West Point, SAAVB, and BBN/AUB Arabic Speech Corpora: A Comparative Survey

Yousef Ajami Alotaibi, Ali Hamid Meftah

Computer Engineering Dept., College of Computer & Information Sciences, King Saud University, Riyadh, Saudi Arabia, Email: {yaalotaibi, ameftah}@ksu.edu.sa

Abstract—The aim of this paper is to evaluate three public Arabic speech corpora, namely the West Point (WP), Saudi Accented Arabic Voice Bank (SAAVB) and the BBN Technologies/American University at Beirut (BBN/AUB) corpus by referring the TIMIT English speech corpus as benchmark. Weaknesses, strengths, and discrepancies of these Arabic corpora regarding their design and content are covered in this paper. This paper is very important to Arabic speech processing because Arabic is one of the under resourced language despite its importance and popularity. Currently, we are considering WP and BBN/AUB corpora to analyse and study Arabic rhythm in our ongoing research project.

Keywords—Arabic language; TIMIT; West Point; SAAVB; BBN/AUB.

I. INTRODUCTION

Standard Arabic has 34 basic phonemes, of which six are vowels that are shared to short vowels (/i/, /u/, and /a/), and long vowels (/ii/, /uu/, and /aa/), and 28 are consonants [1]. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found in Semitic languages like Arabic and Hebrew [2], [3]. Arabic language includes syllable as linguistic unit. The syllables allowed in the Arabic language have the six forms: CV, CVV, CVC, CVCC, CVVC, and CVVV; where V indicates a (long or short) vowel while C indicates a consonant [4]. All Arabic syllables must contain at least one vowel. Arabic vowels cannot appear in syllable-initial position and can occur either between two consonants or in final position in a syllable hence, a vowel cannot be the start phoneme in any Arabic word or Arabic utterances can only start with a consonant [2], [3], [5]. Arabic syllables can be classified as short or long. The CV type is a short one while all others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant.

A. Arabic Dialects

Arabic dialects vary on many levels dimensions but primarily, as to geography and social factors. Depending on geographical linguistics, the Arab world can be divided into many different ways. The following is only one of many that cover the main Arabic dialects: Gulf Arabic (GLF) includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman. Iraqi Arabic (IRQ) which is the dialect of Iraq. In some dialect classifications, Iraqi Arabic is considered a sub-dialect of Gulf Arabic. Levantine Arabic (LEV) includes the dialects of Lebanon, Syria, Jordan, and Palestine. Egyptian Arabic (EGY) covers the dialects of the Nile valley: Egypt and Sudan. Maghreb Arabic covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libya is sometimes included in this class. Yemenite Arabic is often considered as a class on its own. Socially, it is common to distinguish between the regional variants of a major dialect, for example, northern and southern Levantine, for example, Egyptian, Levantine, Gulf, Maghreb, but also between the regional variants of a major dialect, for example, northern and southern Levantine. [6], [7].

II. SPEECH CORPORA BACKGROUND

Speech corpus is usually composed as spontaneously, canonically, or both [8]. A spontaneous (i.e., unstructured) speech corpus is gathered from real world human–human/human–machine communication. In this case, the speaker does not read prompts, rather he or she speaks naturally and freely to express a message and/or get another one. This speech is supposed to be natural and not influenced by the reading habits. A canonical speech corpus, on the other hand, is designed for speakers to stick with certain rules, including the reading of prompts, for the purpose of collecting specific speech sounds and vocabularies. The read speech content can be words, sentences such as application phrases and phonetically rich sentences, or text such as news, lectures, articles, and stories. Although the former is suitable for language understanding and dialogue design, it tends to be made up of unneeded frequently repeated words and utterances but does not essentially include all the sounds of the language under examination. However, a canonical speech corpus tends to be phonetically rich, i.e. all the sounds of the language are offered in various phonetic positions. Most of the recently compiled corpus collections belong to the third type, which contains spontaneous and read speech [8], [9]. Some speech corpora are recorded directly from...
the speakers or through a broadcasting medium such as radio or television; this is usually recorded at a high sampling rate: 16 kHz, 22 kHz, and 48 kHz. Other speech corpora are recorded through a telephone system, and they have a lower sampling rate, 8 kHz, due to the bandwidth of the telephone system. Each of the above categories has its own objectives and applications where the speech content and the recording procedures are different [8].

In this section, we present a brief description of the four corpora used in this research paper. The first corpus is related to the benchmark and the other corpora under evaluation, where the second is related to MSA and the third and fourth are related to colloquial Arabic, Saudi and Levantine dialects.

