accent modeling and disentangling for multi-speaker multi-accent text-to-speech synthesis," *arXiv preprint arXiv:2406.10844*, 2024.

[7] Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li, "Accented text-to-speech synthesis with limited data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1699–1711, 2024.

[8] Chenpeng Du, Yiwei Guo, Xie Chen, and Kai Yu, "Speaker adaptive text-to-speech with timbre-normalized vector-quantized feature," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3446–3456, 2023.

[9] Joun Yeop Lee, Jae-Sung Bae, Seongkyu Mun, Jihwan Lee, Ji-Hyun Lee, Hoon-Young Cho, and Chanwoo Kim, "Hierarchical timbre-cadence speaker encoder for zero-shot speech synthesis," in *Interspeech 2023*, 2023, pp. 4334–4338.

[10] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee, "Diemo-tts: Disentangled emotion representations via self-supervised distillation for cross-speaker emotion transfer in text-to-speech," in *INTERSPEECH*, 2025.

[11] Thanathai Lertpetchpun and Ekapol Chuangsuwanich, "Instance-based temporal normalization for speaker verification," in *Proc. INTERSPEECH*, 2023, vol. 2023, pp. 3172–3176.

[12] Heejin Choi, Jae-Sung Bae, Joun Yeop Lee, Seongkyu Mun, Jihwan Lee, Hoon-Young Cho, and Chanwoo Kim, "Mels-tts : Multi-emotion multi-lingual multi-speaker text-to-speech system via disentangled style tokens," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12682–12686.

[13] Kenichi Fujita, Hiroshi Sato, Takanori Ashihara, Hiroki Kanagawa, Marc Delcroix, Takafumi Moriya, and Yusuke Ijima, "Noise-robust zero-shot text-to-speech synthesis conditioned on self-supervised speech-representation model with adapters," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11471–11475.

[14] Peter Trudgill, *The Dialects of England*, Blackwell, 1999.

[15] James E. Flege, "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, Winifred Strange, Ed., pp. 233–277. York Press, 1995.

[16] Cynthia G. Clopper and David B. Pisoni, "Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization," *Language Variation and Change*, vol. 16, no. 1, pp. 31–48, 2004.

[17] Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, et al., "Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits," *arXiv preprint arXiv:2505.14648*, 2025.

[18] Qiantong Xu, Alexei Baevski, and Michael Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.

[19] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu, "Maskgct: Zero-shot text-to-speech with masked generative codec transformer," in *ICLR*, 2025.

[20] Jinzuomu Zhong, Suyuan Liu, Dan Wells, and Korin Richmond, "Pairwise evaluation of accent similarity in speech synthesis," *arXiv preprint arXiv:2505.14410*, 2025.

[21] Jinzuomu Zhong, Korin Richmond, Zhiba Su, and Siqi Sun, "Accentbox: Towards high-fidelity zero-shot accent generation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[22] Sho Inoue, Shuai Wang, Wanxing Wang, Pengcheng Zhu, Mengxiao Bi, and Haizhou 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)*. IEEE, 2025, pp. 1–5.

[23] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *Interspeech2022*, 2022.

[24] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.