*Natural Language Engineering* (2020), **26**, pp. 691–700
doi:10.1017/S1351324920000091

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

# Spoken Arabic dialect recognition using X-vectors

Abualsoud Hanani* and Rabee Naser

Electrical and Computer Engineering, Birzeit University, Palestine
*Corresponding author. E-mail: abualsoudh@gmail.com

(Received 5 November 2018; revised 13 July 2019; accepted 27 August 2019; first published online 4 May 2020)

**Abstract**
This paper describes our automatic dialect identification system for recognizing four major Arabic dialects, as well as Modern Standard Arabic. We adapted the X-vector framework, which was originally developed for speaker recognition, to the task of Arabic dialect identification (ADI). The training and development ADI VarDial 2018 and VarDial 2017 were used to train and test all of our ADI systems. In addition to the introduced X-vectors, other systems use the traditional i-vectors, bottleneck features, phonetic features, words transcriptions, and GMM-tokens. X-vectors achieved good performance (0.687) on the ADI 2018 Discriminating between Similar Languages shared task testing dataset, outperforming other systems. The performance of the X-vector system is slightly improved (0.697) when fused with i-vectors, bottleneck features, and word uni-gram features.

**Keywords:** X-vectors; Arabic Dialect Recognition

## 1. Introduction

In addition to the linguistic information, the speech signal contains much other information such as speaker identity, gender, emotion, language and accent, age and many others. These inter-speaker and intra-speaker variations can make distortion for some speech processing applications, such as automatic speech recognition (ASR).

Recognizing the language and dialect of the speaker automatically begins gaining more and more interests in the speech community. Dialect recognition has been viewed as more challenging than that of language recognition due to the greater similarity between dialects of the same language.

Dialect recognition can be used for identifying the geographical place/ethnic of the speaker. In addition, successfully recognizing speaker dialect can be used to overcome the dialect variation effect on the ASR performance. This is evident, for example, with the Arabic language, which has multiple dialects, including Modern Standard Arabic (MSA), the formal written standard language of the media, culture, and education, and the informal spoken dialects that are the preferred method of communication in daily life. Written dialectal Arabic has a strong presence in social media applications, such as Facebook, Twitter, WhatApp. These data make a good opportunity to set up statistical learning of the Arabic dialects. However, because all Arabic dialects use the same character set, and furthermore much of the vocabulary is shared among different dialects, it is not an easy job to distinguish and separate the dialects from each other.

Work on dialect recognition in the literature is still traditionally split into acoustic-only, acoustic-lexical, and acoustic-phonetic classification systems.

Most of the Arabic speech recognition applications were developed with a focus on MSA, while Arabic native speakers do not always use it in their daily lives. Historically, using the multilayer deep neural networks (DNNs) was not practical in ASR, because of their high computational needs. But the existence of many cores in Graphical Processing Units made it possible to utilize the DNNs in the speech processing field. DNNs performance is examined in identifying Arabic dialects.

The mentioned reasons motivate us to work on improving the results of recognizing Arabic dialects which help Arabic speech applications to perform better. In the context of the shared task for discriminating between Similar Languages (DSL), Varieties and Dialects[a], dialect identification can be seen as a multi-class sentence classification problem, in which participants must predict a label for each sentence, given several features describing the sentence. Arabic dialect identification (ADI) is a subtask of the DSL shared task. Given a short utterance of Arabic speech, the task is to discriminate MSA and the four main dialects of Arabic: Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR).

## 2. Arabic dialects

Arabic is the official language in more than 20 Arab countries and it is spoken by more than 250 million people. Arabic has different variants including MSA which is the formal written language and formal spoken language in the media, culture, and education. Habash (2010) states that the MSA is the official language of the MSA is syntactically, morphologically, and phonologically based on classical Arabic, the language of the Quran [Islams Holy Book].

However, MSA is not a daily life language. Instead, people use dialectal Arabic which is quite different in different regions. Arabic dialects are traditionally spoken and not written. However, nowadays, dialectal Arabic is heavily used in writing on social media and chatting applications.

