Al-Quds University

Deanship of Graduate Studies

Master of Electronics and Computer Engineering

Thesis Approval

Arabic Text-To-Speech Including Prosody (ATTSIP) for Mobile Devices

Prepared By: Ahmed Ismail Mohammed Elothmany

Registration No: 20913765

Supervisor: Dr. Labib Arafeh

Master thesis submitted and accepted, Date: 11/02/2013

The name and signatures of examining committee members are as follows:

1-	Head of committee	Dr. Labib Arafeh	Signature: .....
2-	Internal Examiner	Dr. Hanna Abdel Nour	Signature: .....
3-	External Examiner	Dr. Radwan Tahboob	Signature: .....

Jerusalem-Palestine

1434-2013

Abstract

This master's thesis describes the construction of an Arabic Grapheme-To-Phoneme system for mobile devices based on the Android operating system. The input to the system is a full Diacritized Arabic Sentences and the output is the Phonetic text for these sentences. The system is enhanced by including prosody to increase the naturalness of the generated voice. Grapheme-To-Phoneme is a substantial process of the Natural Language Processing module of any synthesis, the proposed approach exploits a set of Arabic Letter-To-Sound rules used by this process. These rules include tanween, el Shaddeh, Atta almarbouta, definite article"ال" , the vowel/consonant relationship, and other rules that are used to get the pronunciation of a sentence from its graphemes. A novel algorithm has been developed for the syllabification process which used the Sokoun mark. Tokenization is done at the level of the word before syllabification. So we get a group of syllables and the algorithm searches for pattern syllables within these syllables. Many words have been examined to verify the performance of our algorthim, we found that our algorithm reduces the number of iterations needed to get the syllables for a word against what is used for other algorithm; this means less energy consumption and thus higher efficiency.

In our thesis we have implemented the Grapheme-To-Phoneme in two methods namely: One-By-One call and the Selective call methods. In the first method, all the rules are called serially one by one and the output of one rule is the input for the next rule, in the second only a subset of the required rules are called while other rules are not. The second method saves a lot of processes thus power and this is what any developer for limited resources devices is seeking for. The second method proved to be more efficient by decreasing the execution time for the conversion process by 18.5% compared to the time needed to perform the conversion process for the first method. A training set of eight sentences have been used to train the system for the rules ; a subjective Correctness Test was performed on a Test set of Thirty five sentences to verify the letter-to-sound rules, the obtained result showed a satisfactory and promising result and most of the sentences were converted successfully. This indicates the promising potential of these algorithms. However, more and more sentences must be tested before declaring the correctness of these techniques.

Contents

1 Introduction

With the rapid growth of the economy and market opportunities in the Arab world, spoken language technologies for Arabic have become an area of interest in recent years. Speech synthesis is one of the major areas of Speech Technology. Speech synthesis or Text-To-Speech (TTS) system is the automatic generation of speech from text. In the last few years, several approaches have been developed for TTS systems for English, but for Arabic, TTS system is still in its early development stage.

In this thesis we have introduced an Arabic Text-To-Speech system for mobile devices that is based on Android as its operating system. The application of text-to-speech for mobile devices is getting increasingly interesting because the speech allows users to communicate with minimum effort. For example, people who do not speak the same language can communicate with each other using two systems; Speech-To-Text System is needed to translate the speech into text for the first language then the translation to the second language is done, after that, Text-To-Speech system is used to generate the speech from the translated content. Many techniques are used to produce the TTS system such as Articulatory synthesis, Formant synthesis and Concatenative synthesis [Assaf, 2005]. Less memory and less computations is needed for concatenative synthesis so it is suitable for mobile applications.

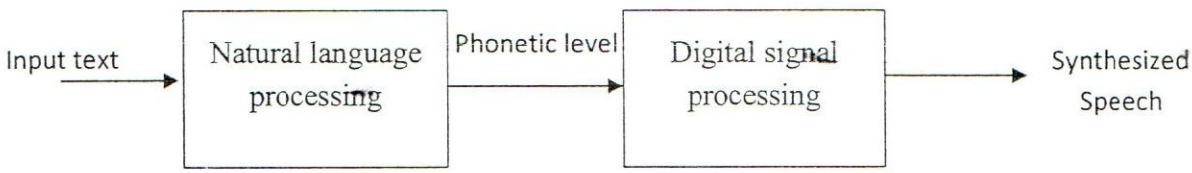

Figure 1.1 A general TTS system [Dutoit, 2008]

