Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects

Georgina Brown[1*], Sam Hellmuth[2]

[1]Department of Linguistics and English Language, Lancaster University, United Kingdom
[2]Department of Language and Linguistic Science, University of York, United Kingdom

[1]g.brown5@lancaster.ac.uk
[2]sam.hellmuth@york.ac.uk

*** Correspondence:**
Georgina Brown
g.brown5@lancaster.ac.uk

Keywords: Arabic dialects, accent, intonation, automatic accent recognition, Support Vector Machines.

Abstract

Dialect variation spans different linguistic levels of analysis. Two examples include the typical phonetic realisations produced and the typical range of intonational choices made by individuals belonging to a given dialect group. Taking the modelling principles of a specific automatic accent recognition system, the work here characterises and observes the variation that exists within these two levels of analysis among eight Arabic dialects. Using a method that has previously shown promising performance on English accent varieties, we first model the segmental level of analysis from recordings of Arabic speakers to capture the variation in the phonetic realisations of the vowels and consonants. In doing so, we show how powerful this model can be in distinguishing between Arabic dialects. This paper then shows how this modelling approach can be adapted to instead characterise prosodic variation among these same dialects from the same speech recordings. This allows us to inspect the relative power of the segmental and prosodic levels of analysis in separating the Arabic dialects. This work opens up the possibility of using these modelling frameworks to study the extent and nature of phonetic and prosodic variation across speech corpora.

1 Introduction

Many recent approaches to automatic accent recognition have depended heavily on machine learning techniques, falling in line with trends across the breadth of speech technology (Najafian, et. al., 2018; Shon et. al., 2018). Usually though, these approaches do not yield accent recognition rates that are comparable with the low error rates we see in related areas like automatic speaker recognition (Snyder et. al., 2017). Additionally, these approaches demand enormous, and therefore often unattainable, datasets to develop working systems. One way of overcoming the need for very large datasets in automatic accent recognition is to be selective in its development and inform the system of the specific features it should use to model speakers' accents. The York ACCDIST-based automatic accent recognition system (Brown, 2015; Brown and Wormald, 2017) is an example of a system that takes this more targeted approach. Based on the ACCDIST metric (Huckvale, 2004, 2007), Y-ACCDIST models encapsulate only a subset of features that are expected to represent a speaker's production of the phoneme inventory. In doing so, Y-ACCDIST has a lowered reliance on machine learning techniques that would otherwise involve the extraction of many features from right across the speech sample, which would then be used to derive a subset that is estimated to comprise the most useful features for the task at hand. As implemented to date, Y-ACCDIST targets the phonetic realisations of the individual vowels and consonant segments in the language and compares one speaker's set of realisations with the corresponding sets of other speakers. This comparison gauges which group of speakers (grouped by accent) the speaker is most similar to. The first experiments in this paper demonstrate the performance of this "segmental" version of the Y-ACCDIST system on speech recordings taken from speakers of eight Arabic dialects. These experiments simultaneously show its use as an automatic dialect classification system and as a way of observing variation among accents and dialects.

While attempting to isolate the segmental level has its advantages (as it is the level of analysis that is expected to be most valuable to dialect classification), we are aware that there are other potentially useful features within the speech signal that this approach overlooks. There is growing evidence of accent- or dialect-specific intonation patterns in a number of languages. For example, computational analysis of data from the Intonational Variation in English (IViE) project in seven different British English varieties, showed differences in the shape and distribution of f0 contours across dialects (Grabe, Kochanski, & Coleman, 2007). A key contribution of this paper is to ascertain whether the modelling procedure in the standard segmental form of the Y-ACCDIST system can also be applied to the prosodic level of analysis. This will then enable us to compare the contribution of segmental and prosodic cues to a specific dialect classification task, while removing other potentially distracting information embedded within the speech signal.

The dataset that has been used in the experiments presented in this work is the *Intonational Variation in Arabic* (IVAr) corpus (Hellmuth and Almbark, 2019). There are other, larger, speech corpora available that would allow for research to be conducted on different Arabic dialects. The Multi-Genre Broadcast (MGB-5) challenge dataset (Ali et. al., 2019) is one such example which consists of hundreds of hours of data from 17 countries, a subset of which has been labelled for dialect group by human annotators. Despite MGB-5's appealing size, there are a number of reasons why the IVAr corpus is better suited to the present study. Firstly, Ali et. al., (2019) concede that there will be labelling errors as a result of their dataset construction method. Much of the metadata is often led by the country of the YouTube channels, for example, from which the speech data have been identified. Dialect labels may therefore be estimations at times, bringing in noise to any dialect research. The IVAr corpus metadata, on the other hand, are extremely controlled and reliable, allowing us to draw more robust findings. Secondly, the speech samples in the MGB-5 dataset are generally too short. MGB-5 speech samples are categorised according to their durations: *short* (<5s), *medium* (5-20s) and *long* (>20s). This distribution of sample durations is insufficient for the methods implemented in the present study, where, ideally, we would be using at least one minute of speech per speaker. Thirdly, the IVAr corpus was collected in such a way that elicited speech for the purpose of prosodic research (i.e. a carefully selected and informed set of sentences and speech tasks that prompt intonation patterns of interest). The present study would not be possible without such control in the data construction. Lastly, the IVAr corpus has already had a substantial amount of prosodic analysis conducted on it (Hellmuth, 2018; Hellmuth, to appear). This enables us to interpret the performance of the modelling procedure in the context of prosodic analysis that has been conducted using more traditional analytical methods. These kinds of analytical procedures have typically involved manual qualitative labelling of samples of data using a system of prosodic annotation such as the Tones and Break Indices (ToBI) system (Beckman & Elam, 1997; Beckman, Hirschberg, & Shattuck-Hufnagel, 2005) or more recent systems proposed for use across languages (Hualde & Prieto, 2016), as well as quantitative approaches such as visualisation and statistical analysis of f0 contour shapes (Hellmuth, 2018).

Recently, a more innovative way of capturing prosodic variation has been proposed. Elvira-García et. al. (2018) introduced the *ProDis* dialectometric tool for measuring prosodic distances between linguistic varieties based on acoustic measurements. *ProDis* involves logging the correlations between the pitch contours of specified sentences produced by speakers, and then comparing these correlations among a speaker set representing a range of languages. This provides a dialectometric method that aims to reveal prosodic similarities and differences between linguistic varieties. The authors motivate their work by pointing out that efforts have been made to measure dialect and language differences by making phonological or lexical comparisons, but that we lack an equivalent that makes use of prosodic information. Their demonstration of using *ProDis* shows its application to a subset of AMPER (Atlas Multimédia Prosodique de l'Espace Roman) (Contini and Romano, 2002), which is an international effort to capture data that represents a full range of Romance linguistic varieties. Within their work, they applied the *ProDis* tool to 7 dialects from across 5 Romance languages. Using *ProDis*, Elvira-García et. al. were able to perform cluster analyses and associated data visualisations on these data, followed by some qualitative evaluation. For example, they produced a dendrogram of their *ProDis* data representations. One of their clusters was neatly made up of varieties that are largely spoken in Sardinia, and they were able to provide an accompanying example of the characteristic intonation contour shape of yes/no-questions produced by speakers of those varieties.

Similarly, one version of the Y-ACCDIST system has been presented as another way to quantify differences among accent varieties, by measuring and modelling phonetic realisational differences of segments, demonstrated in Brown and Wormald (2017). Like Elvira-García et. al.'s study above, Brown and Wormald were able to draw observations from a dendrogram of Y-ACCDIST representations of different speakers in a speech dataset. In their work, they looked at the accent differences between Punjabi-English and Anglo-English speakers in Bradford and Leicester in England. One of the pertinent patterns to emerge was that there were some clusters that grouped the speakers according to the community centre they attended, which perhaps went beyond the types of grouping that the authors originally expected. As well as the cluster analyses, Brown and Wormald were also able to perform some feature selection analyses (using the Y-ACCDIST models as a framework of features) which indicated the vowels and consonants that were estimated to separate the accent varieties in the dataset. This analysis pointed towards the GOAT vowel and /ɹ/ as features that discriminated these accent varieties, which corresponded with some of the more traditional acoustic analysis conducted in Wormald (2016).

Another ACCDIST-based system was demonstrated to observe accent variation among a larger number of accents from across the British Isles in Ferragne and Pellegrino (2010), which also took advantage of the variation in phonetic realisations. In their study, Ferragne and Pellegrino took controlled wordlist data and created an ACCDIST-based model of the vowel systems of 261 speakers who represented 13 accents from the Accents of the British Isles (ABI) corpus (D'Arcy et. al., 2004). They also found that these models yielded linguistically explicable patterns in visualisations of the data. For example, they found a very neat split in a cluster analysis between the Scottish, Irish and English accent varieties in the corpus.

By implementing a Y-ACCDIST-based framework to model speakers' intonational inventories, in this work we apply a similar modelling procedure to that presented in Elvira-García et. al. (2018). However, by implementing a framework that has also been used to capture segmental phonetic realisational differences between different accent varieties, we can draw comparisons between how prosodic information and segmental information distinguish linguistic varieties under investigation. Additionally, by modelling numerous speakers per dialect group, we have an opportunity to train a dialect classification system on the prosodic information alone to be able to observe how much this single level of analysis could contribute to an accent or dialect classification task. Although the dataset used to demonstrate *ProDis* in Elvira-García et. al. was very large, the number of speakers per variety was very small (less than 5), and so did not provide the opportunity for an experiment of the kind presented here.

