International Journal of Speech Technology
https://doi.org/10.1007/s10772-021-09846-8

Balanced Arabic corpus design for speech synthesis

Aissa Amrouche[1,2] · Ahcène Abed[3] · Kamel Ferrat[1] · Khadidja Nesrine Boubakeur[1,2] · Youssouf Bentrcia[1] · Leila Falek[2]

Received: 2 January 2021 / Accepted: 5 April 2021

Abstract
This paper aims to design and validate a phonetically balanced speech corpus for Arabic language. Designing and developing a rich and phonetically balanced corpus in optimal context is one of the key issues in building high quality of text-to-speech synthesis systems. The rich characteristic is in the sense that it must contain all the possible phonemes on the right and left context, whereas the balanced characteristic is in the sense that it respects the phonetic distribution in the language. We propose a new methodology for designing and implementing such corpus for speech synthesis purposes. The paper explains the whole creation process of this corpus, beginning with the design stage, corpus creation, recording phases, and finally the segmentation of the speech corpus. The speech corpus contains 202 sentences with 6174 phonemes. In order to validate the speech corpus, an Arabic speech synthesis system using Hidden Markov Models has been developed.

Keywords Arabic speech corpus · Corpus design · Corpus recording · Phonetically balanced · Text-to-speech synthesis

1 Introduction

Designing a speech corpus is one of the key issues in building high quality text-to-speech synthesis systems (Amrouche et al., 2017a; Itunuoluwa et al., 2014). The richness of its content, the quality of the annotation, the homogeneity of the voices and the conditions of recordings, are parameters that determine the quality of the obtained synthesized speech. Naturalness and fluency become more and more necessary in these systems. The preparation of any type of speech corpus is normally a project on its own. The choice of sentences' word to be recorded is a hard task. In the literature there are many method to select texts for building textual corpus; random selection of sentences from various topics is one of the most frequently used techniques for speech corpora design (Niladri & Ramamoorthy, 2019). But corpus design is a long and difficult task, therefore some means of optimization are necessary (Almosallam et al., 2013). When building open domain applications optimization becomes necessary since recording every possible speech event is practically impossible, we need to build a short corpus, which contains the minimum representations of phonemes. The required corpus is not available for Arabic language. We propose a new methodology for building a corpus. The advantage of the method is that the corpus is produced with full and specific knowledge of its intended use.

The main purpose of the work discussed in this paper is the design and the record of a phonetically balanced speech corpus intended for speech synthesis system, by building an optimal rich corpus for the Modern Standard Arabic (MSA) language.

✉ Aissa Amrouche
amrouche_a@yahoo.fr

Ahcène Abed
abedahcene@gmail.com

Kamel Ferrat
kamelferrat@yahoo.fr

Khadidja Nesrine Boubakeur
boubakeur.khadidja@gmail.com

Youssouf Bentrcia
anoirnaser@gmail.com

Leila Falek
lfalek@hotmail.fr

[1] Scientific and Technical Research Centre for the Development of Arabic Language CRSTDLA, Algiers, Algeria

[2] Laboratory of Spoken Communication and Signal Processing, USTHBAlgiers, Algiers, Algeria

[3] Signal and Communication Laboratory, USDB, Blida, Algeria

Published online: 21 April 2021

This paper is organized as follows: the first section gives a brief overview of the related works and a description of the desired corpus characteristics. The second section shows the Arabic language phonemes and the phonetically balanced criteria. In the third section, a process of the corpus design and its final features are described. Then, in the fourth section explains, the recording phase and the speaker selection process. In the last section we present the validation and some conclusions on the speech corpus.

2 Related works

A corpus-based speech synthesis uses acoustic units resulting from reading a carefully chosen textual corpus. The commonly used criteria mainly assure the coverage of all units (phonemes, diphones, triphones, etc.). The "greedy" method consists of choosing incrementally from a large corpus, a subset of sentences that reached the desired coverage. The percentage of existing units in the initial corpus is present in the built corpus. According to the approaches, the units to be covered are diphones, diphones in context (Alsulaiman et al., 2011), triphones or syllables. For each sentence candidate, a score is calculated to allow the choice of the most useful sentence, which increases the most units' coverage. The selected sentence is then moved from the initial corpus to the created corpus and its units are then removed from the set of units to cover. Many studies (Abdo et al., 2017; Barbot et al., 2012; Chrobaka et al., 2006; Muljono et al., 2019; Novitasari et al., 2020) have used the "Greedy" algorithm for the constitution of the textual corpus.

Other methods, inspired from the "greedy" method, have been proposed including the "reversed greedy method" (or "cracheuse") (Chrobaka et al., 2006) and the method of exchange by pairs (Tian et al., 2005).

The algorithm "cracheuse", in contrast to the greedy algorithm, starts with full coverage, that of the initial corpus. The sentences are deleted one by one until the deletion of a sentence that has units with the total coverage. As for the pairs exchange method, it aims to improve the coverage rather than build, either by increasing the number of units covered or decreasing coverage as a minimum threshold.

The effectiveness of these three methods depends on the chosen criteria for calculating the score of each candidate sentence. Regardless of the method used, the criteria are related to the number of separate units for a sentence and the number of coverage units. To control the length of the selected sentence, the total number of units in the sentence is considered. In Chrobaka et al. (2006), several criteria were presented and evaluated as the criteria based on the number of useful units to coverage in the candidate sentences, or the presence of the rare units in the sentence.

Depending on the goal, the score of each candidate sentence could be calculated differently. To achieve the coverage, the easiest way for calculating the score is to normalize the number of new units for a sentence to the total number of units contained in the same sentence. In addition to get coverage, the aim is to promote the rare events so the score calculation uses the frequencies of the observed units of the initial corpus. In order to obtain a high variability on the phonetic level, Chalamandaris et al. (2010), Thao et al. (2011) propose to calculate the score of each unit (diphones) of the candidate sentence according to his left and right phonetic contexts. The chosen units are those that increase the variability of phonetic corpus. This approach allowed to obtain a better representation of triphones in the corpus. The objective is to achieve coverage and variability of the units, which are partially antagonists. A compromise between the two is necessary. This leads to different scores. One of the difficulties of the constitution of the textual corpus is in the choice of optimal coverage of units in the sense of an application target. In the case of the general speech synthesis, the distribution of the desired units are that limit redundancy of frequent units and maximizes the presence of rare units. Full coverage must be reached at least the elementary units. In addition, adequate representation of the units must be ensured to anticipate the different contexts of possible appearance of these units. For synthesis dedicated to restricted domain, the ideal distribution of the units is one that reflects a particular application context. In this case, the strain on the rare units may be less strong. However, the most frequent units in relation to the domain must be well represented. The corpus can also be smaller and more specific to the domain.