The written form of MSA Arabic varies very little throughout the Arab world. It has been a useful tool for communication of values, for the transaction of business, and for literary artistry over the centuries. Billions of words of written standard Arabic can be found online, and many times as much exist in the libraries of the world. It is the medium in which Arabs learn to read and write, and it is one of the important threads which ties the Arabic community together.

But no native speaker begins with MSA Arabic. His first language, or her mother tongue, is one of the local dialects. And although educated Arabs may be able to express themselves in MSA Arabic, most prefer their local dialect. Few want to sound like a television news announcer.

The individual dialects differ from MSA Arabic in several ways. The most obvious is in the accent, or pronunciation of letter sounds. Each dialect has a series of sound changes that are often or usually applied to common words. For example, in Ramallah, the road to Jerusalem is known as /sharia Al'uds/ (MSA: /sharia Alquds/) because the qaaf is often pronounced as a hamza in an urban Palestinian dialect. Damascus has a street named for the 1967 revolution known as /sharia sawra/ (MSA: /sharia thawra/), because the sound of /thaa/ is replaced in the Shamy (Levantine) dialect by either /s/ or /t/.

There are differences in vocabulary. In Palestine, my husband /jozy/ replaces the MSA /zowjiy/. The numerical vocabulary is streamlined, so that fifteen becomes /khamastash/ instead of /khamas eashara/ (MSA).

There are differences in morphology. For example, in almost every dialect, the verb affixes have changed from MSA ones.

There are differences in grammar. For example, in Shamy dialect, /fiyh/ is one of the most common words. It comes from the MSA phrase in it /fiyhi/ (MSA), which, however, would be pronounced with a long final vowel, /fiyhow/ if it were intended to discuss actual containing of anything; instead, it means in this situation or there is.

Arabic dialects primarily can be divided into major categories based on geography and social class. Furthermore, each of the main dialects can be divided into subcategories and so forth. The following is only one classification (geo-linguistically) of the main Arabic dialects.

[a] http://alt.qcri.org/vardial2018/.

**Table 1.** The ADI data for VarDial 2018 shared task

| Dialect | Training | Development | Testing |
|---|---|---|---|
| Egyptian | 3177 | 315 | 1445 |
| Gulf | 2873 | 265 | 1397 |
| Levantine | 3117 | 348 | 1465 |
| North African | 3205 | 355 | 1324 |
| MSA | 2219 | 283 | 1206 |
| **Total** | **14,591** | **1566** | **6837** |

**Table 2.** The ADI data for VarDial 2017 shared task

| Dialect | Training | Development | Testing |
|---|---|---|---|
| Egyptian | 3093 | 298 | 302 |
| Gulf | 2744 | 264 | 250 |
| Levantine | 2851 | 330 | 334 |
| North African | 2954 | 351 | 344 |
| MSA | 2183 | 281 | 262 |
| **Total** | **13,825** | **1524** | **1492** |

- **Gulf Arabic** (Glf) includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman;
- **Iraqi Arabic** (Irq) is the dialect of Iraq. In some dialect classifications, Iraqi Arabic is considered a subdialect of Gulf Arabic;
- **Levantine Arabic** (Lev) includes the dialects of Lebanon, Syria, Jordan, Palestine, and Israel;
- **Egyptian Arabic** (Egy) covers the dialects of the Nile valley: Egypt and Sudan;
- **Maghrebi Arabic** covers the dialects of Morocco, Algeria, Tunisia, and Mauritania. Libya is sometimes included;
- **Yemenite Arabic** is often considered its own class;
- **Maltese Arabic** is not always considered an Arabic dialect. It is the only Arabic variant that is considered a separate language and is written with Latin script.

Each of these regional dialects can be divided into subdialects based on the social class of speakers. The most common three subdialects are city dwellers, peasants/farmers, and Bedouins.

