IDENTITY LEAKAGE THROUGH ACCENT CUES IN VOICE ANONYMISATION

Rayane Bakari[1,2]*, Olivier Le Blouch*[1]*, Nicolas Gengembre*[1]*, Nicholas Evans*[2]*, Michele Panariello*[2]

[1] Orange Innovation, France
[2] EURECOM, Sophia Antipolis, France
{mohamedrayane.bakari, olivier.leblouch, nicolas.gengembre}@orange.com, evans@eurecom.fr

ABSTRACT

Voice anonymisation is used to conceal voice identity while preserving linguistic content. Even if anonymisation seems strong, non-timbral cues such as accent that remain post-anonymisation can help re-identification and reveal sensitive socio-demographic traits. We report a study of residual accent information involving multiple anonymisation systems. We highlight the role of accent using speaker verification, accent verification, and accent classification using a set of embeddings focusing on timbral, non-timbral and accent-related information and show the extent to which related cues facilitate reidentification post anonymisation. Results show that, while some systems are robust to reidentification attempts using accent cues, others leave residual, speaker-dependent, accent-related cues which can be used to reveal the voice identity. We also highlight accent-dependent variation in anonymisation performance, raising fairness concerns, and show that a system with character-level conditioning can help obfuscate identity-revealing accent cues, reducing accent-identification accuracy by 68% on average and improving overall anonymisation performance by 11% relative.

Index Terms— accent identification, anonymisation, voice privacy, fairness

1. INTRODUCTION

Voice anonymisation involves the processing of a speech signal to conceal the voice identity (privacy) while preserving linguistic content and paralinguistic attributes (utility). The VoicePrivacy Challenge (VPC) [1] proposed standardized evaluation protocols and metrics to guide anonymisation system development. Privacy is typically measured using automatic speaker verification (ASV), with performance estimated via the Equal Error Rate (EER), while utility measures depend on the application, including estimates of intelligibility and the speaker's emotional state, as for the most recent VPC held in 2024 [2].

The VoicePrivacy Attacker Challenge [3] was launched to study vulnerabilities to attacks against anonymisation, to stimulate progress in evaluation, and hence to improve anonymisation robustness. Attack techniques reported in [4, 5, 6, 7] show that previous approaches to evaluation result in exaggerated estimates of anonymisation performance [8]. In this paper, we show how the results of such attack studies can be used directly to improve robustness.

Most anonymisation systems [9, 10, 11, 12] modify speaker identity through the manipulation or substitution of timbral cues [13]. However, attacks that use non-timbral cues such as prosody, rhythm, speaking style, and accent can reveal the voice identity even after anonymisation [13]. Recent work has also shown that context-dependent duration features compromise privacy [14], leading to duration-based anonymisation strategies [15]. The study presented in this paper investigates the influence of a speaker's accent. Since accent conveys socio-demographic information relating to the speaker's regional, ethnicity, and educational profile [16], attackers can utilise accent-related cues that persist post-anonymisation to facilitate re-identification or speaker profiling.

The use of accent-related cues has been largely overlooked in previously reported studies, while our previous work [13] suggests that the leakage of non-timbral cues, including accent, is a consistent vulnerability in some state-of-the-art anonymisation techniques. Real-world anonymisation use case scenarios further motivate our focus; previous work [17] shows that accent can be a primary identifier in small speaker pools or high-stakes contexts.

The role and influence of accent might have been observed previously. Recent work [18] studied the variability in anonymisation performance by categorising speakers as 'sheeps', 'goats', 'lambs', or 'wolves' [19] based on ASV performance and human perception. Lambs are more easily re-identified, suggesting that, for these speakers, there remains some particular cues post-anonymisation that contribute to identity leakage. In our previous work [13], we argue that, with anonymisation typically focusing upon the manipulation or substitution of timbral cues, those remaining post anonymisation are likely to be of non-timbral origin. We furthermore speculated that these cues are accent-related. The work reported in this paper is designed to confirm this hypothesis.

We investigate the extent to which accent-related cues survive anonymisation and the role that accent plays in identity leakage. Specifically, we examine whether current anonymisation systems effectively preserve or obscure accent information, whether certain accents are systematically more or less protected, potentially raising fairness concerns and accounting for observations in [18], and the potential to improve anonymisation robustness by reducing accent leakage. Our contributions are as follows:

- we demonstrate that accent-related cues that persist post anonymisation contribute to identity leakage;
- we propose a simple criterion to qualify the accent preservation in anonymisation systems;
- we show biases in anonymisation performance for speakers of particular accents;
- we investigate an approach to mitigate accent-related identity leakage;
- we advocate for fairness-aware anonymisation evaluation to promote consistent protection across accents.

