Faculty of Computer and Information

Cairo University

Arabic TTS

Team Names:

Samir Mohamed

Mai Ahmed

Ali Abd Elrahman

Amira AbdElNaby

Dr. Hanaa Bayomi

TA. Amr Magdy

Project Documentation

3. Project Document Form

3.1 Background

3.1.1 Introduction

Language is among the mainly most important features that differentiate humans from Other living creatures and speech is the key medium of language.
With the advent of digital electronic technology, the goal of developing machines that Simulates human sounds has come closest to be achieved. It has to be said that no one has really succeeded in synthesizing a voice that is identical to a human voice. Meanwhile, speech synthesizers are now available which produce speech of a quality adequate for many applications.

Text-To-Speech Synthesizer System (TTS) is simply defined as written text transformed into speech; reading or dictating machines; the part of speech technology, which is concerned with automatically generating speech from a computer.
Speech synthesizers used in TTS have been developed over the years as the memory resources become available and cheaper; as a result, large enhancements in the quality and the intelligibility of the synthesized speech have been achieved.

A TTS synthesizer System is a computer-based system that has the ability to read any text, whether it is directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. OCR is the process that allows the transformation of a string of phonetic/syllabic and symbols into an artificial signal, (i.e. the automatic production of speech, through a Grapheme-To-Phoneme transcription of the sentences to utter). Grapheme is the letters in a words' dictionary, while, Phoneme is the smallest unit of speech that differentiates one word from another. TTS system includes mainly two parts: natural language processing and digital signal processing. The general block diagram of TTS system is shown in (figure 1)

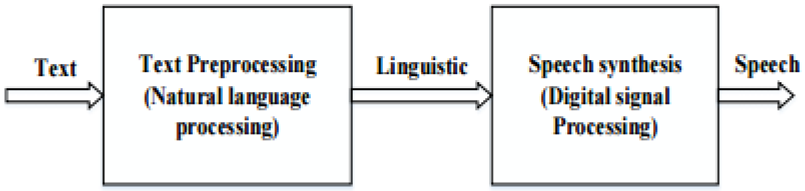

Text
Text Preprocessing (Natural language processing)
Linguistic
Speech synthesis (Digital signal Processing)
Speech

***Figure1** .General block diagram of Text to speech (TTS)*

The TTS technology is becoming inevitable in some businesses that need to provide for their customers with the latest and fundamental information in real time
Converting fundamental data stored in Web sites, databases and files into human voice using the traditional expensive and time-consuming human recordings is becoming a hard .

Arabic is a complex language and it is not like other languages, those languages written in Latin script have vowels, while the Arabic language has special characters called “diacritics” (التشكيل). These diacritics give the Arabic words the correct meaning within a sentence. For example, two Arabic words have different meaning can be written the same and only the diacritics can help the reader to distinguish them. Such as, the word “مدرسه” is for example, pronounced differently in “ سَلِمْتُ عَلَى الْمُدَرِّسَةِ"meaning “I greeted my teacher” and in the other sentence
"ذَهَبْتِ الى الْمَدْرَسَةَ"meaning “I went to school”. This depends on diacritics.

3.1.2. Motivation

The motivation behind building and developing such a system is that the TTS interface can improve the user’s experience on a desktop. It is more relaxing to listen instead of reading large portions of text. It is good for the blind, slow readers, and less straining for the eyes. Arabic TTS Synthesizer System brings benefits especially in the educational field. It assists the research, data collecting and text analyzing. It is very useful for the students, educators, and language researchers. It provides them with an effective way of knowing how to pronounce the words. The following are benefits of the Arabic TTS Synthesizer System:

1. Easy to use – intuitive: Arabic TTS Synthesizer System interface will be designed to be intuitive and easy to use.
2. Efficient: Arabic TTS Synthesizer System reduces costs and increases efficiency.
3. Arabic TTS Synthesizer System will help users to learn Arabic language
4. Provides accessibility: Over the Web and over the phone.
5. Offers adaptability and flexibility: Anytime, anywhere.

