Indonesian Journal of Electrical Engineering and Computer Science
Vol. 41, No. 3, March 2026, pp. 1049∼1059
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v41.i3.pp1049-1059

A rich and balanced phonetics corpus for modern standard Arabic ASR systems

Youssef Boutazart, Naouar Laaïdi, Abderrahim Ezzine, Hassan Satori, Mohamed Taj Bennani
Department of Mathematics and Computer Science, Faculty of Sciences Dhar-Mahraz (FSDM),
Sidi Mohamed Ben Abdellah University (USMBA), Fez, Morocco

Article Info

Article history:

Received Sep 11, 2024
Revised Jan 12, 2026
Accepted Feb 27, 2026

Keywords:

Modern standard Arabic
Phonetically balanced corpus
Phonetically rich corpus
Segmentation grapheme to phoneme
Zipf's law

Abstract

This research delves into the creation of an innovative Modern Standard Arabic corpus, aiming for a comprehensive balance and richness while adhering to Zipf's law. Building a phonetically diverse Arabic sentence collection yields significant advantages in terms of efficiency, cost-effectiveness, and storage capacity compared to conventional corpora. The corpus undergoes meticulous segmentation into graphemes, which are then manually converted into phonemes, resulting in a total of 19769 phonemic units. Among these phonemes, consonants like 'Laam - l' account for 10%, while 'Fatha - A' vowels constitute 20%. Evaluation of this corpus using an automatic speech recognition (ASR) system reveals a sentence error rate (SER) of 30% and a word error rate (WER) of 15%. Furthermore, statistical analysis unveils that diacritic marks encompass 47.59% of the corpus, with graphemes comprising the remaining 52.41%. These diacritized marks provide valuable insights into the precise phonetic transcription of the corpus. Additionally, the study provides detailed breakdowns of consonants based on their place and manner of articulation, enhancing our understanding of phonetic structures.

Corresponding Author:

Hassan Satori
Department of Mathematics and Computer Science, Faculty of Sciences Dhar-Mahraz (FSDM)
Sidi Mohamed Ben Abdellah University (USMBA)
FSDM, USMBA, B.P. 1796, Fez, Morocco
Email: hassan.satori@usmba.ac.ma

1. INTRODUCTION

A corpus is a digital collection of natural language samples, used to discover its structure and patterns. Researchers analyze it to study word order, the arrangement of sentences and clauses, and the use of grammatical structures [1], [2]. The corpus development serves as a pivotal element in various disciplines, including natural language processing (NLP) and automatic speech recognition (ASR) [3]-[5]. Developing any corpus demands considerable resources and effort. Consequently, there's a growing interest in crafting phonetically diverse and balanced text and speech corpora [6], [7]. The rich and balanced corpus is a valuable source for developing ASR systems, as it allows the system to be trained on a variety of speech patterns and accents [8], [9]. Reflecting this interest, in their study, the authors in [10] introduced an automated approach for constructing a phonetically rich and balanced corpus sourced from the web, selecting 6082 phrases to develop a robust recognizer tool. Similarly, Wang in [11] conducted a statistical analysis of various Mandarin acoustic units using an extensive Chinese text corpus gathered from daily newspapers. Following this analysis, Wang proposed an algorithm to automatically extract phonetically rich sentences from the corpus, which are then utilized

for training and evaluating a Mandarin speech recognition system. Radová and Vopálka in [12] address the challenge of phonetically balanced sentence selection, presenting two iterative procedures to choose sentences that accurately reflect the occurrence of phonetic events in natural speech, resulting in a set of 40 phonetically balanced sentences. In the same way, Matoušek and Romportl in [13] propose a method for preparing and recording a phonetically and prosodically rich Czech language corpus for text-to-speech synthesis. They implement an algorithm that selects sentences based on both phonetic and prosodic criteria, including the random selection of paragraphs to capture supra-sentential prosody phenomena. On the other hand, for the English language, Yazawa in [14] developed a set of 720 phonemically balanced phrases for English learners, selecting 50 core vocabulary words based on the Harvard New General Service List (NGSL). However, preparing and selecting suitable sentences and words poses significant challenges in ensuring comprehensive linguistic representation and maintaining desired phonetic diversity. While abundant databases exist for major languages like English, German, French, and Mandarin [15], [16], the task is considerably more complex for underrepresented languages such as MSA. Regarding the MSA, recently, Alqudah *et al.* [17] have developed the Arabic automatic speech recognition (ASR) for speakers with speech disorders (SD), identifying research gaps and highlighting the need for comprehensive ASR systems that address various SD types and continuous speech in Arabic. Alghamdi *et al.* in [18] have proposed a manually written Arabic corpus, based on a phonetically rich and balanced created list of 663 words. They were one of the first works on the production of this type of corpus. The database consists of 367 sentences, 2 to 9 words per sentence. Later, in 2012 Abuschariaa *et al.* [19] described the preparation, recording, analyzing, and evaluation of a new speech corpus for MSA. The sentences used contained all phonemes and preserve the phonetic distribution of the Arabic language. Yuwan and Lestari in [20] have explained that creating a phonetically rich and balanced corpus not only makes the system more robust and intelligent but saves time, cost, and storage capacity. They have collected verses as speech corpus for the Quranic recognition system with special symbols. The selected verses contained 180 verses of 6236 whole verses in the Quran. Our primary goal is to develop a rich and balanced corpus. We prioritize readability and pronunciation by incorporating phonetically rich and balanced, structurally simple sentences. The Corpus collection encompasses diverse Arabic texts from various sources. This deliberate selection adheres to the 50 most prevalent words in the Arabic language, ensuring compliance with Zipf's law, which states that the frequency of a word in a text is inversely proportional to its rank in a frequency table [21], [22].

