Dialect classification using acoustic and linguistic features in Arabic speech

Mohammad Ali Humayun, Hayati Yassin, Pg Emeroylariffion Abas
Faculty of Integrated Technologies, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam

Article Info

Article history:

Received Jan 17, 2022
Revised Sep 27, 2022
Accepted Oct 27, 2022

Keywords:

Acoustic features
Arabic language
Dialect identification
Linguistic features
Spontaneous speech

ABSTRACT

Speech dialects refer to linguistic and pronunciation variations in the speech of the same language. Automatic dialect classification requires considerable acoustic and linguistic differences between different dialect categories of speech. This paper proposes a classification model composed of a combination of classifiers for the Arabic dialects by utilizing both the acoustic and linguistic features of spontaneous speech. The acoustic classification comprises of an ensemble of classifiers focusing on different frequency ranges within the short-term spectral features, as well as a classifier utilizing the 'i-vector', whilst the linguistic classifiers use features extracted by transformer models pre-trained on large Arabic text datasets. It has been shown that the proposed fusion of multiple classifiers achieves a classification accuracy of 82.44% for the identification task of five Arabic dialects. This represents the highest accuracy reported on the dataset, despite the relative simplicity of the proposed model, and has shown its applicability and relevance for dialect identification tasks.

Corresponding Author:

Mohammad Ali Humayun
Faculty of Integrated Technologies, Universiti Brunei Darussalam
Jalan Tungku Link, Brunei Darussalam
Email: mohammadalihumayun@gmail.com; 19h8937@ubd.edu.bn

1. INTRODUCTION

Speech variations sharing the same linguistic characteristics but differing only in pronunciation styles are referred to as accent variations, whereas variations in terms of linguistic as well as pronunciation characteristics of spontaneous speech are referred to as dialect variations. Speech processing task for the identification of speech dialects has multiple applications, including providing assistance in forensic investigation. In the presence of crime-related speech evidence, dialect profiling for forensics can be used to track down criminals. Moreover, automatic speech recognition (ASR) models can also be adapted according to the identified dialects for performance improvement. Adapting the speech recognition models for particular dialects improves the model performance as the universally trained global ASR models considerably underperform when faced with dialect variations. An effective method to handle dialect variations for ASR is to train multiple models for different dialects and then select the suitable speech recognition model after identifying the speaker dialect [1].

Dialect identification requires considerable linguistic differences with discrete boundaries between dialects and is particularly applicable to the Arabic language as Arabic dialects commonly have distinct divisions between the different dialects whilst sharing a common lexical inventory. The most widely considered dialects of the Arabic language are Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African (NOR), and modern standard Arabic (MSA) [2]. The Egyptian dialect is spoken in urban Egypt. North African is the dialect of the North-West African countries: Algeria, Libya, Morocco, and Tunisia. The Gulf dialect prevails in the countries of the Arabian Peninsula: Bahrain, Kuwait, Oman, Yemen, Saudi Arabia, UAE, and Qatar. Levantine is the dialect of Syria, Jordan, Palestine, and Lebanon, i.e., the countries of Levant. Finally, Modern Standard Arabic is the formal Arabic dialect used in the literature and informal settings. The MSA is also, arguably, considered identical to classical Arabic, which is the language of the Quran.

Classification models for spontaneous speech can utilize both acoustic and linguistic features. Most utterance-level acoustic classification models use fix-sized embeddings obtained by temporal modeling of the sequence of short-term spectral or cepstral features [3]–[7]. On the other hand, deep learning models pre-trained on large unlabeled datasets have proved most effective in capturing linguistic representations for sentence classification tasks [8]–[10].

Multiple speech datasets have been collected for Arabic dialects, and a number of dialect classification models have been proposed using these datasets. Ziedan *et al.* [11], [12] have collected a dataset of Arabic spontaneous speech from YouTube videos in three different dialects: the dialects of Levant, Egypt, and Arabian Peninsula, and have named the dataset as the spoken Arabic Regional archive (SARA). Subsequently, the authors have proposed a dialect classification model using acoustic cepstral features together with delta coefficients as input to the universal background Gaussian mixture model (UBM-GMM).