2. METHODOLOGY

2.1. Anonymisation systems

To address the leakage of identity from residual accent cues, we evaluate a set of baselines and competing systems submitted to the 2024 VPC [2]. They include the B3-B5 baselines and the T8-5, T10-2, T12-5 and T25-1 participant systems defined for the VPC attacker challenge [3].

Baseline B3 combines automatic speech recognition (ASR) and text-to-speech (TTS) [20]. B4 uses a neural audio codec (EnCodec) [21], while B5 exploits vector quantization to better disentangle linguistic and voice features [22]. B4* is built upon B4 by replacing the EnCodec decoder with the Vocos vocoder [23], which is trained with character-level conditioning as described in [24].

Participant systems employ specific strategies to preserve expressiveness in anonymised speech. T8-5 [9] randomly selects one of two methods for each utterance: a k-nearest neighbors voice conversion approach (kNN-VC [25]) or a B3-like ASR+TTS approach. T10-2 [10] distills content and expressiveness features within a quantized neural audio codec framework. T12-5 [11] builds upon B5 with additional pitch control. Finally, T25-1 [12] combines vector quantization and style tokens to preserve expressiveness.

2.2. Evaluation Protocols

We adopt three complementary evaluation tasks: *Accent identification (AID)* is performed using the GenAID classifier [26] to assign directly anonymised utterances to one of 13 accent classes. The Weighted Average Recall (WAR), computed both before and after anonymisation, provides a measure of how much accent information is preserved. The WAR is computed according to:

$$R_i = \frac{TP_i}{N_i}, \quad \text{WAR} = \frac{\sum_{i=1}^{13} N_i R_i}{\sum_{i=1}^{13} N_i} \tag{1}$$

where TP_i and N_i are the number of correct classifications and total utterances for accent i and where R_i is the percentage of utterances in a given accent that are correctly classified.

Theoretically, the WAR can be seen as a simple criterion to qualify the accent-removal power of an anonymisation system. To illustrate, different cases may result depending on anonymisation architectures, including a WAR of 100% for a system that perfectly preserves accents. Conversely, what would be the result of a system that perfectly conceals accents? To assess the quality of accent management in anonymisation systems, we propose establishing a theoretical target value of 100 divided by the number of identifiable accents (7.69% for 13 accents). This value describes, by instance, two acceptable cases of 'perfect' accent anonymisation: all source accents converted either to the same accent (e.g. neutral English) or to random accents.

Speaker Verification (SV) is the well-known protocol used to evaluate the performances of speaker identification systems but transposed to anonymisation performance using genuine (same speaker) versus impostor (different speaker) trials. It is performed on embeddings extracted using GenAID, E-VPC, or W-NT, which are used to compute the cosine similarity between utterance pairs. Positive pairs contain speech from the same speaker, while negative pairs contain speech from different speakers.

Accent verification (AV) performs an SV-like evaluation but with a focus on accents. As in SV, embeddings are extracted and used to compute the cosine similarity between utterance pairs. But positive pairs contain speech in *the same accent*, while negative pairs contain speech in *different accents*.

SV and AV performances are estimated using the EER with trials balanced across speakers and accents.

2.3. Datasets and Models

We report experiments performed using the *test-unseen* split of the Common Accent dataset [27], as defined in [26] (hereafter referred to as the COMMON ACCENT). It includes utterances in 13 English accents collected from 10 speakers per accent with 10 utterances per speaker (1,300 total). Each utterance is accent-labelled, enabling controlled construction of accent verification trials. For speaker verification experiments, we also use the Libri-test set as defined for the 2024 VPC [2].

To quantify accent leakage, we leverage an accent identification (AID) model, which aims to classify accent using speech characteristics [28]. In particular, we use GenAID [26][1], which is trained using adversarial learning to suppress speaker-specific information and produce speaker-agnostic accent embeddings. The GenAID model achieves state-of-the-art performance for unseen speakers [26]. We use three off-the-shelf models to probe accent information, all of which produce embeddings: **GenAID**, used for AID, provides accent-sensitive embeddings; **E-VPC**, a variant of the ECAPA-TDNN model [29, 13] which is sensitive to timbral cues; a **W-NT** model [30] designed to capture non-timbral cues. The use of embeddings produced by each model allows us to evaluate and compare accent leakage along accent-specific, timbral, and non-timbral dimensions.

