ROBUST ACCENT IDENTIFICATION VIA VOICE CONVERSION AND NON-TIMBRAL EMBEDDINGS

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

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

ABSTRACT

Automatic accent identification (AID) remains a challenging task due to the complex variability of accents, the entanglement of accent cues with speaker traits, and the scarcity of reliable accent-labelled data. To address these challenges, we propose a speaker augmentation strategy using voice conversion (VC), with which we generate additional training data by converting original training utterances into different speaker voices while preserving accentual cues. For this purpose, we select two recent VC systems and evaluate their capability to preserve accent. Alternatively, we also explore the use of non-timbral embeddings in AID, for their ability to convey accent information among other non timbral cues. The effectiveness of both methods is demonstrated on the GenAID benchmark, achieving a new state-of-the-art F1-score of 0.66, compared to the previous score of 0.55. Beyond AID, we show that non-timbral embeddings enable accent-controlled Text-to-Speech, producing high-fidelity speech with accurate accent transfer.

***Index Terms*—** accent identification, voice conversion, data augmentation, controlled TTS

1. INTRODUCTION

Accent is a key aspect of spoken language, reflecting geographic, social, and cultural variation, encompassing phonemic, phonetic, rhythmic, and structural features [1]. Automatic accent identification (AID) has attracted attention for application in speech recognition, spoken language understanding, and sociolinguistic studies. Despite advances in related areas like language identification (LID) and speaker recognition, accent recognition remains a challenging task.

A major challenge is the complex variability of accents, often linked to a speaker's native language. Unlike speaker differences, accent variation relies on subtle phonetic and phonological cues. Studies [2] show that LID models misclassify non-native (L2) speech as the speaker's native or related language, illustrating the difficulty of separating accent from language and other speaker traits. The lack of large-scale datasets and the entanglement of speaker and accent cues further hinder the learning of robust models that generalise to unseen speakers.

Recent deep learning approaches [3] improve AID accuracy but may struggle to disentangle speaker and accent cues, highlighting the need for speaker-invariant representations and greater data diversity. Data augmentation is a key strategy to address data scarcity in many speech tasks [4]. Conventional approaches [5, 6] improve robustness but may distort accent cues. In this work, we use voice conversion (VC) as a speaker augmentation technique to tackle two key challenges simultaneously. First, by generating additional training data from multiple target speakers, VC directly alleviates data scarcity. Second, speaker augmentation helps disentangle speaker-related and accent information, since the same accent can be expressed in a number of different voices. This property makes VC particularly well-suited for AID tasks, as it allows models to focus on accentual patterns without being confounded by speaker-specific characteristics. Importantly, effective augmentation requires a VC system that preserves accentual traits while modifying speaker identity. Among recent VC systems, Retrieval-based Voice Conversion (RVC)[1] and k-Nearest Neighbors VC (kNN-VC) [7] have shown strong capabilities in disentangling timbre from linguistic content, but their use for data augmentation in AID remains unexplored.

To the best of our knowledge, this is the first work to analyse the behaviour of specific VC systems for designing speaker augmentation strategies to AID. Specifically, we evaluate RVC and kNN-VC for their ability to alter timbre while preserving accent. We also explore the use of specialized embeddings (LID [8], non-timbral [9]) to enhance AID by improving speaker invariance. Experiments on the GenAID benchmark [10] show significant improvements in generalization to unseen speakers.

The main contributions of this work are as follows :

- we present the first systematic analysis of VC systems (RVC, kNN-VC) for speaker diversity and accent-preserving augmentation ;
- we design task-aware augmentation strategies that leverage VC properties ;
- we use specialized embeddings to enhance accent recognition by promoting speaker invariance ;
- we improve state-of-the-art results on GenAID benchmark with unseen speakers by 10 points ;
- we demonstrate the novel use of non-timbral embeddings in accent-controlled Text-to-Speech (TTS), showing their effectiveness for accent transfer.