For the criteria that look for obtaining a coverage, the distribution of units in the final corpus is hard to master. Therefore, we propose a criterion intended to globally control the distribution of units in the built corpus at each stage of the process.

In this paper, we propose a method of textual corpus based on a statistical approach. This approach aims to build a corpus that the units' distribution tends to a prior distribution. The criteria used evaluates the usefulness of a sentence based on all units in the built corpus.

The creation of such corpus can be divided into three main stages:

(1) Constitution of reading script;
(2) Recording of the script;
(3) Post-processing of data.

Before going to the constitution of the script, we must first give some information about phonemes and diphones of the Modern Standard Arabic. These phonetic units are the basic elements for the construction of words and sentences

of a script in the chosen language, therefore it must be defined beforehand because they are language dependent.

3 The Arabic language

Arabic is a Semitic language of the same family as the Syriac, Aramaic and Hebrew. However it has a rich morphology and a flexible word order. Nowadays it is spoken by almost 420 million people in the world and 22 countries as well. The Arabic natural language and automatic signal processing must deal with several complex problems pertinent to the nature and structure of the Arabic language. For example, Arabic is written from right to left. Like Chinese, Japanese, and Korean there is no capitalization in Arabic. In addition, Arabic letters change shape according to their position in the word (Amrouche et al., 2015; Attia, 2008; Farghaly et al., 2009). The research about the automatic processing of Arabic has started in the 1970s. The first studies were primarily focused on lexicons and morphology. We will state some particularities of the Arabic language.

Modern Standard Arabic (MSA) consists of 40 phonemes; 28 consonants, six vowels (three long [ā, ī, ū] and three short [a, i, u]) and six vowel variants emphatic context. The phoneme is the smallest element of the units of speech, it is the difference in meaning, the word and the sentence. Arabic phonemes contain some distinctive classes called uvular, pharyngeal, glottal and emphatic. These classes characterize the Semitic languages like Arabic and Hebrew. The uvular, pharyngeal and glottal phonemes, are also called back phonemes. The MSA presents four emphatic phonemes: two plosives and two fricatives (Ferrat & Guerti, 2017). Table 1 shows the emphatic and back phonemes with their Speech Assessment Methods Phonetic Alphabet (SAMPA) equivalents.

In addition, the MSA is characterized by the gemination phenomenon. This process is very relevant in this language (Ferrat & Guerti, 2013). Indeed, the sentence [had`ara edarsa] (حضر الدرس) (he attended the lesson) presents a different sense, compared to the sentence with gemination [had`d`ara edarsa] (حضّر الدرس) (he prepared the lesson). Also the word [naqaba] (to dig) differs from the word [naqqaba] (to seek) by a gemination of the phoneme [q]. In phonetic writing, gemination is represented by a double consonant [CC].

There are five types of syllables classified by the features: Open/Closed and Short/Long (Table 2) (Satori et al., 2009).

A syllable is called open (respectively closed) if it ends with a vowel (respectively a consonant). All syllables have a single vowel (short or long) and starts with a consonant followed by a vowel. The syllable [CV] can be at the beginning, middle or end of the word. (Abed & Guerti, 2016) Types considered syllables: [CV], [CVC], [CVV], [CVVC] and [CVCC], where [V], [VV] and [C] are respectively a short vowel, long vowel and consonant. As each syllable contains one and only one vowel (short or long), where the number of syllables of a word is equal to the number of vowels;

1. Each syllable must begin with a consonant;
2. The kind of syllable [CV] is called short syllable, this is the most common in the Arabic language;
3. Other types are all long;
4. All syllables can be at the beginning, middle or end of a word;

Table 2 Classification of syllables in Arabic

Syllable	Open	Closed
Short	[CV]	
Long	[CVV]	[CVC] [CVVC] [CVCC]

Table 1 Inventory of the emphatic and back phonemes of MSA

Phoneme (SAMPA code)	Arabic character	Place of articulation	Manner of articulation			
			Voiced	Emphatic	Plosive	Fricative
[t`]	ط	Dental	–	+	+	–
[s`]	ص	Alveolar	–	+	–	+
[d`]	ض	Alveodental	+	+	+	–
[D`]	ظ	Interdental	+	+	–	+
[q]	ق	Uvular	–	–	+	–
[x]	خ	Uvular	–	–	–	+
[G]	غ	Uvular	+	–	–	+
[X\]	ح	Pharyngeal	–	–	–	+
[?`]	ع	Pharyngeal	+	–	–	+
[h]	ه	Glottal	–	–	–	+
[?]	ء	Glottal	–	–	+	–

5. Types of syllables [CV] and [CVV] are called open, types of syllables [CVC], [CVVC] and [CVCC] are called closed.

4 Corpus design

The goal of the development of a phonetically rich and balanced corpus is to assure that all the possible phonemes and some impossible phoneme combinations between words are occurred in the final corpus. To assure that the speech synthesis system produces at least a natural intelligible sound, we will have to design the corpus to guarantee that there are at least all the smallest units. The purpose of the phonetically rich and balanced corpus is to keep the appearance rate of these units in the corpus, as close as possible to their appearance in the language. In this way, usual phonemes will appear several times in the recorded corpus, in multiple contexts, and all rare ones will appear as close as the appearance in the language.

4.1 Textual corpus

The textual corpus, or script, means all the sentences that will be recorded by the speaker. This may be a literary text (like a novel), a succession of independent texts (such as newspaper articles), or a set of isolated sentences without logical link. The latter is most common in the context of a synthesis of style "neutral reading", in which prosodic consistency between successive sentences is neglected. The script has a direct impact on the final quality of the synthesized voice, which depends on the variety of its units in the corpus (Jawaid et al., 2014). The speech synthesis quality will be better if the system has long segments and adapted to the context. The goal of our research is to build a phonetically rich corpus that provides a certain linguistic richness, covering all Arab diphones and several presentations for each diphones. We start from a very rich initial corpus to arrive at a set of words that represents all the diphones to construct sentences using these words. Figure 1 gives the flowchart of the steps for the constitution of the script.

4.2 Initial corpus

The constitution of an initial text corpus is a milestone in the process of creating a script. The initial corpus allows to observe the statistics of different events of the language and thus to establish coverage priorities, for example through a frequency weighting of the optimization criteria. It can also pick corpus office for the selection of the sentences that are added to the script. In all cases, the correction level of the corpus has a direct impact on the quality of the final script. If the constitution of a universal representative corpus for a language is out of reach, the aim is to cover a significant and balanced corpus for most considered applications for the recognition and speech synthesis (Hafte & Sebsibe, 2018). To achieve our corpus, we have collected an initial corpus of 5.67 million words such as newspapers and the Arabic library Shamela. For this, several sources were assembled. Table 3 shows the initial corpus details:

We can see that a large part of our initial corpus is taken from daily used language, which makes 65% of the collected text data, 40% of the corpus is vowelized. This reflects both our commitment to move away from traditional use of synthesis voice to address new application fields closer to our actual modes of communication, and secondly, speakers will be familiar with the sentences to be recorded.

