A Unified Approach for Arabic Language Dialect Detection

Rania R. Ziedan*[1,2], Michael N. Micheal[2], Abdulwahab K. Alsammak[2],
Mona F.M.Mursi[2] and Adel S. Elmaghraby[1]
[1]Computer Engineering and Computer Science, University of Louisville, KY, USA
[2]Communication and Computer Engineering, Benha University, Egypt
*E-mail: rania.ziedan@louisville.edu / rania.ziedan@feng.bu.edu.eg

Abstract

The paralinguistic information in a speech signal includes clues to the ethnic and social background of the speaker. In this paper, we propose a hybrid approach to dialect and accent recognition from spoken Arabic language, based on phonotactic and spectral systems separately then combining both by decision fusion technique. We extract speech attribute features that represent acoustic cues of different speaker's dialect to obtain feature streams that are modeled with the Gaussian Mixture Model with Universal background model (GMM-UBM) in addition to Identity Vector (I-vector) classifier. Moreover, this paper introduces our proposed dataset SARA, which is a Modern Colloquial Arabic dataset (MCA) contains three different Arabic dialects and its common accents, this dataset will be the master dataset for this work. We find our proposed technique with acoustic features achieves a significant performance improvement over the state-of-the-art systems using Arabic dialects in the dataset.

keywords: accent/dialect recognition, I-vector, GMM-UBM, SARA colloquial Arabic dataset.

1 Introduction

Recently, human-machine interaction has received increasing attention from various fields such as artificial intelligence, machine learning, and information retrieval. One of the most important challenges in human-machine interaction is the proper understanding of human speech by automated systems. The recognition of the speech by a machine permits a deeper interaction between both parties.

Figure 1 shows the most popular speech processing research areas. The fundamental challenge for current research is the understanding and modeling of the individual variation in the dialect and accent based on speech. The dialect refers to the linguistic variations of a language; however, the accent refers to the different ways of pronouncing a language within a community.

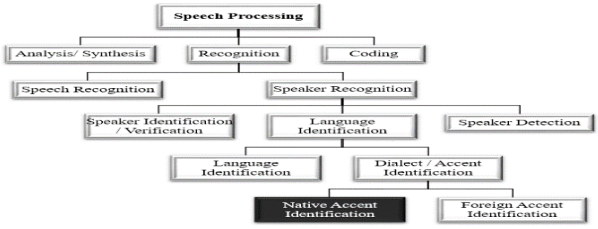

Figure 1: Automatic Speech Processing Research Areas

The nonnative speakers pose many problems for automatic speech recognition systems' (ASR) performance. In addition, the nonnatives adversely affect the speaker verification systems' performance because of the systematic shifts in score distributions relative to the native speakers. Therefore, knowing the nativeness of a speaker would enable the adaptation techniques to mitigate the mismatch between the training and the test data. Moreover, identifying the nativeness of the speaker is useful in many intelligent applications.

After Chinese, Spanish and English, Arabic is the fourth most commonly spoken language around the world with more than 230 million native speakers [2] and it is the official language for more than 22 countries. In addition, the world's Muslims, around one billion people, also use Arabic as a religious language; however, it is not always spoken as the way it is written. The Modern Standard Arabic (MSA) is usually used as a writing form for the official Arabic language, and it is the language of literature, books, newspapers, and the official documents, whereas, the spoken form varies based on the geographical region. Little research has been done on processing Arabic speech especially for Arabic dialects. For Arabic language, in addition to MSA there are a number of regional dialects. These dialects are mainly grouped based on the geographical regions into the Maghrebi group, the Sudanese group, the Egyptian group (EGY), the Arabian Peninsula group (ARP), the Iraqi group (IRQ), and the Levantine group (LEV) each group includes some accents. For Arabic dialects recognition, there are two important challenges that must be considered, which are 1) the

978-1-943436-04-0 / copyright ISCA, CAINE 2016
September 26-28, 2016, Denver, Colorado, USA

dialects have no standards, and 2) the large gap between MSA and the dialects, which led some research studies to classify the Arabic dialects as separate languages. In the dialect recognition systems, the weight of a dialect feature depends on its distance from the standard pronunciation, and the frequency of that feature in the speech. The differences between dialects can be found in two parts, which are the differences in phonetic transcriptions and the differences in acoustical intonations of dialects. The differences in phonetic transcription can be categorized into two classes, differences in the number and identity of the phonemes and differences in phonetic realizations such as phoneme substitution, deletion, and insertion. There are some examples of phonetic transcription differences in different Arabic accents in [7]. In addition to the differences in phonetic transcription, there are four differences of acoustic correlates of dialects, which are formant, pitch prosody correlates, Timing correlates and laryngeal (glottal) correlates. In this paper, we focus on recognizing the Arabic dialects based on the acoustical feature extraction and classification techniques.

