Yet Another Model for Arabic Dialect Identification

Ajinkya Kulkarni
MBZUAI, UAE
ajinkya.kulkarni@mbzuai.ac.ae

Hanan Aldarmaki
MBZUAI, UAE
hanan.aldarmaki@mbzuai.ac.ae

Abstract

In this paper, we describe a spoken Arabic dialect identification (ADI) model for Arabic that consistently outperforms previously published results on two benchmark datasets: ADI-5 and ADI-17. We explore two architectural variations: ResNet and ECAPA-TDNN, coupled with two types of acoustic features: MFCCs and features exratected from the pre-trained self-supervised model UniSpeech-SAT Large, as well as a fusion of all four variants. We find that individually, ECAPA-TDNN network outperforms ResNet, and models with UniSpeech-SAT features outperform models with MFCCs by a large margin. Furthermore, a fusion of all four variants consistently outperforms individual models. Our best models outperform previously reported results on both datasets, with accuracies of 84.7% and 96.9% on ADI-5 and ADI-17, respectively.

1 Introduction

Dialect identification can be viewed as a special case of language recognition (Tong et al., 2006; Vijayan et al., 2018). Both tasks suffer from similar performance issues in the presence of background noise, channel mismatch, prosodic fluctuations, and so on. However, with closely related dialects having a small difference in both acoustic and linguistic feature space, dialect identification tasks are substantially more difficult in nature (Zaidan and Callison-Burch, 2014). The Arabic language is spoken in various dialects across the Arab world, in addition to Modern Standard Arabic (MSA) which is used in official and educational settings. Speech recognition systems trained on MSA data generally don't generalize well to dialectal Arabic and specialized dialectal models may be needed for improving automatic speech recognition (ASR) performance in systems developed for specific populations. Dialect identification could facilitate the development of dialectal speech recognition systems in various ways, such as by identifying dialectal utterances in large multi-dialectal corpora, or online dialect identification for routing utterances to dialect-specific ASR modules.

To enable the development of spoken Arabic dialect identification systems, two benchmark datasets have been developed: ADI-5, which was deployed as part of the MGB-3 challenge (Ali et al., 2017) and ADI-17, deployed as part of the MGB-5 challenge (Ali et al., 2019). For both challenges, the top systems developed and submitted for the initial challenges remain the best performing systems reported in the research literature for these benchmarks. The ADI-5 training set consists of 10 hours of dialectal speech from broadcast news, covering five dialects: Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA), in addition to two hours each for development and test sets. The ADI-17 data set consists of 17 dialectal classes for a total of 3K hours extracted automatically from YouTube. Roughly 58 hours of data were manually verified for the development and test sets.

In this paper, we describe spoken dialect identification models we developed and tested on these benchmarks, and we report results exceeding the best performing models submitted to both challenges. We experimented with the Residual networks (ResNet) (He et al., 2015) and Emphasized Channel Attention, Propagation and Aggregation (ECAPA-TDNN) (Desplanques et al., 2020) architectures. Both architectures have been successfully employed for speaker verification tasks. In addition, ResNet was used in the best performing dialect identification system in the MGB-5 challenge, and ECAPA-TDNN has been recently explored for dialect classification, as in Lonergan et al. (2023) for Irish dialects. In addition, we explored the use of acoustic features extracted from the UniSpeech-SAT (Chen et al., 2021) model, which have been shown to provide improvements in various tasks in the SUPERB benchmark (Yang et al., 2021). We

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 435–440
December 7, 2023

observe large improvements in accuracy by incorporating these features into our models. We also employ data augmentation via additive noise and speed perturbation, which generally help improve the generalization of speech classification models. Our best model result is 84.7% accuracy in the ADI-5 test set, compared to 75% previously reported as the best result in Ali et al. (2017). In ADI-17, our best model achieves 96.9% accuracy compared to 94.9% previously reported as the best model in Ali et al. (2019).

2 Related Work

In this section, we describe the approaches proposed for ADI tasks in MGB-3 and MGB-5 challenges, which are used as baseline systems in this work. We first describe the top two performing systems for the MGB-3 challenge (ADI-5) (Ali et al., 2017), followed by the top two systems in the MGB-5 challenge (ADI-17) (Ali et al., 2019).

The MIT-QCRI ADI system (Shon et al., 2017; Khurana et al., 2017) combines acoustic and linguistic features within a Siamese neural network framework to reduce dimensionality based on i-vectors. They used loss functions involving both Euclidean and cosine distances and employed support vector machines as the backend classifier. In contrast, the University of Texas at Dallas (UTD) submission (Bulut et al., 2017) to the MGB-3 challenge fused five systems, incorporating acoustic and lexical information through various techniques, including i-vectors, Generative Adversarial Networks (GANs), Gaussian Back-end (GB), and BNF i-vector features. The UTD system obtained the second-best performance with an overall accuracy of 70.38% (Ali et al., 2017).