The language recognition and evaluation 2015 challenge (LRE 2015) has provided a dataset of 20 dialects from six different languages, including the Arabic language, as part of the language identification challenge. Gelly *et al.* [13] has proposed the fusion of deep neural network (DNN) on the phonotactic and lexical features, recurrent neural network (RNN) on the acoustic short-term features, and probabilistic linear discriminant analysis (PLDA) on the acoustic i-vector features to classify the different dialects in the LRE 2015 dataset. The fusion of these different classifiers has been shown to achieve a close-set multi-language cost function score of 0.075, an evaluation metric provided by the LRE challenge.

Ali *et al.* [14] have collected an Arabic dataset of spontaneous speech from broadcast news in five different Arabic dialects: EGY, NOR, GLF, LAV, and MSA. Subsequently, a dialect classification model fusing the support vector machine (SVM) outputs of the acoustic i-vectors and singular value decomposition (SVD) of lexical term frequency-inverse document frequency (TF-IDF) has been proposed, which demonstrates 60.2% classification accuracy on the dataset. Najafian *et al.* [15] have used another Arabic corpus [16] with the same five Arabic dialects to evaluate their dialect classification model. The proposed classification model applies convolution neural networks (CNNs) for phoneme sequence, phoneme durations, and phoneme probabilities, before fusing the CNN outputs with SVM output for the ‘i-vector’ to achieve 73.27% classification accuracy.

Ali *et al.* [17] have presented the Arabic dialect identification (ADI) dataset as part of the 3rd multi-genre broadcast (MGB-3) challenge; consisting of the same five Arabic dialects. For benchmark results, the authors have applied SVM to i-vectors as well as bigram-word vector space and obtained a classification accuracy of 57.20%. Many proposals have utilized the ADI dataset in recent literature for various dialect classification techniques. Shon *et al.* [18] has proposed the siamese network to reduce the dimensions of i-vectors by trying to minimize the distance in the reduced dimensions for the same category vectors. The SVM classifiers for the lower-dimensional i-vectors, and the vector space of characters and phonemes, have been subsequently fused for the classification task. Classification accuracy of 75% has been reported on the ADI dataset. Khurana *et al.* [19] have applied CNN directly to the sequence of embedded words, characters, and phonemes as well as the short-term mel frequency cepstral coefficients (MFCC) acoustic features before fusing the CNN outputs with the SVM for i-vector, to give a classification accuracy of 73%. Shon *et al.* [20] have also proposed to use CNNs for acoustic features: MFCC, Filter-Bank (FBANK), and spectrograms by employing volume and speed perturbation for the augmentation of the input speech. The acoustic CNNs are then fused with the CNNs for the character, word, and phoneme sequences, as well as the probabilistic linear discriminant analysis (PLDA) of the ‘i-vectors’, to achieve a final fused classification accuracy of 81.36%. The same ADI dataset has been used in this paper.

This paper analyzes the task of classifying Arabic dialects (EGY, LAV, GLF, NOR, and MSA) from a spontaneous Arabic speech dataset. A classification model based on the fusion of linguistic and acoustic features has been proposed. For the acoustic dialect classification, the method presents a novel ensemble of convolutional neural networks (CNNs), applied to different ranges of the filter Bank (FBANK) energies. To extract the linguistic features, the method uses pre-trained bidirectional encoder representations from transformers (BERT) models [8], without fine-tuning, which is the first time that the BERT model has been analyzed for Arabic speech dialect identification tasks without fine-tuning. The proposed model has demonstrated the highest accuracy, with a much simpler and resource-efficient model, as compared to the existing techniques in recent literature for the same dataset [20]. The next section introduces the proposed method and explains the architecture details for both the linguistic and acoustic classifiers, whilst the third section presents the results and discussion. The last section concludes the paper and lays out possible future directions for this study.

2. METHOD

The proposed method for dialect classification utilizes both acoustic and linguistic speech features. The acoustic model ensembles multiple classifiers for different FBANK frequency ranges, in addition to simple averaging of classifier outputs for multiple fix-sized time frames of acoustic features segmented from the complete utterance. In contrast to the benchmark model [20] for the same acoustic features, the proposed model does not employ data augmentation and thereby considerably reduces the system memory requirement. Additionally, the neural network architecture is also relatively simpler than the benchmark model, with far fewer parameters. For linguistic classification, the proposed method uses two different pre-trained BERT models to obtain sentence representations, which are in turn used to train a simple feed-forward classifier for dialect prediction. The proposed method does not consider word, character, or phone-level linguistic features. Figure 1 illustrates the proposed classification method, which fuses both linguistic and acoustic classifications.