The rest of the paper is organized as follow. In section 2, related work is described. In section 3, we present the proposed dialect and accent recognition approach. The proposed dialected dataset is given with the most popular regional accents in section 4. For the acoustic features, the experimental environment is presented in section 5 then followed by the results in section 6. Finally, section 7 introduces the conclusion and future work presented in section 8.

2 Related Work

This section introduces the challenges in the speech and dialect recognition systems. The first challenge is the selection of the suitable features that represent the dialect differences; various features extraction algorithms are used in speech dialect analysis, which are different in the contexts, the meanings, and the configurations such as:

1- Mel Frequency Cepstral Coefficients (MFCC) is the most popular feature extraction technique in the speech recognition areas, it is based on frequency domain using the Mel scale which is based on the human ear scale [10, 13]. In addition, these coefficients are robust and reliable to variations according to the speakers and the recording conditions.

2- Shifted Delta Cepstral coefficients (SDC), which created by stacking delta cepstra computed across multiple speech frames. It based on the concept that the feature vectors include the temporal information spanning to a large number of frames [16]. Four parameters are used to calculate the SDC features, which are the number of spectral coefficients calculated at each period N, the time delay and advance for calculating the delta d, the number of blocks for which delta coefficients are concatenated k, and the time delay between consecutive blocks P.

3- Perceptual Linear prediction (PLP) model developed by Hermansky [14]. PLP enhances the speech recognition rate by discarding the irrelevant information of the speech.

4- In addition to PLP, Hermansky developed Relative Spectra Filtering (RASTA) where the conventional critical-band short-term spectrum in PLP is replaced with a spectral estimate from frequencies band-pass filtered via a sharp spectral zero at zero frequency in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel [15, 21]

5- RASTA-PLP [15] is a speech feature representation, which is a hybrid from RASTA and PLP steps.

The second challenge is building a classifier model that able to handle and combine efficiently the heterogeneous structure of the acoustic features.

1- GMM-UBM: GMM is one of the most popular classifiers for speaker recognition, due to its capability to represent a large class of sample distributions, its components considered as a model for the underlying broad phonetic sounds that characterize a person's voice [19]. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm, which maximize the likelihood of the GMM given the training data. In speech applications, the adaptation of the acoustic models to new operating conditions is important because of data variability due to different speakers, environments, speaking styles and so on[18]. The universal background model (UBM) which is the M-component of the GMM parametrized by $w_m, m_m, \sum_m, m = 1, ..., M$, where w, m, and $\sum$ are the mixture weight, mean vector, and covariance matrix adapted to a specific speaker using a maximum a posteriori (MAP) scheme. The basic idea in the adaptation approach is when enrolling a new speaker to the system, the speaker's model is derived by updating the well-trained parameters in the UBM via adaptation [20]. This provides a tighter coupling between the speaker's model and UBM, which produces better performance and allows a fast-scoring technique than decoupled models.

2- I-Vector: Recently, Dehak [11] developed a new classifier based on Joint Factor Analysis (JFA) as feature extractor. The idea is finding a low dimensional subspace of the GMM super-vector space, named total variability space that represents the speaker and

channel variability. The vectors in that low-dimensional space are called I-vectors, this low dimension space is classified using PLDA classifier which represented by Kenny [17] in the speaker verification systems. The representation of the I-vector has a small size to reduce the execution time of the recognition time while keeping the recognition rates acceptable. I-vector classifier is tested with different accents in [4, 5, 12], in [6]it uses some Arabic Speakers to test the Finnish language Proficiency. According to our knowledge, the only work used I-vectors classifier for an Arabic dialect is represented by Boulkenafet et al. [9] for the Algerian Arabic dialect, which one of Maghrebi group dialects.

In addition, the proper selection of a speech corpus is a challenge to evaluate the dialect/ accent detection system's performance. Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations, and government research laboratories; it is one of the most common speech/text database provider since1996. In addition to LDC, The European Language Resource Association (ELRA) Agency has a great effort in language resources for the Human Language Technology (HLT) since 2007. Moreover, there are some researchs done to collect Arabic dialected corpus mainly for the academic use in dialect/ accent recognition systems. The summarization in [3], for the Arabic corpus discusses the existing Arabic MSA and Arabic dialected corpus recording conditions and specifications.