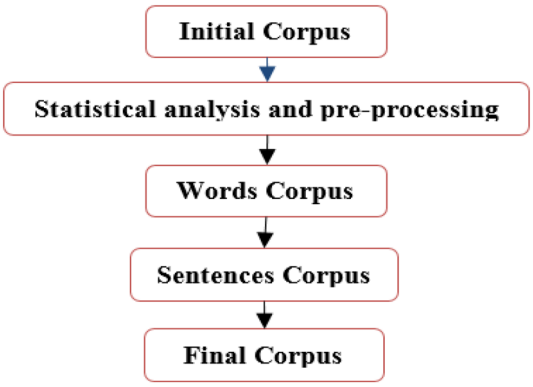

Fig. 1 Flowchart of the development of an optimized textual corpus for Arabic Language

Table 3 Statistics of initial corpus

Sources	Number of words
Prose and poetry	(1.8 million words)
Scientific texts	(1.5 million words)
Newspapers	(1.3 million words)
The Quran and Hadith	(980,000 words)
Recipes	(90,000 words)

4.3 Statistical analysis and pre-processing

This step consists primarily to do general statistics on the initial corpus, to know the distribution of phonemes and diphones in the Arabic language, and to meet this distribution in sentence construction.

Table 4 presents 36 pairs, among 758 possible pairs with the percentage of appearance in the initial corpus. These pairs are divided into two classes: one containing easy diphones, which means the frequent (most used) and the other containing the difficult diphones (less used), which means the least frequent.

Table 4 Representation of some pairs with their appearance percentage in the initial corpus

Pairs	%	Pairs	%	Pairs	%	Pairs	%
(ل، ا)	7.09	(م، ن)	1.13	(ي، ب)	0.57	(س، ت)	0.06
(ا، ل)	2.54	(ل، ع)	1.04	(و، ل)	0.56	(ش، ي)	0.05
(م، ل)	1.76	(أ، ل)	0.86	(ت، ي)	0.56	(ك، ف)	0.04
(ن، أ)	1.67	(ا، ه)	0.85	(ا، ر)	0.56	(ض،ب)	0.04
(ا، م)	1.55	(ه، ن)	0.84	(ر، ي)	0.54	(س، ق)	0.03
(ي،ف)	1.43	(ي، ل)	0.81	(و، أ)	0.54	(خ، ا)	0.03
(ل، و)	1.37	(ا، ق)	0.80	(ا، ذ)	0.54	(ر، ز)	0.02
(ه، ل)	1.34	(ن، ا)	0.74	(ع، د)	0.09	(ش، ك)	0.02
(ا، و)	1.15	(ا، ن)	0.73	(ب، ح)	0.08	(ف، ط)	0.01

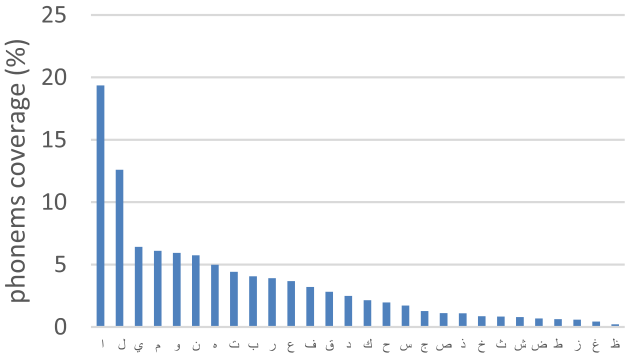

Fig. 2 Distribution of phonemes in the initial corpus

Figure 2 and Table 5 show the phonemes and vowels coverage of the initial corpus.

By analyzing these statistics, we see that the phonemes Alifأ and Lam ل are the most frequent with the percentages of 19.36% and 12.6% respectively, and the 'al' diphoneأل is very frequent. On the other hand the phoneme ظ [D`] is the least used because its acoustic characteristics, which pose conjunction problems with other phonemes. The second step is to make an acquisition by cleaning and formatting the initial corpus to reach a small corpus consisting of non-differentiable words, because we are trying to extract the diphones in lexical units (root). To facilitate and expedite this task, we carried out a program consisting in:

– Eliminating all kinds of proposals, pronouns and conjunctions;
– Applying for a significant word, stemming to detect prefixes and suffixes and keep only the root;
– Calculating the number of occurrences of a significant word after stemming and order them descending.

The stemming algorithm is a computational procedure, which reduces all words having the same root (or, if prefixes are left untouched, the same stem) to a common form. For a given word that represents a number of graphemes greater than three, the program removes the prefixes and suffixes to keep the root. After that it compares this word with a list of registered schemes of the Arabic language (Amrouche et al., 2017b; Tengku & Abu Ata, 2013). If the scheme exist, it removes the prefixes and suffixes and returns the root as output. Table 6 gives an illustrative example.

Table 7 presents some stemming results. The words and their frequencies of appearance in the initial corpus are presented in descending order. This task is very important to select words, it allows to choose the most frequent words, which means the most used and where speakers are familiar with them. This means a perfect pronunciation, which increase the recording quality.

Table 6 Stemming example of the word فسيعلمون

Word	Suffix	Word without suffix	Pattern	Root
فسيعلمون	فس	يعلمون	يفعلون	علم

Table 5 Percentage of vowels in the initial corpus

Fatha /a/	Kasrah /i/	Dammah /u/	Sukun //	Chaddah Consonant doubling	Fathatān /an/	Kasratān /in/	Dammatān /un/
ـَ	ـِ	ـُ	ـْ	ـّ	ـً	ـٍ	ـٌ
%44.32	%18.68	%16.23	%11.57	%6.47	%0.84	%1.078	%0.77

Table 7 Example of output words and their frequencies for Hizb 59 and 60 of Qur'an

Stemmer	Occurrences
كذب	10
قال	9
ذكر	9
خلق	8
علم	7
جعل	5
خرج	5
يوم	5
أمر	5
نفس	5

Table 8 Words chosen by the program

Diphones	Words	Translation
مرشد:مر, رش, شد	مرشد	Guide
نصف:نص, صف	نصف	Half
قرن:قر, رن	قرن	Cycle
زهرة:زه, هر, رة	زهرة	Flower
ضغط:ضغ, غط	ضغط	Pressure
صوت:صو,وت	صوت	Sound
ضرر:ضر,رر	ضرر	Damage
دخول:دخ, خو, ول	دخول	Entrance
سلطة:سل, لط, طة	سلطة	Power
نووي:نو, وو, وي	نووي	Nuclear

4.4 Words corpus

In this step, we made another program that makes the selection of words. These words are chosen such that they ensure at least one for each unit (diphone). For a candidate word containing N phonemes, which consists of n = N − 1 diphones, the selection is as follows:

The units are divided into two classes; a class that contains the most common units C1, and another that contains the least frequent units C2.The program selects only those words, which their units are in both classes at the same time, using (1) and (2).

$$D_i \in C_{1+mod(i+1,2)}, D_{i+1} \in C_{1+mod((i+1)+1,2)} \dots D_n \in C_{1+mod(n+1,2)} \quad (1)$$

$$D_i \in C_{2-mod(i+1,2)}, D_{i+1} \in C_{2-mod((i+1)+1,2)} \dots D_n \in C_{2-mod(n+1,2)} \quad (2)$$

where D_i is the number of diphones in a word; mod is the rest of the Euclidean division.