A. TIMIT corpus

TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different genders and dialects. Each transcribed element has been delineated in time. TIMIT was designed to further acoustic phonetic knowledge and automatic speech recognition systems. It was commissioned by DARPA and worked on by many sites, including Texas Instruments (TI) and Massachusetts Institute of Technology (MIT), hence the corpus' name [10]. TIMIT is a read speech corpus and researchers (especially for the purpose of speech/speaker recognition applications) have been using it since the mid-eighties of the last century. TIMIT contains a total of 6,300 sentences, 10 sentences (about 30 sec of speech) spoken by each of 630 speakers (438 males that account for a percentage of 70%, and 192 females that account for a percentage of 30%) from 8 major dialect regions of the United States (US). A speaker's dialect region is a geographical distribution within the U.S. mainland. Those speakers lived during their childhood years in the same area.

The text material in TIMIT prompts consists of 2 dialect sentences, 450 phonetically-comp act sentences, and 1890 phonetically-diverse sentences. The dialect sentences (attributed as SA sentences) were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences (attributed as SX sentences) and each text was spoken by 7 different speakers. The phonetically-diverse sentences (attributed as SI sentences) were selected from existing text sources - the Brown Corpus (Kuchera and Francis, 1967) and the Playwrights Dialog (Hultzen, et al., 1964) - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts. Each speaker read 3 of these sentences, with each sentence being read only by a single speaker. The speech material was subdivided into portions for training and testing. The test data has a core portion containing 24 speakers, 2 male and 1 female from each dialect region, core test material contains 192 sentences, 5 SX and 3 SI for each speaker, each having a distinct text prompt. The complete test set contains a total of 168 speakers and 1344 utterances, accounting for about 27% of the total speech material which result in dialect distribution of the 168 speaker (112 male and 56 female). This test of material contains 624 distinct texts. The 6300 sentences of TIMIT comprise over five hours of speech. All of the sentences have been segmented and labeled [10].

1) Derivatives of TIMIT corpus

To serve specific applications of digital speech processing, TIMIT derivatives were created from the original TIMIT by using the best and appropriate techniques in order to emulate corpora originally created by that way of digital speech collection. This way of using the derivatives give the researchers the correct option of the same corpus for their specific topic of investigation. There are several TIMIT derivatives already available, explicitly, FFMTIMIT, NTIMIT, CTIMIT, HTIMIT, and STC-TIMIT. The Free-Field Microphone TIMIT (FFMTIMIT) corpus consists of the original TIMIT database recorded with a free-field microphone [11]. Network TIMIT (NTIMIT) serves as a telephone bandwidth adjunct to TIMIT [12]. It contains the original TIMIT speech files transmitted over a telephone handset and the NYNEX telephone network, subject to a large variety of channel conditions [13]. For the Cellular bandwidth speech corpus CTIMIT, the original TIMIT recordings were passed through cellular telephone circuits [14]. The Handset TIMIT (HTIMIT) corpus offers a TIMIT subset of 192 male and 192 female speakers transmitted through different telephone handsets for the study of telephone transducer effects on speech [15]. For the single-channel telephone corpus STCTIMIT, the TIMIT recordings were sent through a real and, in contrast to NTIMIT, single telephone channel. A new TIMIT derivative called Wideband Mobile TIMIT (WTIMIT) containing the recordings of the original TIMIT speech files after transmission over a 3rd generation (3G) AMR-WB mobile network is released [16]. This way finding the supporting derivatives of a specific corpus cannot be found in any Arabic corpora.

B. West Point Corpus

The West Point Arabic speech corpus was produced by the Linguistic Data Consortium (LDC). This corpus contains speech data that was collected and processed by members of the Department of Foreign languages at the United States Military Academy at the West Point Corpus and the Center for Technology Enhanced Language Learning (CTELL) as part of an effort called 'Project Santiago'.