In this study, we introduce an approach for constructing a novel rich and balanced modern standard Arabic corpus, developed at Faculty of Sciences Dhar el Mehraz by University Sidi Mohammed Ben Abdellah (FSDM-USMBA). This approach streamlines the process, saving time, costs, and storage capacity compared to the conventional corpora collection. The corpus adheres to Zipf's law by focusing on the 50 most common Arabic words, includes grapheme-to-phoneme conversion, and conducts a phonetic statistical analysis, all contributing to advancements in Arabic speech recognition technologies.

Apart from the introduction in section 1. The paper is organized as follows: the method is explained in the section 2. The Statistical analysis is discussed in section 3. Section 4, deals with results and discussion. We finished with a conclusion and future research directions.

2. METHOD

In this part, we have noticed that a rich and balanced Arabic corpus is very rare and it is not accessible to the Arabic linguistic researchers. We proposed an approach to creating a rich and balanced Modern Standard Arabic Corpus by University Sidi Mohamed Ben Abdellah called FSDM-USMBA. Indeed, Arabic, is a Semitic language and one of the six official UN languages, is spoken by around 400 million people across 22 countries [23]-[25]. It is categorized into classical Arabic, modern standard Arabic, and dialectal Arabic. Arabic script is written from right to left and consists of two types of symbols: letters and diacritics. These symbols are typically written in a connected form. Additionally, several letters may change shape depending on their position within a word. However, it should be noted that the script alone does not encompass all sounds [26]. It provides valuable information about consonants and vowels, which can be extracted through diverse techniques. The Arabic language consists of 36 phonemes, with 28 of them representing consonantal sounds. Additionally, there are 8 phonemes, including three short vowels, three long vowels and two diphthongs [27]. The Arabic language is characterized by the following diacritics: (تَنْوِين ـٌ ـً ـٍ tanween, - dammatan, fathatan, and kasratan -), (شَدَّ ـّ šaddat) and (سُكُوْنْ ـْ sukun). The diacritical marks are used to indicate vowel sounds and other phonetic

features. It appears on top or below of the graphemes. Here joining all the (ا) diacritic marks is not included in consideration. On the other hand, syllables are units of speech comprising one or more phonemes. In the Arabic language, various syllable structures are allowed. These include CV, CVV, CVC, CVVC, CVCC, and CVVCC [28]. In these structures, C represents a consonant, V represents a short vowel, and VV represents a long vowel. Following a brief overview of the Arabic language, we detail the corpus specifications. Our methodology comprises four phases: corpus initial, corpus text handling, corpus final, and corpus segmentation.

2.1. Corpus initial

We based the initial corpus on written sources. This corpus contained 30 million words and was divided into various text genres of the same size. Each of these text types contained material from all regions of the Arabic-speaking world. The primary source of the corpus was written Arabic, encompassing both its standard form and its dialects. The main objective of this corpus was to generate a frequency count of all Arabic words as they are written, including their prefixes and suffixes [29].

2.2. Corpus text handling

To build a modern standard Arabic rich and balanced corpus, we used the first 50 most frequent words from the initial corpus. Based on these words, we selected and constructed sentences to adhere to the distribution outlined by Zipf's law. We used different sources, including The Holy Quran and Hadith, as well as content related to finance, business, economics, politics, culture, sports, technology, science, weather, art, and others. The goal was to produce simple, short sentences that are phonetically rich and balanced. It is important to note that some expressions may be deleted or replaced with others to better adhere to the grammar rules of the Arabic language.

2.3. Corpus final

In our study, we analyzed a final corpus consisting of 527 sentences and a total of 3,308 words. Table 1 displays the sentence count and word count for each genre, along with the proportion of words within each genre. To understand the relationship between the frequency of a list of 50 words and their rank in the final corpus, Figure 1 illustrates the log-log graph of word frequency in the final corpus. The straight line in the graph represents the average slope of the descending word frequencies. Additionally, Table 2 and Figure 2 provide evidence that the frequency of the words is inversely proportional to their rank. These illustrations confirm the achievement of Zipf's law. We found that the eight most frequent words in the final corpus align with those found in a previous study [30].

Table 1. Statistics of the final 'FSDM-USMBA' corpus

Genre	Number of sentences	Number of words	Percentage %
Healy Quran, Hadith and religion	86	536	0,162
Health and epidemic	51	374	0,113
Finance and Business	20	146	0,044
Technology	29	202	0,061
Literature	49	324	0,098
Economy	22	153	0,046
Politics	58	373	0,113
Arts	13	75	0,023
Tourism and Culture	41	264	0,079
Sports	16	118	0,036
Weather	13	81	0,025
Others	129	662	0,200