3 Proposed Approach

In this research, we propose an approach for recognizing the Arabic dialects from speech. The approach is divided into two systems as shown in figure 2, one based on acoustic features and the other based on phonetic features. The first system is responsible for speech signal analysis to extract the acoustic features that discriminate the Arabic regional dialects. In training stage, these extracted features are used to build a dialect/ accent model in the Acoustic Accent Lexicon. In the test phase, the extracted features are used to determine the speaker nativity using mapping technique. However, the second system is based on phonotactic representation of the speech, a phonetic pattern for the dialect / accent distinctive words is trained, and the phonetic model for each accent word is saved in a lexicon. Then in the recognition phase, the system will search for the accent phonetic pattern for matching and recognize the accent. Finally, the decision is taken by score-level fusion between the acoustic and phonetic systems decisions.

According to our knowledge, the only research recognized the Arabic dialects based on acoustic and

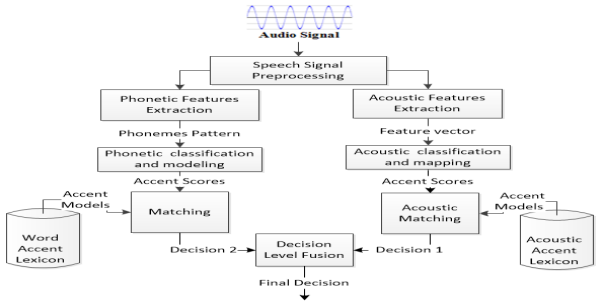

Figure 2: Proposed Approach

phonotactic characteristics of speech separately was proposed by Hynek et al.[8] , which proposed a study that included two parts. The first part identified the acoustic characteristics of the dialect based on the spoken part and the silent part in the speech. In addition, the second part focused on phonotactic dialect modeling that utilizes phone recognizers and support vector machines (PRSVM). However, Hynek et al. used the English and Hindi phoneme recognizers to identify the dialect phonetic characteristics of the Arabic language. However, in our proposed approach we intend to use the Arabic phone recognizer for recognizing the Arabic phonemes.

4 Proposed Dataset

The speech corpus is collected to be spontaneous, canonical, or both [1]. A spontaneous speech corpus is the speech corpus that is collected from the real world, human–human or human machine communication. For this corpus, the speaker does not read prompts, rather he or she speaks naturally to convey a message and/or get information, and the speech is expected to be natural and not affected by the reading habits. However, in the canonical speech corpus, the speakers must follow certain procedures, including the reading of prompts, for collecting specific speech sounds. The speech content can be words or sentences, the sentences may be phrases, phonetically rich sentences. The spontaneous speech corpus is suitable for language understanding and dialogue design; it tends to include unneeded frequently repeated words and utterances but does not necessarily include all the sounds of the language under investigation. However, a canonical speech corpus tends to be phonetically rich; all the sounds of the language are presented in various phonotactic positions.

In this paper, we propose a database consists of a set of spontaneous not pre-specified colloquial phrases in everyday life and life situations that are collected from media shows, episodes and films published on YouTube played by native speakers with three different Arabic dialects and accents. We define the proposed dataset as

Spoken Arabic Regional Archive (SARA). In contrast to Voice over IP (VoIP), which is defined as a real time delivery of the voice packets through the network using internet protocols, the media for SARA are downloaded from YouTube and stored before sampling process. Therefore, the live transmission quality of service (QoS) problems like delay, delay jitter, and unreliable packet delivery, which VoIP suffers from, do not affect our proposed dataset SARA. The subjected database can be used to train speech-processing systems such as automatic speech recognition, speaker verification/ recognition and dialect/ accent recognition. SARA dataset contains only adult speakers to avoid the improper pronunciation of the children that can affect the detection process. The proposed database contains males' and females' speech for three Arabic regional dialects, which are Egyptian dialect (EGY), Arabian Peninsula dialect (ARP) and the Levantine dialect (LEV). In addition, within each dialect group there are a number of different accents. Each utterance contains a one-speaker speech, the number of speakers within the dialect and the number of utterances by a speaker is unknown. The categorization of the proposed dialects and their most popular accents in the dataset is proposed in figure 3 In addition, the three different dialects are explained briefly with numbers in table 1.