3.1.3 Beneficiary

Arabic TTS Synthesizer System designed to be used by beginners – those who have no background or do not speak Arabic – that is the TTS system is to be used mainly by non-
Arabic speaking, audiences, teachers – teachers may use this system to teach their students on the correct pronunciation but the our main target peoples in our project are peoples whose disabled or blind or impaired users and finally This system also benefits students especially in learning and improving their skills and vocabularies of Arabic language. Besides, many different types of people whose jobs require them to search for

knowledge in documents would be useful in the field of linguistics and engineering can use this project. Students who study the words occurrences in text.

3.1.4. Main techniques

There are two techniques:

1**- Supervised Machine Learning**

The majority of practical machine learning uses supervised learning. And we use it too in our project. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

2- Unsupervised Machine Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there are no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems.

3.1.5 Main applications

A Text-To-Speech Synthesizer System is a computer based system that can convert text into speech. Over the last few decades, extensive work has been done on text-to speech synthesis for the English language. Other languages such as Arabic have had limited testing mentioned earlier.

Today we have Text-to-Speech Synthesizer Systems with a very high intelligible level and an adequate level for numerous applications. These high qualities TTS Synthesizer Systems have numerous applications like the examples below:

- Applications for the Blind

By the help of an especially designed keyboard and a fast sentence assembling program, synthetic speech can be produced in a few seconds to remedy the voice handicaps. Also blind people can benefit from TTS Synthesizer Systems which gave them access to written information.

- Educational Applications

Synthesized speech can be used also in many educational situations. They provide a helpful tool to learn a new language known as computer aided learning system. It can also be used with interactive educational applications.

- Applications for Telecommunications

In these systems textual information can be accessed over the telephone.
Mostly they are used when the requirement of interactivity is little and texts range from simple messages. Queries can be given through the user's voice (Needs speech recognition) or through the telephone keyboard.

- Applications for Multimedia

Man-machine communication that can help people with their work and other
Things, for example voice activation systems in the car.

- Fundamental and Applied Research

TTS synthesizers possess a very peculiar feature, which makes them wonderful laboratory tools for linguists. These are completely under control, so that repeated experiences provide identical results (as is hardly the case with human beings). Consequently, they allow investigating the efficiency of intonative and rhythmic models. A particular type of TTS synthesizer that is based on a description of the vocal tract through its resonant frequencies (its formants) and denoted as formant synthesizer has also been extensively used by phoneticians to study speech in terms of acoustical rules.

- Government Services

Government offices receive many calls requesting information. These ranges from tax information from the Internal Revenue Service to the time and place of town meetings. Much of this information can be dispensed by use of speech synthesis over phone lines. To state a few applications, we have tax information, road closing information, lottery results, opening and closing times of public buildings, and unemployment claims processing.

3.2 Problem definition

Information and communication technology (ICT) is rapidly evolving as an effective tool for making information widespread and available online to several communities. The increased use of information technology is enabling people across the world to participate in the knowledge network; however, people in some developing countries are being deprived of the benefits of the use of ICT and the computer system. One of the main reasons for this is the lack of suitable human computer interface for disabled people and the software designed and developed to meet their needs. To design and develop a computer interface for a person, who cannot see what computer displays, is the most challenging task for many software developers.

TTS Synthesizer Systems converts the written input to spoken output by automatically generating synthetic or computer generated speech. Typed text is converted into speech using various algorithms such as formant synthesis, concatenative synthesis or Articulatory synthesis.

As the system being developed, it cannot avoid from having a problem. Primarily, there are problems that are actually faced by the people who develop the program as to make the program works efficiency and fulfills the users requirements. Discusses the major problems that will be rose during the development of the system starting from the stage of designing the system until the stage where it is being implemented and tested.
The problem area in speech synthesis is very extensive. There are quite a few problems in text pre-processing, such as numerals, abbreviations, punctuation. Moreover, the pronunciation of written text is a major problem nowadays as well. For example, when concerning the Arabic words that cannot be translated the same into other languages.