2. RELATED WORKS

2.1. Accent Identification

AID shares similarities with LID [11, 12] and speaker identification [13, 14], but is challenged by data scarcity, limited accent diversity, and imbalanced speaker–accent distributions. Early AID systems used context-dependent HMMs [15] or formant frequency-based GMMs [16]. The AESRC2020 benchmark [3] enabled systematic comparisons but is no longer publicly available. Recent

[1] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

AID approaches leverage large-scale self-supervised speech models to extract accent-discriminative embeddings. For instance, [8] proposed an AID pipeline that uses embeddings extracted from a LID model [12], classifies them with a simple linear layer, achieving strong performance; however, evaluation on unseen speakers was unclear, leaving open the risk of speaker–accent entanglement.

Large-scale accent datasets, such as CommonAccent [17] and GLOBE [18] derived from Common Voice [19], lack reliable speaker-level accent metadata, complicating generalization. In GLOBE, some metadata were auto-generated using HuBERT [20] with ECAPA-TDNN [21], a model that may introduce label noise. To address evaluation bias from overlapping speakers, [10] introduced speaker-disjoint train/test splits based on CommonAccent and released GenAID[2], an accent classifier that achieved a 0.55 F1-score on unseen test speakers, which is claimed to be the state-of-the-art performance.

2.2. Speaker-Invariant Representations and Disentanglement

A key challenge in AID is speaker–accent entanglement, where models inadvertently capture speaker-specific cues instead of accent features. Speaker-invariant representation learning addresses this by extracting embeddings that retain specific information while minimizing speaker identity cues. Adversarial training [22, 10] is commonly used, jointly optimizing an encoder with a speaker classifier to suppress timbre information, improving robustness to unseen speakers [3]. Alternative disentanglement approaches explicitly separate timbral and non-timbral cues. For example, [9] introduced complementary embeddings for timbral and non-timbral features while [23] proposed a multi-level VAE with an accent classifier to disentangle accent and speaker embeddings. [24] applied a TDNN with a bottleneck layer to produce frame-level accent embeddings, facilitating more effective downstream classification.

2.3. Voice conversion for Data Augmentation:

Beyond representation learning, data augmentation is another strategy to improve AID. Traditional augmentations such as SpecAugment [6], SpecMix [5], time-stretching, pitch shifting, additive noise, and synthetic speech generation [25], increase model robustness by introducing variability at the spectral or temporal level, but can distort speaker and accent cues. Voice conversion (VC) offers controlled modification of speaker identity while preserving target attributes [26], making it particularly promising for AID. VC-based augmentation has been applied to multiple contexts: children's speech recognition [27], speaker-independent keyword recognition [28], speaker recognition under adverse conditions [29], etc. RVC and kNN-VC [7] achieve high-quality conversions with minimal distortion of the underlying linguistic attributes. Notably, [4] recently showed RVC benefits for low-resource dialect classification, though their work was limited to a single VC method.

3. TIMBRE-ACCENT DISENTANGLEMENT

While RVC is used as a framework for disentangling different aspects of speech [9, 4], we hypothesize that more recent VC models, such as kNN-VC, may be better suited to this task. Thus, we have used objective metrics to evaluate whether VC systems can modify speaker timbre while preserving non-timbral attributes, and specifically the accent. We therefore compare RVC and KNN-VC along two main axes:

1. **Timbre modification**: We compute cosine similarity between speaker embeddings extracted using ECAPA-TDNN[3]. These embeddings are well-established for capturing timbral characteristics [9], making them a suitable choice for evaluating speaker similarity. We calculate the speaker similarity between the source and converted speech, and between the target and converted speech.
2. **Accent preservation**: We estimate retention of accent using the GenAID model [10]. Since classification accuracy is limited by the accent predictor's imperfections, we also use the Accent Embedding Cosine Similarity (AECS) [30], which provides a continuous, more nuanced measure of similarity between the source and converted speech.

For these evaluations, we perform voice conversion on the *unseen test subset* of the GenAID benchmark, which comprises 13 accents, with 100 utterances per accent from distinct speakers. For each source utterance, four random target speakers are selected from DATASET for conversion.