We start with the most frequent words to less frequent to guarantee the familiarity of these by most community. Which means, the speakers know the exact pronunciation of the word beforehand. Once a word is selected, it will stored and deleted from the initial list as well as the constituting diphones from the two classes. The process stops until the emptying of the two classes.

Table 8 illustrates ten words chosen by the program with their diphones:

At the end of the implementation, we obtained 535 words that contain all possible diphones for the Arabic language. So we provided at least one diphone copy. In all, we received 1605 diphones making up the 535 words, with a repetition of 120 diphones. This repetition is due to take some words containing one diphone from a class and the other was deleted because there was not words contained the diphones from the two classes. In the final step, we organized the words in lexical fields such as verbs, contents names and animals' names. This organization was made to facilitate the construction of sentences.

4.5 Sentences corpus

We built midsize sentences from 4 to 10 words. As needed, we added words, proposals, prefixes/suffixes and we make the prohibited diphones between words. We added also the diacritic marks for all sentences at the same time of their construction. We stopped the construction of the sentences when the 535 words were completed. The added words are most frequent in the Arabic.

The corpus is built of 202 sentences (1254 words). Figure 3 shows the coverage of phonemes in the sentences corpus.

By comparing the phoneme coverage of the initial corpus and corpus sentences, we notice that the order of phonemes is quite respected, from the most frequent phonemes to less frequent.

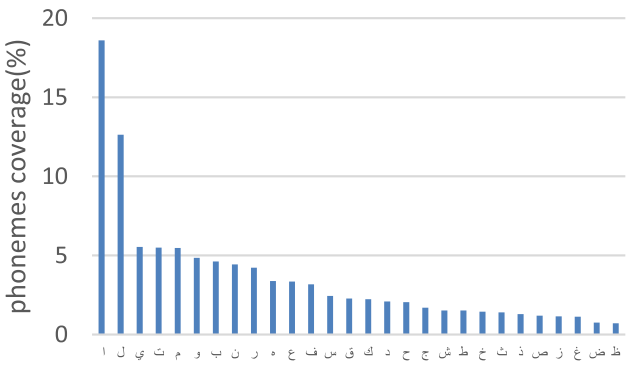

Fig. 3 Distribution of phonemes in the sentences corpus

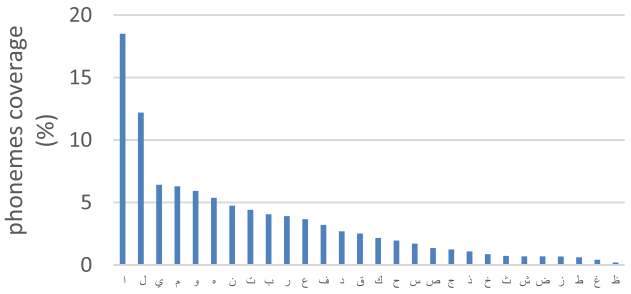

Fig. 4 Distribution of phonemes in the final corpus

Table 9 Final main corpus statistics

Total number of sentences	202
Declarative sentences	180
Interrogative sentences	22
Average of words per sentence	6
Total number of words	1254
Number of different words	975
Total number of phonemes	6174
Total number of diphones	3614

4.6 Final corpus

This last step was to control the phonemes distribution and to respect this distribution in the initial corpus. To do this, we added sentences, words as necessary. In each addition of the words or sentences, we controlled the coverage of phonemes after the addition by analyzing the main corpus statistically, and we compared these results with the results obtained by the analysis of the initial corpus. We stopped the additions once we got a closer phonemes coverage to the initial analysis fixed at the beginning.

The final corpus obtained must contains all Arab diphones and unnecessary diphones between words to insure the distribution of the language fixed before.

Figure 4 and Table 9 show the coverage of phonemes and vowels of the final main corpus respectively.

By analyzing these results, we found that the final corpus is a sample of the language, in which all Arab diphones are presented. Prohibited diphones are shown between words.

Table 10 shows the vowels percentage of the final corpus. We see that the order of fatha, kasra and damma have the same vowels representations in the Arabic language.

4.7 Manual processing

Manual treatments can accompany each phase of creating the script described above.

First, the reading of the initial corpus allows to eliminate aberrant sentences (e.g. Foreign language sentences), correcting some passages (e.g. Misspellings, abbreviations), or even completely rewriting poorly formed texts (e.g. Short Message Service (SMS)). The main objective of these textual corrections is to facilitate the interpretation of the initial corpus, which provides a phonetic transcription and prosodic annotation of the corpus that are consistent with an intuitive reading. The quality of the final script depends on the relevance of linguistic coverage, and the fluent reading of the expected phonetic sequence.

If a complete rereading of the initial corpus is out of reach because it is too costly, we can of course limit ourselves to its critical parts. For example, instant messaging traces will benefit much more from manual intervention than journalistic or literary sources. Throughout the construction of the script, the supervision of selected words can be very beneficial, for example, to reject too complex or phonetically wrong words. This pruning can be carried out as during the construction of the script or later.

It is also possible to correct the phonetic transcription of some sentences to avoid the discrepancies between the expected sequences that actually carried out by the speakers. This type of modification or posteriori deletion can cause holes in the linguistic coverage; a real time treatment (that is to say over the construction of the script), which is preferable, as well the disappearance of some units because the deletion of a sentence or phonetic correction may be automatically compensated in the following sentences.

The monitoring of some indicators such as coverage rates of many units types or the length of sentences, also offers the opportunity to refine or correct the optimization criteria in real time.

Finally, the final script may be annotated or rewritten to facilitate the reading and promote conformity between the expected phonetic sequence and that effectively produced: simplifying of writing complex words, francization of foreign words to remove ambiguities in pronunciation, connection indications, etc. The final corpus answer and respect our conditions fixed at the beginning. In the next section, we present the recording of corpus.