Until the current work, Y-ACCDIST had only been tested on datasets of speech in English. We first demonstrate its performance in distinguishing between dialects of Arabic in its original segmental configuration (i.e. targeting the phonetic realisations of different segments), and we show results on both controlled read speech and spontaneous speech. We then move on to explore the Y-ACCDIST-based framework for modelling the prosodic variation among accents, allowing us to compare the different value that segmental and prosodic levels of speech analysis bring to the dialect recognition task. We also delve into the inner workings of the machine learning within the system to determine whether we can identify particularly useful features within the segmental and prosodic models that can discriminate the Arabic dialects. All the analysis tasks conducted for this study are interpreted in the context of the existing prosodic analysis conducted on these same data.

In summary, this paper addresses the following broad objectives:

- to observe the Y-ACCDIST system's recognition performance on Arabic dialect varieties and interpret the results in the context of existing linguistic analyses of the data.
- to compare the performance of the Y-ACCDIST system on read speech and spontaneous speech on the same dialect classification task.

- to transfer Y-ACCDIST's modelling technique from the segmental level of analysis to the prosodic level and compare dialect classification performance between these two levels of analysis.

2 Arabic Dialects

2.1 Overview of Arabic dialects

Arabic is one of the world's largest languages, spoken as a native language by at least 300 million speakers (Owens, 2013), yet consisting of a diverse array of spoken vernaculars which vary from each other at all levels of linguistic analysis – from phonetics and phonology to morphosyntax and lexis (Retsö, 2013). There is a clear divide between western 'maghreb' dialects spoken in North Africa and eastern 'mashreq' dialects spoken elsewhere (Behnstedt & Woidich, 2013), such that human listeners can distinguish these two broad groups based solely on prosodic information (Barkat, Ohala, & Pellegrino, 1999). A commonly used geographical approach to grouping Arabic dialects, based on shared linguistic features within groups, results in the following five-way grouping, from west to east (Versteegh, 2014): i) dialects of North Africa (including Morocco, Algeria, Libya and Tunisia); Egyptian dialects (including Egypt and Sudan); Levantine dialects (including Jordan, Lebanon, Syria and Palestine); Mesopotamian dialects (including Iraq); and dialects of the Gulf/Arabian Peninsula (including Saudi Arabia, Kuwait, Bahrain, Qatar, Oman and Yemen). This five-way split has been widely implemented in computational approaches to the Arabic dialect classification task (e.g. Biadsy et. al., 2009). Nevertheless, the degree of dialectal variation within each of these five groups is considerable, with additional important dialectal discontinuities due to historical contact and migration, social categories and lifestyle (with a common broad divide between dialects which are sedentary/urban versus nomadic/rural in origin) as well as religious or sectarian affiliation (Behnstedt & Woidich, 2013). As a result of these cross-cutting factors contributing to dialectal variation, Arabic is frequently described as a 'mosaic' of dialects. 'Successful' dialect classification for Arabic would ideally be able to tackle different degrees of granularity, both between and within the broad regional groupings that are usually taken as targets.

2.2 Automatic classification of Arabic dialects

As indicated in the Introduction, many approaches to automatic dialect identification have depended heavily on machine learning approaches, usually inspired by the techniques tested for Language Identification (LID). These approaches have demanded vast quantities of data for training. Biadsy et.

al. (2009) applied a Phone Recognition followed by Language Modelling (PRLM) approach to Arabic dialect classification, which was first introduced by Zissman (1996) for the purpose of LID. As the name suggests, PRLM starts by feeding a speech sample through a phone recognition system to establish an estimated sequence of phones in the sample. This estimated sequence is then compared against the phone sequences and distributions computed for the different linguistic varieties in the reference system (i.e. the training data). PRLM therefore depends on the different varieties we are distinguishing between to have phone sequences and distributions that are separable. For LID, this seems to achieve reasonable performance, but as the varieties we are distinguishing between become more and more similar (i.e. dialects and then accents), this approach is expected to become less effective. In their work, Biadsy et. al. reported that the PRLM approach achieved 81.6% accuracy for an identification task involving speakers of five Arabic dialect groups (using the commonly used grouping described in Section 2.1 above).

The PRLM approach is the more traditional one for this sort of task. Researchers have since applied classifiers based on neural networks to the problem of Arabic dialect recognition (Najafian, et. al., 2018; Shon, et. al., 2018). These works follow in the footsteps of developments in speaker recognition research, where a new method of modelling the variation among different speakers in the form of “embeddings” was proposed, in an effort to improve on the performance of i-vector-based systems (Snyder et. al., 2017). Such methods demand vast amounts of training data (ideally, hundreds of speech samples per dialect group). Both of the studies mentioned above which apply the neural network based approach to Arabic dialect identification used the Multi-Genre Broadcast 3 (MGB-3) dataset, which offers 63.6 hours of training data across the five main Arabic dialect groups. Shon et. al. (2018) achieved 73% accuracy using a neural network based system, outperforming the i-vector systems they compared on the same task.

In this paper, our experiments will be conducted on a corpus of speech recordings taken from 96 speakers spanning 8 Arabic dialect categories. We therefore present ourselves with a dialect classification problem which has a fraction of the data to train a system on. In addition, we assume that this is a more difficult problem in that we have increased the level of similarity between dialects by having 8 dialect categories, rather than 5 broader ones. The Y-ACCDIST-based method we are employing is much better suited to a dataset of this size and nature (as demonstrated in Brown (2016)).

2.3 The IVAr Corpus

The core Intonational Variation in Arabic (IVAr) corpus contains recordings from 12 speakers each in 8 spoken dialects of Arabic (96 speakers in total), collected on location in North Africa and the Middle East (Hellmuth & Almbark, 2019)[1]. IVAr provides at least one dataset from each regional dialect group, with more than one dataset for the more linguistically diverse regional groups (Levantine/Gulf/North Africa). The corpus thus provides for an eight-way dialect classification task, across the geographically defined dialects listed in Table 1.

Table 1. Dialects represented in the Intonational Variation in Arabic Corpus

Code	Dialect	Recording location	Regional group
moca	Moroccan Arabic (Casablanca)	Casablanca, Morocco	North Africa
tuns	Tunisian Arabic (Tunis)	Tunis, Tunisia	
egca	Egyptian Arabic (Cairo)	Cairo, Egypt	Egyptian
joka	Jordanian Arabic (Karak)	Karak, Jordan	Levantine
syda	Syrian Arabic (Damascus)	Amman, Jordan	
irba	Iraqi Arabic (Muslim Baghdadi)	Amman, Jordan	Mesopotamian
kwur	Kuwaiti Arabic (Urban)	Kuwait City, Kuwait	Gulf/Arabian Peninsula
ombu	Gulf Arabic (Buraimi)	Buraimi, Oman	

Use of IVAr allows us to demonstrate the dialect identification task at a more granular level than is typical in the field, since most other work on dialect identification for Arabic attempts at most a five-way regional classification (due, in turn, to the fact that most large Arabic corpora provide datasets defined at a regional level only).

The corpus contains speech elicited in a range of speech styles, from scripted read speech to unscripted semi-spontaneous speech. The scripted materials were presented to participants printed in Arabic script, using the informal spelling conventions of each local dialect (rather than following the norms of standard Arabic); in this paper we use data elicited by means of a scripted dialogue (sd) performed as a role play between pairs of speakers and a monologue narrative folk tale (sto). The spontaneous speech data used in this paper comprise a monologue folk tale retold from memory (ret), an information-gap map task performed in dialogue between pairs of speakers (map), and free

[1] The full corpus comprises 10 datasets across 8 dialects; that is, for one of the 8 dialects, Moroccan Arabic, there are two additional datasets: one with bilingual speakers of Moroccan Arabic and Tashlhiyt Berber aged 18-35 (mobi), and one with Moroccan Arabic speakers aged 40-60 (moco). These two additional datasets are not investigated in the present study.

conversation between pairs of speakers (fco). Further information about the instruments used to elicit the data is available at ivar.york.ac.uk/.

The participants in each location were recruited through a local fieldwork representative, typically through an educational institute such as a university or private language school. Participants ages ranged from 18-35 years. All recordings took place in the city in which participants were resident, and recruitment was carefully monitored to ensure participants were speakers of the target dialect and had been raised in the target city. The only exception was speakers of Syrian and Iraqi Arabic, who were recruited in Amman, Jordan due to the prevailing security situation in Syria and Iraq at the time of recording. Detailed participant metadata is provided with the published corpus. All participants received an information sheet in Arabic and provided informed written consent prior to recording.

Participants were recorded in pairs using head-mounted Shure SM10A dynamic microphones directly to .wav format on a Marantz PMD660/620 digital recorder at 44.1kHz 16 bit, with each speaker recorded to a separate stereo channel which can be split to analyse speakers separately. Recording sessions were run by a local fieldworker who was a native speaker of the same dialect. All of the tasks, scripted and unscripted, were performed in a single recording session, with the same interlocutor and under the same recording conditions.