Duke Kunshan University (DKU) submitted four variants of ResNets with different block sizes and datasets, which were fused to achieve the best performing system in the MGB-5 challenge (Ali et al., 2019). The DKU system employed a ResNet with global statistics pooling and a fully connected layer. They used the Kaldi toolkit for data augmentation, including speed-perturbation and datasets such as MUSAN and RIR. The ResNet system was trained using cross-entropy loss with a softmax layer, taking 64-dimensional mel-filterbank energy features as input. On the other hand, the University of Kent (UKent) MGB-5 system (Miao and Mcloughlin, 2019) used a neural network architecture combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks with Time-Scale Modification (TSM). The UKent system reported an accuracy of 93.1% on the test set.

While the best performing models reported in the original MGB-3 and MGB-5 challenges have not been outperformed in later publications (to the best of our knowledge), several other studies proposed model variants and analyzed the performance in various ways. Regarding the use of pre-trained self-supervised acoustic models, Sullivan et al. (2023) recently utilized the XLS-R model (Babu et al., 2022), which is a multi-lingual pre-trained acoustic model that includes Arabic as one of the languages used in pre-training, and HuBERT (Hsu et al., 2021), which was pre-trained solely in English. They fine-tuned dialect classification models on the ADI-17 dataset, and interestingly, the model based on HuBERT outperformed the XLS-R-based model, in spite of the multi-lingual pre-training of the latter. This indicates that the quality of the features extracted from pre-trained acoustic models may depend more on the self-supervised training details rather than linguistic coverage. A model outperforming HuBERT on several benchmark tasks is the UniSpeech-SAT acoustic model (Chen et al., 2021), which includes additional objectives on top of the HuBERT model to facilitate speaker-aware representations, which also generally embody non-linguistic characteristics of utterances, such as tone and emotion.

3 Proposed Model

As the space of possible architectural or feature variations increases with the increasing volume of developments in the ML field, exhaustively searching all possible architectures is unfeasible. Therefore, we draw inspiration from the best performing models in related literature to reduce the search space and increase the likelihood of finding a best performing model. We selected two neural network architectures, ResNet and ECAPA-TDNN, for their potential in speech classification tasks. For feature extraction, we compare classical MFCC features with the pre-trained UniSpeech-SAT large acoustic model (Chen et al., 2021) that has been shown to provide consistent improvements in various Speech classification benchmarks. Finally, as best models in previous works typically include a form of ensemble, we experimented with fusing all model variants to further improve performance. We describe the details of these parts in this section.

3.1 Feature extraction

We experimented with two types of features: classical acoustic features, namely MFCCs, and modern acoustic features extracted from a large pre-trained acoustic model, namely the Universal Speech representation learning with speaker-aware pre-training (UniSpeech-SAT) (Chen et al., 2021). The large variant of this model demonstrated outstanding performance in various tasks in the SUPERB benchmark (wen Yang et al., 2021), including linguistic and non-linguistic tasks, such as speaker diarization and emotion recognition. UniSpeech-SAT model is built on the HuBERT model (Hsu et al., 2021) with additional self-supervised objectives involving utterance-wise contrastive learning and utterance mixing augmentation. The speaker-aware pre-training enabled the model to improve the discriminating capabilities of embeddings learned under self-supervised learning. In total, the large variant of UniSpeech-SAT was trained on 94K hours of English speech data from various sources, including Audiobooks and YouTube. We extracted 1024-dimensional features from the pre-trained UniSpeech-SAT[1] model and kept model parameters frozen. For MFCCs, we extract 80-dimensional features using a window length of 25 ms with a sliding window of 10 ms and frame-level instance normalization.

3.2 Network architectures

We experimented with two network architectures that have been shown to work well in speech classification tasks: ResNet and ECAPA-TDNN, which we describe below.

3.2.1 ResNet

We use the ResNet architecture (He et al., 2015) as our first model. Our model is composed of four residual networks, each consisting of two convolutional layers in addition a skip connection. We utilize batch normalization and ReLU activation functions. Statistical pooling is implemented to map the variable length feature frames to a time-invariant representation by aggregating frame level mean and variance as statistical parameters. The output of statistical pooling is followed by two feed-forward layers. We employ the original ResNet34 set-up as described in the original paper (He et al., 2015), which has 34 2D-convolutional layers organized into 4 residual network blocks, with each block containing a specific number of layers [3, 4, 6, 3], and the convolutional filters for these layers are [32, 64, 128, 256] respectively. The last feed-forward layer includes the output dimension of a number of dialect classes to identify with Additive Angular Margin (AAM) softmax layer (Deng et al., 2018) with a scale of 30.0 and margin of 0.4, trained with cross-entropy loss function.