Table 1 presents the average speaker similarity scores and accent preservation results for RVC and kNN-VC. Both systems achieve low similarity with the source speaker, indicating a shift away from the original speaker identity. However, kNN-VC demonstrates a significantly higher similarity to the target speaker, suggesting more effective timbre conversion compared to RVC. It also shows a modest improvement in accent retention, as measured by both classification accuracy and AECS. For reference, the mean AECS between random utterance pairs is 0.80, serving as a baseline threshold; values above this indicate that accent cues are preserved. Overall, these results indicate that both RVC and kNN-VC are effective at modifying timbre while preserving accent, with kNN-VC performing slightly better, consistent with our hypothesis.

Table 1. Timbre modification and accent preservation for RVC and kNN-VC on unseen GenAID test set. Lower speaker similarity with the source and higher with the target indicates better timbre conversion. Higher accent accuracy and AECS indicate better retention of the original accent.

VC System	Speaker Similarity		Accent prediction	
	Source↓	Target↑	Acc↑	AECS↑
RVC	0.19 ± 0.10	0.48 ± 0.12	48.19	0.84
kNN-VC	**0.18** ± 0.10	**0.70** ± 0.08	**50.12**	**0.90**

4. PROPOSED APPROACHES

Following the GenAID training setup, we aim to improve the AID model through better input representations and training data diversity. Two complementary strategies were explored: (1) speaker and accent-disentangled data augmentation using VC, and (2) accent-specific input representations via specialized embeddings.

4.1. Voice conversion-Based Data Augmentation

Building upon our previous analysis showing that specific VC systems such as RVC and kNN-VC can effectively mask timbre while preserving accentual cues, we employ one of these methods at a time

[2] https://github.com/jzmzhong/GenAID/tree/GenAID

[3] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

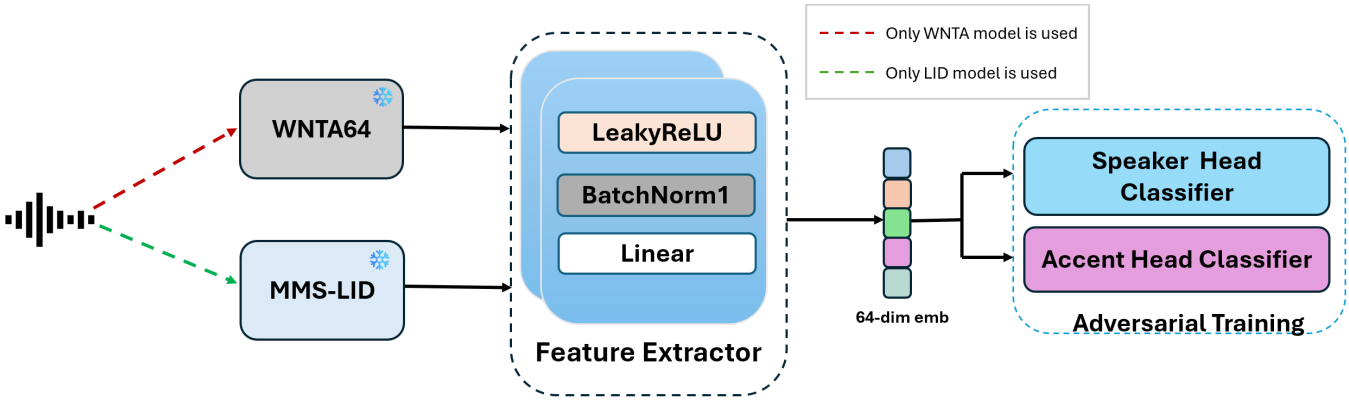

Fig. 1. Architecture of the AID model. The model takes as input either LID embeddings extracted from a pretrained MMS-LID model or non-timbral WNTA64 embeddings. The embeddings are fed into a feed-forward classifier with two heads: an accent classification head and a speaker classification head, the latter being used during adversarial training with KL regularization.

to augment the training data. For each original training utterance, we generate two additional voice-converted versions using the selected VC system, with target speaker identities randomly drawn from the LibriTTS train-clean-100 subset [31]. This process increases speaker variability in the training set while minimally affecting the underlying accent, thereby encouraging the model to learn speaker-invariant accent features.