Speech synthesis has been found also more difficult with female and child voices. Female voice has a pitch almost twice as high as with male voice and with children it may be even three times as high

Text-to-Phonetic Conversion

The first task faced by any TTS system is the conversion of input text into linguistic representation, usually called text-to-phonetic or grapheme-to-phoneme conversion. The difficulty of conversion is highly language includes many problems. In some languages, the conversion is quite simple because written text almost corresponds to its pronunciation. For Arabic and most of other languages the conversion is much complicated. A very large set of different rules and their exceptions is needed to produce correct pronunciation for synthesized speech

Natural language processing contains three steps. They are text analysis, phonetic analysis and prosodic analysis. The text analysis includes segmentation (the input sentence is segmented into token), text normalization, and part of speech (POS) tagger. Phonetic conversion is to assign phonetic transcription to each word and it is a Dictionary based approach. There are two approaches in phonetic conversion. They are rule based and dictionary based approaches. Rule based is applied for unknown words whereas dictionary based is used for known words. Prosodic analysis is to determine intonation, amplitude and duration modeling of speech. It describes speaker‘s emotion.

Text preprocessing is usually a very complex task and includes several language dependent problems. Digits and numerals must be expanded into full words. For example in Arabic, numeral 243 would be expanded as meaning two-hundreds and forty three. Fractions and dates are also problematic. 2/3 would be expanded as meaning (ثلثان)or (الثانى من مارس)in case if it is a date . Abbreviations may be expanded into full words, pronounced as written or pronounced letter-by-letter. There are also some contextual problems. For example, can be pronounced either as, كجمmeaning kilogram (كيلوجرام)or as meaning kilograms depending on preceding number; yet another example . دas, Dr. (دكتور)meaning Doctor and as, الخmeaning etcetera(الى اخره).

Special characters and symbols, such as #, %, &, *, (,), -, /, <, >, [,]) are generally spoken as at symbol, cause also special kind of problems. In some situations, the word order must be changed. For example, $71.55 must be expanded as meaning "واحد وسبعون دولاراً جنيه وخمسون سنتاً"

The second task faced by any text-to-speech synthesizer system is to find correct pronunciation for different contexts in the text.
Some words, called homographs, cause maybe the most difficult problems in TTS systems. Homographs are spelled the same way but they differ in meaning and usually in pronunciation .The word ذهب" "is for example pronounced differently in sentences "ذَهَّبَ الطَّالِبُ الى الْمَدْرَسَةَ. " meaning "the boy went to the school" and
"اِشْتَرَتْ أُمى ذَهْبٌ." meaning "my mother bought gold".

The pronunciation of a certain word may also be different due to contextual effects. Some sounds may also be either voiced or unvoiced in different context. For example, phoneme /س/ in word " الصراطin " اهْدِنَا الصِّرَاطَ الْمُسْتَقِيمَ " meaning "path" the character "ص" is voiced as "س", but unvoiced in word " المستقيم" meaning "straight".

3.3 Project specification

Functional and nonfunctional For Tekrar code:

Functional:

1. The system should remove Tashkeel from words.
2. The system should remove samples from words.
3. The system should read file and splits the file into words.
4. The system should connect to Mysql Database.
5. The system should select all elements from the table of the character we use.
6. The system should identify the word is new if the word does not be in the selected data from database.
7. The system should identify the word is updated if the word is in the selected data and update its counter.
8. The system should identify the word is old if it's not modified.
9. The system should insert the data back to the database based on tags.
10. The system should work on (n var) window size left and right.
11. The system should make key of the hash map word1&word2
12. The system should add the key of the hash map to database if it's not in the selected data.
13. The system should update the counter if it's exists in the selected data.

14. The system should run on each character to insert its data in the associated table.

Non Functional:

1. The data should be Average 600,000 words.
2. Selection of data from database should be less than 10 seconds.
3. Execution of the whole program should be less than 24 hours.
4. Database should be Mysql.
5. This program should be in java.
6. Program will run properly if word is in the database or not.

Functional and nonfunctional for modowana:

Functional:

1. The system should remove Tashkeel from words.
2. The system should remove samples from words.
3. The system should read file and splits the file into words.
4. The system should connect to Mysql Database.
5. The system should select all elements from the table of the character we use.
6. The system should identify the word is new if the word does not be in the selected data from database.
7. The system should identify the word is updated if the word is in the selected data and update its counter.
8. The system should identify the word is old if it’s not modified.
9. The system should insert the data back to the database based on tags.
10. The system should add a word in primary table if it’s new and it’s (مشكلة) version in secondary.

Non Functional:

1. Selection of data from database should be less than 10 seconds.
2. Execution of the whole program should be less than 24 hours.
3. Database should be Mysql.
4. This program should be in java.
5. Program will run properly if word is in the database or not.

