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

Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification

Badr M. Abdullah[1], *Matthew Baas*[2], *Bernd Möbius*[1], *Dietrich Klakow*[1]

[1]Language Science and Technology, Saarland University, Germany
[2]Camb.AI, UAE

Corresponding author badr.nlp@gmail.com

Abstract

Arabic dialect identification (ADI) systems are essential for large-scale data collection pipelines that enable the development of inclusive speech technologies for Arabic language varieties. However, the reliability of current ADI systems is limited by poor generalization to out-of-domain speech. In this paper, we present an effective approach based on voice conversion for training ADI models that achieves state-of-the-art performance and significantly improves robustness in cross-domain scenarios. Evaluated on a newly collected real-world test set spanning four different domains, our approach yields consistent improvements of up to +34.1% in accuracy across domains. Furthermore, we present an analysis of our approach and demonstrate that voice conversion helps mitigate the speaker bias in the ADI dataset. We release our robust ADI model and cross-domain evaluation dataset to support the development of inclusive speech technologies for Arabic.

Index Terms: spoken Arabic dialect identification, voice conversion, cross-domain robustness, Arabic speech technology

1. Introduction

Arabic is the native language of more than 320 million people geographically distributed across the Middle East and North Africa [1]. Throughout the Arabic-speaking world, Modern Standard Arabic (MSA) serves as the official language and the medium of formal communication and news broadcasts. However, MSA is not naturally acquired and functional knowledge of it can only be achieved through formal education. Arabic dialects, on the other hand, are the language varieties that Arabic speakers naturally acquire and use for daily communication. These dialects exhibit considerable geographic variation, with varying degrees of mutual intelligibility across regions [2]. Although spoken dialects are neither standardized nor formally taught, they maintain a strong cultural presence through songs, folktales, and movies [3].

The linguistic and regional variations of Arabic pose significant challenges for the development of Arabic speech technologies. While modern ASR systems work well on MSA speech, they struggle with dialectal speech [4]. The limited dialectal resources and lack of writing standardization hinder the development of dialect-aware ASR models [5]. Additionally, text-to-speech systems require high-quality recordings with known sources of variation with respect to dialect and register. Therefore, robust Arabic dialect identification (ADI) systems are an essential component in large-scale data collection pipelines to enable the development of inclusive speech technologies that serve a wider range of speaker communities [6].

Although the ADI task has gained significant research interest within the speech technology community [7, 8, 9, 10], most evaluations of ADI systems have been limited to in-domain settings where test samples are drawn from the training domain. Recent studies have shown that while pre-trained speech models with self-supervised learning (SSL) significantly outperform conventional acoustic features and earlier neural approaches [6, 11], they still under-perform in cross-domain settings [6]. This limitation highlights the need for training strategies that yield ADI systems that are robust to variability in recording conditions and spoken genres. We argue that developing ADI models that transfer to unseen domains is more practical than attempting to collect diverse datasets covering all possible domains, especially since future use cases cannot be anticipated during model development. Therefore, to reliably evaluate real-world robustness of ADI systems, models should be tested on samples from multiple unseen domains not represented in the training data—that is, a zero-shot cross-domain evaluation. To address these challenges, we make the following contributions:

- To evaluate cross-domain robustness of ADI systems, we create a curated multi-domain dataset of $\sim$12 hours of speech from different public sources (§3.2).
- We propose an effective strategy based on voice conversion for training ADI models (§2) that outperforms strong baselines in both in-domain and cross-domain scenarios (§5).
- We further analyze our model and demonstrate that voice conversion mitigates the speaker bias in the ADI dataset (§6).
- We share our robust ADI model[1] and new evaluation dataset[2] with the research community on HuggingFace Hub.

2. Voice Conversion for ADI

We formalize ADI as a classification problem. Given an Arabic speech sample $\mathbf{x}$, the goal is to predict the speaker's dialect $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a closed set of dialects. To do so, we require a dataset of N natural speech samples, each paired with a dialect annotation:

$$\mathcal{D} = \Big\{(\mathbf{x}_i, y_i)\Big\}_{i=1}^{N} \tag{1}$$

The dataset $\mathcal{D}$ is used to train a model via cross-entropy loss to predict the dialect y from the acoustic input $\mathbf{x}$. In our method, we apply voice conversion (VC) to create a re-synthesized dataset from the training samples. VC transforms a spoken utterance in a way that the generated speech is perceived as if it was spoken by a different speaker while preserving content and intelligibility [12]. Although VC has been applied to a few speech processing tasks in the literature [13, 14, 15], its effectiveness for dialect identification remains unexplored. We for-