4.2. Specialized Embedding-Based Representations

Beyond data augmentation, we investigate the use of specialized embeddings to improve accent classification by leveraging latent accent-relevant cues. Inspired by [8], we first extract embeddings from a pretrained LID model[4] [12], which are known to encode language-specific phonetic patterns potentially useful for accent discrimination. In parallel, we employ WNTA64 embeddings [9] extracted from a pretrained model[5] designed to disentangle timbral and non-timbral cues. These embeddings are taken from an intermediate layer that emphasizes prosody and accentual information while minimizing speaker-specific timbre. Compared to other embeddings like x-vector or HuBERT embeddings, WNTA64 embeddings may offer a more explicit separation between accentual patterns and speaker timbre, making them especially suitable for our accent identification task [9].

The architecture used in our AID models is described in Figure 1. LID or WNTA64 embeddings are extracted from the input signals, and then passed through three linear layers with batch normalization and ReLU activation, with output sizes 256, 128 and 64. As in [10], the resulting 64-dim embedding is fed into two heads dedicated to accent and speaker classification respectively, in an adversarial process. The two model variants are denoted as LID_FF and WNTA64_FF, with reference to their Feed Forward nature.

For adversarial training, the speaker classifier is trained to correctly predict the speaker, while the feature extractor and the accent head are optimized together to predict the accent and to fool the speaker classifier, encouraging the embeddings to be speaker-invariant through the Kullback-Leibler divergence (denoted as KL) as a penalty. The overall training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{accent}} + \lambda \cdot \text{KL}\left(\mathbf{p}_{\text{speaker}} \parallel \mathcal{U}\right) \quad (1)$$

[4]https://huggingface.co/facebook/mms-lid-256
[5]https://huggingface.co/Orange/Speaker-wavLM-pro

where $\mathcal{L}_{\text{accent}}$ is the cross-entropy loss for accent classification, $\mathbf{p}_{\text{speaker}}$ is the predicted speaker distribution, $\mathcal{U}$ is a uniform distribution over speaker classes, and λ is a coefficient controlling the strength of the speaker-invariance constraint.

When optimized, the KL loss tends to bring the speaker prediction distribution close to a uniform distribution, thus reducing the amount of speaker-related information brought to the speaker head, i.e. contained in the 64-dim embedding.

5. EXPERIMENTAL SETUP

All experiments follow the GenAID benchmark protocol [10], which is based on the CommonAccent corpus (derived from Common Voice v17.0). The data is split into training, validation, and test sets, with strict separation of speakers to ensure that evaluation reflects generalization to unseen speakers rather than memorization of speaker–accent mappings.

All proposed AID systems are trained for 10 epochs using distinct learning rates: 1e-4 for the accent classifier and 1e-5 for the auxiliary speaker classifier. This setup stabilizes adversarial training by allowing accent classification to dominate optimization, while still enforcing a gradual removal of speaker-specific information from the representations.

The value of λ in Equation 1, which also controls the adversarial process, is set to 0.1.

6. RESULTS AND DISCUSSION

Table 2 outlines the performance of AID systems on the unseen speaker subset of the GenAID benchmark. We compare the baseline GenAID model, its VC-augmented variants (using either RVC or kNN-VC), and the feed-forward classifiers LID_FF and WNTA64_FF, with and without VC-based augmentation.

Retraining the baseline GenAID classifier with RVC- or kNN-VC-augmented data consistently improves performance on unseen speakers. For example, RVC augmentation increases accuracy from 0.56 to 0.61, while kNN-VC yields an even higher accuracy of 0.66. These results confirm that VC-based augmentation effectively enhances generalization by increasing speaker diversity in the training set while preserving the original accent, thereby encouraging the

model to learn speaker-invariant representations. Moreover, the superior performance of kNN-VC highlights that the effectiveness of augmentation is closely tied to the quality of timbre-accent disentanglement: better VC behavior leads to greater improvements in AID performance.

It is worth noting that a similar and complementary experiment involving both sets of augmented data, from RVC and from kNN-VC, did not result in any further performance enhancement. This indicates that increasing the volume of training data is not necessary, neither is the combination of different voice conversion methods. Furthermore, the adversarial speaker head is incorporated to ensure fair comparison and also promote timbre disentanglement. An incremental ablation study in [10] demonstrates that this component is essential for learning speaker-invariant, accent-consistent representations, thereby justifying its inclusion in the proposed system.