A general TTS system is composed of two main modules (figure 1.1), the natural language processing and the digital signal processing. The natural language processing deals with the conversion of the sentences from the written form into the pronunciation form. The natural language processing is composed of Pre-processing, Grapheme-To-phoneme and prosody generation stages. In the pre-processing stage, analysis to the sentence is done to prepare it to the next stages such as replacing abbreviations, translating numbers from digits to counterpart alphabet form, etc. In the Grapheme-To-phoneme stage, a set of Letter-To-Sound rules and an exceptional words database is needed to translate the text from its grapheme into its phoneme i.e. from the written form into the pronunciation form. The rules set are language dependent and need expertise in the language to specify these rules. The exceptional words database is needed for the sentences that contain words that the rules do not apply on to get its phonemes. The database stores every exceptional grapheme or word and its counterpart phonemes; accessing to the database is done before applying the rules set. Prosody generation stage enhances the TTS system by increasing the naturalness of the generated voice because prosody indicates the mood of the speaker and the meaning of the sentence.

Android is an open source mobile operating system that works on tablet, PDA, or other digital mobile devices. We have chosen the Android for our application because it is now the most dominant mobile OS related to bestselling and developing [Lee, 2011].

1.1 Objective of the thesis

In this thesis we will introduce a Text-To-Speech application for Arabic language that includes the automatic generation of stress to make the generated voice as natural as possible. The application is developed to run on mobile devices that are based on Android operating system. The system will have a user interface which enables the user to input a Diacritized Arabic Sentences then synthesized speech will be generated for that sentence. The Arabic Text-To-Speech Including Prosody system (ATTSIP) introduced in this thesis will be divided in two parts namely: the natural language processing which will be discussed and implemented in this thesis and the Digital signal processing which will be implemented by the master degree student Miss Nuha Ouda. So in this thesis we will discuss the stages of the natural language processing specially the Grapheme-To-Phoneme stage and how it can convert the input text from the written form to the pronounce form.

1.2 Contributions

The proposed Arabic Text-To-Speech including prosody, ATTSIP, system will include the most important prosodic feature which is the stress level to increase the intelligibility and naturalness of the generated voice; stress level will be generated automatically for the input text. At our knowledge, until now no Arabic Text-To-speech application for mobile devices based on Android OS has included the automatic generation for stress level from input Arabic text.

We have developed an algorithm for the syllabification of the sentence that is making use of the Soukon mark (◌̊) instead of neglecting it. [El-Imam, 2003] presented algorithm for syllabification that did not make use of the Sokoun because he considered that it did not have any pronunciation. The numbers of iterations needed for syllabification process for a sentence by our algorithm has been reduced by half against the algorithm which have been developed by El-Imam. We will show that our developed algorithm for the syllabification process gives more performance than El-Imam algorithm.

A database for exceptional words has been included for the Arabic language for mobile devices. This database is small in size but contains the most frequently used exceptional words in Arabic language speech.

1.3 Thesis Organization

The organization of the thesis is as follow: Chapter two introduces the related works that have been done so far by other researchers interested in text-to-speech synthesis.

Chapter three gives a short introduction to speech synthesis and its techniques and applications. The two components of Text-to-Speech synthesis, the Natural Language Processing (NLP) and the Digital Signal Processing, are discussed. Then we discuss Android and its version and structure. The chapter ends with the introduction of the Arabic language, its alphabet and syllables structures.

In chapter four we have introduced the proposed Arabic Text-To-Speech Including Prosody (ATTSIP) System and structures. In this chapter we have presented the letter-to-sound rules and the syllabification algorithm used by our system.

In chapter five, we explore the training set sentences used in the training process for the system, after that a test for the ATTSIP synthesis is performed in order to examine the correctness of the letter-to-sound rules. Chapters six and seven will be dedicated for the discussion and conclusion of the ATTSIP system and future direction and enhancement that can be done to increase its performance.

2 Literature Survey

Many systems have been developed for the generation of voice from text. [ITunes, 2012] has been designed as an application for both iPhone , iPad and iPod. Figure 2.1 shows the graphical user interface for the application.