Table 10 Percentage of vowels in the initial corpus

Fatha	Kasrah	Dammah	Sukun	Chaddah	Fathatān	Kasratān	Dammatān
ـَ	ـِ	ـُ	ـْ	ـّ	ـً	ـٍ	ـٌ
%42.45	%18.10	%15.54	%12.60	%7.37	%1.77	%1.24	%0.93

5 Recording phase

5.1 The choice of speakers

The choice of speaker and voice is essential for the quality of speech synthesis. Project management requires some criteria: the speaker ability to follow sometimes-complex reading instructions, ability to control his vocal organs, reading fluency, motivation, etc. Other criteria may be imposed by mastering the work or jointly identified such as sex, tone of voice, the reading style, the regional focus, the level of articulation. A choice can be made on the basis of simple existing voices recordings, or to carry on dedicated recordings in real conditions, which is preferable.

An Algerian speaker who did not receive any special instructions to avoid the influence that may affect its spontaneity made the recording.

5.2 Recording environment and platform

First, the sound acquisition chain must be constant throughout the recordings. This limitation is inherent in the mechanisms of speech synthesis using the unit concatenation: the units must provide closest acoustic characteristics as possible. In particular, the room, the furniture, the equipment acquisition and the position of the speaker from the microphone must remain unchanged. Then, the reverberation must be extremely low because it harms the speech segmental aspects: the stick units on each other cannot segmented or concatenated properly. Also using the same microphone and recording device. Ideally, all sentences should be recorded at once, since the voice quality of the speaker can vary from time to time (Abushariah et al., 2012b; Burkhardt et al., 2005).

Finally, the ambient noise must be as "white" as possible, that is to say that it must not exhibit temporal coherence (successive uncorrelated samples). Non-one-off sound events such as a background discussion, creaking doors, passing vehicles, ringing phones, even weak ones, are to be avoided. The white noise model is theoretical; in practice, it suffices for the ambient noise to present constant characteristics and for the events, which compose it to have a duration of evolution, which is significantly shorter than the phonemes. Under this condition, it is not incompatible with corpus-based synthesis.

To achieve a high audio quality, the recordings took place in a soundproof room at SESTEK (Speech Enabled Software Technologies) studio in Istanbul, Turkey. The soundproof room system consisted of the following subsystems:

- One Professional Microphone: Rote-NT1A;
- An AudioBox USB Studio PreSonus.

The sampling rate is 48 Ksample/s with 16 bit resolution (Janyoi & Seresangtakul, 2020). The speaker were standing in front of the microphone, with a fixed distance along the recording of all sentences about 20 cm (see Fig. 5). The recorded file are monophonic signal and the 202 sentences must be read at a normal speed (from 10 to 12 phonemes/s) by the speaker.

6 Phonetic segmentation

At this stage, each sentence is associated with its phonetic transcription. The Segmentation of the speech to phonemes for each sentences was made in a semi-automatic way. The segmentation phase was divided in three steps. Firstly we developed an automatic Arabic phonemes boundary detection system. This system is mainly used to perform an automatic speech corpus labeling, because the manual labeling is a hard task and consuming time. We have used the HTK (Hidden Markov Tools Kit) model to solve this problem. The Hidden Markov Models implementation is used to detect phonemes boundaries with the textual information given by the transcription file (Abed et al., 2017). Secondly, we proposed an automatic segmentation of speech in phoneme for the Arabic language. This segmentation is based on two different techniques: Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) (Abed et al., 2016). Finally, in order to check the quality of speech segmentation we heard all the extracted units, then we make some alterations to the misunderstood phonemes. The segmented units have been analyzed by Praat software (Boersma & Weenink, 2019).

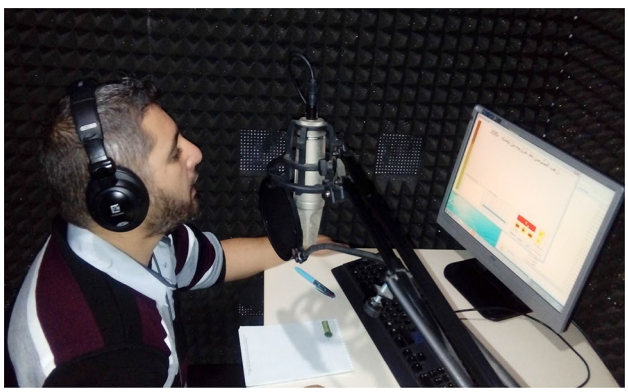

Fig. 5 Recording studio

7 Comparison of the achieved corpus with other works

In this section we will compare our corpus with three other corpus for the Arabic language that are:

7.1 Twenty lists of ten arabic sentences for assessment

Boudraa et al. have developed a corpus contains two hundred sentences distributed on twenty lists of ten sentences (Boudraa et al., 2000). They have based on Pierre Combescure's work applied to the French language (Combescure, 1981). The phonological balance of their corpus was considered as the presence of phonemes in the roots of the Arabic lexicon. Their hundred sentences were formed from these roots. However, this study was not void of some disadvantages such as: Pre-estimate of the number of sentences in two hundred: If the French language needs this number to represent all phonetics aspects, this does not necessarily apply to the Arabic language. In addition, the methodology for the preparation of their set of lists (words or sentences) belongs to speech therapy. Also the concept of phonological balanced is incompatible with the purpose of the study and the absence of phonological richness criteria.

7.2 KACST database of Arabic sounds

King Abdulaziz City of Science and Technology (KACST) created a database for Arabic language phonemes, on the basis of a least number of phonetically rich Arabic words. As a result, a list of 663 phonetically rich words containing all Arabic phonemes based on Arabic phonotactic rules was produced (Abushariah et al., 2012a; Alghamdi et al., 2003). The authors built 367 phonetically rich and balanced sentences. The database contains all phonetic rules of Arabic, which means coverage of all Arabic phoneme clusters. The presence of the least possible number of words to avoid possible redundancy of Arabic phoneme clusters. Although the details given in this study, no objective method for selecting the vocabulary, providing the required conditions on the word vocabulary level and they did not differentiate between lexical and speech usage statistics. The authors considered that the linguistic balance is the presence of all Arabic phonemes in equal proportions, which means each linguistic sound is given an equal score to test it and practice it.

7.3 Phonetically rich and balanced text and speech corpora for Arabic language

Abushariah et al. (2012b) described the preparation, recording, analyzing, and evaluation of a new speech corpus for Modern Standard Arabic (MSA). The speech corpus contains a total of 415 sentences recorded by 40 (20 male and 20 female) Arabic native speakers from 11 different Arab countries representing three major regions (Levant, Gulf, and Africa). They took the three hundred and sixty seven sentences of 7 database of Arabic sounds developed above, The remaining 48 sentences are created for testing purposes, which are mostly foreign to the training speech corpus, which is an important requirement for developing any ASR system.