In this paper, we did all of our experiments presented in this paper on the four main Arabic dialects (shown in Tables 1 and 2); Gulf (includes Iraqi dialect), Egyptian, Levantine, North African (Maghrebi and Libyan dialects), as well as MSA. Yemenite and Maltese dialects are not included in this study.

Although Arabic dialects and MSA differ phonologically, orthographically, morphologically, lexically, and syntactically, we focus on phonological differences in this paper. Since the Arabic dialects are not standardized, there is no one standard orthography for them. However, this does not affect our work presented in this paper since our interest is to use acoustic and phonotactic features to discriminate between the five main dialects without the dialectal transcription. Phonologically, MSA includes 28 consonants, 3 long vowels, 3 short vowels, and 2 diphthongs (/aw/and /ay/).

## 3. Related work

The work on dialect recognition in the literature is traditionally split into acoustic-only, acoustic-lexical, and acoustic-phonetic classification systems. Most of the state-of-the-art systems focus on acoustic methods. In the last decade, the dialect recognition work capitalized on the i-vector technique which represents each utterance by a low-dimensional vector estimated from the variability subspace (DeMarco and Cox 2013). In 2014, there is a clear move toward DNNs for modeling acoustic features (Du *et al.* 2014; Tüske *et al.* 2014). Moreover, in 2017 and 2018, the data augmentation is used to improve the performance of DNN embeddings for speaker and language recognition. The DNN, which is trained to discriminate between speakers/languages, maps variable-length utterances to fixed-dimensional embeddings that are called X-vectors (Snyder *et al.* 2018b).

Recently, spoken and written dialectal Arabic has been significantly increased with the spreading use of social media. In 2013, Elfardy and Diab (2013) built a sentence-level ADI system which identifies the dialect from the input text dialectal Arabic sentence. Zaidan and Callison-Burch (2014) used Arabic Online Commentary for training and evaluating a system that recognizes the Arabic dialect from the given sentence.

The increasing interest in Arabic dialect recognition motivates related Natural Language Processing tasks such as ADI (Zaidan and Callison-Burch 2014). To tackle this challenge, from 2016 the DSL shared task has proposed a dialect identification subtask with multidialectal Arabic data based on audio files accompanied with dialect labels. Best performances have so far been reached by support vector machine (SVM), kernel ridge regression (KRR), and other sophisticated classifiers such as ensemble methods. However, in this section, we focus on describing the performance of neural networks in previous editions.

In the 2016 edition (Malmasi *et al.* 2016), each utterance is represented by a word sequence transcription obtained by an automatic Arabic speech recognition system described in Ali *et al.* (2014). The best presented F1 scores were ranging from 0.495 to 0.513.

In our previous work submitted to the DSL 2016 shared task, we combined word-level features (uni-grams) with different kinds of acoustic and phonotactic features at features level and at the system level for recognizing Arabic dialects. Concatenating different features together into one feature vector prior to system training improved performance compared with the system that uses one kind of these features. On the other hand, a training dialect identification system using only one kind of features and then combining scores of different systems together outperformed system trained on combined features.

The best performing systems obtained F1 scores ranging from 0.495 to 0.513. Three teams reported experiments with neural network architectures. However, the systems finally submitted by these teams were based on machine learning that obtained higher accuracy scores, QCRI (Eldesouki *et al.* 2016), GW LT3 (Zirikly, Desmet, and Diab 2016), and tufbasfs (Çöltekin and Rama 2017).

## 4. Data description

The dataset used in all of the presented experiments in this paper had been taken from the free available datasets for both shared task VarDial 2017[b] and VarDial 2018[c].