2.4. Attacker scenarios

We consider two standard VPC attack scenarios [1]. For the *ignorant* (**I**) scenario, the adversary compares original enrolment utterances with anonymised trial utterances without any compensation for anonymisation. For the *lazy-informed* (**L**), the adversary compensates for the use of anonymisation by anonymising the enrolment utterances (using the same anonymisation system but without knowing its exact configuration) so that comparisons are made between anonymised enrolment and anonymised trial utterances [2]. These scenarios are particularly informative because they test model robustness without requiring retraining on anonymised data,

3. RESULTS

We evaluate identity leakage through accent cues in anonymised speech using E-VPC, W-NT, and GenAID. Our analysis considers both objective privacy in terms of EER for speaker verification (SV) and accent verification (AV). In addition, we estimate the persistence of accent cues, quantified through the WAR, which reflects the degree to which anonymisation systems obfuscate accent-related information.

3.1. Speaker Verification

SV results are shown in Table 1. Results confirm that the use of E-VPC embeddings consistently yields the highest EERs for most anonymisation systems and both attack scenarios. However, higher EERs do not necessarily indicate strong anonymisation. E-VPC embeddings capture predominantly timbral cues which are less

[1] Available at `https://github.com/jzmzhong/GenAID`

Table 1. Speaker verification EER (%) under Ignorant (I) and Lazy-informed (L) scenarios for the Libri-test dataset. Lower values indicate weaker anonymisation.

Model	E-VPC		W-NT		GenAID	
	I	L	I	L	I	L
B3	47.4	45.7	38.2	34.7	46.3	44.1
B4	47.8	49.5	34.2	32.0	44.6	44.2
B4*	49.1	49.8	35.4	38.6	44.0	44.4
B5	49.1	48.7	42.5	42.0	46.8	48.3
T8-5	45.5	48.2	32.8	36.3	46.1	44.9
T10-2	36.2	35.9	23.6	22.1	40.9	38.6
T12-5	49.1	51.1	44.4	43.2	45.5	47.1
T25-1	48.8	49.5	44.7	44.1	45.2	47.8

reliable after anonymisation. Non-timbral information, such as prosody, rhythm, accent, and speaking style, which are generally less impacted by anonymisation, are then a more reliable source of information using which the original speaker can be re-identified. This information is better captured using W-NT embeddings, resulting in lower EERs, indicating weaker anonymisation. EERs for GenAID embeddings are generally lower than those for E-VPC, but higher than those for W-NT. These results show that attacks against anonymisation systems using accent cues instead of timbral cues can be more effective and that, for some systems, there is little difference in SV performance using either accent or non-timbral cues. In these cases, accent cues appear to be the dominant source of reliable speaker information that remains after anonymisation.

Across anonymisation systems, differences in EERs for I and L attack scenarios are generally modest (1-3%). As expected, EERs are generally lower under the L attack scenario, particularly for W-NT and GenAID embeddings. For example, T10-2 and GenAID embeddings show a notable reduction from the I to L scenario. Interestingly, some systems exhibit the opposite trend. B5 with GenAID embeddings yields a higher EER under the L than the I scenario. This suggests that B5 is slightly more accent-resistant in its anonymisation space than when comparing anonymised speech with original speech. In other words, this system still has room for improvement in terms of anonymisation, but this is not related to accents.

3.2. Accent Verification

AV results are shown in Table 2. With no access to T8-5, T10-2, T12-5 and T25-1 models or codes with which to anonymise the COMMON ACCENT dataset, we show results for baseline systems only. With only one exception, EERs for B3, B4* and B5 are higher than those for B4 under all attack scenarios, indicating stronger suppression of accent cues. For B4, the EER for GenAID and the L scenario is lower than that for E-VPC, reinforcing our finding that accent information persists post-anonymisation and that, by focusing on timbral cues, E-VPC gives a potentially misleading interpretation of anonymisation performance.

Of particular interest is the substantially lower EER for B4 and GenAID under the L attack scenario than for the I scenario. This observation suggests that, under the L scenario, B4 tends to map homogeneous source accent clusters to homogeneous anonymized accent clusters. In other words, speakers with similar original accent are often projected similarly in anonymisation space, making identification easier for an attacker with partial system knowledge. This also explains why B5 leads to a lower EER in the L scenario than in the I scenario, contrary to what was observed for SV experiments. These

Table 2. Accent verification EER (%) under Ignorant (I) and Lazy-informed (L) scenarios on COMMON ACCENT. Higher values indicate stronger anonymisation. Systems not available for anonymization are excluded.