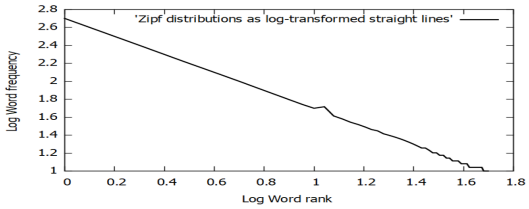


Figure 1. The log-log graph of word frequency in the final corpus

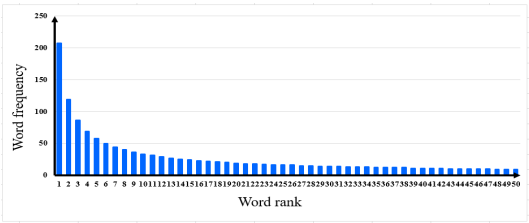


Figure 2. Distribution of word in the final corpus

Table 2. Empirical evaluation of Zipf's law in 'FSDM-USMBA' corpus

Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank	Word	Freq.	Rank
أَلْ	207	1	عَنْ	25	14	لَمْ	15	27	أَوَّل	11	39
وَ	119	2	قَالَ	24	15	مَا	15	28	غَيْر	11	40
يِ	86	3	هَذَا	23	16	إِنَّ	14	29	إِذَا	11	41
مِنْ	69	4	مَعَ	22	17	بَيْنَ	14	30	نَفْس	11	42
لِ	58	5	الَّتِي	21	18	هِيَ	14	31	عَرَبِيّ	10	43
بِ	50	6	كُلّ	20	19	بَعْدَ	13	32	أَيّ	10	44
عَلَى	44	7	هُوَ	19	20	يَا	13	33	رَئِيس	10	45
أَنَّ	40	8	فَ	18	21	ذٰلِكَ	13	34	عَمَل	10	46
إِلَى	36	9	هٰذِهِ	18	22	قَدْ	12	35	عَرَفَ	10	47
كَانَ	33	10	أَوْ	17	23	آخَر	12	36	بَعْض	9	48
لاَ	31	11	الَّذِي	16	24	شَيْء	12	37	دَوْلَة	9	49
اللَّه	29	12	أَنَا	16	25	عِنْدَ	12	38	كَمَا	9	50
أَنْ	27	13	يَوْم	16	26						

2.4. Corpus segmentation

To create a rich and balanced corpus of Arabic, it is essential to encompass all the phonemes of the Arabic language while preserving its phonetic distribution. To accomplish this, we adopt a two-step method (Algorithm 1). Firstly, we segment the text into graphemes.

Algorithm 1:	**text to grapheme**
1.	Determine the path of FSDM by USMBA corpus
2.	Iterate through each character of the Arabic text and print it
3.	Create a text field with Arabic font and right-to-left orientation
4.	Create a table to display the characters and their corresponding Unicode codes
5.	Create a JFrame to display the text field and table
6.	Set JFrame properties such as the size and default close operation
7.	Create an instance of the class (starts the application)

Secondly, we meticulously convert the graphemes into phonemes, adhering to the phonological rules to the Arabic language. This manual conversion ensures the accurate representation of Arabic phonetics in the resulting corpus:

a. Convert (ّ šaddat) to two consecutive ones,

b. Convert (ٰ) to (أا) it is found in the text,

c. Convert tanween (اً اٌ اٍ) to (اَنْ اُنْ اِنْ),

d. Pronunciation of all types of the Hamza (ء إ أ ئ and ؤ) are ء.

3. STATISTICAL ANALYSIS

In this section, we conducted a statistical analysis of syllables, graphemes, and phonemes using the rich and balanced corpus of Modern Standard Arabic. Our primary objective was to gain valuable insights into the morphology and phonology of MSA.

3.1. Statistical analysis of syllables

After segmenting our corpus, we conducted an extraction process to identify and compute the various types of syllables present in Modern Standard Arabic. This step enabled us to determine the frequencies and percentages of each syllable type. The distribution of syllabic structures in the corpus shows CV and CVC syllables as the most frequent, accounting for 55.60% (3360 occurrences) and 25.00% (1511 occurrences) respectively. CVV syllables follow with 990 occurrences (16.38%). Less frequent are CVCC (149 occurrences, 2.46%), CVVC (33 occurrences, 0.55%), and CVVCC (1 occurrence, 0.01%).

3.2. Statistical analysis of graphemes

Regarding the investigation of the frequency and distribution of individual graphemes within the rich and balanced corpus, Table 3 displays the count and percentage of occurrences for each grapheme. The graphemes "ل" and "ا" occur most frequently, accounting for 10.09% and 8.32% respectively. Conversely, the graphemes "ظ" and "ء" are the least frequent, with respective occurrence rates of 0.18% and 0.47%.

Table 3. Repetitions and percentage for each grapheme

Arabic Grapheme	Grapheme repetitions	in %	Arabic Grapheme	Grapheme repetitions	in %	Arabic Grapheme	Grapheme repetitions	in %
ل	1090	10.09	أ	337	3.12	ط	115	1.07
ا	899	8.32	ف	323	2.99	خ	113	1.04
م	751	6.95	ك	321	2.97	ص	113	1.04
ن	711	6.58	ة	296	2.74	ش	105	0.97
ي	599	5.54	آ	276	2.55	ض	84	0.78
ر	588	5.44	ق	268	2.48	ز	67	0.62
ع	461	4.27	س	254	2.35	ث	65	0.60
و	454	4.20	ح	208	1.92	غ	59	0.55
ت	417	3.86	ج	186	1.72	ئ	57	0.53
ه	390	3.61	إ	132	1.22	"	51	0.47
ب	368	3.57	ذ	125	1.16	ؤ	22	0.21
د	343	3.17	ى	121	1.12	ظ	19	0.18

Furthermore, our analysis revealed that diacritic marks constitute 47.59% of the corpus, while graphemes make up the remaining 52.41%, as detailed in Table 4. Consequently, the absence of diacritic information in the grapheme-based (non-diacritized) transcription means that approximately 47.59% of the details required for an accurate phonetic transcription are unavailable. In the analysis of diacritic marks frequencies, presented in Table 5, Fatha emerged as the phoneme most frequent with 20.62%. It's worth noting that tanween (nunation) is restricted to appearing solely on the last letter of a word.