[1]badrex/mms-300m-arabic-dialect-identifier
[2]badrex/MADIS5-spoken-arabic-dialects

10.21437/Interspeech.2025-1809

malize VC as a parametric function

$$\tilde{\mathbf{x}} = \mathcal{C}_\theta(\mathbf{x}, \mathbf{v}) \tag{2}$$

where $\mathbf{v}$ is a speech sample from a target speaker, and $\tilde{\mathbf{x}}$ is a re-synthesized segment that preserves the linguistic content of $\mathbf{x}$ but with the acoustic-phonetic features that characterizes the speaker of $\mathbf{v}$. Modern VC techniques offer efficiency and naturalness while handling unseen speakers effectively [16, 17, 18, 19]. The transform function $\mathcal{C}_\theta$ can be viewed as a generalized form of audio data augmentation. For instance, SpecAugment [20] represents a non-parametric version of $\mathcal{C}_\theta$ that does not require a target voice. Using VC, we obtain a re-synthesized dataset

$$\tilde{\mathcal{D}} = \Big\{ \big(\mathcal{C}_\theta(\mathbf{x}_i, \mathbf{v}_i), y_i\big) \Big\}_{i=1}^{M} \tag{3}$$

Here, the target voice $\mathbf{v}_i$ can be fixed for all training segments or uniformly sampled from a pool of T target speakers: $\mathbf{v}_i \sim \{\mathbf{v}_1, \ldots, \mathbf{v}_T\}$. When each training sample is converted once, $M = N$. Using a single target speaker yields a dataset where all training segments sound as if spoken by the same speaker. We train our ADI model on the combined dataset $\mathcal{D}_{\text{train}}$ formed by concatenating the natural and re-synthesized datasets

$$\mathcal{D}_{\text{train}} = \mathcal{D} \cup \tilde{\mathcal{D}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \cup \{(\tilde{\mathbf{x}}_i, y_i)\}_{i=1}^{M} \tag{4}$$

In our experiments, we show that VC significantly improves ADI performance both in-domain and cross-domains, achieving gains that traditional data augmentation techniques cannot match.

3. Datasets

3.1. Training Dataset: MGB-3 ADI-5

As our training dataset, we use the MGB-3 ADI-5 dataset, which is a widely-used ADI resource with coarse-grained dialect labels derived from Aljazeera TV broadcast [7]. It consists of approximately 14.6k samples (∼53.6 hours) containing speech segments of MSA as well as four Arabic dialect groups based on geography: Gulf Arabic (spoken in the Arabian peninsula), Levantine Arabic, Maghrebi Arabic (spoken in North Africa), and Egyptian Arabic. The speech segments come from diverse content including news reports, panel discussions, and interviews. The dataset is relatively balanced across dialects and features a wide range of speakers and acoustic conditions. The validation and test splits of ADI-5 consists of 10 hours and 10.1 hours of speech, respectively.

3.2. Evaluation Dataset: MADIS-5 Benchmark

We manually curate a dataset for multi-domain ADI in speech (MADIS-5) to facilitate evaluation of cross-domain robustness of ADI systems. Our dataset comprises ∼12 hours of speech (4854 utterances) collected from four different public sources with varying similarity to the TV broadcast domain of ADI-5. The recordings were manually segmented and labeled by a native Arabic speaker with linguistic expertise (PhD in Computational Linguistics) and extensive exposure to Arabic language variation. The dialect labels were then verified by another native Arabic speaker with competence and keen interest in different Arabic dialects. The data sources are:

- **Radio broadcast.** Similar to previous research [21], we harvested radio broadcasts using the radio.garden website to access local radio stations throughout the Arab world during the time period December 2024 - January 2025.
- **TV dramas.** We compiled 5-7 second speech segments from the Arabic Spoken Dialects Regional Archive (SARA) on the Kaggle platform.
- **TEDx Talks.** We created short segments and added dialect annotations to the Arabic portion of the TEDx dataset [22].
- **Theater.** We used YouTube to collect Arabic dramatic and comedy plays performed in theaters in different time periods, which we then segmented into short samples.