Model	E-VPC		W-NT		GenAID	
	I	L	I	L	I	L
B3	50.5	53.7	47.5	49.7	51.5	50.6
B4	48.7	49.9	38.7	40.8	48.5	38.8
B4*	50.6	53.8	40.5	44.9	52.1	43.4
B5	50.5	49.7	46.3	48.7	50.2	49.9

results demonstrate that accent cues can persist after anonymisation, particularly in the case of B4, compromising privacy and highlighting the need to account for accent features in anonymisation system design.

3.3. Accent Classification

Accent classification results using the GenAID classifier are presented in Table 3 in terms of the WAR and individually in terms of recall, for each of the 13 accents. The first line of results is for original speech (no anonymisation). Key observations include:

1. **Degraded accent classification**: the WAR drops from 57% for original speech to 8% for B5 and 28% for B4. This finding confirms the variability in accent obfuscation for different systems. Obfuscation is highest for B5 for which the WAR reaches the theoretical 'perfect' value of 1/13. In contrast, with a WAR of 28%, B4 leaves residual accent information.
2. **Accent-dependent leakage**: Certain accents, such as English (ENG) or South Asian (SA), remain relatively identifiable post-anonymisation (e.g., 62% recall for ENG using B4, and 21% for SA), indicating that these accents are either more resistant to anonymisation or biased in their training to generate dominant accents. Conversely, accents such as Hong Kong (HK) and Malaysian (MYS) are effectively obfuscated, with zero recall for B5 and similarly low recall for B3.
3. **System-dependent patterns**: B5 consistently achieves the lowest per-accent recall across most groups, indicating robust anonymisation of accent features. B4 leaves some accents partially exposed, particularly those that are subjectively similar, such as US, ENG, AUS, and CAN, highlighting a degree of accent bias in anonymisation effectiveness. B3 shows overall strong accent suppression (10% WAR) but exhibits slightly higher recall for US (32%) and CAN (67%) accents. This bias stems from the mapping by B3 and B4 of source accents toward US/CAN accents, reflecting the distribution of their training data.

To ensure faireness, future anonymisation methods should be designed to render all accents equally difficult to identify. While the mapping of all speech to a single neutral accent (e.g., US or ENG) could, in principle, provide strong protection, anonymisation solutions should balance privacy with fairness.

3.4. Enhanced Accent Suppression with B4*

To address the leakage of voice identity through accent cues, especially in the case of voice conversion-based systems such as B4, we investigated means to improve the obfuscation of accent-related

Table 3. Accent identification results for the **COMMON ACCENT** dataset before and after anonymisation. Recall as defined in 1, is reported for each accent. Lower WAR indicates better accent suppression. B2–B5 denote anonymisation systems applied to the dataset.

Dataset	WAR	HK	SA	ENG	SCO	US	SAF	PH	MYS	AUS	IRL	CAN	SG	NZ
Original	56.77	44.0	88.0	78.0	82.0	20.0	57.0	80.0	15.0	81.0	52.0	76.0	15.0	50.0
B5	**7.69**	0.0	0.0	0.0	0.0	24.0	0.0	10.0	0.0	7.0	0.0	56.0	0.0	3.0
B4	27.85	16.0	21.0	62.0	41.0	46.0	14.0	34.0	2.0	47.0	12.0	53.0	2.0	14.0
B4*	18.39	3.0	5.0	25.0	25.0	33.0	4.0	39.0	1.0	42.0	5.0	46.0	1.0	10.0
B3	9.77	2.0	0.0	4.0	2.0	32.0	0.0	13.0	0.0	4.0	1.0	67.0	0.0	2.0

Accent abbreviations :
HK = Hong Kong, SA = South Asian, ENG = English, SCO = Scottish, US = American, SAF = Southern African, PH = Filipino, MYS = Malaysian, AUS = Australian, IRL = Irish, CAN = Canadian, SG = Singaporean, NZ = New Zealand.

information. We propose a new system B4*, an improved version of B4, to improve accent obfuscation. Recall and WAR results for B4* shown in Table 3 show consistently better accent obfuscation compared to B4. Although the WAR for B4* (18%) remains far above the theoretical minimum of 7.69%, this represents a notable improvement. This is reflected by gains observed for all accents preserved using B4, except for PH and AUS. These improvements are also supported by a 5% increase in EER (11% relative) under the L scenario in Table 2, confirming that B4* enhances speaker anonymisation through stronger accent obfuscation. These results demonstrate that leveraging textual conditioning, as in B4* described in 2.1, can significantly improve anonymisation performance by explicitly targeting accent-related information.