7.4 A standard Arabic single speaker corpus SASSC

The text corpus included a variety of genres and writing styles from multiple sources to insure a representative sample. Furthermore, diversity of pronunciation was also taken into consideration during the text selection process to capture the entire phoneme spectrum that exist in the Arabic language. The corpus is divided into sub-categories to achieve a wide variety of vocabulary and to enable users of the data set to perhaps extract "exact" relevant phrases if necessary. For example, the use of numbers, date and time are very common in most applications (Almosallam et al., 2013). The text corpus consists of 51,432 words, it contains 627 unique syllables out of a total of 333,981 syllables. Almosallam did not include several phonemes presented in this work, which are essential in MSA. The phoneme set used in the newly-developed corpus included all of theirs and added phonemes that appeared to have been missed. Also they ignored the discussion of geminated consonants and did not describe how they converted diacritised MSA text to phoneme sequences before alignment (Halabi, 2016).

7.5 What distinguishes our corpus from the others

In addition to the acoustic richness of our corpus, which means the presence of all phonemes, diphones, and unnecessary diphones between words of Arabic language, the phonetically balanced is into preserve the phonetic distribution of Arabic language. The chosen words are taken from

Table 11 Samples of the phonetically rich and balanced sentences

Sample 1	فَرَغْ هِشَامٌ ضُحًى مِنْ طَبْخِ غَذَائِهِ
Sample 2	قَهَرُ عَزِيزِ أُمَّةٍ صَعْبٌ تَحمُّلُهُ
Sample 3	جَاءَتْ عَبْدَةٌ فَأَخَذَتْ كُنُوزَ كَعْبٍ مِنْ بَيْتِنَا
Sample 4	إِذَا أَتَاكَ عَظِيمٌ قَوْمٍ فَأَكْرِمْهُ

the used Language, and the sentences are easy to read and understand. Our corpus have: minimum word repetitions, an average of 4–10 words in a single sentence, simple structure of sentences and a minimum number of sentences. Table 11 shows four samples sentences of our corpus.

The corpus consists of 202 Sentences, which 180 are declarative and 22 are interrogative, with an average of 6 words per sentence. It totals 1251 words, 2940 syllables, 6174 phonemes whose 2258 short vowels, 421 long vowels, 395 semi-vowels/w/and/y/, 990 fricative consonants, 940 plosive consonants, 650 liquid consonants/r/ and/ l/, and 520 nasal consonants/m/ and/ n/.

8 Validation of the corpus

In this section, we have tested the corpus for the developed systems, which are: an HMM-based speech synthesis system (HTS) for the Arabic language (Amrouche et al., 2019), design and implementation of a Diacritic Arabic, Text-To-Speech System (Amrouche et al., 2017a) and a contribution to the improvement of the signal synthesis in a TTS system for Arabic (Amrouche et al., 2014). To assess the intelligibility and natural aspects of the obtained synthesis voice, we applied two types of tests for each system.

A listening test was offered to subjects via a web interface made with HTML/PHP. Beforehand, instructions have been given to participants to explain what was expected of them during the evaluation. Each participant fills a set of personal information: name, gender, age group, educational level, mother tongue (Arabic or other) and the listening device used (speakers or headphones). 40 auditors have participated in the evaluation, divided into three age groups (13 with T1: 15–20; 14 with T2: 21–40; 13 with T3: 41–60), including 20 women and 20 men. Participants have different occupations, and they know the Arabic language. Their education level is divided into four levels (10 with A: primary; 10 with B: secondary; 10 with C: high school student; 10 with D: university). The criteria of acceptance a candidate for the system output evaluation were: the understanding of the Arabic language, the absence of hearing disorders, unfamiliar to listen to the synthetic voice. The 40 participants are native speakers, so all participants understand Arabic (written and spoken).

We applied two tests for comprehensibility (clarity), where listeners will check their answers to multiple choice. The choice responses for each subset is forced (1 or 0). The listener chooses one of the possible answers or opts for a third choice, which is "not clear speech". The correct answers for each participant were coded with a score of 1 and incorrect answers with a score of 0. We get an 'x' if speech is not clear, but when calculating the results for each test, we change the 'x' by 0. The answer is considered correct if it matches the target heard.

8.1 Testing the overall intelligibility of the synthetic speech

In this part of the test, we chose 36 words and 18 sentences that contain all Arabic phonemes.

Both lists allowed the judgment of the synthetic voice clarity in general. Auditors should identify from the elements list of each subset with words and sentences randomly drawn. The answers were rated on a multiple-choice test, which allow to easily automate the counting stage results. The final result is expressed as the percentage of words or sentences correctly identified.

8.2 Test of the intelligibility of the consonants and vowels

As for the second part of intelligibility, the DRT (Diagnostic Rhyme Test) has allowed to test vowels as well as consonants no matter their position in a word or sentence. The scoring was the same as the first part of the test but it was conducted so as to find a good clarity of synthetic speech. The test items were classified into subsets (words or sentences) that were distinguished only by a consonant or a vowel. The number of subsets should be sufficient to test systematically all consonants, associated with at least several different vowels. We chose 24 words and 12 sentences for this test. At the end of the tests, we asked listeners to judge the quality of the sentence or word heard, based on listening and give a quality measurement as follows: (5 - Excellent 4 - Very good, 3-good, 2-Fair, 1-Poor). After recovering the intelligibility test scores for groups of words and sentences, we calculated the percentage of the correct answers in each group. Figure 6 and Fig. 7 show the obtained results.

The auditors acknowledged the majority of words and sentences with percentages of 99.25%, 99.58% respectively.

For the DRT test, the changes words and sentences are carried out at the consonants and vowels as follows: The words offered differ in consonants and vowels that have been replaced by other consonants and vowels, introducing confusion, which prevents the transparency of the sound to the place of articulation, the position of consonants and vowels in the sentences and words.

The results obtained are illustrated in Figs. 8 and 9:

The auditors acknowledged the majority of changes of consonants and vowels in words and sentences with percentages of 98.35%, 93.06% respectively. The average percentage of these tests is 97.53%, the auditors understood the majority of the words and sentences and have distinguished the words, the changes are made to create confusion and impede comprehensibility, and in addition, there is