Table 4. Frequency of graphemes and diacritics

Type	Frequency	Percentage
Graphemes	10806	52.41
Diacritics marks	9813	47.59
Total	20619	100

Table 5. Arabic diacritics and their frequency of occurrence

Type	Frequency	Percentage
Fatha	4253	20.62
Kasra	1859	9.02
Damma	1028	5.14
Shadda	611	2.96
Sukun	1265	6.14
Tanween Fatha	69	0.34
Tanween Kasra	360	1.75
Tanween Damma	337	1.63

3.3. Statistical analysis of phonemes

After applying the grapheme-to-phoneme approach with phonological rules, we conducted a statistical analysis to examine the occurrence of phonemes in 3,308 words. This analysis encompassed the positions of phonemes at the beginning, middle, and end of the words. Table 6 displays the statistics for each phoneme. Our results are in accordance with those of the researchers in [31], for Arabic phonemes frequencies in the final corpus.

According to the findings presented in Table 6 and Figure 3(a), the deductions concerning phoneme statistics can be summarized as:

- In the case of short vowels: Fatha, (اَ) is the most frequent, followed by kasra (اِ) and damma (اُ),
- In the case of long vowels: The long vowels آى and اا are counted as one and pronounced Aaa (أى). They are the most frequent, followed by Aii (اِ يْ) and Oue (اُوْ),

– In the case of Consonants: Noon, (ن) is the most frequent phoneme, this is explained by that it also comes from the tanween (fathatan - اً >, dammatan - اٌ, and kasratan - اٍ), followed by Laam (ل) and Hamza (ء), all hamza types (أ ئ ء إ and ؤ) are counted as one (ء). The following consonants have a frequency lower than 0.5%, thaa (ث), zain (ز), dhaad (ض), and ghayn (غ). Dhaa (ظ) is the least frequent phoneme,

– The difference in the percentages of the two diphthongs is exceedingly small (around 0.07%).

Table 6. Arabic phonemes statistics in the FSDM-USMBA corpus

Conson.	Arpa.symbols	IPA symbols	Description and Syllables	Repetitions			Percentage %
				Start	Inside	End	
ء			Hamza - همزة - CVC - CVC				
أ			Alif - ألف - CVC - CVC				
إ ؤ ئ	E	P	Alif+Hamza below Waw+Hamza above Ya+Hamza above	667	176	33	4.43
ب ت	B	b	Baa - با - CV Taa - تا - CV	136	221	49	2.05
ة	T	t	Taa marbuta	95	485	144	3.66
ث	TH	T	Thaa - ثا - CV	13	55	6	0.37
ج	JH	g	Jeem - جيم - CVC	49	136	10	0.99
ح	HH		Haa - حا - CV	79	114	17	1.06
خ	KH	x	Khaa - خا - CV	36	73	4	0.55
د	D	d	Daal- دال - CVVC	69	218	93	1.92
ذ	DH	D	Thaal - ذال - CVVC	21	99	6	0.64
ر	R	r	Raa - را - CV	66	452	100	3.12
ز	Z	z	Zaiy - زي - CVC	11	53	5	0.35
س	S	s	Seen - سين - CVC	48	190	36	1.39
ش	SH	S	Sheen - شين - CVC	44	71	4	0.60
ص	SS	s	Saad - صاد - CVC	27	99	1	0.64
ض	DD	d	Dhaad - ضا - CVC	7	58	21	0.43
ط	TT	t	TTaa - طا - CV	18	88	5	0.56
ظ	DH2	D	Dhaa - ظا - CV	5	14	2	0.11
ع	AI	Q	Ayn - عين - CV - CVC	182	235	46	2.34
غ	GH	G	Ghayn - غين - CV - CVC	27	28	4	0.30
ف	F	f	Faa - فا - CV	154	146	28	1.66
ق	Q	q	Qaaf - قاف - CVC	81	187	21	1.46
ك	K	k	Kaaf - كا - CVC	124	153	63	1.72
ل	L	l	Laam - لام - CVC	147	914	140	6.08
م	M	m	Meem - ميم - CVC	321	358	91	3.89
ن	N	n	Noon - نون - CVC	76	410	1063	7.84
ه	H	h	Haa - ها - CV	87	117	186	1.97
و	W	w	Waw - واو - CVC	166	289	37	2.49
ي	Y	y	Yaa - يا - CV	143	381	213	3.73
			Short vowel				
اَ	AE	A	Fatha - فتح	0	2542	520	15.49
اُ	UH	u	Damma - ضمّ	0	959	398	6.86
اِ	IH	i	Kasra - كسر	0	1692	428	11.23
			Long vowel				
اَى ,اَا	AE:	a:	Aaa - أَى	0	670	350	5.16
اُوْ	UW	u:	Oue - أُوْ	0	190	23	1.08
اِيْ	IY	i:	Aii - إِيْ	0	179	339	2.62
			Diphthong				
اَوْ	AW	aw	Aoue - أَوْ	0	105	8	0.57
اَيْ	AY	ay	Aye - أَيْ	0	126	1	0.64

The Arabic consonants are classified based on their place and manner of articulation. Tables 7 and 8 present corresponding statistics for these categories, following the classifications established in the literature [32], [33]. The place

of articulation refers to where in the vocal tract the airflow is obstructed, leading to distinct sounds. Meanwhile, the manner of articulation describes how the airflow is modified or obstructed, further distinguishing consonant sounds. The utilization of this dual classification system provides linguists and phoneticians with a structured framework to methodically examine and comprehend the variety of consonant sounds present in Modern Standard Arabic.