4. DISCUSSION

Our experiments show that current anonymisation systems effectively suppress timbral cues, but accent, a critical non-timbral feature, remains partially exposed. This leakage is both system- and accent-dependent. For instance, B5 achieves the highest overall WAR, reducing recall to zero for most accents, whereas B4 leaves accents such as ENG, US, AUS, and CAN partially identifiable. Accents like US and CAN consistently remain more detectable across systems, while HK and MYS are effectively obfuscated, reflecting the influence of training data distributions that include mostly US-accented speakers. These results indicate that anonymisation effectiveness is uneven and may introduce residual accent bias.

The comparison between B4 and its variant B4* highlights the impact of system design. B4* improves accent suppression by conditioning the vocoder on character-level transcriptions, enforcing more canonical pronunciation and reducing speaker-specific traits. While this design improves recall across most accent groups, the observed gains may result from a combination of vocoder replacement and character-level conditioning. Future ablations are needed to disentangle the contribution of each factor. Similarly, speech-to-text-to-speech-based systems like B3 remove most accentual information by resynthesizing speech from text, inherently eliminating accent cues. These observations suggest that leveraging textual conditioning or transcription-based approaches is a promising strategy to mitigate residual accent information.

The persistence of accent-related cues has privacy implications. Accents resistant to anonymisation could allow attackers to re-identify speakers or infer socio-demographic traits. This is especially relevant for distinctive or minority accents, where residual accent information may reveal identity more easily. Users belonging to more detectable accent groups may perceive themselves as less protected, highlighting equity concerns in privacy guarantees. Our findings underscore the need for evaluations that consider both objective metrics and subjective assessments by human listeners to fully capture privacy risks.

From a system design perspective, B5 emerges as the strongest system for anonymising accent-related cues, but it fails to provide fair protection across all accents, particularly for North American varieties (e.g., US, CAN). The partial confusion observed among these closely related accents lowers direct re-identification risk but still reflects a systematic bias. In contrast, the substantial residual leakage across multiple accent groups in B4 demonstrates the importance of integrating accent-aware anonymisation techniques. Stronger accent suppression may come at a cost. For example, resynthesis-based approaches or character-conditioned vocoders could reduce naturalness. Quantifying this trade-off with metrics such as Word Error Rate (WER) or Mean Opinion Score (MOS) is essential to balance privacy and utility. A careful evaluation framework can help identify methods that suppress accent cues while maintaining intelligibility and user satisfaction. More broadly, our findings motivate fairness-aware anonymisation: systems should provide consistent privacy protection across accents to mitigate bias.

5. CONCLUSIONS

This work investigates how accent cues contribute to identity leakage in voice anonymisation, using a multi-perspective evaluation that captures accent-specific, timbral, and non-timbral information. Our experiments show that accent information persists after anonymisation in some contexts, particularly when evaluated with GenAID, which outperforms a speaker-identification-focused model (E-VPC from VPC 2024) in detecting residual speaker-specific cues. Additionally, we show that anonymisation systems rarely afford uniform protection to speakers of different accents, highlighting fairness concerns. As a first step toward accent-aware anonymisation, we explored B4*, a technique which improves accent obfuscation by enforcing more canonical pronunciation, confirming that privacy can be enhanced at the accent level. Future work could explore accent-agnostic anonymisation strategies, expand evaluations to cover a broader set of sociolectal traits, and integrate accent verification as a standard metric in privacy assessments. By addressing these gaps, anonymisation systems can better balance privacy, fairness, and usability across diverse populations.

The research reported here was partly supported by the ANR-23-CE23-0018 EVA project.

6. REFERENCES

[1] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al., "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, pp. 101362, 2022.

[2] Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco, "The voiceprivacy 2024 challenge evaluation plan," 2024.

[3] Natalia Tomashenko, Xiaoxiao Miao, Emmanuel Vincent, and Junichi Yamagishi, "The first voiceprivacy attacker challenge evaluation plan," *arXiv preprint arXiv:2410.07428*, 2024.

[4] Yanzhe Zhang, Zhonghao Bi, Feiyang Xiao, Xuefeng Yang, Qiaoxi Zhu, and Jian Guan, "Attacking voice anonymization systems with augmented feature and speaker identity difference," in *ICASSP*, 2025, pp. 1–2.