The original purpose of this corpus was to train acoustic models for automatic speech recognition that could be used as an aid in teaching Arabic to the West Point Corpus cadets. The corpus consists of 8,516 speech files, totaling 1.7 gigabytes or 11.42 hours of speech data. Each speech file represents one person reciting one prompt from one of four prompt scripts. The files were recorded as 16-bit PCM low-byte-first ("little-Endian") raw audio files, with a sampling rate of 22.05 KHz. They were then converted to The NIST SPHERE format. Approximately 7,200 of the recordings are from native informants and 1200 files are from non-native informants [17].
The West Point Corpus consists of collections of four main Arabic scripts. First is Collection Script 1, which contains 155 sentences, spoken by all 75 native Arabic speakers and it has a total of 1152 tokens and 724 types, while in Collection Script 2, which contains 40 sentences, spoken by 23 of the non-native speakers and it has a total of 150 tokens and 124 types. Third, is Collection Script 3, which contains 41 sentences, spoken by 4 of the non-native speakers. It has a total of 138 tokens and 84 types. Finally Collection Script 4, which contains 22 sentences, used by 9 of the non-native speakers, all of them third year Arabic speakers. It has a total of 72 tokens and 59 types. The total number of distinct Arabic words is 1131 [17]. A summary of the statistics of this database is given in Table I.

### Table I. WET Point Corpus Summary [17]

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of speakers</td>
<td>native</td>
<td>41</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>66</td>
<td>44</td>
</tr>
<tr>
<td>hours of data</td>
<td>native</td>
<td>6</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td>0.74</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>6.74</td>
<td>4.68</td>
</tr>
<tr>
<td>megabytes of data</td>
<td>native</td>
<td>913</td>
<td>663</td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td>111</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>1024</td>
<td>705.4</td>
</tr>
<tr>
<td>number of speech files</td>
<td>native</td>
<td>4107</td>
<td>3163</td>
</tr>
<tr>
<td></td>
<td>Non-native</td>
<td>883</td>
<td>363</td>
</tr>
<tr>
<td></td>
<td>total</td>
<td>4990</td>
<td>3526</td>
</tr>
</tbody>
</table>

C. SAAVB Corpus

SAAVB is a telephony and noisy speech corpus that was collected by KACST during 2002 and 2003 [8]. This corpus is an example of normal life speech corpus. This data was acquired from 1,033 native speakers of Arabic with Saudi accent (51% males and 49% females). Around 51% of those speakers are aged 16 to 30 years and 49% of them are above 30 years old. 70% of the corpus was recorded over the Saudi mobile network while the remaining 30% was acquired through the fixed-line network [9]. Each speaker read 59 prompts that consisted of the following: numbers, phonetically rich words, sentences and pronunciation of the Arabic and English alphabets. The average number of words in each prompt is 5 words. The duration of the total recorded speech in SAAVB is 96.37 hours distributed among 60947 audio files (1033 speakers x 59 audio files). This means that the average duration for each speaker is 5.60 minutes and the average duration of each audio file is 5.70 seconds [8]. The corpus was digitized at 8000 Hz sampling rate with A/D conversion with 8 bits/sample. The size of SAAVB is 2.59 GB. It contains 1,033 directories with 183,518 files. Every directory corresponds to a speaker and contains 178 files distributed as follows [8], [9]: The first is a text document that has all the information required about the speaker (gender, telephone type, age and acoustic environment). The second has 59 text files of the prompts. Finally, there are 59 text files of the speech transcription.

D. BBN/AUB Corpus

Levantine Arabic is the dialect of Arabic spoken by ordinary people in Lebanon, Jordan, Syria, and Palestine. It is significantly different from MSA Arabic, in that it is a spoken rather than a written form. It includes different word pronunciations, and even different words compared to MSA Arabic and the written "official" form of Arabic. This corpus was developed with funding from the Defense Advanced Research Project Agency (DARPA), as part of the Babylon program. Approximately 20% of the corpus was recorded by BBN using paid subjects recruited in the Boston area from May 2002 to September 2002. This portion of the corpus was the first to be collected. Subsequently, the remaining 80% was recorded by the American University of Beirut (AUB), under subcontract to BBN, from July 2002 to November 2002. AUB students and staff served as both experimenters and subjects. This portion of the corpus was recorded in Beirut, Lebanon, on the AUB campus [18]. The subjects in the corpus were responding to refugee/medical questions (such as “where is your pain?”, how old are you?). Each subject was given a part to play, that prescribed what information, they were to give in response to the questions, but were told to express themselves naturally, in their own way, in Arabic. The BBN/AUB dataset consists of 164 speakers, 101 males and 63 females. It is a set of spontaneous speech sentences, recorded from 164 subjects speaking in Levantine colloquial Arabic. BBN/AUB Speech data has been recorded using a close-talking, noise-cancelling, headset microphone (the Andrea Electronics NC-65). A Java-based data-collection tool, developed by BBN, was used to do the collection of speech. The audio was recorded in MS WAV, signed PCM. Sampling rate was 16 KHz, with 16 bits resolution. The duration of the total recorded speech in BBL is 45 hours distributed among 75,900 audio, the total audio size: 6.5 GB. The total text size is 3.1 MB, Vocabulary is 15K words and Total words are 336K words [18].