The spontaneous speech data were orthographically transcribed by native speaker research assistants using a romanised phonetically transparent transliteration system adapted for each dialect; these transcriptions are available as part of the published corpus. For the read speech, the script used during data collection was transcribed into the same transliteration system, and are also made available with the corpus. For the present project we created a merged dictionary of all of the dialect-specific forms used in transcripts for read and spontaneous speech across all dialects; a native speaker of Arabic proficient in Modern Standard Arabic (MSA) created a transcription of each dialect-specific form using a common MSA phone set to create the merged dictionary. This was based on the accepted cognate sound in MSA of dialect-specific variants. For example, the name of the main character in the folk tale retold from memory is variously produced in the dialects as [ʒuħa], [dʒuħa] or [ɡuħa] <جحا> and appears in the merged dictionary as pronounced in MSA i.e. as [ʒuħa]. We intend on publication of the present paper to make this merged dictionary available as an appendix to the main published IVAr corpus.

As already discussed in the Introduction, larger speech datasets of Arabic dialects exist, such as MGB-3 and MGB-5. However, such datasets have not been collected in a way that allows us to explore the specific research questions in this paper that involve analysing prosodic variation as well as segmental variation.

3 The Y-ACCDIST System

Y-ACCDIST is a text-dependent system, which requires a transcription to be processed alongside the audio sample we are classifying. However, a text-dependent system here is defined as one that requires a transcription, but the speech can be spontaneous (as discussed in Brown (2018)). In some works, text-dependent systems only refer to those where the spoken content of the test samples and the training samples match. This is one of the key features that separates Y-ACCDIST from other ACCDIST-based recognisers found in Huckvale (2004, 2007) and Hanani et. al., (2013). The initial experiments will allow us to compare the performance of this approach on the IVAr dataset on both read speech (where the spoken content is matched across training and test data), and spontaneous speech (where the spoken content does not match across speech samples).

3.1 System Description

For each speaker in the IVAr dataset, we take a speech sample and a transcription and pass them through a forced aligner (developed in-house using the Hidden Markov Model Toolkit (HTK) (Young et al, 2009)) to estimate where each phone in the sequence is produced in the sample. Given a speech recording and a phonemic transcription of that recording, the aligner extracts acoustic features from across the speech sample and estimates where each phone is in the signal, i.e. producing an estimated time alignment of the phone sequence. Some forced aligners, particularly those that are widely available, have ready-trained acoustic models for a given language that may provide multiple options for a phonemic transcription of a given word. The specific phone labels attributed to a speech sample will therefore be partly determined by the acoustics of the segments in the speech sample, and how they compare against the pre-trained acoustic models of the forced aligner. For the present study, however, we created a bespoke lexicon containing all lexical items in the analysed data subset (described in section 4.1), based on the phoneme inventory of Modern Standard Arabic (MSA). To achieve this, we generated a cross-dialectal lexicon from the dialect-specific transcripts made available with the IVAr corpus, which was manually edited by an Arabic speaker to replace dialect-specific phoneme labels with MSA phoneme labels; for example, a dialect-specific entry for the word ‘heart’ such as [galb] or [2alb] appears in the bespoke lexicon as [qalb]. We then used the IVAr dataset itself to train speaker-specific acoustic models for the MSA phoneme categories in the bespoke lexicon, which the aligner used to estimate where each phoneme is in the sample. We initialised the models by “flat-starting”; that is, we imposed evenly spaced notional phoneme boundaries on the speech samples as a starting point. We then repeatedly applied an Expectation-Maximization algorithm which iteratively adjusted the placement of these boundaries to more accurately segment the sample according to phone segments. More reliable boundaries should be reflected in the production of increasingly stable acoustic models during this process. Performing forced alignment in this way was possible because we had enough speech per speaker to do so. This allows us to impose just one set of MSA symbols on the range of different productions that different speakers may produce. This lays the foundations for our method of dialect classification.

Using these estimated time boundaries between phones in the sequence, a vector of Mel Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980) was extracted at the midpoint to acoustically represent each phone. The MFCCs used in this work consist of 12 coefficients. Larger MFCC vectors have been trialled in past work (Brown, 2014), but 12 coefficients were shown to provide sufficient information. An average MFCC was calculated for each phoneme category in MSA from these midpoint acoustic features. The result of this is that we have the phoneme inventory represented by average acoustic features (one per phoneme) for the speaker. By using midpoint features, this approach overlooks temporal differences that might exist between dialects. This is a factor to keep in mind when interpreting the results.

Using this set of averaged acoustic features, we calculated the Euclidean distance between all phoneme-pair combinations that are possible within the phoneme inventory. This was achieved by computing the Euclidean distances between all the possible pairs of average MFCC vectors that represent each phoneme. We can organise this in a matrix (for clarity, this is illustrated below in Figure 1).

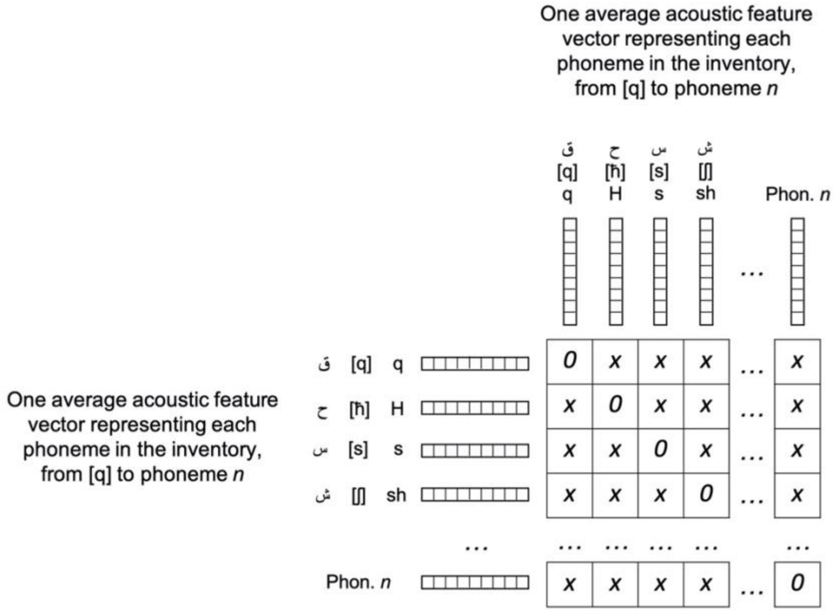

Figure 1. A demonstration of how a speaker-specific matrix is calculated with the whole segmental inventory from [q] to phoneme *n*. The ‘0’/‘x’ symbols represent the Euclidean Distance for that pair.

The resulting set of Euclidean distances is expected to encapsulate the range of phonetic realisations that are associated with a speaker’s pronunciation system (or accent). This matrix of distances is our model of a speaker’s accent. Using British English accents as an example, typical speakers in Northern England will produce similarly realised vowels for FOOT and STRUT (both realised as [ʊ]), whereas typical speakers in the South of England will produce different vowel realisations (FOOT would be produced as [ʊ], while STRUT would be more likely to be produced as [ʌ]). A parallel example for Arabic would arise for consonants; an Arabic speaker from Egypt will typically realise the target sound < ق> [q] in the same way as target < ء> [ʔ], whereas an Arabic speaker from Morocco will more frequently produce these two target sounds ([q]~[ʔ]) as two separate categories. One of the Euclidean distances in the matrix is expected to reflect this accent-specific feature of the speaker. An entire matrix is therefore expected to contain numerous accent-specific features of this kind. Simultaneously, by computing intra-speaker distances in this way, we should eliminate other information embedded within the speech signal that does not necessarily assist in the accent classification task. For example, the distance between the FOOT and STRUT vowels for a typical Northern male speaker and a typical Northern female speaker should be equally small; similarly, the distance between targets [q]~[ʔ] will all be equally small for a typical Egyptian female speaker and a typical Egyptian male speaker.

We performed the above procedure on the speech samples and transcriptions of all our training speakers. The resulting speaker-specific matrices are then fed as features into a Support Vector Machine (SVM) classifier (Vapnik, 1998). It is also possible to make use of Deep Neural Networks (DNNs) as classification mechanisms in these sorts of experiments. However, DNNs are much more suited to extremely large datasets of thousands of data samples. SVMs also tend to require larger datasets, but they are not as "data-hungry" as DNNs.

Within the SVM, which acts as multi-dimensional space, a *one-against-the-rest* rotation is implemented for classification. In turn, each accent group of training speakers becomes the 'one', while the speakers for all other accent categories are collapsed into a single group that form 'the rest'. An optimal hyperplane (i.e. a separating boundary within multidimensional space) is computed on each rotation to achieve the best separation between these two groups of speakers[2]. To classify an unseen speaker, we form a matrix model for that speaker as described above, and this model is presented to the SVM on each rotation. The accent category of the unseen speaker is determined by the clearest margin it forms with the hyperplane in each of these rotations.

4. Experiments

A sequence of experiments was conducted in the commonly implemented *leave-one-out cross-validation* setup, where each speaker became the test speaker, in turn, while the remaining speakers in the dataset were used to train the Y-ACCDIST system. This was in an effort to maximise the number of training speakers.