These sources are characterized by varying recording conditions and diverse themes, which makes the dataset ideal for evaluating cross-domain robustness of ADI systems beyond TV broadcast. It is worth noting that for native Arabic speakers with exposure to multiple dialects through media, a coarse-grained classification of Arabic dialects is a relatively trivial task. This stems mainly from the abundance of discriminating linguistic features at both the phonetic and lexical levels that signal regional dialects. Although our two annotators achieved perfect agreement in categorizing regional dialectal speech, they disagreed on the classification of 2.3% of radio broadcast segments as dialect or MSA. This disagreement reflects an ongoing research challenge, as MSA in some scenarios exists on a continuum with dialectal Arabic [23, 24].

4. Experimental Setup

To evaluate the effectiveness of voice conversion for improving ADI, we compare it against several strong text and speech baselines as well as various audio augmentation techniques.

4.1. Text and Speech Baselines

For text baselines, we transcribe the datasets using two publicly available models: (1) a universal phoneme recognizer for phonetic transcriptions[3], and (2) a Whisper-based model for orthographic transcriptions[4]. We then train SVM classifiers for each transcription system with hyperparameters optimized using the scikit-learn library. We chose SVMs because they have demonstrated superior performance over neural classifiers in discriminating closely related languages from text [25]. Additionally, we use the orthographic transcriptions to fine-tune a BERT model pre-trained on a mixture of MSA and Arabic dialectal [26] to perform ADI[5].

For speech baselines, we fine-tune multilingual pre-trained speech models based on the wav2vec2 architecture; XLSR-53 [27], XLSR-128 [28], and MMS [29]. We use natural segments (10 seconds max) from the ADI-5 training dataset for fine-tuning. While all these models have 300M parameters, they differ in their pre-training data volume, with MMS having been pre-trained on more than 1,000 languages. We experimented with different learning rates (1×10^{-4}, $\{1, 3, 5, 7\} \times 10^{-5}$), out of which 5×10^{-5} was optimal for the all pre-trained models. We trained our models for 6 epochs using the default parameters of the Hugging Face trainer.

4.2. Data Augmentation

We compare our voice conversion method against several common data augmentation techniques in speech processing. These include SpecAugment [20], which applies time stretching, time masking and frequency masking to the spectrogram; pitch shifting to alter the fundamental frequency; room impulse response

[3] facebook/wav2vec2-lv-60-espeak-cv-ft

[4] speechbrain/asr-whisper-large-v2-commonvoice-ar

[5] CAMeL-Lab/bert-base-arabic-camelbert-mix

Table 1: *Results comparison across different baselines and data augmentation techniques measured in accuracy (%). The Δ columns show relative improvement over our strongest speech baseline (MMS) trained on natural speech only. $|\mathcal{D}_{train}|$ is the size of the training dataset for each model where N is the number of natural training samples in ADI-5 dataset. The results for models trained with VC are averaged across four runs with different target speaker sets. Chance-level accuracy is ∼20%.*

			In-domain		Cross-domain (MADIS-5)					
Model	Training Data	$\lvert\mathcal{D}_{\text{train}}\rvert$	ADI-5	Δ (%)	Radio	TEDx	Dramas	Theater	Avg.	Δ (%)
Text baselines										
SVM	phone n-grams	N	66.82		69.80	64.56	50.62	49.32	58.58	
SVM	character n-grams	N	50.00		55.09	52.56	47.23	37.20	48.02	
Arabic BERT	sub-words	N	62.73		62.53	64.03	56.95	54.69	59.55	
Speech baselines										
XLSR-53	natural speech	N	68.57		46.75	67.80	41.36	49.72	51.41	
XLSR-128	natural speech	N	70.58		52.38	62.67	48.36	48.76	53.04	
MMS	natural speech	N	75.94	–	67.63	69.23	54.12	49.88	60.22	–
MMS with audio augmentations										
MMS	natural + SpecAugment	$2 \times N$	78.75	+3.70	70.23	77.60	56.50	60.81	66.29	+10.08
MMS	natural + Pitch Shift	$2 \times N$	80.63	+6.18	70.45	76.70	57.97	61.37	66.62	+10.64
MMS	natural + RIR	$2 \times N$	77.88	+2.55	71.82	76.40	52.32	50.20	62.69	+04.10
MMS	natural + Additive Noise	$2 \times N$	79.42	+4.58	77.31	78.66	51.30	59.78	66.76	+10.87
MMS	natural + All Augmentations	$5 \times N$	81.64	+7.51	78.76	80.39	63.73	60.10	70.75	+17.49
MMS with voice conversion (VC)										
MMS	natural + VC with 1 voice	$2 \times N$	82.17	+8.21	77.68	82.79	67.80	62.62	72.72	+20.77
MMS	natural + VC with 2 voices	$3 \times N$	84.28	+10.99	85.17	83.84	72.51	65.06	76.65	+27.29
MMS	natural + VC with 4 voices	$5 \times N$	**85.32**	**+12.35**	**87.86**	**86.73**	**77.85**	**70.47**	**80.73**	**+34.07**

(RIR) simulation to emulate different acoustic environments; and additive noise augmentation using both music and background noise [30] at various signal-to-noise ratios.