Functional:

1. The system should implement Naive Bayes Algorithm.
2. The system should connect to 2 databases (Modawana & gpFrequency)
3. The system should get the id for specific word in primary table.
4. The system should get all words associated to a certain id from secondary table.
5. The system should get the number of times 2 words repeated.
6. The system should remove samples from words.
7. The system should remove tashkeel from words.

3.4 Related works

There is some of the commercial TTS Synthesis Systems available today.
More than 28 TTS Synthesizer Systems currently existing in the market.

ARABTALK

The ARABTALK TTS Synthesis System was developed at Research and Development International (RDI), for Arabic language. ARABTALK is a state-of-the-art corpus based Concatenative TTS System. The system uses Artificial Neural Networks (ANN) Statistical prosody based models.
In addition, it has a real time synthesis by selection algorithm to explore large speech Corpus. ARABTALK has a Hidden Markov models (HMMs) based procedure to automatically time-align new voices transcriptions to their acoustic phoneme boundaries.
The RDI Product is based on morphological analysis (التحليل الصرفي)

The speed of automatic diacritics for Arabic text reaches to 100 words/sec and memory of 64 MB and runs on Microsoft's Windows system.

The Automated diacritics accuracy is more than 95% measured at the level of words.

http://www.rdi-eg.com/

Cimos French Company

This company produced system for automated formation of the Arabic text, have been buying this program and testing them. The correct formation ratio is equal to almost 70% at the level of the word and protects it, as it does not allow using it in more than one device.

This program add the diacritics to Arabic text, and it is on three copies

First one works on computer device, second one cooperate with another system through API and third one connected to the server to work on a world network.

http://www.cimos.com/

Aramedia company (Diacritizer)

It has developed the formation of a system where it seems the fastest and most accurate (according to the producing company) since the formation at high speed and accuracy up to 98% and gives the option of forming end of words (اعراب نهاية الكلام او عدمه) or not. This system is part of the office software tools "صخر" Office Tools
http://www.sakhr.com

https://www.aramedia.com/diacritizer.htm

Sakhr TTS

Sakhr TTS engine converts any Arabic/English text into a human voice. Sakhr has been focusing in the last 5 years on creating an Arabic TTS engine that can match in its quality the human voice.

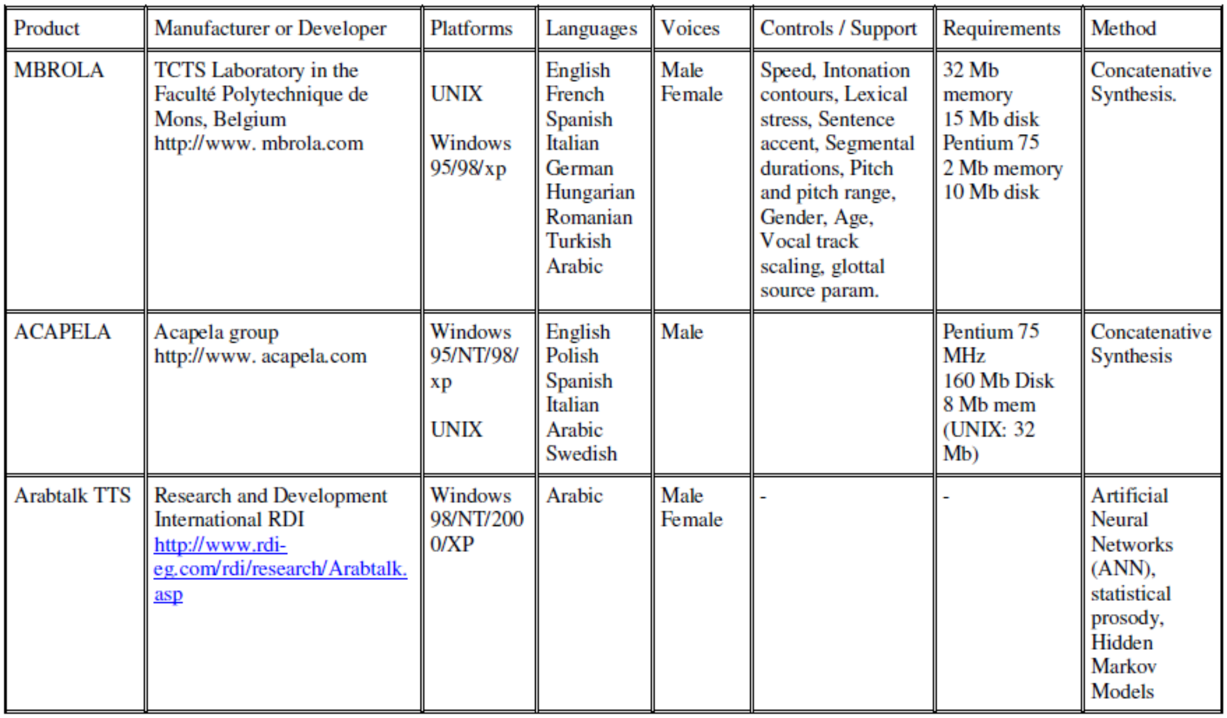
Sakhr developed the Diacritizer engine .This engine can put the diacritics needed in Arabic texts automatically. The Diacritizer is the main component in Arabic TTS. Without the Diacritizer, the output quality of the TTS engine would be inaccurate and not clear. Since Arabic native speakers write Arabic text without diacritics, the TTS engine should handle the non-diacritized text. The Diacritizer will convert the non-diacritized text into a diacritized text and then the TTS engine will convert it to a clear and human Arabic voice. Moreover, Text-To-Speech technology Software Development Kit (SDK) converts any computer readable text into a human sounding synthetic speech. Arabic is at least one order of magnitude difficult than other common languages due to the lack of diacritics.