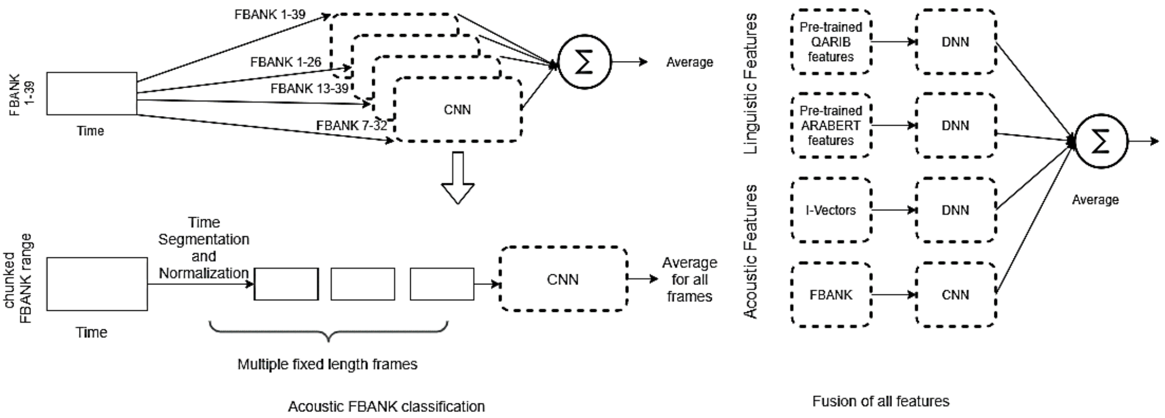

Figure 1. The proposed dialect classification model

2.1. Acoustic model

For the acoustic classification, CNNs are used on the time series of Mel-scaled FBANK energies for the speech utterances. The convolution kernel runs across the time axis of the FBANK vector frames. To focus on different frequency ranges, the classification uses an ensemble of four CNNs, with each CNN considering different bands of the frequency bins. Separate CNNs are trained using the lower, middle, and higher ranges of the FBANK energies, as well as using the complete range of the FBANK energies. As such, each classifier focuses on accent-specific characteristics from different frequency ranges of the speech signal. The outputs of the multiple classifiers focusing on different frequency ranges are then averaged to get the ensemble output.

The sequences of FBANK vectors for each range have been further segmented into multiple, smaller frames of fixed time durations, with the average of output predictions from the CNN classifiers for all the frames of a speech utterance giving the utterance level classification. Smaller duration frames allow the classifier to focus on accent-specific characteristics from a localized time segment. Moreover, segmentation also increases the number of samples to train the classifier.

The speech signals for the proposed model are initially transformed into the log FBANK energies with 39 filters applied to the frequency spectrum for 30 milliseconds sliding windows with a hop size of 20 milliseconds. The FBANK hop length is double the 10ms hop length used by the benchmark model [20] for the same FBANK features, resulting in half the length of the feature vector series and thereby reducing the system memory requirement. Subsequently, frequency bins of 1-26, 7-32, 13-39, and 1-39 from the FBANK energies are used as input for training the four different classifiers. The final predictions for each speech utterance are calculated by averaging the classifier outputs for the four different FBANK ranges.

For each of the CNN inputs, the FBANK feature series is segmented into multiple frames, with the length of each frame equal to the length of the smallest utterance within the training set. The fixed length of the smallest frames for experiments in this paper was 81 samples, with a speech sampling rate being 16 kHz. The frames are extracted with a sliding hop length of almost half of the frame length, i.e. after every 40 samples, with each feature frame normalized by its mean and variance. During testing, the classification outputs of all frames for a particular speech utterance are averaged to get the utterance level classification.

The segmented frames are fed into the CNN classifiers for dialect classification. Each CNN consists of two 1 dimensional (1D) convolutional layers, with the convolutions running across the time axis of feature frames. The two 1D convolutional layers have 256, and 512 filters and kernel sizes 4 and 2, respectively. The stride length for both the convolutional layers is one. Each convolutional layer is followed by rectified linear unit (ReLU) activation and max-pooling from two samples. The two convolutional layers are followed by a global average pooling layer, which collapses the time axis into the 512 filters from the 2nd convolutional layer. Output from the global average pooling is fed to a dense, fully connected layer with 256 units and ReLU activations. The dense layer is then connected to the output layer for classification with five softmax units corresponding to the five dialects to be classified. The model is trained with Adamax optimizer [21], with training stopped when the training error does not reduce over 1 epoch. Generally, the employed CNN architecture has far fewer parameters than the benchmark FBANK model [20], making it more resource-efficient.