4.1 Segmental Modelling

The above process was conducted for read speech recordings from the speakers (where speakers were asked to read the same scripted dialogue and story) and also spontaneous speech as a comparison of performance on the two modes of speech. As we pointed out above, a transcription must accompany the recordings. Most, but not all, speakers' spontaneous speech samples have been orthographically transcribed. For these experiments, we have therefore used data for both read speech and spontaneous speech experiments from a subset of the speakers (reduced from 96 speakers to 86 speakers). This

[2] While SVMs can be very useful in classification problems, they are susceptible to 'overfitting', particularly on moderate-sized datasets. In these experiments, while overfitting is a risk, we have used a linear kernel, and have also set the regularization parameter to tolerate some errors during training. The controlled nature of the dataset also mitigates against overfitting as it provides less "noise" and therefore fewer overfitting opportunities.

results in an imbalance in the number of speakers for different dialect groups, though an even gender balance was retained within each group. Table 2 shows the number of speakers per dialect group in our analysed subset, along with the volume of data in minutes used in the experiment (with silences removed), by dialect, gender and speech style.

Table 2. Number of speakers per accent category in the data subset, with total duration (rounded up to the nearest whole minute) of speech data used in training and/or testing (silences removed).

		Speakers		Scripted data (mins)			Unscripted data (mins)		
Dialect Group	**Code**	**Female**	**Male**	**Female**	**Male**	**Total**	**Female**	**Male**	**Total**
Egyptian (Cairo)	egca	4	4	15	12	26	15	7	20
Iraqi (Muslim Baghdadi)	irba	6	6	15	13	28	21	15	36
Jordanian (Karak)	joka	6	6	15	15	30	21	17	37
Kuwaiti (Urban)	kwur	6	6	14	14	28	18	23	41
Moroccan (Casablanca)	moca	6	6	14	14	29	21	44	65
Gulf (Buraimi, Oman)	ombu	6	6	16	15	31	18	12	30
Syrian (Damascus)	syda	3	3	17	15	32	12	17	29
Tunisian (Tunis)	tuns	6	6	14	13	27	23	21	44

Table 3 provides the means and standard deviations of the amount of speech (in seconds) per speaker used in model training and/or testing in this study.

Table 3. Mean/standard deviation of speech in seconds per speaker in the data subset by speech task.

	Speech task	**Mean amount of speech per speaker (seconds)**	**Standard deviation per speaker (seconds)**
Read Speech (scripted)	Story	72.09	10.74
	Read sentences	73.36	9.64
	Total (read)	**145.45**	**16.54**
Spontaneous Speech (unscripted)	Free conversation	77.28	46.72
	Map task	70.01	59.64
	Retold folk tale	65.11	17.26
	Total (spontaneous)	**212.41**	**102.48**

We present the overall results and their corresponding confusion matrices in the subsections below.

4.1.1 Read Speech

The read speech data used for these experiments come from a scripted role-play dialogue which was designed to elicit a number of different sentence types, including declarative statements (*dec*), yes/no-questions (*ynq*), wh-questions (*whq*) and coordinated questions (*coo*, also known as alternative questions, of the form “is it X or Y?”). The sentences were designed to control the segmental content and prosodic structure of the last lexical item in each utterance, so that it contained mostly sonorant sounds (to facilitate pitch tracking) and the position of the stressed syllable was systematically varied over the last three syllables of the word. A set of sample yes/no-questions elicited in one dialect (here, Jordanian Arabic) are provided in Table 4.

Table 4. Sample set of yes/no questions (in joka) elicited using the scripted dialogue.

Code	**Target sentence**	
ynq1	ruħt l-nnaːdi l-ʻjamani	*Did you go to the Club Yemeni?*
Ynq2	l-zawaːʒ l-madani raħ jkuːn fi-l-mabna l-ʻbaladi	*Will the civil wedding be in the municipal office?*
Ynq3	gaːbalu baʕidˤ ʕan tˤariːg 'zeːna	*Did they meet each other through Zena?*
ynq4	jaʕni raħ tzuːr ʕuxutha la'jaːli	*Do you mean she will visit her sister Layali?*
ynq5	yaʕni tʕarrafit ʕaleː fi-l-matˤʕam illi fi-l-'moːl	*Do you mean they met in the restaurant in the mall?*
ynq6	waːlid nabiːl raħ jkuːn maw'ʒuːd	*Will Nabil’s parents be present?*

For the story task participants read a monologue narrative folk-tale ‘Guha and the banana seller’, adapted from a story in Abdel-Massih (2011) and adjusted to contain appropriate lexical and grammatical forms for each target dialects. The story is typically realised by speakers in 40-45 prosodic phrases or breath groups. As noted above all scripted material was presented in Arabic script using local spelling conventions.

Although considerable effort went into making the reading material as comparable as possible across dialects, the scripts read by speakers across dialects did not necessarily match word-for-word. This is because certain words are simply not shared across dialects so there is some lexical and grammatical variation across speech samples. We acknowledge that this may have weighted the result to some extent in that a small number of the phones were produced in specific phonological environments and this varies according to dialect. However, we do not expect this to be the leading factor in determining accent as we are cutting out individual phonemes and taking acoustic values from the midpoints of these segments.

Using the read speech data, the Y-ACCDIST system achieved 95.3% correct. The accompanying confusion matrix is presented in Table 5.

Table 5. Confusion matrix for Arabic dialect classification task using the segmental models on the full read speech dataset (scripted dialogue/story).

		Predicted Labels								
		egca	**irba**	**joka**	**kwur**	**moca**	**ombu**	**syda**	**Tuns**	**TOTAL**
True Labels	**egca**	**12 (100%)**	0	0	0	0	0	0	0	**12**
	irba	0	**11 (91.7%)**	1 (8.3%)	0	0	0	0	0	**12**
	joka	0	0	**8 (100%)**	0	0	0	0	0	**8**
	kwur	0	0	0	**12 (100%)**	0	0	0	0	**12**
	moca	0	0	0	0	**12 (100%)**	0	0	0	**12**
	ombu	0	0	0	0	0	**12 (100%)**	0	0	**12**
	syda	0	0	3 (50%)	0	0	0	**3 (50%)**	0	**6**
	tuns	0	0	0	0	0	0	0	**12 (100%)**	**12**

The least successful result in this experiment was for the Syrian group of 6 speakers, 3 of whom were identified as Jordanian. We note that all of the Syrian speakers were resident in Jordan at time of recording.

4.1.2 Spontaneous Speech

Using the spontaneous speech data for modelling, the Y-ACCDIST system achieves 77.9% correct (67/86). The accompanying confusion matrix is presented in Table 6.

Table 6. Confusion matrix for Arabic dialect classification task using the segmental models on spontaneous speech.

		Predicted Labels								
		Egca	**irba**	**joka**	**kwur**	**moca**	**ombu**	**syda**	**tuns**	**TOTAL**
True Labels	**egca**	**6 (75%)**	0	0	0	1 (12.5%)	0	1 (12.5%)	0	**8**
	irba	1	**8 (66.7%)**	1 (8.3%)	2 (16.7%)	0	0	0	0	**12**
	joka	0	1 (8.3%)	**8 (66.7%)**	1 (8.3%)	0	1 (8.3%)	1 (8.3%)	0	**12**
	kwur	0	1 (8.3%)	0	**9 (75%)**	0	2 (16.7%)	0	0	**12**
	moca	0	0	0	0	**12 (100%)**	0	0	0	**12**
	ombu	0	0	1 (8.3%)	3 (25%)	0	**8 (66.7%)**	0	0	**12**

	syda	1 (16.7%)	0	1 (16.7%)	0	0	0	**4 (66.7%)**	0	**6**
	tuns	0	0	0	0	0	0	0	**12 (100%)**	**12**

We should also note that for the spontaneous speech condition, speakers did not necessarily produce the same quantity of speech (durations of speech per speaker generally ranged from 4 to 8 minutes), and so there is variability across the dataset in this respect.

From the above two results, we can get an indication of the detriment to performance that content-mismatched data has. This is because the different phone tokens are produced in different environments which is likely to introduce an additional element of variability that is not present in the read speech condition. We should also bear in mind that there is a smaller number of speakers available for training the system for some dialect categories which is also likely to impact on the result.

4.2. Prosodic Modelling

As discussed above, the Y-ACCDIST-based approach has allowed us to isolate the segmental level of analysis and ignore other information embedded within the acoustic signal that might distract away from cues useful to the accent recognition task. In this part of the study, we aim to transfer the principles of the Y-ACCDIST modelling procedure to the prosodic level of analysis to see whether we can confirm previous prosodic analysis of these same data, which indicated that there are prosodic patterns that are typical of speakers of one or more Arabic dialects but different from patterns observed in a parallel context in one or more other dialects (Hellmuth 2018).

4.2.1 Organisation of Prosodic Data

The read speech data from the IVAr scripted dialogue include the sentence types presented in Table 7, elicited because these may be characterised by different prosodic patterns between sentence types within one dialect, and/or by different prosodic patterns between dialects within one sentence type.

Table 7. Sentence types elicited using the IVAr corpus role-play scripted dialogue.