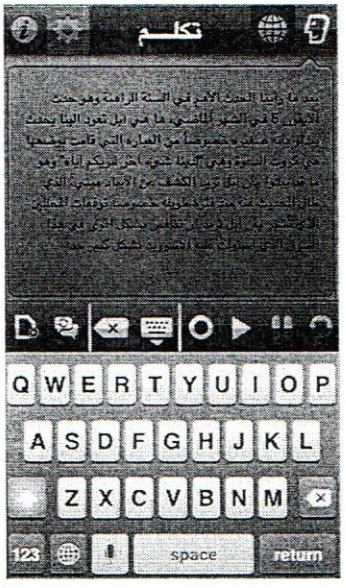

Figure 2.1 Graphical user interface for Takalam application [ITunes, 2012]

The application enables the users to enter a standard Arabic sentence and listen to it aloud. The application did not support any prosody features and according to the customers reviewers the generated sound is very weak [ITunes, 2012]. [ISpeech, 2012] developed an application for a Variety of voices in many different languages including Arabic language for iPhone, iPad, iPod Touch devices.

Figure 2.2 Graphical user interface for ISpeech application [ISpeech, 2012]

No customers review for this application, but according to the website the specification for their application include:

- Human quality speech.
- Variety of voices in many different languages.
- Fast text-to-speech conversions.
- For any iOS device: iPhone, iPad, iPod Touch.

iSpeech can convert text to speech, convert documents to speech, convert web content to speech, or convert blogs to speech.

[Sakhr, 2009] provides software for Arabic Text-to-Speech, Sakhr TTS converts Arabic text into a natural, human-sounding synthetic voice for iPhone devices. Sakhr system features include:

- Embedded TTS engine for mobile devices
- Rules-based powerful diacritizer with 97% accuracy
- Normalizes ambiguous text such as dates, time, currencies, abbreviations
- Unlimited vocabulary, text size, raw text, phonetic and prosodic input.

They did not speak about what prosodic features their system support, no mention for stressed level feature at all.

[SVOX, 2011] have been implemented a text-to-speech Malik application for Arabic language based on Android OS. Figure 2.3 shows the user interface for the Malik Application from SVOX.

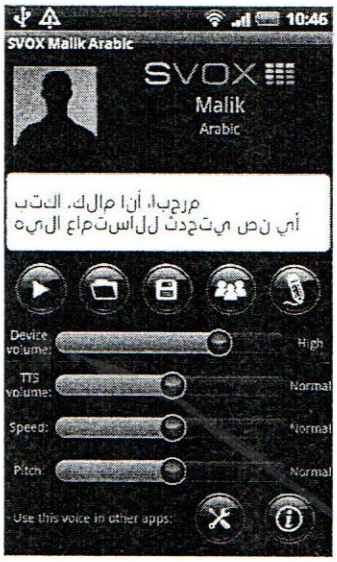

Figure 2.3 Graphical user interface for Malik application [SVOX, 2011]

When combined with TTS-enabled apps, Malik can give directions, read a notification containing the caller / sender's name, and read the content of an incoming call / sms / mms / email. Many customers' reviewers have expressed their dissatisfaction. From the user interface of Malik application we can detect that, No stress level automatic generation for prosody because the user is required to enter the values for many parameters such as pitch and speed.

Different approaches have been developed for TTS system and it uses different mechanisms. [Al-Muhtaseb, Elshafei, Al-Ghamdi2, 2000] proposed diphone/sub-syllable method for Arabic Text-to-speech systems. They define a set of Arabic diphones/sub-syllables for concatenative Arabic text-to-Speech synthesis. In this method the boundaries of the speech segments are chosen to occur only at the sustained portion of vowels. The speech segments consist of consonants, half vowels, half vowel-consonants, middle portion of vowels, and suffix consonants. They also proposed extension of the Arabic diphones/sub-syllables set to improve the quality of the speech and to incorporate the common co-articulation effects found in modern standard Arabic.

[Rashad, El-Bakry, Isma'il, 2010] have proposed Diphone Speech Synthesis System for Arabic using MARY TTS. They choose the Diphone unit because Diphones have the advantage of modeling coarticulation by including the transition to the next phone inside the diphone itself. The modular design of the MARY system has many advantages: it is easy to modify a certain module without affecting other modules; any module with a similar interface can be used instead of an employed module. To integrate two modules, the only requirement is that the output data types of the first module must match the input data types of the second module. The MARY TTS accepts two types of input text plain text and markup text (such as SABLE-annotated text or SSML-annotated text). The use of SABLE-annotated text for example as input gives the users the ability to add information to the text that improve the way it is spoken such as pausing at the right places or emphasizing on certain words. MARY TTS has been evaluated by two types of tests which are the Diagnostic Rhyme Test (DRT) that measures the intelligibility of the synthesized speech. In this test, twenty pairs of words that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which word of each pair of the words they think is correct. The listener identifies 89.3% of the test words in the 1st listening session and this percent increases slightly and becomes 90.9% in the 2nd listening session. The other evaluation test was the Categorical Estimation (CE) test that measures the overall quality of the synthesized speech, in this test the listeners were asked a few questions about several attributes such as the speed, the pronunciation, the stress of the speech and they were asked to rank the voice quality using a five level scale. The second test shows that the results of the CE need more improvements.