3.2.2 ECAPA-TDNN

The ECAPA-TDNN architecture (Desplanques et al., 2020), based on the x-vector architecture (Snyder et al., 2018), utilizes a Squeeze-excitation (SE)-Res2Net module in each block. These modules consist of 1-dimensional convolutional layers, ReLU activation, batch normalization, and 1-dimensional Res2Net modules with impactful skip connections and SE blocks. This design allows the model to extract hierarchical and global information from the input features. Additionally, the architecture incorporates attentive statistical pooling by calculating channel-dependent frame attention-weighted statistics (mean and variance). This process transforms variable-length hidden outputs into a time-invariant representation. The representation is further processed through feed-forward layers. Similar to the ResNet architecture, we use the AAM-softmax as the final layer and train it with the cross-entropy loss criterion. The model uses 512 channels in 1-dimensional convolutional layers, 128 dimensions for SE-Block and attention, and a scaling factor of 8 for each Res2Block. The output dimension for feed-forward layers is set to 192, and the last feed-forward layer's dimension corresponds to the number of dialect classes.

3.3 Inference Scheme

In our model, we integrate a similarity measure with our learned classifiers to enhance classification performance (Lee et al., 2012; Nguyen et al., 2013; Roul and Arora, 2017). ResNet and ECAPA-TDNN are optimized for dialect identification via softmax, which we augment with a similarity-based measure based on the final embeddings produced by the network. For each dialect class, we randomly extract a cohort of 500 samples from the training set, and we calculate the average cosine similarity score between the test utterance and the cohort representing each class. After normalizing the scores, we combine them with the softmax scores by averaging them with equal weight (0.5) and selecting the class with the maximum score.

[1] https://github.com/microsoft/UniSpeech

4 Experimental setup

4.1 Datasets

We evaluate the dialect identification model on two Arabic dialect identification tasks: the MGB-3 ADI-5 dataset (Ali et al., 2017), and the fine-grained MGB5 ADI-17 dataset (Ali et al., 2019). ADI-5 training set consists of 13,825 utterances (53.6 hours), and the test and development sets consist of 1,524 (10 hours) and 1,492 (10 hours) utterances, respectively, with each set having approximately 2 hours of data per dialect class: Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA). In ADI-17, approximately 3,000 hours of training data were labeled via distant supervision into 17 dialect classes using the origin country of the YouTube videos from which they were extracted. The testing and development sets contain ∼25 and ∼33 hours of speech, respectively, manually verified by human annotators.

4.2 Data Augmentation

For data augmentation, we apply additive noise drawn from the Music, Speech, and Noise corpus (MUSAN) (Snyder et al., 2015) and the QMUL impulse response dataset (Stewart and Sandler, 2010). We also apply speed perturbation, where the tempo is modified by factors of 0.9 and 1.1. All noise augmentation was implemented using the Kaldi toolkit (Povey et al., 2011).

4.3 Training settings

During the training phase, each model was initially trained with randomly selected 5-second segments from training utterances for the first 50 epochs. Subsequently, the duration of the training segments was reduced to 4 seconds for a total of 100 epochs to enable the model to generalize to short-duration utterances. All systems were trained using the Adam optimizer with a triangular learning scheduler policy and a batch size of 256.

5 Results

Tables 1 and 2 show the performance of our model variants in ADI-5 and ADI-17 test sets, respectively. *Fusion* refers to an ensemble model where scores from all four variants are combined, each with an equal weight of 0.25. We also show the performance of the best performing models from the original challenges, which have not been previously outperformed to the best of our knowledge.

Table 1: Performance evaluation on MGB-3 ADI-5 test set (in %) with baseline systems submitted to MGB-3 challenge. UniS denotes the UniSpeech-SAT feature extraction.

System	Features	Accuracy	Precision	Recall
Best systems from (Ali et al., 2017)				
MIT-QCRI	—	75.0	75.1	75.5
UTD	—	70.4	70.8	71.7
ResNet	MFCC	74.2	74.1	74.4
ECAPA	MFCC	75.3	75.1	75.3
ResNet	UniS	80.4	80.4	80.5
ECAPA	UniS	82.5	82.6	82.7
Fusion	—	**84.7**	**84.8**	**84.9**

Table 2: Performance evaluation on MGB-5 ADI-17 test set (in %) with baseline systems submitted to MGB-5 challenge. UniS denotes the UniSpeech-SAT feature extraction.