III. EVALUATION

In this section we try to evaluate three corpora West Point, SAAVB and BBN/AUB depending on estimation based on the situation seen TIMIT corpus and through specific elements.

A. Corpus Type

TIMIT is a canonical corpus [10], where every speaker read ten sentences. On the other hand, West Point is also a canonical corpus [17]. It has four scripts divided between all speakers to read them, BBN/AUB is done spontaneously recorded from 164 subjects speaking in Levantine colloquial Arabic [18], while SAAVB is both canonical and spontaneous [8], [9]. In SAAVB, the number of utterances for each speaker is 59. They consist of 49 read items (83%) and 10 elicited spontaneous responses (17%) [8].
Spontaneous speech is harder to work with than read speech and has its own characteristics and problems [9]. It has some challenges like recovering hidden punctuation, coping with hesitations, allowing for realistic turn-taking, hearing non-speech such as coughs and lips clicks, and hearing more than words [19].

On the other side, when the speakers read sentences under formal recording conditions, and not all were fluent readers, then these samples do not convey the actual speech features that are normally used between humans to communicate messages. That is, though this is connected speech, the speech style is far from the casual setting used in normal communication [20]. More sophisticated and recent corpora contain both spontaneous and read speech parts [8].

**B. Speakers**

Speakers’ attributes are one of the most important parameters of any speech corpus. How to reach to the speakers, select the speakers, determine numbers, gender, ages, and distribute them across the targeted area of dialects are very important examples of employing these attributes. There are several options available to approach the target speakers and record their utterances. One of them is through advertisements that can be done on television, the Internet or in newspapers or other available media. Another option is to search for speakers with the target dialects in large cities where thousands of people have moved from native cities, regions, and communities [8].

<table>
<thead>
<tr>
<th>TABLE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
</tr>
<tr>
<td>No. of speakers</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
</tbody>
</table>

Normally, speech corpora are different in the number of speakers as we depicted in Table II in our investigated corpora. In this table, the range between SAAVB corpus and other two corpora West Point and BBN/AUB is highly noticeable, if we are comparing them with the benchmark corpus. To be more specific, the number of speakers in SAAVB is almost twice that of TIMIT; while the BBN/AUB corpus represents around a quarter the number of TIMIT speakers, and the West Point represent (17.5%) from the number of TIMIT speakers.

Regarding speakers’ gender distribution, the TIMIT corpus is very clear on this point, where the number of males and females from the eight major dialectal regions of the US is as clearly in TIMIT documented, in totally males represents (70%) and females (30%) as we see in Table II. For our investigated corpora, West Point is very clearly on this point and stable as we see in Tables I and II where males account for 66 that represent (60%), natives acts 62.12% (41 speakers) and non-native 37.87% (25 speakers), and females account for 44 that represent (40%), where the native speakers acts 77.27% (34), and non-native 22.72% (10 speakers), also the gender of speakers in SAAVB corpus is clearly distributed, balanced, and explained in the catalogue and it is as follows: First, males account for 523 speakers (50.63%); second, females account for 510 speakers (49.37%). On the other hand, the BBN/AUB corpus implicitly reported speaker gender distribution as follows: First, males account for 101 speakers (61.58%); second, female account for 63 speakers (38.41%). Although the gender distribution is well defined in the sum of all the speakers in each database, SAAVB is the only one of the four databases where the gender is well defined (one half is male and the other half female) in all the subcategories such as age, telephone type and background noise.

The TIMIT corpus does not give information about the speakers’ ages. In the case of the SAAVB corpus, the ages are distributed and documented for each speaker in a well-defined manner. It is as follows: First, there are 512 speakers (50.53%) who are 16–30 years old; second, there are 364 speakers (35.24%) who are 31–45 years old; finally, 147 speakers (14.23%) are 46–60 years old. Regarding the ages of the speakers of the West Point and BBN/AUB corpora, unfortunately no reference to this was reported in their catalogs, and this kind of important speaker feature was totally overlooked.

All speakers of the TIMIT corpus are native speakers of English, and they lived during their childhood in the region of their dialect. In our inspected corpora, speakers of West point Corpus are divided into native (68.18%), and non-native (31.81%) speakers. all speakers of the SAAVB corpus are native Arabic speakers, while the BBN/AUB corpus does not refer to this important information, which helps digital speech and language investigators. The only information related to this fact is 20% of the corpus was recorded in Boston and the remaining 80% was recorded in AUB [8], [17], [18].