Table 7. Percentage of consonants classes based on their place of articulation

Place of articulation	Consonants	in %
Alveolar	T,D,R,Z,S,SS, DD,TT,L,N	25.99
Glottal	E,H	6.40
Bilabial	B,M	5.94
Palatal	Y	3.75
Velar	JH,K	2.72
Uvular	KH,GH,Q	2.31
Post-Alveolar	JH,SH	1.59
Labiodental	Q	1.46
Interdental	TH,DH,DH2	1.12

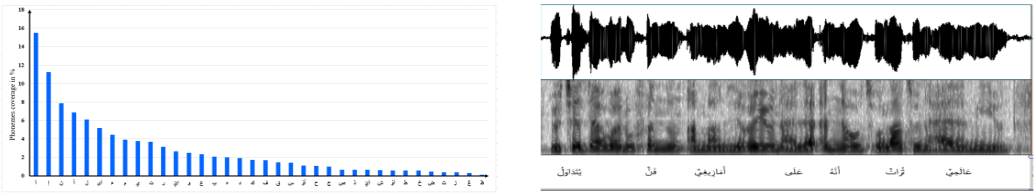
Table 8. Percentage of consonants classes based on their way of articulation

Way of articulation	Consonants	in %
Stop	E,B,T,D,Q,K	15.24
Nasals	M,N	11.73
Fricative	TH,DH,HH,KH Z,S,SH,AI,GH,H	11.23
Glide	W,Y	2.31
Lateral	L	6.08
Trill	R	3.12
Affricative	JH	0.99
Emphatic stop	DD,TT	0.99
Emphatic fricative	SS,DH2	0.75

4. RESULTS AND DISCUSSION

4.1. Speech corpus

The FSDM-USMBA database was established for this study, comprising a speech corpus and transcriptions from 130 Moroccan speakers (63 males and 67 females) aged between 17 and 50 years. During the recording sessions, speakers were asked to utter the 527 sentences with 10 repetitions of every sentence. Voice clarity is fundamental for successful recording. Factors like recording environment, equipment, and speaker-microphone distance influence sound quality. Optimal microphone placement at 10 cm proved effective after testing. For accurate capture, recordings should occur in quiet environments with noise levels below 30 dB, closed windows, and weather impacts such as wind and rain must be avoided. To streamline the process, speakers recited each sentence 10 repetitions consecutively, resulting in 25-50 second audio files. Using WaveSurfer, each recitation was isolated in (.wav) format by removing the unnecessary parts of the audio signal. Audio file names encode multiple details about the speakers. For instance, "XY18ZW21_10.wav" reveals the following information: the initials X and Y representing first and last names respectively, followed by the age 18, city Y, gender W, sentence number 21, and 10 denoting the number of repetitions. Thus, the task of segmenting speech is easy. These recordings have a sampling rate of 16 kHz and a resolution of 16 bits. In the recording sessions, the waveform and spectrogram of each phrase were reviewed to verify the inclusion of the entire sentence in the recording, as illustrated in Figure 3(b). Only correctly pronounced utterances were retained. Our dictionary contains symbolic representations for all the sounds used in the sentences of our corpus.



(a) (b)

Figure 3. Arabic speech corpus illustration: (a) Arabic phonemes of the final corpus and (b) clean speech waveform of an example of an Arabic sentence spoken by a female speaker, is referred to MB18FF21_01 in our audio database

4.2. Speech test

To test our corpus a set of experiments were conducted. A subset of the final corpus, we selected 130 sentences spoken by 60 speakers (30 male and 30 female). This resulted in a vocal corpus of 78,000 audio files. To optimize system performance, we divided the corpus for training (70%) and testing (30%) and adjusted the parameters of Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). This information is stored is stored in the MCA-USMBA.dic file for symbolic representation of each word. The Baum-Welch algorithm, which is a special case of the Expectation-Maximization (EM) method, is used to estimate transition probabilities during training. The acoustic model is trained with

a continuous state probability density, using between 2 and 16 Gaussian mixture distributions and 3 to 5 HMMs. Table 9 shows the achieved Sentence Error Rate (SER) and Word Error Rate (WER). Figures 4 and 5 present the decoding results and the influence of HMM and GMM parameters on SER and WER performance, respectively. The system is evaluated based on three types of errors: insertion, deletion, and substitution, which can occur at both the word and sentence levels. Figures 6 presents concrete examples of sentence recognition errors, illustrating the different types of errors: insertions, deletions, and substitutions at the sentence level. The best configuration used 3 HMMs and 8 GMMs, resulting in a SER of 30.00% and a WER of 15.00%. Our results are in accordance with the study of Abushariah and colleagues [18], who demonstrated a word error rate (WER) of 13.48% for Arabic speech recognition on different sentences spoken by different speakers.