i-vectors computed in [14] has also been used to aid the classification. The 400-dimensional i-vectors are fed to a feed-forward neural network for the dialect classification task, with the feed-forward neural network consisting of a single hidden layer with 192 units, ReLU activations, and batch normalization, followed by the softmax output layer with 5 units corresponding to the five dialects to be classified. Similar to the CNN employed for the FBANK features, the training uses an Adamax optimizer, with early stopping if training error stops reducing over 1 epoch.

2.2. Linguistic model

Linguistic classification is based on sentence-level embeddings obtained using the two different pre-trained transformer models for the Arabic language: QARIB-base [22] and AraBERTv0.2-base [23]. Both the models use the base BERT architecture [8] and have been trained on different Arabic datasets. As such, the models capture different linguistic feature representations of the sentences. The models consist of 12 encoder blocks with 12 self-attention heads followed by hidden layers of size 768, with the base model having a total of 110 million parameters and allowing a maximum sequence length of 512 for a sentence. Each Bert model has a particular tokenizer that segments the sentences to obtain a numeric embedding for each segment by considering its vocabulary index and position in the sentence. A classification (CLS) token is added at the start of each sentence for sentence-level embedding.

A mix of formal and informal Arabic datasets has been utilized by the QARIB model for its pre-training. The formal data was extracted from news websites and movie subtitles, whereas the informal data comprises tweets extracted using Twitter streaming application programming interface (API) with the language filter set to Arabic. The pre-training dataset comprises a total of 420 million tweets and 180 million sentences, resulting in 14 billion tokens and a vocabulary size of 64k. The model has been pre-trained by the unsupervised task of predicting masked words, with 15% of the words masked for prediction during pre-training. The next sentence prediction has not been considered for pre-training the QARIB model as the shorter length of the tweets somehow makes the sentences less correlated to each other. On the other hand, the AraBERT model has been pre-trained for both masked word prediction and next sentence prediction tasks on Arabic texts, which have been collected from Wikipedia and news websites. Similar to the pre-training of the QARIB model, 15% of the words have been masked for prediction during the masked word prediction training. Fine-tuning the BERT models for the Arabic dialect identification task has been experimented with but has not shown performance improvement and hence, has not been implemented in this paper.

Arabic sentences are first tokenized using the relevant tokenizer for the particular transformer model, such that out-of-vocabulary words are segmented down to sub-words or alphabets, which are available in the vocabulary for the particular model. The tokens for the sentences from the training as well as test partitions are transformed into a 768-dimensional representation in the last hidden layer of the transformer model. Since the first CLS token represents the sentence, the transformer output for only the first token is retained as the representation or features for the sentence-level dialect classification. Extracted transformer features from the training and test sets using both the QARIB and AraBERT models are used to train two separate, supervised feed-forward neural network classifiers for the dialect classification task. Both feed-forward neural networks have a single hidden layer of 192 units and linear activations followed by a Softmax output layer. Adamax optimizer has been used, with training stopped when optimization error stops decreasing.

2.3. Fusion system

The final result is obtained by simple averaging of the outputs from the four different classifiers: one for the FBANK features, the second for i-vector, the third for AraBERT based linguistic representation of sentences, and the fourth for the QARIB sentence representations, as shown in Figure 1. The final class prediction for each speech utterance is taken as the class with the maximum probability value in the averaged output.

2.4. Dataset

To analyze the performance of the proposed fusion method, the Arabic dialect identification (ADI) dataset [17] has been used. The dataset had been compiled from audio recordings of Al-Jazeera programs, and it contains a mix of spontaneous and scripted speech, with the samples belonging to one of the five main Arabic dialects: EGY, NOR, GLF, LAV, or MSA. Table 1 gives the number of sentence samples for each of the dialect categories in the training and test sets. The partitions named as 'training' and 'development' from the original dataset [17] have been merged to constitute the extended training set for experiments in related works [19], [20] as well as for this paper.