[Zeki, KhAlifa, Naji, 2010] proposed a rule based text- to- speech Hybrid synthesis system which is a combination of formant and concatenation techniques. Their system has the flexibility of changing the speaker from male to female and other sound variants like whispering and has the ability to enrich the exception dictionary by listing the exact pronunciation of the common words. Their system features include small size, high accuracy, and vocabulary independence.

Many grapheme-to-phoneme algorithms have been developed. A possible model topology for grapheme-to-phoneme conversion is Hidden Markov Models [Taylor, 2005]. The model is designed to be based on the use of hidden states representing phonemes which generate graphemes as observations. The model topology is different from that used in automatic speech recognition in that there are no looping states, but instead all the states have an arc to the final,

non-emitting state. The observation probabilities are represented as a discrete quantity, since the observations are letters. Taylor identifies a number of advantages to the use of HMMs as a model for grapheme to phoneme conversion. While data driven techniques rely on a separate model to align graphemes to phonemes, for HMM-based models, alignment and decoding can be performed on the same model. HMMs are flexible in that a topology can be chosen to optimize the model for the problem domain, and the powerful Baum-Welch training algorithm can be used to optimize the model parameters.

Other techniques have been developed to achieve the conversion process, [Damper, Marchand , Adamson, Gustafson,1999] have proposed driven technique called Pronunciation by Analogy Grapheme-to-Phoneme for English language . In this method, a directed graph is built from substrings in the input word and their partial pronunciations. A letter to phoneme alignment is created automatically from the lexical database. The input word is then scanned for substrings that match the alignment database, and a pronunciation lattice is created. The decision function finds a possible pronunciation for the word by traversing the lattice and concatenating the phoneme labels on the arcs. The path chosen is shortest, and in the case of a tie, a heuristic is used. The result scored 71.8% words correct conversion. A statistical method have been developed by [Chen, 2003], this method uses a model which is similar to that used in machine translation. The problem is framed as follows: given a letter sequence G, find the phoneme sequence S that maximizes P(S | G). He claims that word error rates are significantly lower using this model compared with any of the rule-based approaches.

[El-Imam,2003] has been presented transcription of written text into sounds based on a set of letter-to-sound rules for Arabic language. He divides the rules into types namely: phonemic and phonetic rules. Phonemic rules operate on the graphemes to convert them into phonemes. Phonetic rules operate onto the phonemes and convert them into phones or actual sounds. He also presented an algorithm for syllabification process, every phonemic word is scanned in a backward manner (right-to-left scan starting from the end of the phonemic word) looking for vowels. Whenever a vowel is detected, another more localized scan is performed looking for a syllable match for any of the six syllable types. He deletes the Sokoun mark before applying the letter-to-sound rules because this mark has no pronunciation. His letter-to-sound transcription system was tested using a list of most frequently used Arabic words and proper names. His letter-to-sound system reported few transcription errors on both words and phonemes/phones. The overall score of the system is over 98% phonemes correct while the percentage of correctly pronounced words is around 92% correct words.

Recent work have been done for Arabic stressed syllable, [Chentire,Guerti and Hirst,2009] proposed discriminant analysis classifier for syllables detection for standard Arabic words that have the structure (CVCVCV). They use the fundamental frequency and energy parameters for the classifier. After segmenting and transcribing manually the corpus, they applied their algorithm based on discriminant analysis. The result shows that 89.81% in the learning phase and 83.33% in the recognition phase scored. And a percentage of detection equal to 85 % of the primary stressed syllable was obtained.

[Hamiti, Dika, 2010] develop algorithm for Albanian language for speech generation based on written text through using the most frequent words and syllables. They create a list for the words that are used often. For the words that do not exist in the list they used a list of syllable used to

get these words. During the conversion, for words that cannot convert as complete words, or by using syllables from the list of those most frequent, they extend the conversion to be letter-by-letter. [Zemirli , Mosteghanem ,2007] proposed an algorithmic approach for the automatic generation of the stressing in Arabic language. They adapted diagrams, generated for the text processing that act on the size of the sentences to reading with intonative contours of natural speech. The performance of syllabification information is improved in [Chen, 2003] by using n-gram model. He inserts syllable boundaries using a model which is trained on supervised data and dynamic programming is used for placement of syllables.[Braham, Belghith, 2001] specify Letter-To-sound rules in a neural network based Diphone system. Phonemes are represented by a feature vectors. These vectors are composition of their constituent phonemes.