Among the systems #1, #4 and #7, the latter (WNTA64_FF) achieves the best overall performance, achieving **a 10-point improvement** (absolute, in percent) on accuracy over the baseline #1, and nearly the same improvement when considering the f1-score. When combined with any a posteriori VC augmentation, it maintains the same high performance. This shows that isolating non-timbral or accent-relevant features, whether using WNTA64 or LID embeddings, is highly effective for accent generalization. VC-based augmentation adds only marginal gains for LID_FF or WNTA64_FF, as these embeddings already encode accent-specific information while minimizing speaker-specific timbre. As a result, the additional speaker variability introduced by VC-based augmentation has a negligible impact, since the model is already largely invariant to speaker identity. In other words, when using specialized embeddings, the benefits of VC augmentation are naturally reduced compared to models relying on raw features.

Table 2 also shows that the LID-based models do not perform as well as the WNTA64-based ones, suggesting that some of the accent information is not captured in these embeddings.

Overall, these results indicate that both VC-based data augmentation and specialized, non-timbral embeddings lead to similar improvements on unseen speakers, validating the effectiveness of our two approaches.

Table 2. Accent identification results. Tick (✓) indicates which VC-based augmentation is applied.

AID Systems	**VC Aug**		**Unseen Spks↑**			
	RVC	kNN-VC	prec	rec	f1	acc
#1 GenAID$_{baseline}$ [10]	-	-	0.63	0.56	0.55	0.56
#2 GenAID$_{RVC}$	✓	-	**0.72**	0.61	0.60	0.61
#3 GenAID$_{knnVC}$	-	✓	0.70	**0.66**	0.65	**0.66**
#4 LID_FF	-	-	0.57	0.58	0.57	0.58
#5 LID_FF$_{RVC}$	✓	-	0.57	0.58	0.58	0.58
#6 LID_FF$_{knnVC}$	-	✓	0.59	0.58	0.57	0.58
#7 WNTA64_FF	-	-	**0.66**	**0**.66	**0.66**	**0.66**
#8 WNTA64_FF$_{RVC}$	✓	-	0.65	0.65	0.65	0.65
#9 WNTA64_FF$_{knnVC}$	-	✓	0.66	**0.66**	**0.66**	**0.66**

7. ACCENT-CONTROLLED TTS

In addition to our main experiments, we evaluate the practical utility of non-timbral embeddings compared to dedicated accent representations in an accent-controlled TTS. Building on the AccentBox framework [10], we implemented two TTS systems: the original AccentBox using GenAID accent embeddings, and a modified version where we replaced the GenAID embeddings with WNTA64 embeddings, which may isolate accent-specific features without speaker information. For timbre control, both are also conditioned on WTA embeddings[6] from [9] to avoid potential confounding effects related to speaker identity.

We evaluated the quality of accent propagation in generated speech using the GenAID model, measuring how accurately the system produces the intended accent. While this approach provides a consistent evaluation, it is important to note that using GenAID to assess the generated speech may introduce a bias in favor of GenAID embeddings, since the classifier is inherently optimized for the features and representations produced by the GenAID model itself. The evaluation dataset contains 720 generated utterances, built from 5 phonetically rich sentences and conditioned on 12 VCTK speakers, 2 per each of 6 accents. It includes 144 unique voice–accent combinations (12 timbres × 12 accents).

Results in Table 3, show that using WNTA64 embeddings improves accent control across all accents. This indicates that WNTA64 embeddings effectively encode accent information and can be used to generate speech with high accent fidelity. Overall, these findings demonstrate that these embeddings are not only useful for accent recognition but also valuable for controlling accent in speech synthesis, broadening their application in speech technology.

Table 3. Accent classification performance on TTS outputs conditioned on GenAID vs WNTA embeddings.