4.3. Voice Conversion

In our study, we use nearest neighbor voice conversion (kNN-VC), a simple yet effective method for VC that does not require text transcriptions of any sort [16], making it ideal for untranscribed dialectal speech. k-NN VC transforms a speech segment into a target voice using only a few examples of the target speaker as a reference. We experimented with k-NN VC on Arabic speech and observed that it produces high-quality output in the target voice while preserving intelligibility and relevant cues for dialect classification. For our study, we used native Arabic target voices from LibriVox audio books, with approximately one minute of speech per speaker. All models using data augmentation or VC are trained for 3 epochs on the combined natural and re-synthesized datasets. Note that the test segments remain unmodified and were not subjected to any transform.

5. Experiments and Results

5.1. Baselines

Text baselines. On the in-domain test set, the phone-based SVM (66.82%) outperforms both the character-based SVM (50.00%) and Arabic BERT (62.73%). This shows that text-based classifiers trained on ASR transcripts are not reliable for the ADI task since ASR models are trained on MSA speech and normalize dialect-specific lexical features in their output. However, the phone-based SVM performs worse than Arabic BERT on the cross-domain setting (57.90% vs 59.40%).

Speech baselines. MMS with 75.94% in-domain accuracy demonstrates clear advantages over both XLSR variants, outperforming XLSR-128 (70.58%) and XLSR-53 (68.57%). This indicates that MMS's massive multilingual pre-training provides better feature representations for dialect identification. However, all models perform poorly on out-of-domain data, with XLSR variants even falling behind two of the text baselines. This consistent performance drop on out-of-domain data highlights a critical limitation: fine-tuning pre-trained speech models yields ADI systems that transfer poorly to unseen domains or "in-the-wild" speech that differs from the training data. Given that MMS performance is superior compared to XLSR variants, we perform our main experiments with data augmentation and voice conversion by fine-tuning MMS.

5.2. ADI with Data Augmentation

Audio augmentation techniques improve upon the MMS model trained only natural speech, with the combination of all augmentations achieving 81.64% accuracy (+7.51% in-domain relative improvement over MMS baseline). When applying a single augmentation in isolation, pitch shifting proves most effective (80.63%, +6.18%), followed by additive noise (79.42%, +4.58%). We also observe consistent improvements in the cross-domain setting across all data augmentation techniques.

5.3. ADI with Voice Conversion

Fine-tuning MMS on a combination of natural and re-synthesized speech using VC yields the most substantial improvements in our study. Even with just one target speaker for VC, our approach achieves 82.17% in-domain accuracy, surpassing all data augmentation methods. Increasing the number of target speakers consistently enhances performance, reaching 85.32% with four target speakers with a 12.35% relative improvement over the baseline MMS model. Additionally, we observe consistent cross-domain improvements with the model trained on four target voices achieves state-of-the-art results across all domains (overall relative improvement of up to 34.07%). Our best model shows exceptional performance on

Table 2: *Our best model compared to the SoTA in the literature. Δ shows relative improvement over MIT-QCRI system (%).*

	ACCURACY		PRECISION		RECALL	
Model	%	Δ	%	Δ	%	Δ
MIT-QCRI [7]	75.0	–	75.1	–	75.5	–
UTD [7]	79.8	+6.4	79.9	+6.4	80.3	+6.4
ResNet (R) [11]	80.4	+7.2	80.4	+7.1	80.5	+6.6
ECAPA (E) [11]	82.5	+10.0	82.6	+10.0	82.7	+9.5
Fusion (R + E) [11]	84.7	+12.9	84.8	+12.9	84.9	+12.5
MMS-VC (Ours)	**85.3**	**+13.7**	**85.4**	**+13.7**	**85.3**	**+13.0**

radio (87.86%) and TEDx (86.73%) domains that matches in-domain performance. The strong cross-domain performance suggests that voice conversion is indeed an effective strategy for learning representations that are robust to domain shifts.