[A1-Dakkak, Ghneim, Abou Zliekha, Al-Moubayed, 2006] have proposed Arabic text-to-speech system based on Orthographic-Phonetic Transliteration language. They generate speech using the MBROLA diphones, because MBROLA permits the control of some prosodic features (fundamental frequency F0, duration) so they construct their prosodic models and test it. For prosody they focus their sentences for the four types (informative, interrogative, exclamation, ban) of sentences. These sentences were analyzed and we have the curves for the fundamental frequency F0 and the intensity versus time, after analyzing the sentences they notice that F0 for all types of the sentences is increasing in the beginning, decreasing at the end, except for the interrogative sentences. An automated tool has been developed for emotional Arabic synthesis. It is based on an automatic prosody rough generation model, which performs intonation modification taking into account some punctuation marks (',', '.', '!', and '?') and the number of phonemes in the sentence.

[Jong, Zawaydeh, 1999] presents the results of a study of the expression of word-level prosody in Jordanian Arabic. They examine the phonetic correlates of Arabic stress, an apparent prominence marker at the level of the word, and also examine similar effects of word final juncture. Their study investigates the extent to which word-level prosodic effects are sensitive to higher levels of prosodic structure. The study focuses on the durational, spectral, and fundamental frequency correlates of stress and word final juncture in the speech of four speakers. Stress lengthening correlates with higher first formants, while penultimate lengthening does not. Analyses of fundamental frequency patterns support an analysis in terms of pitch accents associated with stressed syllables and juncture-marking phrasal pitch specifications. They found extensive word final lengthening effects, on the order of 60% for word (and phrase) final, to 100% for utterance final syllables.

تحويل النص العربي إلى مقطع صوتي متضمناً إيقاع الجملة خاص بالأجهزة النقالة

إعداد: أحمد اسماعيل محمد العثماني

إشراف: د. لبيب عرفة

الملخص

تصف اطروحة الماجستير بناء نظام يهدف لتحويل نص مكتوب باللغة العربية الى نص يحدد طريقة لفظ هذه الجملة خاص بالاجهزة النقالة التى تستند على نظام التشغيل اندرويد. المدخل الى هذا النظام هو جملة عربية معربة تعريباً كاملاً والمخرج منه الفونات التى تمثل طريقة لفظ الجملة. تم تحسين اداء هذا النظام عن طريق تضمين انماط اللحن وتقطيع المقطع الصوتى المولد ليصبح اكثر قرباً من طبيعة الصوت البشرى . عملية تحويل الكتابة من نص مكتوب الى نص يحدد طريقة لفظ الجملة عملية هامة و اساسية لوحدة المعالجة اللغوية للنص و لذلك نستعرض هنا بعض قواعد النطق التى تستخدم فى هذه العملية. تشمل هذه القواعد التنوين و الشدة و التاء المربوطة و(الاف و اللام ال) و المصمتة و اللينة والقواعد الاخرى التى تستخدم فى عملية الحصول على النطق الصحيح للنص المكتوب. تم تطوير خوارزميه جديدة للحصول على المقاطع اللفظية للجملة من النص المكتوب و ذلك بإستخدام علامة السكون. اولا عملية الحصول على المقاطع تبدأ بتقطيع الكلمة المكتوبة على اساس علامة السكون ثم نبحث داخل هذه المقاطع النصية على انماط معينة للحصول على جميع المقاطع التى تشكل المقطع اللفظى للكلمة و لذلك نحن هنا نبحث على الانماط داخل المقاطع و ليس داخل الكلمة نفسها وقد تم فحص كفاءة الخوارزميه على العديد من الكلمات ووجدنا انه يحتاج الى عمليات اقل من العمليات المطلوبة لتقطيع نفس الكلمة بواسطة استخدام خوارزميات اخرى و هذا يؤدى الى تقليل استهلاك الطاقة وبالتالى رفع كفاءة النظام. فى عملية تحويل النص المكتوب الى نص يحدد طريقة اللفظ استخدمنا طريقتين لانجاز هذه العملية، فى الطريقة الاولى كل قواعد النطق تستدعى بطريقة متتالية واحدة بعد اخرى لتفيذ العملية و الطريقة الاخرى تعتمد على استدعاء فقط القواعد المطوبة لتفيذ العملية وهناك قواعد لا تستدعى على الاطلاق حيث ان الجملة المكتوبة لا تحتوى على مثل تلك القواعد. الطريقة الثانية اثبتت كفائتها عن الطريقة الاولى بنسبة 18.5% بالنسبة للزمن المطلوب لتنفيذ عملية التحويل و هى الطريقة المفضلة للاجهزة المحمولة حيث انها تحتاج الى معالجة اقل و بالتالى كمية اقل من الطاقة و هذا ما يطمح اليه اى مطور. تم تجريب النظام فى مراحل تطويره على ثمانى جمل تشمل كل قواعد النطق المتضمنة فى هذه الاطروحة ، بعد الانتهاء من تطوير النظام تم اختباره على خمسة و ثلاثون جملة للتحقق من التطبيق الصحيح لهذه القواعد و أظهرت النتائج نجاح كبير لعملية تحويل الجملة من نص مكتوب الى نص مسموع وهذا يدل على إمكانات واعدة من هذه الخوارزميات و مع ذلك، لا بد من اختبار الجمل أكثر وأكثر قبل أن يعلن عن صحة هذه التقنيات.