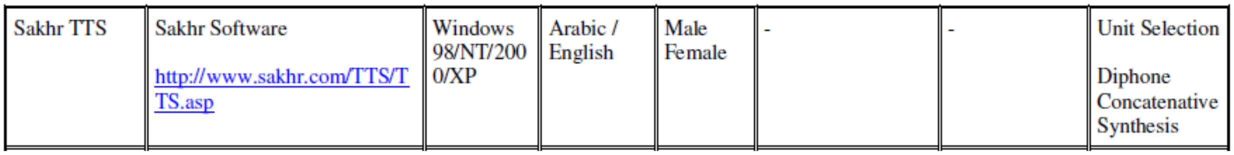
MBROLA – PROJECT

MBROLA-project is one of the main systems that have an Arabic voice.
The main goal of the project is to have a speech synthesis for as many languages as possible. MBROLA is used for non-commercial purposes. Another purpose with it is to increase the academic research, especially in prosody generation(علوم نحوية).
The MBROLA speech synthesizer is based on diphone concatenation.
The diphone databases are currently available for English, Arabic Brazilian Portuguese, Dutch, French, German, Romanian, Spanish, Greek, Turkish,.etc.
Some of these languages exist with male and female voice (MBROLA).

ACAPELA – GROUP

Acapela group constitutes all speech technologies that have been developed over the last
20 years. Speech synthesis and speech recognition have been created and improved by Acapela. Acapela Group evolves from the strategic combination of three major European companies in vocal technologies: "Babel Technologies" created in Mons, Belgium, "Infovox" created in Stockholm, Sweden and "Elan Speech" created in Toulouse, France.
Acapela owns currently three technologies, TTS by diphone, TTS by Unit Selection and Automatic Speech Recognition. Acapela is currently available for US English, UK English, Arabic, Belgian Dutch, Dutch, French, German, Italian, Polish, Spanish and Swedish.





Product	Manufacturer or Developer	Platforms	Languages	Voices	Controls / Support	Requirements	Method
MBROLA	TCTS Laboratory in the Faculté Polytechnique de Mons, Belgium http://www. mbrola.com	UNIX Windows 95/98/xp	English French Spanish Italian German Hungarian Romanian Turkish Arabic	Male Female	Speed, Intonation contours, Lexical stress, Sentence accent, Segmental durations, Pitch and pitch range, Gender, Age, Vocal track scaling, glottal source param.	32 Mb memory 15 Mb disk Pentium 75 2 Mb memory 10 Mb disk	Concatenative Synthesis.
ACAPELA	Acapela group http://www. acapela.com	Windows 95/NT/98/ xp UNIX	English Polish Spanish Italian Arabic Swedish	Male		Pentium 75 MHz 160 Mb Disk 8 Mb mem (UNIX: 32 Mb)	Concatenative Synthesis
Arabtalk TTS	Research and Development International RDI http://www.rdi-eg.com/rdi/research/Arabtalk.asp	Windows 98/NT/2000/XP	Arabic	Male Female	-	-	Artificial Neural Networks (ANN), statistical prosody, Hidden Markov Models
Sakhr TTS	Sakhr Software http://www.sakhr.com/TTS/TTS.asp	Windows 98/NT/2000/XP	Arabic / English	Male Female	-	-	Unit Selection Diphone Concatenative Synthesis