Table 9. SER and WER in percentages for different values of the HMM and GMM

HMM	3				5			
GMM	2	4	8	16	2	4	8	16
WER	22.00	17.75	15.50	23.50	22.20	19.00	18.00	27.50
SER	40.00	37.50	30.00	40.50	50.20	40.00	30.00	40.50

```
Baum Welch starting for 4 Gaussian(s), iteration: 6 (1 of 1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Normalization for iteration: 6
Current Overall Likelihood Per Frame = -138.148867313916
Convergence Ratio = 0.101254599459139
Baum Welch starting for 4 Gaussian(s), iteration: 7 (1 of 1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Normalization for iteration: 7
Current Overall Likelihood Per Frame = -138.148867313916
Split Gaussians, increase by 4
Convergence Ratio = 0.0722613822761673
Baum Welch starting for 8 Gaussian(s), iteration: 1 (1 of 1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Normalization for iteration: 1
```

Figure 4. Optimizing model training with the Baum-Welch algorithm

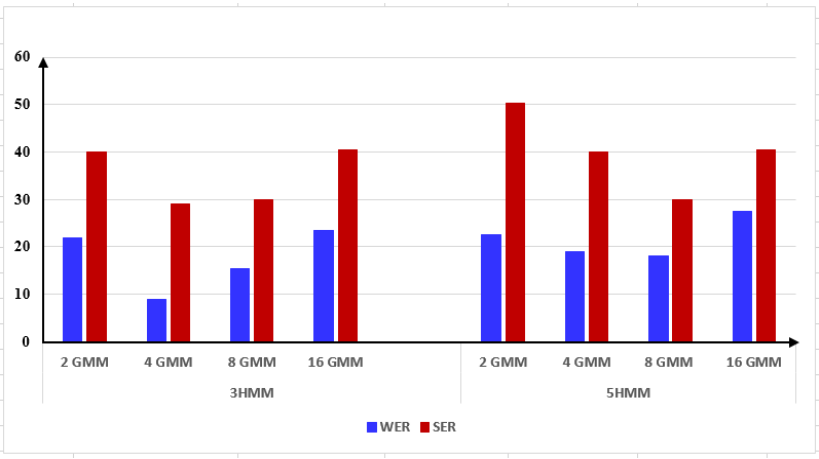


Figure 5. Impact of HMM and GMM values on SER and WER performance

Error Types	Spoken Sentences	Obtained Sentences
Insertions	يُشَارُ إِلَى أَنَّ أُسْتَاذَهُ ذَكِيٌّ جِدًّا	كَسَبَ يُشَارُ إِلَى أَنَّ أُسْتَاذَهُ ذَكِيٌّ جِدًّا
Deletions	وُضِعَ عَلَى مِحَكِّ حَقِيقِيّ	كَسَبَ عَلَى مِحَكِّ حَقِيقِيّ
Substitutions	جِئْتُكَ عَلَى أَنْ تُسَاعِدَنِي وَ لَا تُؤْذِينِي	جِئْتُكَ عَلَى أَنْ تُسَاعِدَنِي وَ *** قَاتِلَةٍ

```
***     يَشَارُ إِلَى أَنَّ أُسْتَاذَهُ ذَكِيٌّ جِدًّا   (speaker129-MBFF18244_11_6)
كَسَبَ يُشَارُ إِلَى أَنَّ أُسْتَاذَهُ ذَكِيٌّ جِدًّا   (speaker129-MBFF18244_11_6)
Words: 6 Correct: 6 Errors: 1 Percent correct = 100.00% Error = 16.67% Accuracy = 83.33%
Insertions: 1 Deletions: 0 Substitutions: 0
وُضِعَ عَلَى  مِحَكِّ حَقِيقِيّ   (speaker101-MBFF18214_11_10)
***     أَوْضَاعُ مِحَكِّ حَقِيقِيّ   (speaker101-MBFF18214_11_10)
Words: 4 Correct: 2 Errors: 2 Percent correct = 50.00% Error = 50.00% Accuracy = 50.00%
Insertions: 0 Deletions: 1 Substitutions: 1
جِئْتُكَ عَلَى أَنْ تُسَاعِدَنِي وَ  لَا تُؤْذِينِي   (speaker114-MBFF18229_11_6)
جِئْتُكَ عَلَى أَنْ تُسَاعِدَنِي وَ  مَا تُؤْذِينِي   (speaker114-MBFF18229_11_6)
Words: 7 Correct: 6 Errors: 1 Percent correct = 85.71% Error = 14.29% Accuracy = 85.71%
Insertions: 0 Deletions: 0 Substitutions: 1
```

Figure 6. Possible errors, for recognition of Arabic sentences examples

5. CONCLUSION

This paper introduces an innovative and efficient method for acquiring a comprehensive and balanced Modern Standard Arabic corpus. The method, meticulously outlined from initial corpus selection to sentence curation, adheres closely to Zipf's law and principles of phonetic distribution equilibrium. The resulting corpus comprises 527 meticulously selected sentences, ensuring the representation of diverse Arabic phonemes across various linguistic contexts, encompassing consonants, vowels, diphthongs, and syllables. The study evaluates an Arabic continuous speech recognition system using 25% of the final corpus. Fine-tuning hidden Markov model (HMM) and Gaussian mixture model (GMM) parameters notably enhances system performance. The findings indicate that employing 3 HMM and 8 GMM achieves optimal sentence error rate (SER) and word error rate (WER) at 30.00% and 15.00%, respectively. In future endeavors, we aim to expand recordings to diverse speaker groups independently, leveraging the entirety of the final comprehensive and balanced corpus. Then, the results obtained in this study are very satisfactory, to the development of a continuous Arabic speech recognition system, which encourage us to extend our research scope to spontaneous Arabic language. Additionally, expanding the corpus, exploring various ASR system architectures, and developing an automatic continuous speech recognition system for the Moroccan dialect.