[5] Xiang Lyu, Yuxuan Wang, Tianyu Zhao, and Huadai Liu, "Fast adaptation of pretrained speaker verification system for source speaker tracking," in *ICASSP*, 2025, pp. 1–2.

[6] Candy Olivia Mawalim, Aulia Adila, and Masashi Unoki, "Fine-tuning titanet-large model for speaker anonymization attacker systems," in *ICASSP*, 2025, pp. 1–2.

[7] Henry Li Xinyuan, Ashi Garg, Zexin Cai, Kevin Duh, Leibny Paola García-Perera, Sanjeev Khudanpur, Nicholas Andrews, and Matthew Wiesner, "Hltcoe submission to the voiceprivacy attacker challenge," in *ICASSP*, 2025, pp. 1–2.

[8] Natalia Tomashenko, Xiaoxiao Miao, Emmanuel Vincent, and Junichi Yamagishi, "The first voiceprivacy attacker challenge," in *ICASSP*, 2025, pp. 1–2.

[9] Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola García-Perera, Sanjeev 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*, 2024, pp. 61–66.

[10] Jixun Yao, Nikita Kuzmin, Qing Wang, Pengcheng Guo, Ziqian Ning, Dake Guo, Kong Aik Lee, Eng-Siong Chng, and Lei Xie, "Npu-ntu system for voice privacy 2024 challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 67–71.

[11] Nikita Kuzmin, Hieu-Thi Luong, Jixun Yao, Lei Xie, Kong Aik Lee, and Eng-Siong Chng, "Ntu-npu system for voice privacy 2024 challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79.

[12] Wenju Gu, Zeyan Liu, Liping Chen, Rui Wang, Chenyang Guo, Wu Guo, Kong Aik Lee, and Zhen-Hua Ling, "Ustc-polyu system for the voiceprivacy 2024 challenge," in *SPSC 2024*, 2024.

[13] Rayane Bakari, Olivier Le Blouch, Nicholas Evans, Nicolas Gengembre, Michele Panariello, and Massimiliano Todisco, "The influence of non-timbral cues in voice anonymisation and evaluation," in *5th Symposium on Security and Privacy in Speech Communication*, 2025.

[14] Natalia Tomashenko, Emmanuel Vincent, and Marc Tommasi, "Exploiting Context-dependent Duration Features for Voice Anonymization Attack Systems," in *Interspeech*, 2025, pp. 5128–5132.

[15] Carlos Franzreb, Arnab Das, Tim Polzehl, and Sebastian Möller, "Private kNN-VC: Interpretable Anonymization of Converted Speech," in *Interspeech*, 2025, pp. 3224–3228.

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

[17] Sarina Meyer and Ngoc Thang Vu, "Use cases for voice anonymization," 2025.

[18] Jennifer Williams, Karla Pizzi, Natalia Tomashenko, and Sneha Das, "Anonymizing speaker voices: Easy to imitate, difficult to recognize?," in *ICASSP*, 2024, pp. 12491–12495.

[19] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas A. Reynolds, "Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *5th International Conference on Spoken Language Processing (ICSLP 1998)*, 1998, p. paper 0608.

[20] Sarina Meyer, Florian Lux, Julia Koch, Pavel Denisov, Pascal Tilli, and Ngoc Thang Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP*, 2023, pp. 1–5.

[21] Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP*. IEEE, 2024, pp. 4725–4729.

[22] Pierre Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *Ph.D. dissertation, Université de Lorraine,*, 2023.

[23] Hubert Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[24] Michele Panariello, Massimiliano Todisco, and Nicholas Evans, "Preserving spoken content in voice anonymisation with character-level vocoder conditioning," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 12–16.

[25] Matthew Baas, Benjamin van Niekerk, and Herman Kamper, "Voice conversion with just nearest neighbors," in *Interspeech*, 2023, pp. 2053–2057.

[26] Jinzuomu Zhong, Korin Richmond, Zhiba Su, and Siqi Sun, "Accentbox: Towards high-fidelity zero-shot accent generation," in *ICASSP*, 2025, pp. 1–5.

[27] Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan, "Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice," in *Interspeech*, 2023, pp. 5291–5295.

[28] Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie, "The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," in *ICASSP*, 2021, pp. 6918–6922.

[29] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 3830–3834, ISCA.

[30] Nicolas Gengembre, Olivier Le Blouch, and Cédric Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech*. International Speech Communication Association, 2024.