Table 4.1Summary of Text-To-Speech Existing Products

3.4 Suggested solution

Study the diacritization system (نظام التشكيل) in Arabic text and hence build a system that would be able to diacritize Arabic text automatically. Such a system can be integrated into other systems such as text-to-speech and speech-to-text systems.

4- REFERENCES AND RESOURCES

[1] Yasser H., Shady Q., Salah H., & Mohsen R. (2000). ARABTALK® An Implementation for Arabic Text To Speech System. www.nemlar.org/ARAB-TALK-RDI.doc.

[2] MBROLA. The MBROLA project towards a freely available multilingual speech synthesizer. http://tcts.fpms.ac.be/synthesis/mbrola.html.

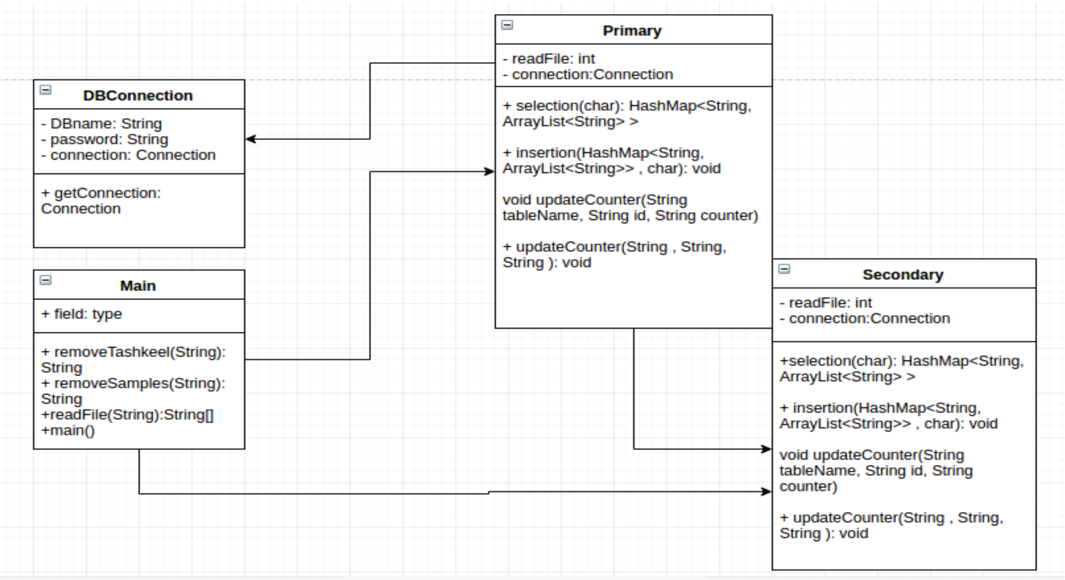
[3] Wael H. & Mohsen R. (2000). Concatenative Arabic speech synthesis using large database, In Proceedings of ICSLP2000, vol. 2, pp. 182-185, Beijing, China. http://repository.um.edu.my/142/1/Arabic%20TTS%20Synthesizer.pdf