Code	**Sentence type**	
dec	declarative	response to an open question (e.g. ‘what’s new?’)
whq	wh-question	question using wh-word such as who or what
ynq	yes/no-question	polar question inviting a yes or no answer

coo	coordinated question (or alternative question)	question between two alternatives (e.g. ‘is it X or Y?’)
inf	information focus	statement produced in response to a wh-question
con	contrastive focus	statement produced in response to a yes/no-question
idf	identification focus	statement produced in response to a coordinated question

For this reason, it is only these sentences extracted from the scripted dialogue (sd) that are being used in these experiments that compare the prosodic Y-ACCDIST system with the segmental Y-ACCDIST system, with the read story (sto) data removed from the dataset. A set of results for each of these system configurations are therefore presented within this section, where each has been trained and tested on only the sentences extracted from the scripted dialogue.

4.2.2 Integration of Prosodic Data into the Y-ACCDIST System

To provide the system with prosodic information, we calculated Euclidean distances between the f0 contours of all the possible pairs of read sentences available for each speaker. It is this collection of Euclidean distances between f0 contours that is expected to characterise the intonational patterns of a speaker. While it may not be immediately obvious what sorts of Euclidean distances are likely to occur between these f0 contours, it is expected that logging the similarities and differences between f0 contours in this way will express any systematic similarities and differences in intonational patterns within and between dialects. For example, for Syrian speakers, who more frequently use a rising contour in declarative sentences than is observed in other Arabic dialects (Hellmuth, 2020), the f0 contour between a particular declarative sentence X and a particular polar interrogative sentence Y might be reasonably expected to be more similar to one another than for a speaker of another dialect.

We used the read portion of the corpus in which all speakers produced more or less the same sentences. F0 contours were extracted by marking out 50 equally distributed points throughout an utterance and extracting the f0 at those points in the signal. Of course, at the points in the signal where there is no voicing, extracting f0 was not possible. This therefore reduced these vectors down to a size which was slightly smaller than 50, and the same reduction was performed on the f0 vector that it was being compared against. The result of this was a Y-ACCDIST matrix that reflected the intonation realisations of the “prosodic contour inventory” that the dataset allowed. Like the default segmental configuration explained in Section 3.1, this modelling method has the advantage of normalising against factors such as gender. By making intra-speaker calculations in this way, the model only characterises the shapes of a speaker’s f0 contours, and so, regardless of whether the speaker has a relatively high or low f0 range on average, any dialect-specific contour shapes should be expressed in the matrix. Figure 2 provides an illustration of the prosodic modelling procedure that can be compared with the segmental modelling procedure.

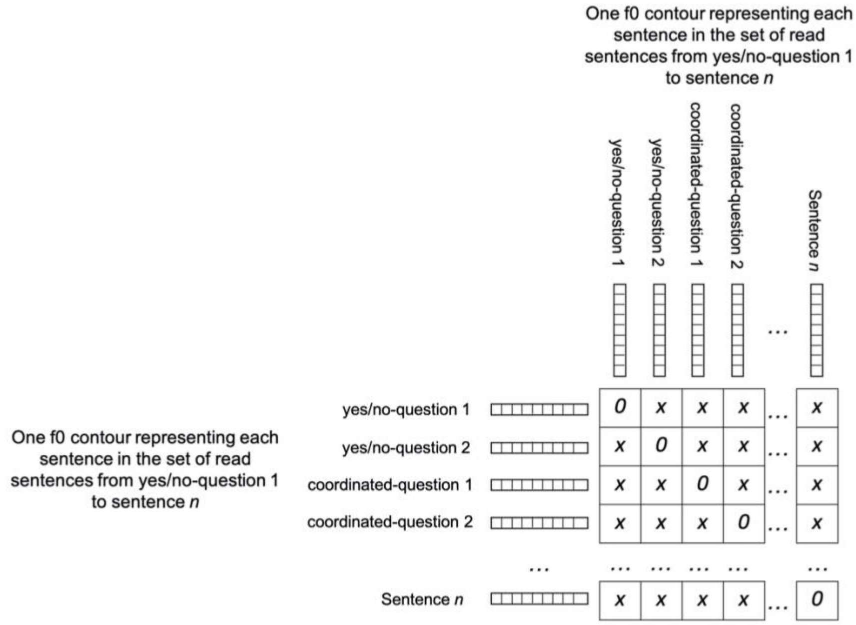

Figure 2. A diagram to demonstrate how a matrix is formed using f0 contours, rather than acoustic feature vectors, with the set of sentence types in the IVAr dataset, from yes/no-question 1 to sentence *n*. The ‘0’/‘x’ symbols represent the Euclidean Distance for that pair.

One key difference between these prosodic models and the segmental models is that there is no averaging of vectors in the construction of the matrices. While it is expected that intonation is affected by sentence type in Arabic, it is also affected by a range of other discourse factors, such as information structure (topic, givenness and focus) (Krifka 2008) as well as the interactional context (Walker 2014). The read speech sentences in the corpus were elicited at different points throughout a scripted dialogue, and thus appear in subtly different discourse contexts with varying information structure. This meant that it would be artificial to try to model an average f0 contour for each whole sentence type. We have therefore treated each individual sentence that was produced as a single category in the construction of the speaker-specific matrices. Each individual sentence was elicited in the same discourse context (i.e. position in the scripted dialogue) from each speaker in each dialect. Overall, this means that we are restricted to performing these experiments on a dataset where speakers produce the same spoken content. We return to this point further below in the Discussion.

Using these prosodic matrices to represent each speaker, we followed the same experimental procedure as for the segmental experiments described above to achieve a recognition rate and confusion matrix. The recognition rate we achieved in this configuration was 52.1% correct. The confusion matrix for this task is shown in Table 8.

Table 8. Confusion matrix of Y-ACCDIST's performance on the IVAr dataset using the prosodic models on the read speech data subset (scripted dialogue only).

		Predicted Labels							
		egca	**irba**	**joka**	**kwur**	**moca**	**ombu**	**syda**	**tuns**
True Labels	**egca**	**8 (66.7%)**	2 (16.7%)	0	0	0	0	1 (8.3%)	1 (8.3%)
	irba	1 (8.3%)	**6 (50%)**	3 (25%)	0	0	1 (8.3%)	1 (8.3%)	0
	joka	1 (8.3%)	1 (8.3%)	**2 (16.7%)**	3 (25%)	0	2 (16.7%)	2 (16.7%)	1 (8.3%)
	kwur	0	0	1 (8.3%)	**9 (75%)**	1 (8.3%)	0	0	1 (8.3%)
	moca	0	0	1 (8.3%)	1 (8.3%)	**7 (58.3%)**	2 (16.7%)	1 (8.3%)	0
	ombu	0	0	1 (8.3%)	0	3 (25%)	**5 (41.7%)**	2 (16.7%)	1 (8.3%)
	syda	1 (8.3%)	1 (8.3%)	2 (16.7%)	0	2 (16.7%)	3 (25%)	**2 (16.7%)**	1 (8.3%)
	tuns	0	0	0	0	0	0	1 (8.3%)	**11 (91.7%)**

We can compare these results with the segmental system's results that were produced using the same subset of read data, which was 93.75% correct, and the confusion matrix for this is shown in Table 9.

Table 9. Confusion matrix of Y-ACCDIST's performance on the IVAr dataset using the segmental models on the read speech data subset (scripted dialogue only; this is the same dataset that was used to build and train the prosodic system).

		Predicted Labels								
		egca	**irba**	**joka**	**kwur**	**moca**	**ombu**	**syda**	**tuns**	**TOTAL**
True Labels	**egca**	**12 (100%)**	0	0	0	0	0	0	0	**12**
	irba	0	**11 (91.7%)**	0	0	0	0	1 (8.3%)	0	**12**
	joka	0	0	**11 (91.7%)**	1 (8.3%)	0	0	0	0	**12**
	kwur	0	0	0	**11 (91.7%)**	0	0	1 (8.3%)	0	**12**
	moca	0	0	0	0	**12 (100%)**	0	0	0	**12**
	ombu	0	0	0	0	0	**12**	0	0	**12**

							(100%)			
	syda	0	0	3 (25%)	0	0	0	**9 (75%)**	0	**12**
	tuns	0	0	0	0	0	0	0	**12 (100%)**	**12**

For the results from the prosodic system, although accuracy in the classification task varies considerably, all dialects are recognised by prosodic contour alone at above chance levels (if chance is 12.5% correct). The four ‘best’ recognised dialects are tuns (91%), kwur (75%), egca (67%) and moca (58%). The four ‘worst’ dialects are irba (50%) and ombu (42%), followed by syda and joka (both at 16%). These best and worst groupings resemble those observed in the segmental classification task on spontaneous speech data (Table 6): tuns and moca (100%), kwur and egca (75%), then irba, ombu, syda and joka (all on 67%).

5 Feature contributions to Arabic dialect classification

One obvious area of interest is identifying the specific linguistic units (i.e. phonemes or sentences) that are contributing most to distinguishing between the dialect varieties. This section presents an attempt to access this information within the inner workings of the systems. It builds on a similar attempt to achieve this in Brown and Wormald (2017), which simply applied ANOVA to Y-ACCDIST models of different British English speakers to reveal which phoneme-pairs were estimated to distinguish between four accent varieties. The work here makes use of the machine learning mechanisms implemented in this study to help identify any linguistic units or categories which might be key features in separating the varieties.