Another important item as to speakers attributes is the distribution of the speakers within each investigated region of the targeted dialects. TIMIT corpus divided the 630 speakers among eight regions [10]. One criticism is that they are mostly white and male, and they are mostly in their 20s and 30s [20]. In our investigated corpora, the SAAVB report shows a statistical model based on the population of all the cities in Saudi Arabia according to the official Saudi census data of 1999; then the model predicts the number of speakers to be selected per city as a function of the population of that city. On the other hand, the developers of West point and BBN/AUB have not shown how they chose the group of speakers form the Arabic countries, although Arabic language users are distributed among 22 country [5], and Levantine dialects users are distributed among four countries: Syria, Lebanon, Palestine, and Jordan [17].

**C. Data Sources**

In the TIMIT corpus the utterances are recorded in a soundproof chamber at 16-kHz sample rate [8]. In derivative versions of the TIMIT corpus, other types of TIMIT were created including the telephone version NTIMIT [13], the free-field microphone FFIMIT [11] and the cellular bandwidth CTIMIT [14]. In our investigated corpora, West Point data for Scripts 1, 2 and 3 were collected using Pentium 133 MHz laptop
computers running Windows NT. Recordings were captured at a sampling rate of 16 bit @ 22050 Hz using a Shure SM10A microphone and a RANE Model MS1 pre-amplifier. The speech data for Script 4 was recorded on Pentium 166 MHz desktop computers running Red Hat Linux 6.2. Recordings were captured at the same sampling rate with the same microphone as Scripts 1, 2 and 3[17]. SAAVB corpus was recorded through a telephone system, and it has a lower sampling rate, 8 kHz, due to the telephone system bandwidth, Tables III and IV Show recording environments for mobile network and fixed line network on SAAVB database.

TABLE III

<table>
<thead>
<tr>
<th>Recording environment</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiet</td>
<td>34.76</td>
</tr>
<tr>
<td>Noisy</td>
<td>34.76</td>
</tr>
<tr>
<td>Moving vehicle</td>
<td>30.48</td>
</tr>
</tbody>
</table>

TABLE IV

<table>
<thead>
<tr>
<th>Recording environment</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiet</td>
<td>75.32</td>
</tr>
<tr>
<td>Noisy</td>
<td>23.68</td>
</tr>
</tbody>
</table>

BBN/AUB corpus is recorded at a sampling rate of 16 kHz. BBN/AUB corpus used a close-talking, noise-cancelling, headset microphone (the Andrea Electronics NC-65). A Java-based data-collection tool was used [18]. It was recorded through a microphone and using a kind of switch (or mouse click) which induced bursts or spikes in all the wave files. The recording speech amount was not of the same size for all the speakers, as some speakers spoke for a long time compared to others. A ratio of 1-5 was sometimes noticed; some speakers had many "euh" sounds at each answer, producing a non-spontaneous manner of talking. There was too much silence at the start and at the end of the files, in the whole dataset. Silence showed a ratio of approximately 1/3 from the full dataset, and some speech was not usually clear due possibly to the distance from the microphone [4], [18].

D. Lexicon and Labeling

The TIMIT corpus symbolization for the lexical information uses abstract quasi-phonemic marks representing the underlying sound. It typically corresponds to a variety of different sounds in the actual recordings. The term quasi-phonemic is used because some differences represented in the lexicon are often found in speech, even though they are not really phonemically distinctive in English, such as the /e/~/æ/ in which /e/ co-occurs with stress. All entries have been converted to lowercase. Stress markers are given as 1 for primary stress and 2 for secondary stress, tacked on to the end of the vowel symbol. Hyphenated words such as "head-in-the-clouds" can be found as a single entry and as individual words "head", "in", "the", and "clouds". One pronunciation is provided per entry, except in the case where the same orthography corresponds to different parts of speech with different pronunciations, and both forms exist in the TIMIT scripts. To differentiate these words, multiple entries are given, with the syntactic class following the symbol ~. Also TIMIT clearly explains vowel variability, stress differences, and Syllabics.