REFERENCES

[1] K. Shaalan, A. E. Hassanien, and F. Tolba, eds. *Intelligent natural language processing: trends and applications: Springer*. vol. 740, 2017.

[2] Kennedy, *An introduction to corpus linguistics*. Routledge, 2014.

[3] A. A. M. Alqudah *et al.*, "Modern Standard Arabic speech disorders corpus for digital speech processing applications," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 157–170, 2024, doi: 10.1007/s10772-024-10086-9.

[4] Z. Oumaima, and A. Meziane, "Modern Arabic speech corpus for text to speech synthesis," *In : 2020 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, IEEE, pp. 1–6, 2020- November, doi: 10.1109/ICTMOD49425.2020.9380606.

[5] U. Kamath, J. Liu, and J. Whitaker, *Deep learning for NLP and speech recognition. Cham, Switzerland: Springer*. vol. 84, 2019.

[6] J. Egbert, and P. Baker, eds. *Using corpus methods to triangulate linguistic analysis*. London: Routledge, 2019.

[7] M. Weisser, *Practical corpus linguistics: An introduction to corpus-based language analysis,* vol. 43, John Wiley and Sons. 2016.

[8] H. Satori, O. Zealouk, K. Satori, and F. ElHaoussi, "Voice comparison between smokers and non-smokers using HMM speech recognition system," *International Journal of Speech Technology* , vol. 20, no. 4, pp. 771–777, 2017, doi: 10.1007/s10772-017-9442-0.

[9] H. Satori, and F. ElHaoussi, "Investigation Amazigh speech recognition using CMU tools," *International Journal of Speech Technology* , vol. 17, no. 17, pp. 235–243, 2014, doi: 10.1007/s10772-014-9223-y

[10] L. Villaseñor-Pineda, M. Montes-y-Gómez, D. Vaufreydaz, and J. F. Serignat, "Experiments on the Construction of a Phonetically Balanced Corpus from the Web," *In Conference on Intelligent Text Processing and Computational Linguistics*, vol. 2945, pp. 416–419, 2004- February, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-540-24630-5-50.

[11] H. M. Wang, "Statistical analysis of mandarin acoustic units and automatic extraction of phonetically rich sentences based upon a very large chinese text corpus," *In International Journal of Computational Linguistics and Chinese Language Processing* , vol. 3, no. 2, pp. 93–114, 1998- August, doi: 10.30019/IJCLCLP.199808.0005.

[12] V. Radová, and P. Vopálka, "Methods of Sentences Selection for Read-Speech Corpus Design," *In International Workshop on Text, Speech and Dialogue*, vol. 1692, pp. 165–170, 1999- September, Springer Berlin Heidelberg, doi: 10.1007/3-540-48239-3-30.

[13] J. Matoušek, and J. Romportl, "On building phonetically and prosodically rich speech corpus for text-to-speech synthesis," *In: Proceedings of the second IASTED international conference on Computational intelligence: ACTA Press* , pp. 442–447, 2006-