5.4. Comparison to state-of-the-art on ADI-5

The ADI-5 dataset serves as a standard benchmark for Arabic dialect identification in the literature [7, 11]. Here, we compare the in-domain performance of our model to previously reported results as illustrated in Table 2. Our model, trained with re-synthesized speech using voice conversion, sets a new state-of-the-art performance, outperforming all previous approaches across all metrics. Even the strong ECAPA-TDNN system trained on SSL representations and its fusion with ResNet architecture falls short of our approach by at least 0.5% across all metrics. While the fusion of ResNet and ECAPA models demonstrates the benefits of model combination (84.7%), our single model still outperforms this ensemble approach (85.3%). These results provide convincing evidence of the effectiveness of voice conversion for ADI.

6. Model Analysis

In the previous section, we established that voice conversion is an effective method for training robust ADI systems. Here, we investigate why voice conversion yields such substantial improvements and better cross-domain generalizations. Our hypothesis is that re-synthesizing the training data using voice conversion helps normalize speaker variations in the dataset, which otherwise introduce a significant bias. This bias stems from an inherent limitation in dialect identification datasets: each speaker typically speaks only one dialect at the native level, therefore the training segments for each dialect are drawn from a disjoint speaker set. This lack of speaker overlap between dialects creates a strong association between speaker identity and dialect label. Consequently, neural networks can exploit speaker identity as an easier shortcut for dialect classification, rather than learning the more subtle but relevant dialectal features. While collecting training samples from a large native speaker pool could mitigate this bias, there is still no explicit incentive that prevents the model from simply memorizing the speakers rather than learning robust dialect representations.

To test this hypothesis, we conduct a controlled experiment by training models on re-synthesized speech in two conditions: unbiased and biased. In the unbiased setting, we re-synthesize the training speech samples using a unified set of 12 target speakers across all dialects, ensuring no association between the speaker and the spoken dialect (i.e., target speakers for VC are uniformly distributed across dialects). This setting is similar to our best performing model, but excludes the natural speech data from the final training dataset. In the biased setting, we deliberately introduce speaker bias by using dialect-specific target speakers: we re-synthesize the training data using a pool of 60 target speakers, with a disjoint set of 12 speakers for each dialect. This setting resembles the natural training data but with limited speaker variation within each dialect. As in our previous experiment, the evaluation samples for this experiment are not modified and remain in their natural form. The results of this experiment, shown in Table 3, are quite surprising. Fine-tuning MMS on re-synthesized speech only (no natural speech) in the unbiased condition yields substantial improvements, achieving 83.38% in-domain accuracy (+9.80%) and, more importantly, 76.61% cross-domain accuracy (+28.54%). In contrast, the accuracy of the model trained on the biased re-synthesized dataset drops dramatically to 27.33% in-domain and 24.32% cross-domain, performing close to random chance. These results strongly support our hypothesis that neural networks indeed exploit speaker-predictive features when available, and that voice conversion can effectively reduce this bias, leading to more robust dialect identification systems.

Table 3: *Model accuracy in unbiased and biased conditions. The Δ columns show relative improvement over MMS baseline.*

	In-domain		Cross-domain	
Model	ADI-5	Δ	Avg.	Δ
MMS	75.94	–	59.60	–
MMS-VC (unbiased)	**83.38**	**+09.80**	**76.61**	**+28.54**
MMS-VC (biased)	27.33	−64.01	24.32	−59.19

7. Discussion and Conclusion

Our experiments demonstrated that voice conversion significantly improves the generalization of ADI systems, particularly in cross-domain scenarios. The remarkable performance gain achieved through voice conversion (+34.07% cross-domain) cannot be matched by traditional data augmentation techniques. Our analysis with controlled experiments revealed that voice conversion is effective because it eliminates the speaker bias in ADI datasets, which current models are likely to exploit as shortcuts for dialect classification.

When we initiated this research, we hypothesized that voice conversion could serve as a data augmentation technique to expand our training data. However, our investigation revealed that the benefits of voice conversion extend far beyond simple data augmentation as it serves as an effective technique for bias mitigation in language and dialect identification datasets. This insight opens up new avenues for handling dataset biases in speech technology as we demonstrate the value of taking a data-centric approach to model development, where understanding and mitigating dataset biases can lead to more substantial improvements than model refinements alone. Furthermore, our proposed approach could be extended to other tasks where speaker identity is strongly correlated with output labels, such as accent identification and speech classification for healthcare applications.