In our scrutinized corpora, West Point Each waveform file has a monophone and word level master label file (*.mlf) transcription in HTK-format. These files contain a multi-line entry for every speech file in the corpus -- the first line of each entry gives the file name, and the phones are provided in sequence on the following lines, one phone per line. Master label files are provided at both the word level and the phone level. Phone level labels are provided both with and without "sp". All sentence transcripts begin and end with the "sil" (silence) segment. The label data without the "short pause" (sp) segment represents a direct phonemic transcription of the prompting text, replacing each Arabic orthographic form with the exact phoneme sequence provided for the word in the dictionary file, whereas the +sp version involves the addition of "short pause" segments and hand labeling of some utterances. For example, a phonological rule that deletes a word initial glottal stop and coalesces the preceding and following vowels into a single phone was applied in some cases. /iy # Q ah l l/ -> /ih l/. That is, the sequence of word-final high-front tense vowel followed by the definite article "al" is pronounced as a single syllable with a high-front lax vowel. This hand labeling is not standardized and was applied in some instances and not in others [17].

SAAVB corpus used Arabic orthography to transcribe the audio files. Arabic vowels are written as diacritics above or below the letters. Arabic diacritics include symbols other than vowels. When an Arabic text is fully diacritized, every letter must be followed by a diacritic except for vowels that are represented by alphabets and not diacritics [8]. For the SAAVB there is no information in transcription about lexical cues in utterances. On the other side, the transcription of BBN/AUB does not include diacritics such as short vowels and gemination.

From the above information we can see that the West Point is labeled as TIMIT corpus while SAAVB and BBN/AUB are not labeled and segmented as in the case of TIMIT. This is even more pronounced for dialectal Arabic where a single word can be pronounced quite differently based on the speaker’s nationality, level of education, and region[21]. In TIMIT all phonemes and words boundaries are labeled and incorporated in the audio files with usable format to most speech processing systems. Unfortunately, these features are not provided by SAAVB or BBN/AUB corpora.

E. Training and Testing Subsets

TIMIT has been subdivided into suggested training and testing subsets using the following criteria: First, roughly 20% to 30% of the corpus should be used for testing purposes, leaving the remaining 70% to 80% for training. Second, no speaker should appear in both the training and testing portions. Third, all the dialect regions should be represented in both subsets, with at least 1 male and 1 female speaker from each dialect. In addition, the amount of overlap of text material in the two subsets should be minimized; if possible no texts should be identical. Finally, all the phonemes should be covered in the test material; preferably, each phoneme should occur multiple times in different contexts.
The core test set consists of 2 male speakers and 1 female speaker from each dialect and 192 unique texts. Those wishing to perform more extensive testing, should use the complete test set consisting of a total of 1,344 sentences from 168 speakers [10]. In our scrutinized corpora, SAAVB left it open for researchers and application developers to select the training and testing sets according to their need, while West Point and BBN/AUB do not refer to that. We think that by not specifying training and testing subsets in West Point, SAAVB and BBN/AUB there is a major drawback. In fact, this will erase the standardization and the logic of comparing the different research results that use the same corpus due to differences in selecting training and testing subsets.

IV. CONCLUSION

After studying these corpora, we can conclude that West Point is identical to TIMIT in some aspects like type of corpus where, the two corpora are canonical, speakers’ gender attribute are clearly included in the two corpora, and they ignored speakers’ age attribute. Also, the two corpora labeling and number of speakers in TIMIT are greater than West Point by five times. West Point is not considering the speakers’ attribute of specifying distribution among large Arab country contrariwise to TIMIT corpus, and does not specify or suggestion data parts for testing and training subset as supplied by well-prepared corpora such as TIMIT. SAAVB and BBN/AUB corpora are different if compared with the TIMIT corpus. The differences are in the targeted language, used dialects, type of intended applications, and degree of resources in order to create, use, and maintain the corpus. The SAAVB corpus is presented and documented in a clearer manner than even TIMIT regarding speakers’ distribution. On the other hand, contrary to the TIMIT corpus, SAAVB and BBN/AUB do not have test and training subsets. They differ in data source, and they are not labeled. Compared to other major languages of the world, the Arabic language suffers from absence of research initiatives and research resources. In addition to this and as we noted from TIMIT, none of the three Arabic corpora has any kind of derivative but the original version of them. As a result, Arabic language lacks reliable speech corpora for research in speech processing. Robust Arabic speech corpora must consider the different Arab countries and dialects, different speakers’ ages, genders and good distributions. Training and testing subsets are too important additionally to the phonemes labeling. In this context, we consider that the investigated Arabic corpora are a big step forward towards enhancing Arabic speech and language processing.

ACKNOWLEDGMENT

This work was supported by the National “NPST” key project # 10INF1325_02, Saudi Arabia.

REFERENCES