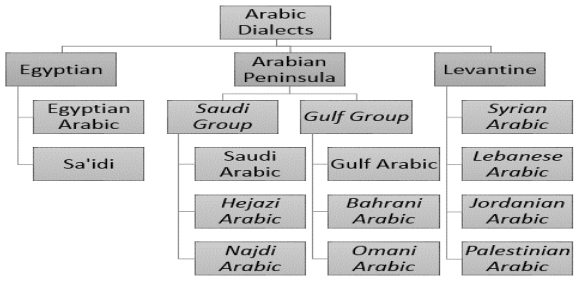

Figure 3: Proposed corpus distribution

Table 1: Corpus Utterances' Dialects Distribution

Dialect	Male	Female	Total
EGY	877	488	1365
ARP	657	573	1230
LEV	583	662	1245

For EGY, the number of males' and females' speech samples is distributed according to the accent as in table 2. Tables 3, 4 and 5the distribution for the ARP dialect and its common accents in the proposed corpus is shown. In table 6 the LEV dialect and accent distribution is shown. In addition, table 7proposes that the dataset samples are variant in length in order to verify the minimum time in which we can determine the speaker dialect or accent when the speaker speaks in free talk.

Table 2: EGY Accents Distribution

Accent	Male	Female
Egyptian	623	221
Sa'idi	254	267

Table 3: ARP Accents Distribution

Accent	Male	Female
Saudi	180	207
Gulf	477	366

Table 4: Saudi Accents

Accent	Male	Female
Saudi	143	203
Hejazi	29	0
Najdi	8	4

Table 5: Gulf Accents

Accent	Male	Female
Bahrani	335	282
Gulf	64	46
Omani	78	38

Table 6: LEV Accents Distribution

Accent	Male	Female
Syrian	46	261
Lebanese	76	341
Jordanian	362	42
Palestinian	99	18

Table 7: Dialected utterances' length distribution

Length	EGY	ARP	LEV	Total
3 sec	468	388	398	1254
4 sec	377	409	354	1140
5 sec	271	269	246	786
6 sec	166	117	158	441
7 sec	83	47	89	219
Total	1365	1230	1245	3840

5 Experiment Environment

In this paper, the proposed acoustic system shown in figure 2 is used to identify the different dialects of SARA. The samples used in this experiment are divided into approximately 70% and 30% for training and testing. These samples are gender-independent to identify the dialect features regardless of the speaker gender features. The dialect acoustic features are extracted using 12 MFCC coefficients with the log energy

component (a), delta MFCC (c), double delta MFCC (e) and SDC with parameters (7,1,3,7) (g), and these techniques are used with and without cepstral mean and variance normalization. These features normalization form are denoted as (b, d, f and h) respectively. In addition, we also use PLP (w), double delta PLP (x), RASTA (y) and Rasta with PLP coefficient (z) techniques with PLP order =12. Moreover, these features are mainly classified using the classification technique GMM-UBM with 1024 mixtures. In addition, the I-vector is used with total variability equals 60 and classified by the PLDA classifier. The result is evaluated by computing EER, which is the rate at which both acceptance and rejection errors are equal.

6 Experimental Results

From the results, the best recognition rates are achieved when the utterance length is more than 4 seconds. In the most features techniques, when increasing the sample size, the EER decreased. In the 3 seconds case, the EER increased up to 2% using GMM-UBM and 4% using I-vector depending on the feature extraction technique. As shown in table 8 and table 9 when using GMM-UBM, the best accuracy found when using normalized delta MFCC, EER varies from 14.2% to 5.6% depending on sample size. From the experiments, we noticed that in the case of the utterances longer than 6-seconds with MFCC features, in the GMM classifier case the classifier misclassified all the test samples as Egyptian dialect. Moreover, using the GMM-UBM produces the same classification result. Table 8 shows that using the normalized features solve this problem in long utterances. However, table 10 and table 11 show the EER of the I-vector classifier with the MFCC features, delta MFCC and double delta MFCC give comparable EER values, which varies from 20% to 11% depending on sample size. Like GMM-UBM, using RASTA and Rasta-PLP gives the worst results. EER varies from 28% to 14% depending on sample size.

It is noticed that RASTA and Rasta-PLP features gave worse results than PLP features in continuous dialected speech application in contrast to speech recognition application.

Table 8: GMM-UMB classifier EER %

	a	b	c	d	e	f
3sec	14.2	16.3	12.5	13.8	12.5	12.3
4sec	14.8	16.7	14.2	14.2	13.8	15.4
5sec	10	13.3	10.7	13	11.3	13
6sec	12.2	15	11.1	11.1	11.7	12.2
7sec	66.7	13.9	66.7	5.6	66.7	9.7

Table 9: GMM-UMB classifier EER %

	g	h	w	x	y	z
3sec	22.9	20	16.9	14.6	25	26.3
4sec	21.7	19.6	17.1	14.6	25	25
5sec	21	15.7	14.7	13.7	26	24.7
6sec	23.3	21.1	15.6	14.4	26.7	23.3
7sec	16.7	19.4	16.7	12.5	15.3	19.4