7 Conclusion and Future works

The purpose of this thesis is to explore the full steps for designing an Arabic text-to-speech system for mobile devices based on Android operating system. The overall architecture and general features of ATTSIP Text-To-Speech system for Arabic language has been presented. We have discussed the natural language process of the ATTSIP system and its modules; the Pre-processing module has been presented to handle abbreviations and detecting the sentence type. We have presented a small database which stored exceptional words used by the Text Analysis module to handle exceptional Arabic words. Many methods for developing the Grapheme-to-Phoneme module have been explored, Rule-based method has been chosen because it is suitable for mobile application, it needs a small database for exception words and a letter-to-sound rules for the conversion process. To implement this method, letter-to-sound rules have been presented to get the pronunciation of the sentence. The phonemic output of Grapheme-to-Phoneme module is a function of the precision of the letter-to-sound rules and the completeness of the exceptional dictionary. A syllabification algorithm have been developed to extract the syllables for the words which form the sentence, these syllables are used by another developed stress level algorithm to put the prosody information between the syllables. The syllabification algorithm search for a pattern matching within each word of the sentence for the specified Arabic language syllables patterns (cv,cvv,cvc,cvvc,cvcc and the rare cvvcc) . Every Arabic syllable starts at Sokoun mark, we have used this Arabic linguistic rule in our algorithm to speed up the algorithm by reducing the number of iterations for extracting the syllables comparable with what is needed for the El-Imam work. El-Imam removes the Sokoun mark before syllabification because he consider it does not have any pronunciation and start scan left to right or right to left within the word looking for pattern matching. In our algorithm we have been kept it in the words , we have been started the syllabification process by tokenizing the word based on the Sokoun mark, after we get the initial syllables produced from tokenization process, we search for a pattern matching within these syllables instead of the word itself. We need syllabication for putting prosody data between these syllables. The generated voice without prosody sounds robotic. A stress level prosody feature algorithm have been developed according to the rules have been discussed in section (3.1.3) to enhance the generated voice to give it the human being nature.

It has been described why we choose the Android for our mobile application, Android nowadays the most dominant and effective OS for mobile devices. We train the system on eight sentences to ensures that all the letter-to-sound rules have been covered, after training the system we used thirty five sentences in testing the correctness of the letter-to-sound rules. We think the result was satisfied but more letter-to-sound rules is needed to cover the special cases as we have seen in sentence(11) of the Testing set , but more rules means more processing and this will affect on the overall system speed performance.

The grapheme-to-phoneme module has been implemented in two ways, the one-by-one call and the selective call methods. We have show the difference between the two methods and show that the second method is better than the first by 18.5 % so it is suitable for limited resources devices and deserve more future developments.

Future development can be done to the system to increase its performance and capabilities. Reading incoming SMS or reading the directions in a map while driving the car are features that can be added to the system. Now ATTSIP only accept full diacritic plain text, but the system can be developed to allow modern standard Arabic to be processed. This needs a new module to add the diacritic to the text "Al-Tashkeel". In our thesis, we have been used a set of static rules to

implements the grapheme- to- phoneme conversion process; another approach could be designed to use a dynamic rule based system. In such a system, you will need a small library of sounds and the corresponding text; both will be used to derive the required rules dynamically to be used in the conversion process.
Continued work is required to develop the system further and to increase its performance.