These speech data were collected from the Broadcast news domain Aljazeera channel, in four Arabic dialects (EGY, LEV, GLF, and NOR) as well as MSA. The speech was recorded at 16 KHz sampling frequency. Nonspeech segments, such as music and background noise, have been removed. In addition, speaker overlap was avoided by segmentation recordings into short chunks. This dataset has been divided into three subsets: train subset, development subset, and testing subset. Although all recordings of these subsets were collected from the same broadcast domain,

[b] http://ttg.uni-saarland.de/vardial2017/index.html.
[c] http://alt.qcri.org/vardial2018/.

https://doi.org/10.1017/S1351324920000091 Published online by Cambridge University Press

the recording setup is different. The testing subset recordings were downloaded directly from Aljazeera high-quality video server (bright-cove) in the period between July 2014 to January 2015, as part of QCRT advanced transcription service (QATS) (Ali, Zhang, and Vogel 2014).

The data were labeled using the crowdsource platform CrowdFlower, with the criteria to have a minimum of three judges per file and up to nine judges, or 75 percent inter-annotator agreement (whichever comes first). More details about the dataset and crowdsourcing experiment can be found in Wray and Ali (2015). The test subset has been collected from different channels, and the recording setup is different from the training data. This makes the experiments less sensitive to channel/speaker characteristics.

As shown in Table 1, the training dataset consists of 14,591 utterances, the development dataset consists of 1566 utterances, where the testing dataset consists of 6837 utterances; 5435 testing utterances were added to the original 1492 testing utterances from the mgb_3 dataset.

## 5. System description

Similar to any traditional classification system, our Arabic dialect recognition system consists of two main components: feature extraction followed by analysis carried out by the model. Given a short speech sample, a set of representative features (acoustic, lexical, and phonetic) are extracted and then used to find the best match with pretrained dialect-dependent models. The dialect model which gives the best match is the recognized dialect of the speaker. Various types of features and modeling techniques are employed in our proposed system for the Arabic dialect recognition task. The following subsections describe these features and modeling techniques.

### *5.1 Acoustic i-vectors*

The prior knowledge of the speaker dialect is not required for the acoustic features. The traditional and well-studied method for extracting effective acoustic features is the i-vector approach (Dehak *et al.* 2010). I-vectors are based on the Gaussian mixture model and universal background model (GMM-UBM) system described in Hanani, Qaroush, and Taylor (2017).

The acoustic feature vectors are based on 19 mel-frequency cepstral coefficients (MFCCs), derived from the log power output of 19 filters, with a frame length of 20ms window processed at 10-ms frame rate.

Shifted delta cepstra with 7-3-1-7 configuration (Torres-Carrasquillo *et al.* 2002) is computed and appended to the 19 MFCC feature vectors resulting in feature vectors with dimension equal to 68. RASTA filtration is applied to the power spectra. A simple energy-based voice activity detection was performed to discard the nonspeech frames. Cepstral mean and variance normalization was applied to the resulting 68-dimensional feature vectors.

A UBM is 2048-component GMM trained on the acoustic features (68 feature vectors) extracted from all training dataset of all Arabic dialects. The K-means clustering algorithm is used for finding initial parameters of UBM GMM (means, diagonal covariance matrices, and weights). The extracted i-vectors are 400-dimensional vectors.

### *5.2 Bottleneck features*

DNN consisting of multiple interconnected layers between the input and output layers were used in speech recognition tasks (Najafian *et al.* 2018). DNNs had not been used just as classifiers but they were also used as feature extractors (Ali *et al.* 2015). Neural networks can be configured to have multiple layers with large numbers of neurons, and between these layers, a much smaller layer can be added as illustrated in Figure 1. The state of this layer can be used as a feature vector to represent the original data. The resulted bottleneck features can be used to feed another learning model or another DNN.

A five-layer DNN was used to extract 60-dimensional bottleneck features from each utterance.