Table 1. Sentence samples in the ADI dataset

Dialect	Training	Testing
EGY	3391	302
GLF	3008	250
LAV	3181	334
MSA	2464	262
NOR	3305	344
Total	15349	1492

For the linguistic features, sequences of words from transcripts corresponding to speech utterances have been used. Transcripts for the ADI dataset is available at an online repository [24], with the transcripts generated using an Arabic speech-to-text translation model trained on a separate dataset from the MGB-2 challenge [25]. Similarly, the i-vectors for the speech utterances have also been obtained from the online repository [24]. The i-vectors have been extracted by a model trained on another speech dataset [14]. On the other hand, the FBANK features have been computed for this paper using the available audio files at the links provided in the repository [24].

3. RESULTS AND DISCUSSION

Classification performance has been analyzed in terms of classification accuracy by individual classifiers as well as the overall fusion model. Table 2 shows classification accuracies by the CNN classifiers for the different FBANK ranges, as well as the overall accuracy of the ensemble of these CNN classifiers. Amongst the different individual classifiers, CNN with inputs covering the complete range of FBANK filters gives the highest accuracy of 59.58%. This is followed by CNN with inputs covering FBANK ranges of 1-26, 7-32, and 13-39, with accuracies of 57.98%, 55.43%, and 53.82%, respectively. The classification accuracy of the ensemble of these CNNs is 64.54%, which is considerably higher than the accuracy for each of the individual CNNs. This indicates that focusing separately on the smaller ranges of FBANK allows the capture of complementary information, which, when combined, gives an overall improvement in results.

Table 2. Classification accuracy by CNNs for different FBANK ranges

Features	Accuracy %
FBANK 1_39	59.58
FBANK 1_26	57.98
FBANK 13_39	55.43
FBANK 7_32	53.82
Ensemble	64.54

Table 3 indicates the classification accuracies of the different classifiers using the different acoustic and linguistic features, as well as the accuracies obtained by fusing these different classifiers. The feed-forward DNN using the 'i-vectors' achieves the highest individual accuracy of 70.17%. The fusion of the two classification models utilizing acoustic features: FBANK and i-vector, is referred to as acoustic fusion and is shown to give 74.8% accuracy. For the classification by linguistic features, the DNN utilizing transformed features from the pre-trained QARIB and AraBERT transformers give classification accuracies of 61.93% and 66.62%, respectively. Fine-tuning the pre-trained transformer models on the ADI training dataset has not demonstrated an improvement in classification accuracy, and hence, only results obtained from pre-trained QARIB and AraBERT without fine-tuning have been reported. Classification accuracy by fusing classification models which utilize linguistic features is 68.63%. Fusing all the four classification models via simple averaging gives an overall classification accuracy of 82.44%. The higher classification accuracy from the fusion of the classification models, in contrast to the accuracies of the individual classifiers, indicate that the different features used capture complementary information. Subsequently, combining the outputs from these different classification models result in higher overall accuracy.

Table 4 compares the classification accuracy of the proposed fusion system with the classification accuracies of other fusion methods on the same ADI dataset. It can be seen that the proposed fusion model gives the best accuracy of 82.44% on the ADI dataset. This is an improvement of 1.08% over the fusion method given in [20], despite the proposed method having a simpler architecture and far fewer parameters. The proposed models in [26] and [15] also have much larger CNN architectures and have reported final fusion accuracies of 79.76% and 73.27%, respectively, which are considerably lower than the proposed model.

Table 3. Classification accuracy for linguistic and acoustic features

Features	Accuracy %
'QARIB'	61.93
'AraBERT'	66.62
'FUSION_LINGUISTIC'	68.63
'FBANK'	64.54
'IVEC'	70.17
'FUSION_ACOUSTIC'	74.80
'FUSION_ALL'	82.44

Table 4. Classification accuracy comparison for ADI dataset

Proposal	Accuracy %
Khurana *et al*. [19]	73.00
Shon *et al*. [18]	75.00
Najafian *et al.* [15]	73.27
Bulut *et al*. [26]	79.76
Shon *et al*. [20]	81.36
The proposed model	82.44

Figure 2 illustrates the f-score values achieved by the four individual classifiers, the fusion of classifiers utilizing the linguistic and acoustic features, as well as the fusion of all the classifiers over the five Arabic dialects. F-score gives the harmonic mean of precision and recall, where precision is defined as the ratio of true predictions to total predictions, whilst recall is defined as the ratio of true predictions to the total number of test samples for each class. It can be seen that the fusions of classification models utilizing either the acoustic or linguistic features result in higher f-score values as compared to the f-score values of the individual classification models, for all Arabic dialects, with the exception of classification models utilizing linguistic features for MSA. Subsequently, the overall fusion of classification models utilizing both the acoustic and linguistic features gives the highest f-score value for all the Arabic dialects.