System	Features	Accuracy	Precision	Recall
Best systems from (Ali et al., 2019)				
DKU	—	94.9	94.9	94.9
UKent	—	91.1	91.1	91.1
ResNet	MFCC	90.1	90.1	90.1
ECAPA	MFCC	92.2	92.2	92.2
ResNet	UniS	95.7	95.7	95.7
ECAPA	UniS	96.1	96.1	96.2
Fusion	—	**96.9**	**96.9**	**96.9**

We observe consistent results in both datasets: ECAPA-TDNN network consistently outperforms ResNet, and the models using UniSpeech-SAT features consistently outperform those using MFCC features. Incorporating these pre-trained features results in 4% to 5% absolute improvement in accuracy for both models. We observe additional gains of 0.8% to 2% improvement in absolute accuracy by fusing all four model/feature combinations. The highest performance gain is observed by using UniSpeech-SAT features as input, which leads to outperforming all previous baselines.

6 Conclusions

This paper described variations of model architectures, namely ResNet and ECAPA-TDNN, employing two acoustic features: classical MFCCs and self-supervised UniSpeech-SAT, leading to state-of-the-art performance in two spoken Arabic dialect identification benchmarks: ADI-5, and ADI-17. UniSpeech-SAT features, which are extracted from a large pre-trained model optimized for acoustic and speaker variability, consistently demonstrated superior performance compared to MFCC features. Despite being pre-trained solely in English speech, UniSpeech-SAT illustrates transfer learning capability by extracting suitable feature representations for this discriminative task in the Arabic language. This may also indicate that non-linguistic acoustic variability (such as speaking tone, for example) could play a role in dialect identification. Consistent with previous models from the MGB-3 and MGB-4 challenge, fusing multiple models results in consistent improvements of overall performance.

7 Limitations

In this work, we limited our analysis and exploration to two network architectures and two types of acoustic features. We based our choice on observations from the current literature on dialect identification, speech classification, and self-supervised acoustic models. However, many additional features and architectural variations could have been explored, with additional detailed analysis of the different combinations. Furthermore, we did not analyze the acoustic features that are most discriminative in these datasets, which is a complex analysis that eludes us at this stage, but future work could explore more on which aspects of an utterance (linguistic, tonal, other) are most useful for dialect identification.

References

Ahmed M. Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James R. Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033.

Ahmed M. Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2278–2282. ISCA.

Ahmet Emin Bulut, Qian Zhang, Chunlei Zhang, Fahimeh Bahmaninezhad, and John H. L. Hansen. 2017. Utd-crss submission for mgb-3 arabic dialect identification: Front-end and back-end advancements on broadcast speech. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 360–367.

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, and Xiangzhan Yu. 2021. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6152–6156.

Jiankang Deng, J. Guo, and Stefanos Zafeiriou. 2018. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Sameer Khurana, Maryam Najafian, Ahmed M. Ali, Tuka Al Hanai, Yonatan Belinkov, and James R. Glass. 2017. Qmdis: Qcri-mit advanced dialect identification system. In *Interspeech*.

Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, and Dino Isa. 2012. An enhanced support vector machine classification framework by using euclidean distance function for text document categorization. *Applied Intelligence*, 37:80–99.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023. Towards spoken dialect identification of irish. *arXiv preprint arXiv:2307.07436*.

Xiaoxiao Miao and Ian Mcloughlin. 2019. Lstm-tdnn with convolutional front-end for dialect identification in the 2019 multi-genre broadcast challenge. *ArXiv*, abs/1912.09003.

Tam T Nguyen, Kuiyu Chang, and Siu Cheung Hui. 2013. Supervised term weighting centroid-based classifiers for text categorization. *Knowledge and information systems*, 35:61–85.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit.

In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Rajendra Kumar Roul and Kushagr Arora. 2017. A modified cosine-similarity based log kernel for support vector machines in the domain of text classification. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 338–347.

Suwon Shon, Ahmed M. Ali, and James R. Glass. 2017. Mit-qcri arabic dialect identification system for the 2017 multi-genre broadcast challenge. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 374–380.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Rebecca Stewart and Mark Sandler. 2010. Database of omnidirectional and b-format room impulse responses. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 165–168. IEEE.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the Robustness of Arabic Speech Dialect Identification. In *Proc. INTERSPEECH 2023*, pages 5326–5330.

Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li, and Eng Siong Chng. 2006. Integrating acoustic, prosodic and phonotactic features for spoken language identification. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Karthika Vijayan, Haizhou Li, Hanwu Sun, and Kong Aik Lee. 2018. On the importance of analytic phase of speech signals in spoken language recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5194–5198.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdel rahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Omar Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40:171–202.