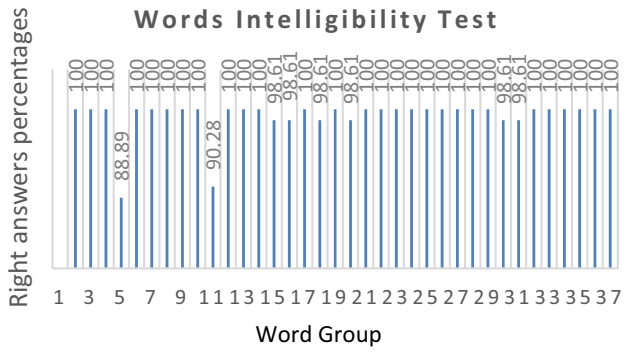

Fig. 6 Words test results

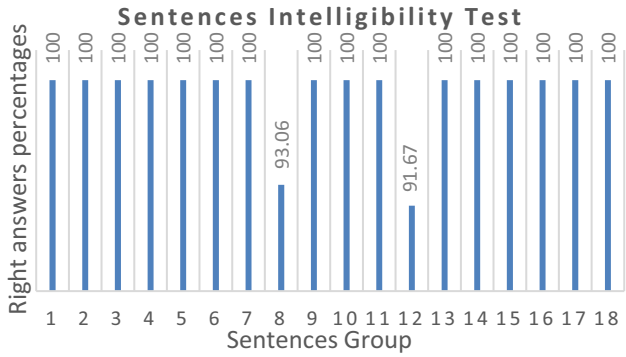

Fig. 7 Sentences test results

no discontinuity along the tested sentences. In the regards of the assessment, the analysis shows the lack of effect of age and gender, as well as a significant effect of education where academics have achieved better results compared with other secondary schools, medium and primary. Words in context are more intelligible than isolated words. The auditors noted the naturalness with an average of 4.259. Analysis of the results of these tests suggested that, when it comes to

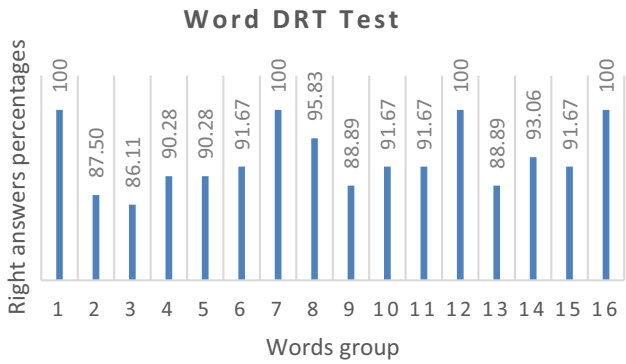

Fig. 8 DRT test results for words

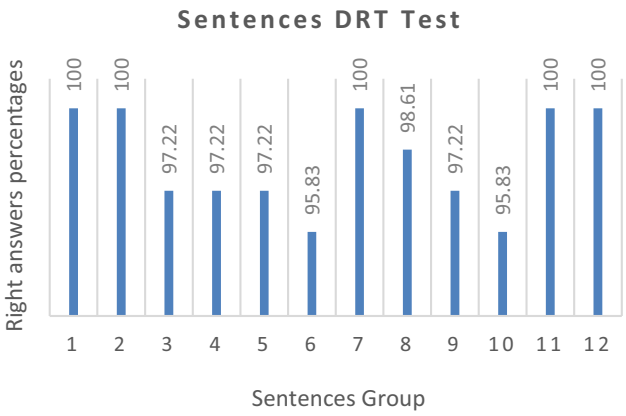

Fig. 9 DRT test results for sentences

the system intelligibility, our system is performing. In these tests, the words in context are more intelligible than isolated words with percentages of 97.74%, 93.56 respectively.

An objective evaluation is performed using the perceptual model PESQ (Perceptual Evaluation of Speech Quality) programmed in C language by the International Telecommunications Unit (ITU 2001). We give the two input signals (original and synthetic), the value of the sampling frequency and the bandwidth. The test was performed using the same signals as the previous tests. After the comparison between the original signal and the synthesis, the PESQ gives two ratings: the first ranging from − 0.5 to 4.5 and the other MOS-LQO (Mean Opinion Score - Listening Quality Objective) is obtained after processing in the MOS (Mean Opinion Score) scale of 1 to 5 (Boros et al., 2014; Qiong et al., 2014; Tadashi et al., 2015).

The similarity was assessed by comparing the actual recording and synthesis of the sentence, using the PESQ test. Figure 10 shows the results of MOS and PESQ tests. The results have a mean of 4.259/5 (standard deviation $\sigma = 0.501$) for the MOS and equal to 3.734/4.5 (standard

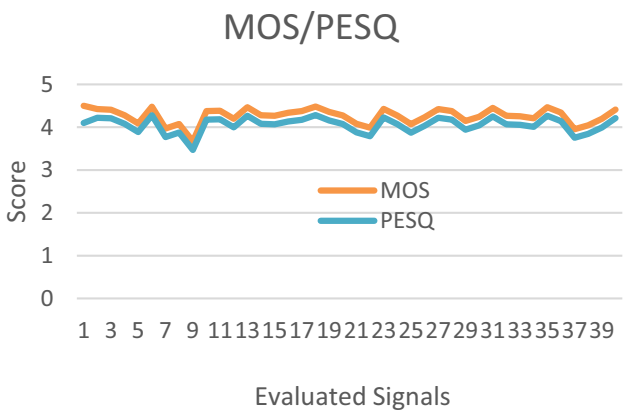

Fig. 10 Results of tests and PESQ MOS

deviation $\sigma = 0.341$) for the PESQ. The MOS gave a good correlation coefficient (0.924) with the objective PESQ test.

The results of these tests showed that the participants can hear what was said and recognize the changes of the synthesized speech. The majority of the words and sentences were correctly recognized and perceived by the majority of listeners. The overall evaluation quality of the system was satisfactory. The objective tests rating is lower than subjective tests.

9 Conclusion

This paper gives an account of the designing and recording a rich balanced Arabic corpus for speech synthesis purposes. The evidence from this study suggests a statistical method for creating a phonetically balanced and rich script. The steps that we have followed are detailed from the collection of the initial corpus to the sentence construction. The final corpus assure the conditions set beforehand, which contains all phonemes and diphones of the Arabic language. We have obtained comprehensive results demonstrating that our phonetically rich and balanced speech corpus have positive impact on the performance of our speech synthesis systems for Arabic language. Nevertheless, our work clearly has some limitations. The speech corpus can be used just for Arabic speech synthesis applications. To further our research we plan to record the script by different speakers for use it in several applications including speaker recognition, speaker identification and different accents of the Arabic language.

Acknowledgements We would like to thank Professor Levent Arslan for giving us the opportunity to visit SESTEK and BUSIM laboratory, the excellent working environment, allowing using its recording studio, his help and guidance during our internship and the exceptional exchanges with BUSIM and SESTEK teams.

References

Abdo, O., Abdou, S., & Fashal, M. (2017). Building audio-visual phonetically annotated Arabic corpus for expressive text to speech. *Interspeech 2017*, Stockholm.

Abed, A., Amrouche, A., & Boubakeur, K. N. (2017). Investigation of HTK for Arabic phonemes boundary detection. *International Conference on Engineering Research and Applications (ICERA-17)*, pp. 17–18

Abed, A., Amrouche, A., Delmadji, A., & Boubakeur, K. N. (2016) Segmentation Automatique des Signaux Sonores par HMM et RNA pour la langue Arabe. *Conférence Internationale en Sciences et Technologies Electriques au Maghreb CISTEM"2016*, Marrakech.