For both the segmental and prosodic systems, SVMs were used as the classification mechanism. SVMs assign different weights to the features of the models or representations they are learning. These weights help with the separation of the groups in the SVM which, in turn, should help to achieve better classification results. The weights can also be used for a feature selection process called *Recursive Feature Elimination* where they are used to rank the features by their weights, and then the weakest features are removed (either iteratively or as a specified amount, *n*). By removing features expected to be less useful to the task, it is thought that the classification performance of a system could be improved. One option is to run experiments that iteratively remove the weaker features, and observe what the effect is on classification performance. However, a more efficient, and perhaps more direct, way of accessing information about the estimated value of the different features in a feature set is to look at the full set of weights that the SVM has assigned.

In the present case, the Euclidean distances that make up the Y-ACCDIST matrices are assigned different weights, according to those that are estimated to discriminate the different dialect groups among the training data. Because we have applied a linear kernel in the SVM in these experiments, it is possible to look more closely at the weights that the different Y-ACCDIST features are assigned by the SVM, and then use them to make estimations around which features are most useful to the task of discriminating Arabic dialects. This was carried out for both the segmental system and the prosodic system to assess whether this method allows us to pick out any particular phonemes or sentence types that are particularly useful in distinguishing between the dialects. We were keen to use as much data as possible in order to observe the most reliable indications of sociophonetic variation within this dataset, so we chose to include the read speech from both the scripted dialogue and story tasks in this analysis using the segmental system.

Using the Y-ACCDIST modelling method, however, this process of drawing on SVM weights to observe individual feature contribution is not wholly straightforward. The speaker representations that we feed into the SVM are values that are computed between pairs of phonemes or pairs of sentence types (i.e. the values represented by "x" in Figures 1 and 2), rather than a single value mapping directly on to an individual phoneme or sentence type. While the modelling method has been shown to be very strong, these pairs are very difficult to disentangle to be able to observe the contribution of individual phonemes and sentence types. Clear and obvious patterns may therefore not emerge. Nevertheless, it is still of interest to see whether this method yields any insight into feature contributions and so we observe the values that we can in this section. To estimate feature contribution, we have accumulated all of the absolute weight values assigned to all the pairs of features that a single feature belongs to, and we reflect these values in boxplots.

5.1 Segmental feature contributions

Figure 3 shows this tentative measure of which phonemes appear to contribute the most weight to the task of distinguishing the eight dialect groups.

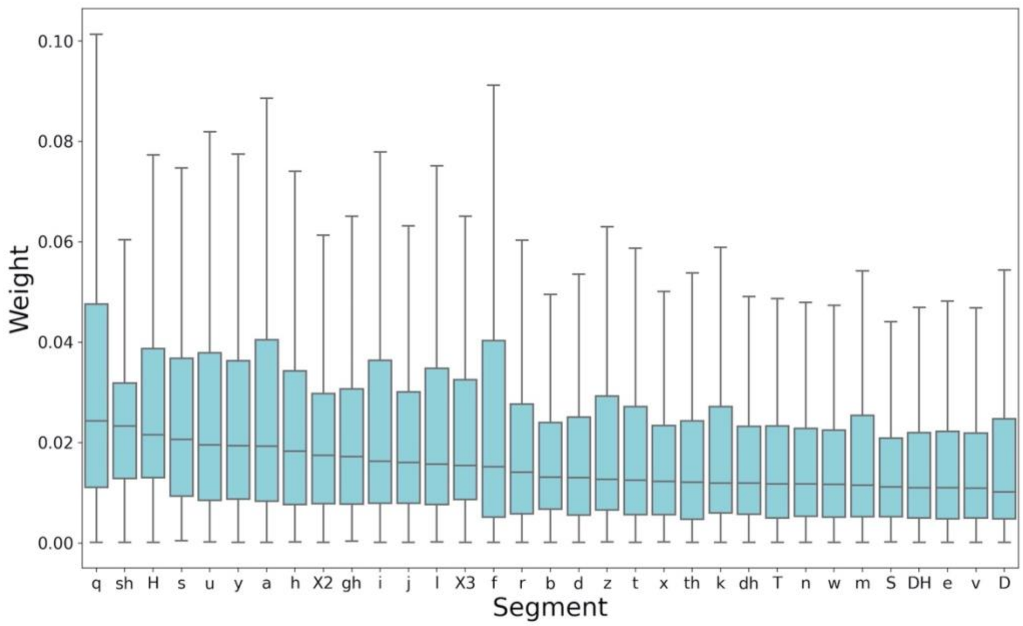

Figure 3. Boxplots to represent the distributions of feature weights associated with each phoneme segment[3].

The segments have been ordered according to the median, where we find those segments that are estimated to have the greatest contribution overall to the left, and the segments that are estimated to make the least contribution are positioned to the right. To reiterate, because of the pairwise nature of the modelling method, the visual evidence of an individual segment's contribution to a classification task is somewhat diluted. We should also keep in mind that segments are represented by midpoints and so segments that differ in terms of temporal characteristics (rather than quality characteristics) are less likely to emerge in this analysis.

The highest ranked phoneme in the feature weights boxplot is (q) /q/, matching systematic variation in the realisation of this sound across Arabic dialects (Al-Essa 2019). Similarly, the relatively high ranking of (j) /dʒ/ matches the status of that sound as a known locus of variation between dialects. For both these sounds, variation between dialects in their realisation is well-documented in the

[3] A key for the symbols used can be found here: https://reshare.ukdataservice.ac.uk/852878/15/transliteration.pdf

research literature, and they frequently appear as a variable in variationist sociolinguistic studies of individual dialects.

In contrast, most of the other highly ranked sounds in Figure 3, by feature weight value, are classes of sound which have received little attention in the research literature on Arabic sociolinguistic or dialectal variation. These include both fricatives, notably (sh) /ʃ/, (s) /s/ and (H) /ħ/, and vowels /a/ and /u/ (with /i/ not far behind). Variation across Arabic dialects in the gradient phonetic realisation of fricatives and vowels is under-researched and as a result not yet fully documented, but those few studies that exist are nevertheless consistent with the patterns seen here. For vowels, Alghamdi (1998) reports a complex pattern of differences in values of the first formant (reflecting vowel height) in data from Saudi, Egyptian and Sudanese speakers, for [a, a:, i, i:, u, u:]. Al-Tamimi & Ferragne (2005) similarly report a difference in the size of the i~a~u vowel space between Moroccan and Jordanian Arabic (with comparison also to French); furthermore, they show that Principal Component Analysis on a simple measure of vowel space size (using between-vowel-vectors for the first and second formants) yielded 88% correct classification of the three languages. For fricatives, Alsabhi et al (2020) report a main effect of *dialect* in models of standard acoustic measures of overall spectral shape (centre of gravity and peak Hz) for /s/ and /sˤ/, in experimental data elicited alongside the IVAr corpus from the same speakers and dialect groups examined in the present study.

In addition, we note that dialectal variation in realisation of (q) and (j) in Arabic can be characterised as sociolinguistically salient: the variation is above the level of awareness among speakers and may serve as a stereotype of particular dialects (Ateek & Rasinger, 2018). To our knowledge no studies have systematically investigated the relative sociolinguistic salience of different linguistic features in Arabic dialects. However, these feature weights suggest that there may be gradient sociophonetic features related to fine-grained phonetic realisation of fricatives and/or vowels, which may be below the level of phonological awareness among speakers and thus not perceived as dialect stereotypes, but which nevertheless contribute to automatic dialect classification.

We have already indicated that what draw can be drawn from the feature weights for this particular modelling method is rather limited. Having said this, there are patterns emerging that align with some expectations based on previous research, as well as patterns that have perhaps previously gone virtually unnoticed. The method has therefore focussed attention on potential features of interest and motivated future research objectives in relation to Arabic dialects.

5.2 Prosodic feature contributions

We also produced the equivalent boxplot visualisations for the prosodic system models to determine whether any sentence types were particularly influential in distinguishing between the dialects. Our conclusion here is that there seems to be very little to report, as we found very flat and invariable distributions among the different features. This will in part be due to the particular modelling method (as we have already said, a pairwise modelling approach is not the best foundation for reporting the value of individual segments). This will also be due to the fact that these features are not particularly powerful dialect discriminators, as the classification results have already demonstrated.

The combination of the lower classification result and the invariable boxplots indicates that we can expect to find a lot of variability among the intonation contours within the dialect groups. As noted earlier, however (section 4.2.2) this is exactly what we would expect, since prosodic contour realisation varies not only according to semantic categories such as question versus statement, but also due to the information structure and the wider discourse and interactional context. The lack of feature weight information thus supports the methodological choice to have Y-ACCDIST use individual sentences (realised at the same position in the dialogue sequence) as the unit of analysis.

6 Comparison of visualisations of segmental and prosodic models

It is also possible to compare visualisations of the two modelling methods for the IVAr speakers in the read speech scripted dialogue data subset. Having modelled the speakers in this data subset, we performed multi-dimensional scaling (MDS) on the data, once under the segmental configuration and once under the prosodic configuration. This allows us to observe any interesting clusters of speakers for each of the levels of analysis in isolation. These are presented in Figure 4 and Figure 5 respectively.