Table 10: PLDA classifier with I-vector EER %

	a	b	c	d	e	f
3sec	19.2	20.8	19.6	19.2	18.8	20
4sec	18.8	20	18.3	18.8	17.5	18.8
5sec	15	18	16	16.7	16	16.7
6sec	16.1	11.1	14.4	13.3	13.3	15
7sec	18.1	13.9	16.7	13.9	13.9	13.9

Table 11: PLDA classifier with I-vector EER %

	g	h	w	x	y	z
3sec	26.3	28.8	22.1	20.4	27.7	28
4sec	22.5	19.6	19.4	19.2	23.8	23.3
5sec	20	21.3	16.3	18	23	23.3
6sec	23.3	21.1	18.9	17.8	22.8	21.1
7sec	22.2	19.4	19.4	15.3	13.9	18.1

7 Conclusion

In this paper, we introduced a new approach for dialect/ accent recognition based on the fusion between the acoustic and phonetic features. In addition, we explained the specifications of our proposed dialected Arabic dataset SARA. In addition, The I-vector technique that is used with English, British, and Finnish languages is used to classify the Arabic dialects and compared to GMM-UBM.

8 Future Work

We believe that there is a significant room for improvement by including some factors that we did not explicitly model in this paper. These factors may include gender dependent recognition, recognizing the speaker nationality rather than only their region. Additionally we plan to apply fusion techniques at the feature level to improve the recognition rate. In addition, we will make our dataset SARA available to the researchers for comparative studies. We also plan to expand this archive with other dialects.

References

[1] Mansour Alghamdi, Fayez Alhargan, Mohammed Alkanhal, Ashraf Alkhairy, Munir Eldesouki, and Ammar Alenazi. Saudi accented arabic voice bank.

Journal of King Saud University-Computer and Information Sciences, 20:45–64, 2008.

[2] Khalid Almeman and Mark Lee. A comparison of arabic speech recognition for multi-dialect vs. specific dialects,". In *Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), Cluj-Napoca, Romania*, pages 16–19, 2013.

[3] Khalid Almeman, Minhung Lee, and Ali Abdulrahman Almiman. Multi dialect arabic speech parallel corpora. In *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*, pages 1–6. IEEE, 2013.

[4] Mohamad Hasan Bahari, Rahim Saeidi, David Van Leeuwen, et al. Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7344–7348. IEEE, 2013.

[5] Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen. Factors affecting i-vector based foreign accent recognition: A case study in spoken finnish. *Speech Communication*, 66:118–129, 2015.

[6] Hamid Behravan, Ville Hautamauki, Sabato Marco Siniscalchi, Tomi Kinnunen, and Chin-Hui Lee. Introducing attribute features to foreign accent recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 5332–5336. IEEE, 2014.

[7] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pages 53–61. Association for Computational Linguistics, 2009.

[8] Hynek Boril, Abhijeet Sangwan, and John HL Hansen. Arabic dialect identification-'is the secret in the silence?'and other observations. In *INTERSPEECH*, pages 30–33, 2012.

[9] Z Boulkenafet, Messaoud Bengherabi, Farid Harizi, Omar Nouali, and Cheriet Mohamed. Forensic evidence reporting using gmm-ubm, jfa and i-vector methods: Application to algerian arabic dialect. In *Image and Signal Processing and Analysis (ISPA), 2013 8th International Symposium on*, pages 404–409. IEEE, 2013.

[10] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1(6):1–4, 2013.

[11] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Interspeech*, volume 9, pages 1559–1562, 2009.

[12] Andrea DeMarco and Stephen J Cox. Native accent classification via i-vectors and speaker compensation fusion. In *INTERSPEECH*, pages 1472–1476, 2013.

[13] Taabish Gulzar, Anand Singh, and Sandeep Sharma. Comparative analysis of lpcc, mfcc and bfcc for the recognition of hindi words using artificial neural networks. *International Journal of Computer Applications*, 101(12):22–27, 2014.

[14] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[15] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, 1994.

[16] Herman Kamper and Thomas Niesler. A literature review of language, dialect and accent identification. Report, 2012.

[17] Patrick Kenny. Bayesian speaker verification with heavy-tailed priors. In *Odyssey*, page 14, 2010.

[18] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[19] Douglas Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, pages 827–832, 2015.

[20] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[21] Urmila Shrawankar and Vilas M Thakare. Techniques for feature extraction in speech recognition system: A comparative study. *arXiv preprint arXiv:1305.1145*, 2013.