	WTA + GenAID			**WTA + WNTA**		
Accent	Prec	Rec	F1	Prec	Rec	F1
US	0.56	0.48	0.52	0.62	0.51	**0.56**
Australian	0.42	0.50	0.45	0.50	0.46	**0.48**
South Asian	1.00	0.06	0.11	1.00	0.11	**0.20**
English	0.41	0.81	0.55	0.47	0.92	**0.62**
Scottish	0.77	0.53	0.63	0.97	0.50	**0.66**
Irish	0.67	0.33	0.44	0.74	0.64	**0.69**

8. CONCLUSIONS

In this work, we demonstrate the effectiveness of using voice conversion and specialized embeddings to improve robust accent identification for unseen speakers. By systematically analysing VC systems like RVC and kNN-VC, we showed that targeted data augmentation or the use of non-timbral, speaker-invariant embeddings significantly enhances accent recognition performance, achieving state-of-the-art performance with a 0.66 F1 score on 13-accent classification for unseen speakers. Additionally, our downstream experiments with accent-controlled TTS confirmed that these embeddings can effectively encode accent information, enabling high-fidelity accent synthesis. Future work will focus on developing TTS systems aimed at creating more inclusive and adaptable speech technologies, capable of accurately reproducing a wide range of accents for diverse applications.

[6] https://huggingface.co/Orange/Speaker-wavLM-tbr

9. REFERENCES

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

[2] Niyati Bafna and Matthew Wiesner, "LID Models are Actually Accent Classifiers: Implications and Solutions for LID on Accented Speech," in *Interspeech*, 2025, pp. 1488–1492.

[3] 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.

[4] Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek, "Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion," in *Interspeech*, 2025, pp. 2780–2784.

[5] Gwantae Kim, David K. Han, and Hanseok Ko, "Specmix : A mixed sample data augmentation method for training with time-frequency domain features," in *Interspeech*, 2021, pp. 546–550.

[6] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617.

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

[8] Dan Lyth and Simon King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," 2024, arXiv preprint.

[9] Nicolas Gengembre, Olivier Le Blouch, and Cédric Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Interspeech*, 2024, pp. 2765–2769.

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

[11] Nur Safitri, Amalia Zahra, and Mirna Adriani, "Spoken language identification with phonotactics methods on minangkabau, sundanese, and javanese languages," *Procedia Computer Science*, vol. 81, pp. 182–187, 12 2016.

[12] V. Pratap, A. Tjandra, Bowen Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, Wei-Ning Hsu, A. Conneau, and M. Auli, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[13] Sreenivas S. Tirumala, Seyed R. Shahamiri, Abhimanyu S. Garhwal, and Ruili Wang, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250–271, 2017.

[14] Douglas A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91–108, 1995.

[15] Carlos Teixeira, Isabel Trancoso, and António Serralheiro, "Accent identification," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 3, pp. 1784–1787.

[16] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, 2005, pp. 139–143.

[17] 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.

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

[19] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. May 2020, pp. 4218–4222, European Language Resources Association.

[20] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP Processing*, vol. 29, pp. 3451–3460, 2021.

[21] 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.

[22] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," in *ICASSP*. IEEE, 2018, pp. 4854–4858.

[23] Jan Melechovsky, Ambuj Mehrish, Berrak Sisman, and Dorien Herremans, "Accent conversion in text-to-speech using multi-level vae and adversarial training," in *TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON)*, 2024, pp. 473–476.

[24] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Interspeech*, 2018, pp. 2454–2458.

[25] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *ICASSP*, 2020, pp. 7069–7073.

[26] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM TASLP Processing*, vol. 29, pp. 132–157, 2020.

[27] S. Shahnawazuddin, Nagaraj Adiga, Kunal Kumar, Aayushi Poddar, and Waquar Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," in *Interspeech*, 2020, pp. 4382–4386.

[28] Yeshanew Ale Wubet and Kuang-Yow Lian, "Voice conversion based augmentation and a hybrid cnn-lstm model for improving speaker-independent keyword recognition on limited datasets," *IEEE Access*, vol. 10, pp. 89170–89180, 2022.

[29] Ruijie Tao, Zhan Shi, Yidi Jiang, Tianchi Liu, and Haizhou Li, "Voice conversion augmentation for speaker recognition on defective datasets," 2024, arXiv preprint.

[30] 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, pp. 1–5.

[31] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, 2019, pp. 1526–1530.