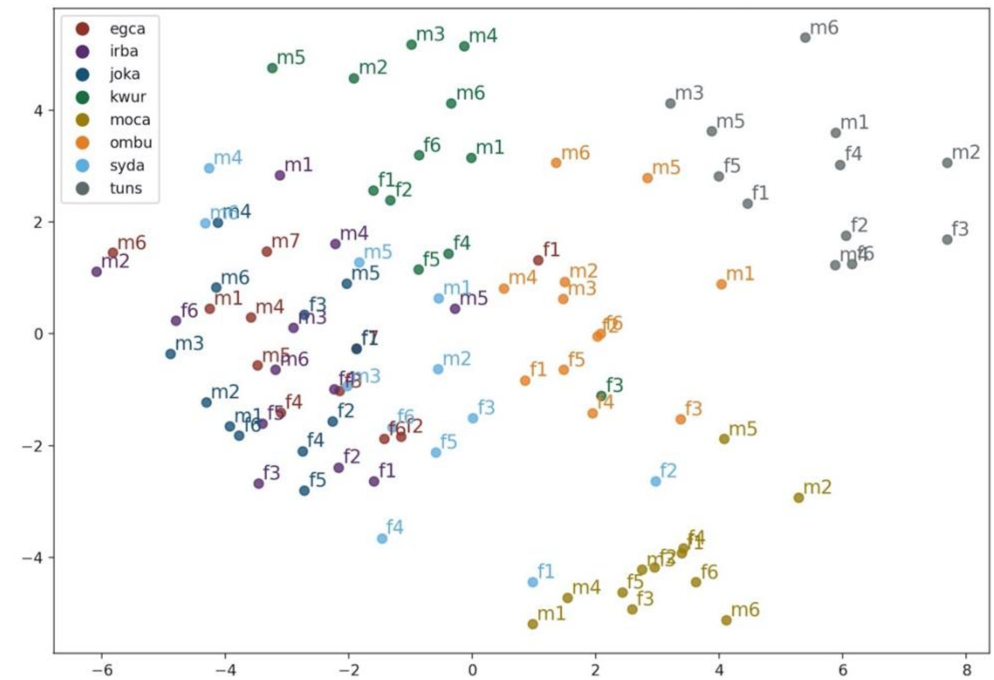

Figure 4. Multidimensional Scaling of the IVAr dataset based on segmental Y-ACCDIST modelling of speakers

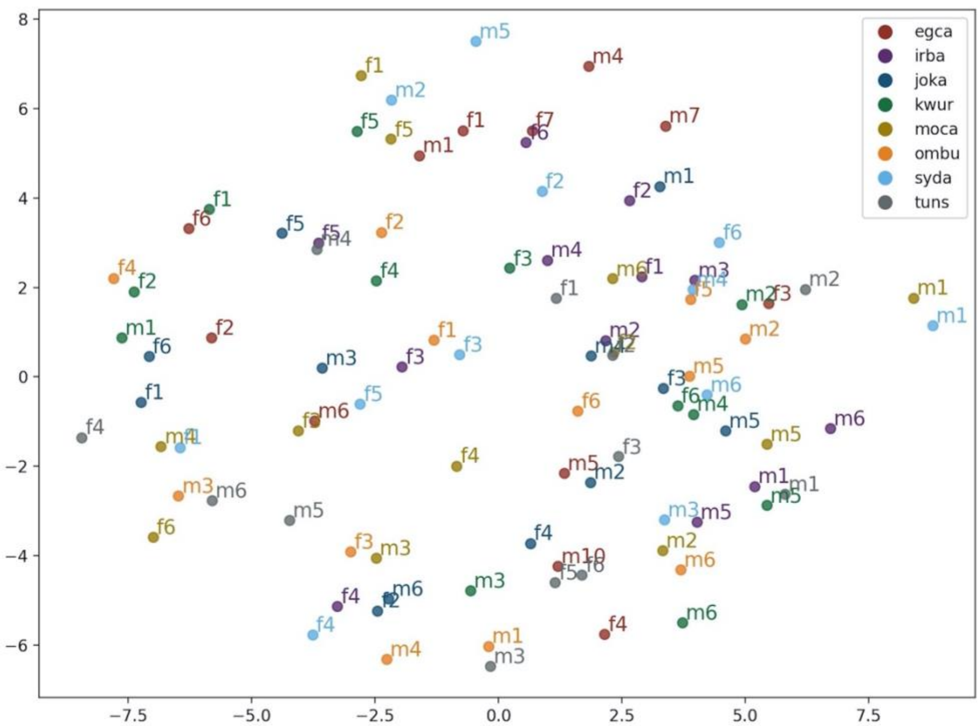

Figure 5. Multidimensional Scaling of the IVAr dataset based on prosodic Y-ACCDIST modelling of speakers.

Figure 4 shows clear groupings for most of the individual dialect groups, showing that the segmental level of analysis is a good unifier of speakers of the same dialect. This is consistent with the very high classification result that this version of the system achieved. The clustering reflects both the geographical spread of dialects and their position in the Arabic dialect continuum. The clusters of speakers from Egypt, Iraq and the Levant (Jordan and Syria) overlap somewhat, towards the left of the plot, but corresponding to their more central geography and position in the middle of the dialect continuum. The Gulf dialects (Kuwait/Oman) are distinct from each other, matching their positions at extreme north and southern ends of the Gulf Arabic dialect group, but both equally separate and distinct from the central dialects. Similarly, the North African dialects (Tunisia/Morocco) are clearly separated from each other, again matching their geographical and dialectal separation within the Maghreb group, but are both equally separate and distinct from the central dialects, and placed at a greater distance from the central group than the Gulf dialects, reflecting the clear east/west divide noted in section 2.1.

Figure 5 for the prosodic model shows a less clear clustering of individual dialect groups. The relatively tight cluster of Tunisian speakers just right of centre of Figure 5 perhaps aligns with the very high classification rate achieved for Tunisian speakers in the prosodic experiments, and interestingly, the Egyptian speakers seem to form a somewhat more consistent cluster compared to the other dialect groups. Such a clustering for these two dialects would be consistent with some of the more distinctive prosodic patterning in these dialects found in previous work by the second author, namely a distinctive rise-plateau contour in yes/no-questions in Tunisian Arabic (Hellmuth, 2018) and overall higher frequency in the distribution of prosodic peaks in Egyptian Arabic compared to other dialects (Hellmuth 2007, 2020).

7 Discussion

This paper has demonstrated approaches to analysing dialect variation that take into account whole collections of features, rather than just focussing on a single feature and seeing how it varies across different dialects. These approaches are reliant on there being an “inventory” of categories for the models to work with. In the case of the segmental system, this is the phoneme inventory, and in the case of the prosodic system, this is a range of different sentences produced at different points in a scripted dialogue. These aggregate approaches therefore currently only offer a broad-brush account of the variation in a dialect dataset on either one of these levels of analysis, rather than a detailed account of exactly which feature is discriminating the different varieties.

In the case of the prosodic version of the system, we have presented the modelling approach only on a subset of read speech data in which we could guarantee balance and control in the different sentence types that we used. The corpus was originally designed for prosodic research to be conducted on it and so other analysis on the prosodic variation in this dataset had already been done which opened up the opportunity to corroborate results or to even uncover surprising findings. Having tested the approach on these very controlled data and having discovered that it appears to have some value in capturing the variation and distinguishing between the dialects, it is natural to now consider how it could be transferred to spontaneous speech data. Given a dataset of spontaneous speech recordings that have been tagged for key features such as sentence type and information structure, we could evaluate the prospect of using this approach on spontaneous recordings. In addition, it could be that a larger dataset than the one used in this work would be required to achieve a more stable representation of the specific variation that exists in Arabic prosody.

There has been some other work that has actively sought to integrate prosody's potential role in the automatic classification of Arabic dialects. Biadsy and Hirschberg (2009) modelled spontaneous speech utterances using a combination of features relevant for intonation and rhythm. They captured a selection of pitch and rhythm values to represent whole utterances (e.g. capturing the variation in pitch and vocalic proportion of an utterance). These were termed more "global" measurements. They then went on to characterising utterances with "sequential prosodic features", which logged various characteristics of the pitch and intensity contours of utterances. On a broad four-way Arabic dialect classification task, the "global" features alone achieved 60% accuracy, whereas when more sophisticated sequential features were combined with them, they achieved 72% accuracy. Between Biadsy and Hirschberg's study and the present one, there are many differences to do with the size and the dimensions of the datasets used, but it may be of interest in future to compare these two methods like-for-like. One key difference is that Biadsy and Hirschberg applied their prosodic modelling method to spontaneous speech, a natural next step for the Y-ACCDIST modelling method implemented in the present study.

Although the Y-ACCDIST modelling approaches themselves are very adaptable and can feasibly be used on large datasets of speech recordings, there is some manual preprocessing of the data (i.e. either broad transcription or tagging) that is required before modelling, classification and visualisation can take place. It could be possible to overcome this preprocessing by either automatically transcribing or tagging a corpus, but this will inevitably introduce errors. Work on this less labour-intensive version is currently ongoing.