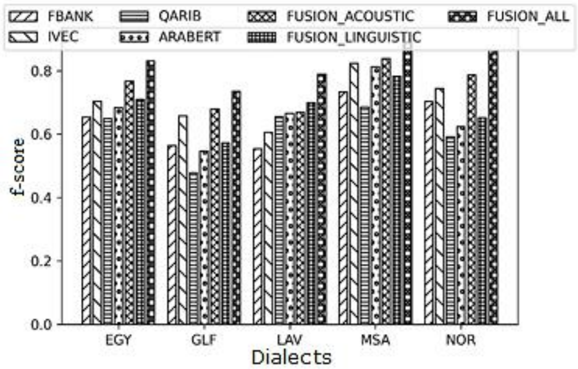

Figure 2. Class specific f-scores for different classifiers

4. CONCLUSION

Dialect classification requires both acoustic and linguistic differences in the different dialects. This paper proposes a dialect classification model based on the fusion of different acoustic and linguistic features-based classification models. The acoustic classification utilizes a novel ensemble of CNN classifiers focusing on different ranges in the FBANK energies. It has been shown that the ensemble of smaller FBANK range CNNs achieves considerably higher accuracy as compared to a single complete range CNN. For the linguistic classification, two different transformer models, pre-trained on different Arabic datasets, have been used without any fine-tuning to extract complementary input features for dialect classification. The proposed fused model has been demonstrated on a widely-used Arabic dialects dataset. The higher performance metric scores by the fusion systems as compared to the individual classification models confirms that the proposed model is capable of extracting complementary information to differentiate the different dialects. Accuracy by fusing the linguistic as well as acoustic-based classification models is obtained at 82.44%, which represents the highest reported accuracy on the ADI dataset. This demonstrates the applicability of the proposed method for the dialect identification task. In the future, different speech datasets shall be considered to study the applicability of the model on different datasets.

ACKNOWLEDGEMENTS

The first author would like to acknowledge Universiti Brunei Darussalam for providing the University Graduate Scheme (UGS) which enabled this study. This work was supported by UBD Research Grant: UBD/RSCH/URC/NIG/3.0/2022/002.

REFERENCES

[1] S. Singh, “The role of speech technology in biometrics, forensics and man-machine interface,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, p. 281, 2019, doi: 10.11591/ijece.v9i1.pp281-288.

[2] K. Alrifai, G. Rebdawi, and N. Ghneim, “Arabic tweeps dialect prediction based on machine learning approach,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1627–1633, 2021, doi: 10.11591/ijece.v11i2.pp1627-1633.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011, doi: 10.1109/TASL.2010.2064307.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN embeddings for speaker recognition,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol. 2018-April, pp. 5329–5333, 2018, doi: 10.1109/ICASSP.2018.8461375.

[5] A. I. Abdurrahman and A. Zahra, “Spoken language identification using i-vectors, x-vectors, plda and logistic regression,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2237–2244, 2021, doi: 10.11591/EEI.V10I4.2893.

[6] M. A. Humayun, H. Yassin, and P. E. Abas, “Spatial position constraint for unsupervised learning of speech representations,” *PeerJ Computer Science*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.650.

[7] M. A. Humayun, H. Yassin, and P. E. Abas, “Native language identification for Indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions,” *Speech Communication*, vol. 139, pp. 92–101, 2022, doi: 10.1016/j.specom.2022.03.007.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference*, vol. 1, pp. 4171–4186, 2019.

[9] M. A. Fauzi, “Word2Vec model for sentiment analysis of product reviews in Indonesian language,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, p. 525, 2019, doi: 10.11591/ijece.v9i1.pp525-530.

[10] D. E. Cahyani and I. Patasik, “Performance comparison of tf-idf and word2vec models for emotion text classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021, doi: 10.11591/eei.v10i5.3157.

[11] R. R. Ziedan, M. N. Micheal, A. K. Alsammak, M. F. M. Mursi, and A. S. Elmaghraby, “Improved dialect recognition for colloquial Arabic speakers,” *2016 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2016*, pp. 16–21, 2017, doi: 10.1109/ISSPIT.2016.7886002.