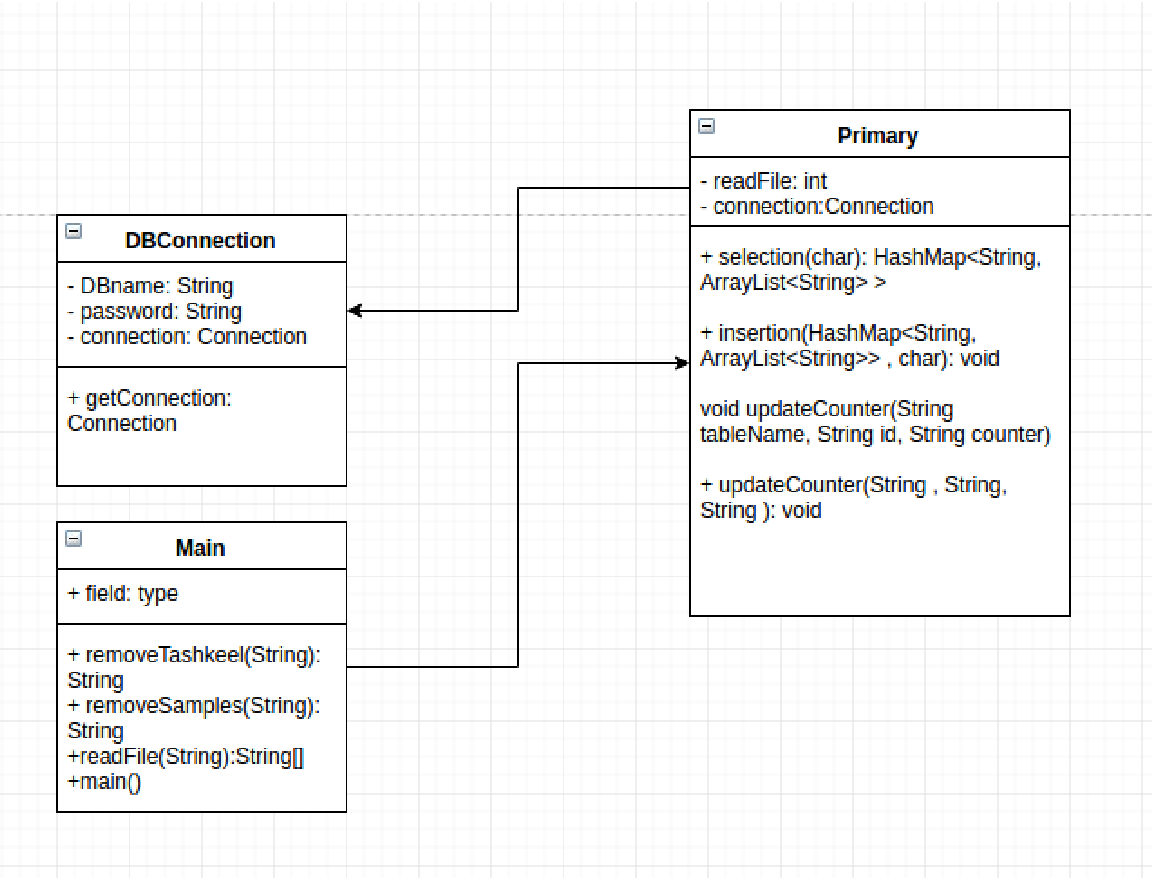
[4] http://www.rdi-eg.com/Downloads/Scientific%20Papers/PaperOnArabTalk.pdf

[5] http://link.springer.com/article/10.1007/s10772-015-9304-6

[6]https://www.researchgate.net/publication/221429664_ARAB_TTS_An_Arabic_Text_To_Speech_Synthesis

Class diagram:


Primary
- readFile: int
- connection:Connection
+ selection(char): HashMap<String, ArrayList<String> >
+ insertion(HashMap<String, ArrayList<String>> , char): void
void updateCounter(String tableName, String id, String counter)
+ updateCounter(String , String, String): void
DBConnection
- DBname: String
- password: String
- connection: Connection
+ getConnection: Connection
Main
+ field: type
+ removeTashkeel(String): String
+ removeSamples(String): String
+readFile(String):String[]
+main()
Secondary
- readFile: int
- connection:Connection
+selection(char): HashMap<String, ArrayList<String> >
+ insertion(HashMap<String, ArrayList<String>> , char): void
void updateCounter(String tableName, String id, String counter)
+ updateCounter(String , String, String): void


Primary
- readFile: int
- connection:Connection
+ selection(char): HashMap<String, ArrayList<String> >
+ insertion(HashMap<String, ArrayList<String>> , char): void
void updateCounter(String tableName, String id, String counter)
+ updateCounter(String , String, String): void
DBConnection
- DBname: String
- password: String
- connection: Connection
+ getConnection: Connection
Main
+ field: type
+ removeTashkeel(String): String
+ removeSamples(String): String
+readFile(String):String[]
+main()