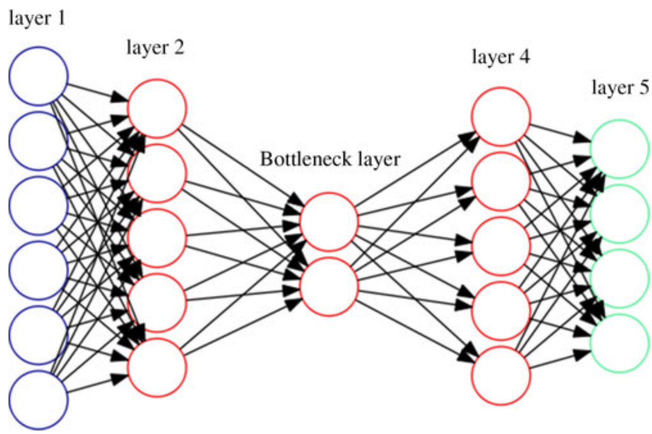


**Fig. 1.** Bottleneck features.

### *5.3 Phonetic features*

Phoneme sequence modeling has been successfully used for both language and dialect recognition which focuses on the distribution of allophone sequences (Hanani, Russell, and Carey 2013). A phone recognizer is used to convert each utterance into a sequence of phones. Phone-level n-gram phonotactic features are computed for each utterance and then used to train SVMs for discriminating between target dialects. Four phone recognizers, English, Russian, Hungarian, and Czech, free available from speech processing group at Brno university of technology[d], are tried and the Czech phone recognizer was found to work slightly better than the others, hence, used in all of our experiments presented in this paper. A weighting technique, used in our previous work (Hanani *et al.* 2017), is applied to the resulted n-gram probabilities to emphasize the most discriminate components (i.e., those which are common in one dialect but not in the other dialects). Uni-gram features are concatenated with bi-grams to form 6252-dimensional feature vectors used to train a multi-class SVM classifier.

### *5.4 GMM tokenization*

We used the same 2048-Gaussian UBM model described in Section 5.1, as a tokenizer, that converts a sequence of acoustic features (MFCCs) into a sequence of the Gaussian components indices which gives the highest probability for each frame. Comparing with the phonotactic system, the Gaussian component with the highest probability is recognized as the phoneme of the frame. N-gram probabilistic vector is extracted for each utterance in the same way described for the phonetic features above. A uni-gram and bi-gram vectors are concatenated together and then used to train a multi-class SVM model.

### *5.5 Word-level n-gram features*

The text corresponding to each audio utterance in the dataset is extracted using Arabic ASR, as described in Ali *et al.* (2014), trained on Gale corpus available at Linguistic Data Consortium[e]. We gathered all the text files from training and development data to extract a vocabulary of 68,707 unique words. Then, we extracted a word-level uni-gram vector from each text file. The Term Frequency - Inverse Document Frequency weight is applied to the generated vectors. The term

[d] https://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.
[e] https://www.ldc.upenn.edu/.

**Table 3.** The standard X-vector DNN configurations

| Layer | Layer context | Tot. context | In $\times$ Out |
|---|---|---|---|
| Frame 1 | $[t-2; t+2]$ | 5 | $5F \times 512$ |
| Frame 2 | $t-2; t; t+2$ | 9 | $1536 \times 512$ |
| Frame 3 | $t-3; t; t+3$ | 15 | $1536 \times 512$ |
| Frame 4 | $t$ | 15 | $512 \times 512$ |
| Frame 5 | $t$ | 15 | $512 \times 1500$ |
| Stats pooling | $[0; T)$ | $T$ | $1500T \times 3000$ |
| Segment 6 | 0 | $T$ | $3000 \times 512$ |
| Segment 7 | 0 | $T$ | $512 \times 512$ |
| Softmax | 0 | $T$ | $512 \times L$ |

frequency represents how frequent the word is in the text, the word which occurred the most will have the highest score. But the words that are frequent in all the documents will have little information for the classifier. The inverse document frequency is used to reduce the score of the words that are frequent in most of the documents.

### *5.6 X-vectors*

Recently, X-vectors are successfully applied to the speaker and language identification tasks (Snyder *et al.* 2018a,b). X-vector is a fixed-dimensional embedding extracted from a sequence of speech features using a DNN. A temporal pooling layer in the network, which aggregates information across time, is used to capture long-term characteristics of the dialects. By this, each utterance is represented by one X-vector. The generated X-vectors are then used to build a dialect identification system in the same way as the i-vectors, described earlier.