8 Conclusion

In this paper we have focussed on automatically classifying speakers of Arabic into different dialect groups and considered the systems' outputs in the context of an interest in the variation among Arabic dialects. We have demonstrated how these kinds of system can both reinforce what we know about a set of linguistic varieties, but also how it could possibly illuminate new questions to pursue around certain features. Previous work has shown that we can do this on one level of analysis, but part of this work has demonstrated that sometimes performance might be too high for us to learn about a set of dialects from the errors that a system makes. However, this paper has demonstrated that it is possible to transfer similar modelling principles that have been used for one level of analysis across to another. By isolating the segmental level of analysis and then the prosodic level in a similar modelling framework, we can observe the contribution of each level of analysis to distinguishing between the classification of a particular set of varieties. It is probably no surprise that the segmental level outperforms the prosodic level in a simple dialect classification task, but the prosodic version of the system showed a performance that sat well above the level we would expect if the system were working by chance. This difference in performance between the two levels of analysis is likely to be down to the fact that one forms models based on a full and well-established phoneme inventory and the other makes use of a (partly arbitrary) list of target sentences. The former is both more fine-grained and more controlled than the latter.

In the context of Arabic dialects, we were able to corroborate some of the findings surrounding the prosodic system's outputs with past prosodic analyses conducted on the same data. The work here was also able to indicate that Arabic has a wealth of sociophonetic variation to discover at the segmental level, which is arguably under-explored in Arabic dialects. The detail of this segmental variation cannot be accurately uncovered by the macro-level computational method implemented in this work, but would require other more detailed methods to gain a richer understanding.

Data availability

The corpus that was used for this study can be found at the following reference:

Hellmuth, S., & Almbark, R. (2019). *Intonational Variation in Arabic Corpus (2011-2017)*. Retrieved from: http://reshare.ukdataservice.ac.uk/852878/

Author contributions

GB took the lead in the technology development, computational methods and generating results. SH took the lead in data collection, data processing and interpretation of results. The authors contributed to the planning and writing of this article in equal measure.

Acknowledgments

The authors would like to thank Rana Alhussein Almbark for her valuable input towards the beginning of the project when planning and executing the data processing required for this project.

References

Abdel-Massih, E.T. 2011. *An introduction to Egyptian Arabic*. Ann Arbor, University of Michigan.

Al-Essa, Aziza (2019) Phonological and morphological variation. In E. Al-Wer, & U. Horesh (Eds.), *Routledge Handbook of Arabic Sociolinguistics*. Routledge. Pp151-168.

Al-Tamimi, J.-E., & Ferragne, E. (2005). Does vowel space size depend on language vowel inventories? Evidence from two Arabic dialects and French. Paper presented at the 9th European Conference on Speech, Communication and Technology (Interspeech 2005), Lisbon, Portugal.

Alghamdi, M. M. (1998). A spectrographic analysis of Arabic vowels: A cross-dialect study. *Journal of King Saud University*, 10(1), 3-24.

Ali, A., Shon, S., Samih, Y., Mubarak, H., Abdelali, A., Glass, J., Renals, S. and Choukri, K. (2019). The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. *Proceedings of Automatic Speech Recognition and Understanding.* (Sentosa, Singapore). 1026-1033.

Alsabhi, M., Bailey, G and Hellmuth, S. (2020) Cross-dialectal variation in phonetic realisation of the emphatic contrast in Arabic fricatives. Paper presented at the 4th Arabic Linguistics Forum, 30th June-2nd July 2020, Leeds.

Ateek, M., & Rasinger, S. M. (2018). Syrian or non-Syrian? Reflections on the use of LADO in the UK. In I. M. Nick (Ed.), *Forensic Linguistics: Asylum-seekers, Refugees and Immigrants* (pp. 75-93): Vernon Press.

Barkat, M., Ohala, J., & Pellegrino, F. (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. *Eurospeech 99*, 395-398.

Beckman, M., & Elam, G. A. (1997). *Guidelines for TOBI Labelling (version 3.0 1997)*.

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic Typology: The phonology of intonation and phrasing* (pp. 9-54). Oxford: Oxford University Press.

Behnstedt, P., & Woidich, M. (2013). Dialectology. In J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics* (pp. 300-325). Oxford: Oxford University Press.

Biadsy, F., & Hirschberg, J. (2009). Using Prosody and Phonotactics in Arabic Dialect Identification. *Proceedings of Interspeech 2009 (Brighton, UK)*, 208-211.

Biadsy, F., Hirschberg, J., & Habash, N. (2009). Spoken Arabic Dialect Identification using Phonotactic Modeling. *Proceedings of the EACL Workshop on Computational Approaches to Semitic Languages. (Athens, Greece),* 53-61.

Brown, G. (2014). Y-ACCDIST: An automatic accent recognition system for forensic applications. Master Thesis, University of York, UK.

Brown, G. (2015). Automatic recognition of geographically-proximate accents using content-controlled and content-mismatched speech data. *Proceedings of the 18th International Congress of Phonetic Sciences. (Glasgow, UK)*. Paper 458.

Brown, G. (2016). Automatic accent recognition systems and the effects of data on performance. *Proceedings of Odyssey: the speaker and language recognition workshop*. *(Bilbao, Spain).* Paper 29.

Brown, G. (2018). Segmental content effects on text-dependent automatic accent recognition. *Proceedings of Odyssey: the speaker and language recognition workshop. (Les Sables d'Olonne, France*). 9-15.

Brown, G., & Wormald, J. (2017). Automatic Sociophonetics: Exploring corpora with a forensic accent recognition system. Journal of the Acoustical Society of America. 142. 422-433.

Contini, M. and Romano, A. (2002) Atlas Multimédia Prosodique de l'Espace Roman. URL http://dialecto.u-grenoble3.fr/AMPER/new.htm.

D'Arcy, S., Russell, M., Browning, S. and Tomlinson, M. (2004). The Accents of the British Isles (ABI) corpus. *Proceedings of Modélisations pour l'Identification des Langues. (Paris, France).* 115-119.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic Word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing. 28.

Elvira-García, W., Balocco, S., Roseano, P., & Fernández-Planas, A. M. J. S. C. (2018). ProDis: A dialectometric tool for acoustic prosodic data. *Speech Communication, 97*, 9-18.

Ferragne, E., & Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*. 38. 526-539.

Grabe, E., Kochanski, G., & Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech, 50*(3), 281-310.

Hanani, A., Russell, M., & Carey, M. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech and Language*. 27. 59-74.

Hellmuth, S. (2007). The relationship between prosodic structure and pitch accent distribution: Evidence from Egyptian Arabic. *The Linguistic Review, 24*(2-3), 289-314.

Hellmuth, S. (2018). *Variation in polar interrogative contours within and between Arabic dialects. Proceedings of the 9th International Conference on Speech Prosody*. *(Poznań, Poland).* 989-993.

Hellmuth, S. (2020). Contact and variation in Arabic intonation. In C. Lucas & S. Manfredi (Eds.), *Arabic and contact-induced change: a handbook.* Berlin: Language Science Press.

Hellmuth, S. (to appear). *Intonation in spoken Arabic dialects*. Oxford: Oxford University Press.

Hellmuth, S., & Almbark, R. (2019). *Intonational Variation in Arabic Corpus (2011-2017)*. Retrieved from: http://reshare.ukdataservice.ac.uk/852878/

Hualde, J., & Prieto, P. (2016). Towards an International Prosodic Alphabet (IprA). *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 7*(1).

Huckvale, M. (2004). ACCDIST: a metric for comparing speakers' accents. *Proceedings of the International Conference on Spoken Language Processing. (Jeju, Korea).* 29-32.

Huckvale, M. (2007). ACCDIST: an accent similarity metric for accent recognition and diagnosis. In C. Müller (Ed.) *Lecture Notes in Computer Science: Speaker Classification*. Vol 2. Berlin Heidelberg: Springer-Verlag. 258-274.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica, 55*(3-4), 243-276.

Najafian, M., Khurana, S., Shon, S., Ali, A., & Glass, J. (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). *(Calgary, Canada.)* 5174-5178.

Owens, J. (2013). A house of sound structure, of marvelous form and proportion: An introduction. In J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics.* (pp. 1-22). Oxford: Oxford University Press.

Retsö, J. (2013). What is Arabic? In J. Owens (Ed.), *The Oxford Handbook of Arabic Linguistics* (pp. 433-450). Oxford: Oxford University Press.

Shon, S., Ali, A., & Glass, J. (2018). Convolutional neural network and language embeddings for end-to-end dialect recognition. In *Proceedings of Odyssey: the speaker and language recognition workshop.* Les Sables d'Olonne, France. 98-104.

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., & Carmiel, Y. (2017). Deep neural network embeddings for text-independent speaker verification. Proceedings of Interspeech. (Stockholm, Sweden). 165-170.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Versteegh, K. (2014). *The Arabic Language*. Edinburgh University Press.

Walker, T. (2014). Form (does not equal) function: The independence of prosody and action. *Research on Language and Social Interaction*, 47(1), 1-16.

Wormald, J. (2016). Regional Variation in Punjabi-English. PhD thesis. University of York, UK.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2009). *The HTK Book for HTK* Version 3.4. Cambridge University Engineering Department, Cambridge.

Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing.* 4. 31-44.