8. Acknowledgements

We thank the anonymous reviewers for their positive feedback. We sincerely thank Aravind Krishnan for his valuable comments on the work presented in this paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

9. References

[1] K. Brown and S. Ogilvie, *Concise encyclopedia of languages of the world*. Elsevier, 2010.

[2] G. Khan, M. P. Streck, and J. C. Watson, *The Semitic languages: An international handbook*. Walter de Gruyter, 2011, vol. 36.

[3] N. Y. Habash, *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers, 2010.

[4] B. Talafha, K. Kadaoui, S. M. Magdy, M. Habiboullah, C. M. Chafei, A. O. El-Shangiti, H. Zayed, M. C. Tourad, R. Alhamouri, R. Assi, A. Alraeesi, H. Mohamed, F. Alwajih, A. Mohamed, A. El Mekki, E. M. B. Nagoudi, B. D. M. Saadia, H. A. Alsayadi, W. Al-Dhabyani, S. Shatnawi, Y. Ech-chammakhy, A. Makouar, Y. Berrachedi, M. Jarrar, S. Shehata, I. Berrada, and M. Abdul-Mageed, "Casablanca: Data and models for multidialectal Arabic speech recognition," in *Proceedings of EMNLP*. Association for Computational Linguistics, Nov. 2024.

[5] A. Waheed, B. Talafha, P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "VoxArabica: A robust dialect-aware Arabic speech recognition system," in *Proceedings of ArabicNLP 2023*. Singapore (Hybrid): Association for Computational Linguistics, Dec. 2023.

[6] P. Sullivan, A. Elmadany, and M. Abdul-Mageed, "On the robustness of arabic speech dialect identification," in *Interspeech 2023*, 2023, pp. 5326–5330.

[7] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic mgb-3," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 316–322.

[8] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *arXiv preprint arXiv:1803.04567*, 2018.

[9] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1026–1033.

[10] W. Lin, M. Madhavi, R. K. Das, and H. Li, "Transformer-based arabic dialect identification," in *2020 International Conference on Asian Language Processing (IALP)*. IEEE, 2020, pp. 192–196.

[11] A. Kulkarni and H. Aldarmaki, "Yet another model for arabic dialect identification," in *Proceedings of ArabicNLP 2023*, 2023, pp. 435–440.

[12] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[13] M. Baas and H. Kamper, "Voice conversion can improve asr in very low-resource settings," in *Proc. Interspeech 2022*, 2022, pp. 3513–3517.

[14] E. Casanova, C. Shulby, A. Korolev, A. C. Junior, A. d. S. Soares, S. Aluísio, and M. A. Ponti, "Asr data augmentation in low-resource settings using cross-lingual multi-speaker tts and cross-lingual voice conversion," in *Proc. Interspeech 2023*, 2023, pp. 1244–1248.

[15] Y. A. Wubet and K.-Y. 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. 89 170–89 180, 2022.

[16] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," in *Interspeech 2023*, 2023, pp. 2053–2057.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[18] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech 2021*, 2021, pp. 3615–3619.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, p. 2613, 2019.

[21] P. Foley, M. Wiesner, B. Odoom, L. P. Garcia Perera, K. Murray, and P. Koehn, "Where are you from? geolocating speech and applications to language identification," in *Proceedings of NAACL*. Association for Computational Linguistics, Jun. 2024.

[22] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "Multilingual tedx corpus for speech recognition and translation," in *Proceedings of Interspeech*, 2021.

[23] A. Keleg, W. Magdy, and S. Goldwater, "Estimating the level of dialectness predicts inter-annotator agreement in multi-dialect Arabic datasets," in *Proceedings of ACL*, Bangkok, Thailand, Aug. 2024.

[24] A. Keleg, S. Goldwater, and W. Magdy, "ALDi: Quantifying the Arabic level of dialectness of text," in *Proceedings of EMNLP*, Singapore, Dec. 2023.

[25] M. Medvedeva, M. Kroon, and B. Plank, "When sparse traditional models outperform dense neural networks: the curious case of discriminating between similar languages," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 2017, pp. 156–163.

[26] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Online): Association for Computational Linguistics, Apr. 2021.

[27] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLSR-53: Universal cross-lingual speech representations," in *Proc. Interspeech*, 2021, pp. 3429–3433.

[28] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and A. Conneau, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2634–2650, 2022.

[29] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.