The DNN network is implemented using the nnet3 neural network library in the Kaldi speech recognition toolkit (Povey *et al.* 2011). The recipe is based on the SRE16 v2 recipe available in the main branch of Kaldi[f]. The training classes and features have been modified for dialect recognition. Table 3 includes main configurations of the DNN network used for extracting X-vectors, as used and explained in Snyder *et al.* (2018a,b).

X-vectors are extracted at layer segment6, before the nonlinearity. The input layer accepts F-dimensional features. The L in the softmax layer corresponds to the number of training dialects. The input to the DNN is a sequence of T speech frames. The first five layers process the input at the frame level, with a small temporal context centered at the current frame $t$. For example, the input layer, frame1, splices together the $F$-dimensional features at frames $t-2$, $t-1$, $t$, $t+1$, and $t+2$, which gives it a total temporal context of five frames. The input to the next layer, frame 2, is the spliced output of frame 1 at $t-2$, $t$, and $t+2$. This builds on the temporal context established by the previous layer, so that frame 2 sees a total context of nine frames. This process is continued in the following layers and results in frame 5 seeing a total context of 15 frames.

The statistics pooling layer aggregates information across the time dimension, so that subsequent layers operate on the entire segment. The input to the pooling layer is a sequence of $T$ 1500-dimensional vectors from the previous layer, frame 5. The output is the mean and standard deviation of the input (each 1500-dimensional vectors). These statistics are concatenated together

[f]https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2.

https://doi.org/10.1017/S1351324920000091 Published online by Cambridge University Press

**Table 4.** System performance (F1 micro) when using different types of features. All systems conform to the fixed training condition

| Feature type | Dimension | 2017 test data | 2018 test data |
|---|---|---|---|
| i-vectors | 400 | 0.58 | 0.595 |
| Bottleneck features | 60 | 0.608 | 0.613 |
| Phonetic | 6252 | 0.344 | 0.358 |
| GMM-tokens | 2048 | 0.468 | 0.478 |
| Word-level uni-grams | 68,707 | 0.443 | 0.514 |
| X-vectors | 1500 | 0.653 | 0.687 |
| **Fused**(i-vec+BT+X-vec) | | 0.663 | 0.697 |

(to produce a 3000-dimensional vector) and passed through the segment-level layers and finally the softmax output layer. The nonlinearities are rectified by linear units.

## 6. Classifier and fusion

In all of our experiments, a multi-class SVM (Fan *et al.* 2008) was trained on the earlier described features extracted from the training subset for ADI. The testing subset was used for systems evaluation. Confusion matrix and F1 micro were used for system performance presentation.

A fused system was introduced to fuse the results of a subset of the previously described systems. The raw output of each of the previous models is a vector of five features that represent the scores of the dialect-specific models for each utterance. The vectors of the development data were recorded and concatenated.

We employed fusion based on linear logistic regression using the FoCal toolkit (Brümmer 2007). The development subset was used for system development and for estimating fusion coefficients.

## 7. Experiments and results

The earlier described features that ranged from traditional acoustic i-vectors to recently proposed X-vectors are considered good representation for the dialects characteristics. Some of these features are acoustic features, which exploit differences between the distributions of sounds in different dialects, and some others are phonotactic features that exploit dialect-dependent differences in the sequences in which these sounds occur.

The feature vectors, extracted from both ADI training utterances of DSL 2017 and 2018 shared tasks, as described in Tables 1 and 2, are used to train a multi-class SVM. The feature vectors extracted from the testing set are used for system evaluation (both 2017 and 2018 DSL data). F1 micro is used as a performance measure and to compare different systems.

Table 4 shows system performance when using different types of features as described in Section 5.