20-22 November, San Francisco, USA.
[14] K. Yazawa, "Harvard-NGSL Sentences for English Learner Speech Corpora," *In 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA),* IEEE, pp. 1–5, 2022, doi: 10.30019/IJCLCLP.199808.0005.
[15] H. Schwenk, and X. Li, "A corpus for multilingual document classification in eight languages," *arXiv preprint*, arXiv:1805.09821. 2018, doi: 10.48550/arXiv.1805.09821.
[16] E. Grave *et al.*, "Learning word vectors for 157 languages," *arXiv preprint.*, arXiv:1802.06893, 2018, doi: 10.48550/arXiv.1802.06893.
[17] A. A. M. Alqudah *et al.*, "Arabic Automatic Speech Recognition for Speakers With Speech Disorders: A Comprehensive Review," *2023 International Conference on Information Technology (ICIT), Amman*, pp. 667–673, 2023, doi:10.1109/ICIT58056.2023.10225965.
[18] M. Alghamdi, A.H., Alhamid, and M.M., Aldasuqi, *Database of Arabic sounds: sentences*, in Arabic, Technical report, King Abdulaziz City of Science and Technology (KACST), Riyadh, Saudi Arabia, 2003.
[19] M. A. Abushariah *et al.*, "Phonetically rich and balanced text and speech corpora for Arabic language," *Language resources and evaluation.*, vol. 46, pp. 601–634, 2012, doi: 10.1007/s10579-011-9166-8.
[20] Y. Yuwan and D.P. Lestari, "Automatic extraction phonetically rich and balanced verses for speaker-dependent quranic speech recognition system," *In: Hasida, K., Purwarianti, A. (eds) Computational Linguistics.*, vol 593, pp. 65–75, 2015, doi: 10.1007/978-981-10-0515-2-5.
[21] D. Qi, and H. Wang, "Zipf's Law for Speech Acts in Spoken English," *Journal of Quantitative Linguistics*, pp. 231–258, 2024, doi: 10.1080/09296174.2023.2202470.
[22] A Ech-Charfi, "Frequency and text coverage in Standard Arabic based on Arabic Internet Corpus," *Journal of Applied Language and Culture Studies*, vol. 6, no 3, pp. 1-19, 2023.
[23] R. Bassiouney and E. G. (Eds.). Katz, *Arabic language and linguistics*. Georgetown University Press, 2012.
[24] A. Hussein, S. Watanabe and A. Ali, "Arabic speech recognition by end-to-end, modular systems and human," *Computer Speech and Language*, vol. 71, p. 101272, 2022, doi: 10.1016/j.csl.2021.101272.
[25] I. Guellil, H Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 497-507, 2021, doi: 10.1016/j.jksuci.2019.02.006.
[26] Y. A. El-Imam, "Phonetization of Arabic: rules and algorithms," *Computer Speech and Language* , vol. 18, no. 4, pp. 339–373, 2004, doi: 10.1016/S0885-2308(03)00035-4.
[27] F. Sindran, F. Mualla, T. Haderlein, K. Daqrouq, and E. Nöth, "Automatic phonetization-based statistical linguistic study of standard Arabic," *Int. J. Comput. Linguist.(IJCL)*, vol. 7, pp. 38–53, 2016.
[28] M. Elmahdy, R. Gruhn and W. Minker, *Novel techniques for dialectal arabic speech recognition.* Springer Science and Business Media, 2012.
[29] T. Buckwalter, and D. Parkinson, A *frequency dictionary of Arabic: Core vocabulary for learners. Routledge*. 2014.
[30] A. Masrai, and J. Milton, "How different is Arabic from other languages? The relationship between word frequency and lexical coverage," *Journal of Applied Linguistics and Language Research*, vol. 3, no. 1, pp. 15–35, 2016.
[31] A. Amrouche, A. Abed, K. Ferrat, , K. N. Boubakeur, Y. Bentrcia, and L. Falek, "Balanced Arabic corpus design for speech synthesis," *International Journal of Speech Technology* , vol. 24, no. 3, pp. 747–759, 2021, doi: 10.1007/s10772-021-09846-8.
[32] J. C. Watson, *The phonology and morphology of Arabic*, Oxford University Press, USA, 2002.
[33] F. Sindran, Automatic Phonetic Transcription of Standard Arabic with Applications in the NLP Domain (Doctoral dissertation, Friedrich-Alexander-Universitaet Erlangen-Nuernberg (Germany)). 2021.

BIOGRAPHIES OF AUTHORS

Youssef Boutazart received the engineer degree in Automation from the Belarusian state Agrarian Technical University of Minsk – Belarus and the Bachelor in electronics from Moulay Ismail University of Meknes – Morocco. Since 2009, he has been administrator of the Presidency by Sidi Mohamed ben Abdellah University. Currently He is a Ph.D. student in the LISAC of the Dhar Mehrez Faculty of sciences of Fez. His research interests are focused on the development of the rich and balanced speech corpus for high- performance speech recognition systems. He can be contacted at email: youssef.boutazart@usmba.ac.ma.

Naouar Laaïdi got her Master in Electronics, Automatics and Signal Processing Faculty of Sciences, Chouaib Doukkali University, El-Jadida. Currently, she is a Ph.D. student at LISAC Laboratory at University Sidi Mohamed Ben Abdellah Faculty of Sciences of Fez. Specialist in many disciplines among Clustering, Machine Learning, Classification, Automatic speech recognition. He can be contacted at email: naouarlaaidi@gmail.com.

Abderrahim Ezzine is a Ph.D student in the LISAC laboratory of the Faculty of Sciences of Fez. He obtained a master's degree in industrial computer science and a bachelor's degree in electronic engineering from the Faculty of Science of Fez-Morocco. His research interests include automatic speech recognition system in clean and noisy environments, natural language processing machine, feature extraction signal, machine learning and deep learning for speech applications. He can be contacted at email: ezzine2abderrahim@gmail.com.

Hassan Satori is Associate Professor, Department of Computer Science, Faculty of Sciences, Dhar El Mahraz, Sidi Mohammed Ben Abdellah University Fez/Morocco. He is Ph.D in Speech recognition and Signal Processing in 2009. He received a MSc in Physics from the Mohamed Premier University, and a Ph.D in Nanotechnology and computer simulation, also from the same University, in 2001. From 2001-2002 he was a postdoctoral fellow in Physics, mathematical modeling and computer simulation at the Chemnitz University of Technology, Germany, and was then a postdoctoral fellow in computer simulation at Department of Interface Chemistry and Surface Engineering, Max-Planck-Institut, Düsseldorf, Germany, (2003-2005). Dr. Hassan SATORI has large academic teaching and research experience in the fields. He can be contacted at email: hassan.satori@usmba.ac.ma.

Mohamed Taj Bennani received his Master's degree in Computer Science and Networking from Science Faculty of Tangier, Tangier in 2011. He received his Phd in 2019 (Computing Science and Networking) from Science Faculty of Tangier. At present, he is working as Prof in Faculty of Science Of Dhar el Mahraz Fez since 2019. He is qualified university professor since June 2023. He can be contacted at email: taj.bennani@usmba.ac.ma.