Abed, A., & Guerti, M. (2016). HMM/GMM classification for articulation disorder correction among Algerian children. *The International Arab Journal of Information Technology*, *13*(4).

Abushariah, M., Ainon, R., Roziati, Z., Elshafei, M., & Khalifa, O. (2012a). Arabic speaker independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *The International Arab Journal of Information Technology*, *9*(1), 84–93.

Abushariah, M., Ainon, R., Roziati, Z., Elshafei, M., & Khalifa, O. (2012b). Phonetically rich and balanced text and speech corpora for Arabic language, *Lang Resources and Evaluation,* Springer, pp. 601–634, 46.

Alghamdi, M., Alhamid, A. H., & Aldasuqi, M. M. (2003). Database of Arabic sounds: Sentences. *Technical Report, Saudi Arabia: King Abdulaziz City of Science and Technology (in Arabic).*

Almosallam, I., Alkhalifa, A., Alghamdi, M., Alkanhal, M., & Alkhairy A. (2013). SASSC: A standard Arabic single speaker corpus. In *8th ISCA Speech Synthesis Workshop*.

Alsulaiman, M. M., Ghulam, M., Bencherif, M. A, Awais, M., Zulfiqar, A., & Aljabri, M. (2011). Building a rich Arabic speech database. In *5th Asia International Conference on Mathematical Modelling and Computer Simulation*.

Amrouche, A., Abed, A., & Boubakeur, K.N. (2017b). New method for stemming of Arabic language text. *International Conference on Engineering Research and Applications (ICERA-17)*, pp. 17–18.

Amrouche, A., Falek, L., Teffahi. (2014). Contribution à l'amélioration du signal de synthèse dans un système TTS pour la langue arabe. *Fifth International Conference on Arabic Language Processing (CITAL2014)*, Oujda, Morocco.

Amrouche, A., Falek, L., & Teffahi, H. (2015). Text-to-speech synthesis system for the Arabic language. In *International Conference on Automatic control, Telecommunications and Signals (ICATS15)*.

Amrouche, A., Falek, L., & Teffahi, H. (2017a). Design and implementation of a diacritic arabic text-to-speech system. *The International Arab Journal of Information Technology*, *14*(4).

Amrouche, A., Falek, L., & Teffahi, H. (2019). Arabic speech synthesis system based on HMM. In *Sixth International Conference on Electrical and Electronics Engineering (ICEEE 2019)*.

Attia, M. (2008). Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation. PhD Dissertation, University of Manchester.

Barbot, N., Boeffard, O., & Delhay, A. (2012). Comparing performance of different set-covering strategies for linguistic content optimization in speech corpora. *International Conference on Language Resources and Evaluation (LREC'12)*.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. *Version 6.0.46, from* http://www.praat.org/.

Boros, T. et al. (2014). RSS-TOBI: A prosodically enhanced romanian speech corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 316–320.

Boudraa, M., Boudraa, B., & Guerin, B. (2000). Twenty lists of ten Arabic sentences for assessment. *Acta Acustica united with Acustica*, pp. 870–882.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). *A database of German emotional speech*. Interspeech.

Chalamandaris, A., Karabetsos, S., Tsiakoulis, P., & Raptis, S. (2010). A unit selection text-to-speech synthesis system optimized for use with screen readers. *EEE Transactions on Consumer Electronics*, *56*(3).

Chrobaka, M., Kenyonb, C., & Younga, Y. (2006). The reverse greedy algorithm for the metric K-Median problem. *Information Processing Letters, 97*(2), 31, 68–72.

Combescure, P. (1981). 20 listes de 10 Phrases Phonétiquement Equilibrées. *Revue d'Acoustique*, *14*(56), 34–38.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, *8*(4).

Ferrat, K., & Guerti, M. (2013). Classification of the Arabic emphatic consonants using time delay neural network. *International Journal of Computer Applications*, *80*(10), 1–6.

Ferrat, K., & Guerti, M. (2017). An experimental study of the gemination in Arabic language. *Archives of Acoustics, 42*(4), 571–578.

Hafte, A., Sebsibe, H. M. (2018). Design of a tigrinya language speech corpus for speech recognition. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pp. 78–82.

Halabi, N. (2016). *Modern Standard Arabic Phonetics for Speech Synthesis*. Thesis for the degree of Doctor of Philosophy: University of Southampton.

Itunuoluwa, I., Jelili, O., & Olufunke O. (2014). Design and implementation of text to speech conversion for visually impaired people. *International Journal of Applied Information Systems (IJAIS) Foundation of Computer Science FCS*.

Janyoi, P., & Seresangtakul, P. (2020). F0 modeling for isarn speech synthesis using deep neural networks and syllable-level feature representation. *The International Arab Journal of Information Technology*, *17*(6).

Jawaid, B., Kamran, A., & Bojar O. (2014). A tagged corpus and a tagger for Urdu. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*.

Muljono, H. A., Winarsih, N. A. S., & Supriyanto, C. (2019). An evaluation of sentence selection methods on the different phone-sized units for constructing Indonesian speech corpus. *International Journal of Speech Technology*, *23*(1), 141–147.

Niladri, S. D., & Ramamoorthy, L. (2019). *Utility and application of language corpora*. Springer pp. 1–16.

Novitasari, S., Tjandra, A., Sakti S., & Nakamura, S. (2020). Cross-lingual machine speech chain for javanese, sundanese, balinese, and bataks speech recognition and synthesis. In *Language Resources and Evaluation Conference (LREC 2020)*.

Qiong, H., Yannis, S., Ranniery, M., Korin, R., Junichi, Y., & Javier, L. (2014). An investigation of the application of dynamic sinusoidal models to statistical parametric speech synthesis. In *Interspeech 2014*. pp. 780–784.

Satori, H., Hiyassat, H., Harti, M., & Chenfour N, (2009). Investigation arabic speech recognition using CMU sphinx system. *The International Arab Journal of Information Technology*, *6*(2), 186–190.

Tadashi, I., Sunao, H., Masanobu, A., Yusuke, I., Noboru, M., & Hideyuki, M. (2015). Sub-band text-to-speech combining sample-based spectrum with statistically generated spectrum. In *16th Annual Conference of the International Speech Communication Association*.

Tengku, M. T. S., & AbuAta, B. (2013). Arabic word stemming algorithms and retrieval effectiveness. In *Proceedings of the World Congress on Engineering WCE 2013,*.

Thao, V. D., Do-Dat, T., & Thu-Trang, T. N. (2011). Non-uniform unit selection in Vietnamese Speech Synthesis. In *Proceedings of the 2011 Symposium on Information and Communication Technology*, SoICT 2011.

Tian, J., Jani, N., & Imre, K. (2005). Optimal subset selection from text databases. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp 305–308.