The best F1 micro achieved by the participants of the ADI task in the DSL 2018 shared task is 0.5892. As shown from the results presented in Table 4, the X-vector system outperforms the other systems which use different features. This underscores the findings of Snyder *et al.* (2016); Garcia-Romero *et al.* (2017) that DNNs may be capable of producing more powerful representations of dialects from short speech segments.

https://doi.org/10.1017/S1351324920000091 Published online by Cambridge University Press

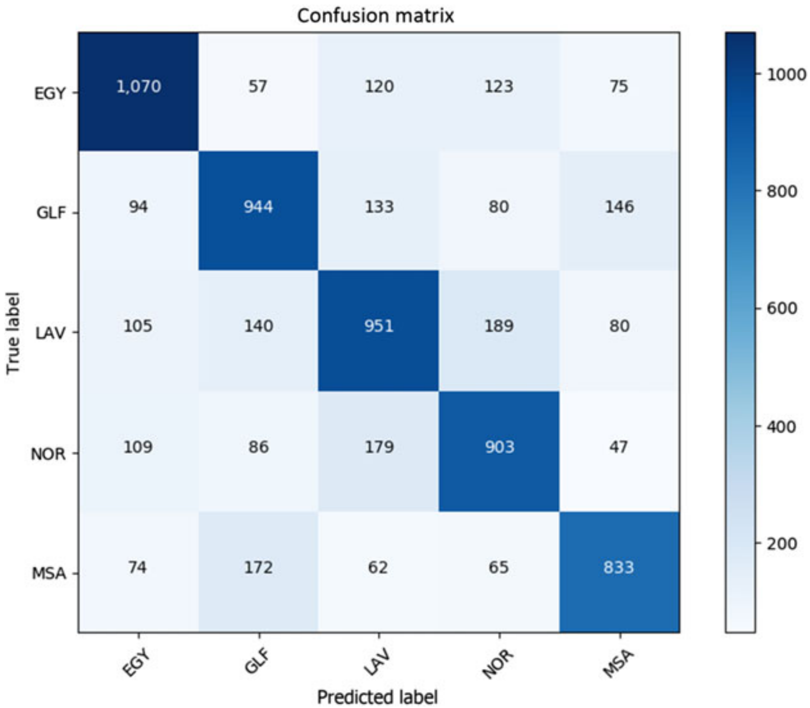


**Fig. 2.** Fusing results.

A slight improvement (1.5%) is achieved when the X-vector system is fused with the i-vector, bottleneck, and word uni-gram systems. The confusion matrix of the X-vector system is shown in Figure 2.

## 8. Conclusion and future work

In this paper, we adapted the X-vector framework, which was originally developed for speaker recognition, to the task of Arabic dialect recognition. We found X-vectors achieved good performance (0.687) on the ADI 2018 DSL shared task testing dataset, outperforming several systems that use other state-of-the-art features. Compared with the best result presented for the ADI shared task, the X-vector system achieved 16.6% relative improvement. The performance of the X-vector system is slightly improved (0.697) when fused with i-vectors, bottleneck features, and word uni-grams. We saw that X-vectors outperformed i-vectors, suggesting that they may be more robust to this domain mismatch.

The results of the X-vector framework for Arabic dialect recognition are promising. In future work, we will use the DNN described in this paper as pretraining for the full end-to-end approach as in Snyder *et al.* (2016), so that a more appropriate similarity metric is learned along with the X-vectors. Also, we will use X-vectors for training the Gaussian classifier.

## References

**Ali A., Dehak N., Cardinal P., Khurana, S., Yella, S.H., Glass, J., Bell, P. and Renals, S.** (2015). Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

**Ali A., Zhang Y., Cardinal P., Dahak N., Vogel S. and Glass J.** (2014). A complete Kaldi recipe for building Arabic speech recognition systems. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 525–529. IEEE.

**Ali A., Zhang Y. and Vogel S.** (2014). QCRI advanced transcription system (QATS). *Proceedings of SLT*.

**Brümmer N.** (2007). Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scorestutorial and user manual. Software. Available at http://sites.google.com/site/nikobrummer/focalmulticlass/

**Çöltekin Ç. and Rama T.** (2017). Tübingen system in vardial 2017 shared task: Experiments with language identification and cross-lingual parsing. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 146–155.