[12] R. R. Ziedan, M. N. Micheal, A. K. Alsammak, M. F. M. Mursi, and A. S. Elmaghraby, “A unified approach for Arabic language dialect detection,” *29th International Conference on Computer Applications in Industry and Engineering, CAINE 2016*, pp. 165–170, 2016.

[13] G. Gelly, J. L. Gauvain, L. Lamel, A. Laurent, V. B. Le, and A. Messaoudi, “Language recognition for dialects and closely related languages,” *Odyssey 2016: Speaker and Language Recognition Workshop*, pp. 124–131, 2016, doi: 10.21437/Odyssey.2016-18.

[14] A. Ali *et al.*, “Automatic dialect detection in Arabic broadcast speech,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, pp. 2934–2938, 2016, doi: 10.21437/Interspeech.2016-1297.

[15] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, “Exploiting convolutional neural networks for phonotactic based dialect identification,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, vol. 2018-April, pp. 5174–5178, 2018, doi: 10.1109/ICASSP.2018.8461486.

[16] S. Wray and A. Ali, “Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-January, pp. 2824–2828, 2015, doi: 10.21437/interspeech.2015-594.

[17] A. Ali, S. Vogel, and S. Renals, “Speech recognition challenge in the wild: Arabic MGB-3,” *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017-Proceedings*, vol. 2018-January, pp. 316–322, 2018, doi: 10.1109/ASRU.2017.8268952.

[18] S. Shon, A. Ali, and J. Glass, “MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge,” *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017 - Proceedings*, vol. 2018-January, pp. 374–380, 2018, doi: 10.1109/ASRU.2017.8268960.

[19] S. Khurana, M. Najafian, A. Ali, T. Al Hanai, Y. Belinkov, and J. Glass, “QMDIS: QCRI-MIT advanced dialect identification system,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*,

vol. 2017-August, pp. 2591–2595, 2017, doi: 10.21437/Interspeech.2017-1391.

[20] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *Speaker and Language Recognition Workshop, ODYSSEY 2018*, pp. 98–104, 2018, doi: 10.21437/Odyssey.2018-14.

[21] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[22] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training BERT on Arabic tweets: Practical considerations," 2021, [Online]. Available: http://arxiv.org/abs/2102.10684.

[23] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, [Online]. Available: http://arxiv.org/abs/2003.00104.

[24] A. Ali, "ADI dataset," 2017, [Online]. Available: https://github.com/qcri/dialectID.

[25] S. Khurana and A. Ali, "QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge," *2016 IEEE Workshop on Spoken Language Technology, SLT 2016-Proceedings*, pp. 292–298, 2017, doi: 10.1109/SLT.2016.7846279.

[26] A. E. Bulut, Q. Zhang, C. Zhang, F. Bahmaninezhad, and J. H. L. Hansen, "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 2017-Proceedings*, vol. 2018-January, pp. 360–367, 2018, doi: 10.1109/ASRU.2017.8268958.

BIOGRAPHIES OF AUTHORS

Mohammad Ali Humayun is a PhD Student in the Faculty of Integrated Technologies at Universiti Brunei Darussalam with research focusing on deep learning for speech and language processing. He holds an MSc in Telecommunication Engineering from the University of Engineering and Technology Peshawar and a BSc in Electronics Engineering from Ghulam Ishaq Khan Institute of Engineering, Topi, Pakistan. He can be contacted at email: mohammadalihumayun@gmail.com.

Hayati Yassin holds a joint appointment at the Faculty of Integrated Technologies and Faculty of Science, Universiti Brunei Darussalam; with research focusing on Biometrics, Neural Networks, Renewable Energy, and Image Processing. She holds a BSc in Computer Network Security from Birmingham City University, UK, as well as an MSc in Information Security and Biometrics, and a PhD in Electrical Engineering, from the University of Kent, Canterbury, UK. She can be contacted at email: hayati.yassin@ubd.edu.bn.

Pg Emeroylariffion Abas is an Assistant Professor at the Faculty of Integrated Technologies, Universiti Brunei Darussalam; with research focusing at Energy Efficiency, Internet of Things (IoTs), Image Processing, Signal Processing, Resource Allocation of Networks, Information Security, Photonic Crystal Fibre, and Data Analytics. He holds a BEng in Information System Engineering and a PhD in Communication from Imperial College, London. He can be contacted at email: emeroylariffion.abas@ubd.edu.bn.