**Dehak N., Dehak R., Glass J.R., Reynolds D.A., Kenny P.**, et al. (2010). Cosine similarity scoring without score normalization techniques. In *Odyssey*, p. 15.

**DeMarco A. and Cox S.J.** (2013). Native accent classification via i-vectors and speaker compensation fusion. In *INTERSPEECH*, pp. 1472–1476.

**Du J., Wang Q., Gao T., Xu Y., Dai L.-R. and Lee C.-H.** (2014). Robust speech recognition with speech enhanced deep neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*.

**Eldesouki M., Dalvi F., Sajjad H. and Darwish K.** (2016). Qcri@ dsl 2016: Spoken arabic dialect identification using textual features. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pp. 221–226.

**Elfardy H. and Diab M.** (2013). Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, pp. 456–461.

**Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R. and Lin C.-J.** (2008). Liblinear: A library for large linear classification. *Journal of Machine Learning Research* **9**(Aug), pp. 1871–1874.

**Garcia-Romero D., Snyder D., Sell G., Povey D. and McCree A.** (2017). Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4930–4934. IEEE.

**Habash N. Y.** (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies* **3**(1), 1–187.

**Hanani A., Qaroush A. and Taylor S.** (2017). Identifying dialects with textual and acoustic cues. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pp. 93–101.

**Hanani A., Russell M.J. and Carey M.J.** (2013). Human and computer recognition of regional accents and ethnic groups from british english speech. *Computer Speech & Language* **27**(1), 59–74.

**Malmasi S., Zampieri M., Ljubešić N., Nakov P., Ali A. and Tiedemann J.** (2016). Discriminating between similar languages and arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pp. 1–14.

**Najafian M., Khurana S., Shan S., Ali A. and Glass J.** (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5174–5178. IEEE.

**Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P.**, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Number EPFL-CONF-192584. IEEE Signal Processing Society.

**Snyder D., Garcia-Romero D., McCree A., Sell G., Povey D. and Khudanpur S.** (2018a). Spoken language recognition using x-vectors. In *Odyssey: The Speaker and Language Recognition Workshop, Les Sables dOlonne*.

**Snyder D., Garcia-Romero D., Sell G., Povey D. and Khudanpur S.** (2018b). X-vectors: Robust DNN embeddings for speaker recognition. *Submitted to ICASSP*.

**Snyder D., Ghahremani P., Povey D., Garcia-Romero D., Carmiel Y. and Khudanpur S.** (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170. IEEE.

**Torres-Carrasquillo P.A., Singer E., Kohler M.A., Greene R.J., Reynolds D.A. and Deller J.R.**, Jr. (2002). Approaches to language identification using gaussian mixture models and shifted delta Cepstral features. In *Seventh International Conference on Spoken Language Processing*.

**Tüske Z., Golik P., Schlüter R. and Ney H.** (2014). Acoustic modeling with deep neural networks using raw time signal for LVCSR. In *Fifteenth Annual Conference of the International Speech Communication Association*.

**Wray S. and Ali A.** (2015). Crowdsource a little to label a lot: Labeling a speech corpus of dialectal arabic. In *Sixteenth Annual Conference of the International Speech Communication Association*.

**Zaidan O.F. and Callison-Burch C.** (2014). Arabic dialect identification. *Computational Linguistics* **40**(1), 171–202.

**Zirikly A., Desmet B. and Diab M.** (2016). The GW/LT3 vardial 2016 shared task system for dialects and similar languages detection. In *COLING*, pp. 33–41. The COLING 2016 Organizing Committee.

**Cite this article:** Hanani A and Naser R (2020). Spoken Arabic dialect recognition using X-vectors. *Natural Language Engineering* **26**, 691–700. https://doi.org/10.1017/S1351324920000091