Text-to-Speech Synthesis State-Of- Art

Prof. Dr. Hilal M. Yousif
AL-Rafidain University College

Dr. Mouyad A. Fadhil
Informatics Institute
for Postgraduate Studies
Technology university

Yahya M. Hadi
AL-Rafidain University College
Ph.D Student (Research Stage) At
Informatics Institute for Postgraduate Studies
Technology university

(December 2004)

Abstract

Speech synthesis is one of the major areas of Digital Signal Processing (DSP)field. Synthesis refers to generation of speech computational model based on human speech system. An important area of Speech synthesis is the Text-To-Speech (TTS). In the past two decades years, many studies have focused on Text-To-Speech (TTS) systems for different languages such as: English, Franch, Italyain, Japanise, and Chinese. In particular, Arabic (TTS)systems have made significant progresses in the last few years especially in Egypt and Algeria.

In Iraq there is a limited work in Synthesis area , Where the most DSP work that has been done in speech Recognition area.

TTS synthesis can be used in many areas, such as: Telecommunication services, language education, vocal monitoring, multimedia, and as an aid to handicapped people.

Generally speaking a TTS system can be divided into two major components: natural language processing (NLP) and digital signal processing (DSP).

This paper gives an overview of general TTS systems and the Arabic TTS with some implementation using concatenation approach.

بحث عن تطور تحويل النص المكتوب الى كلام

الأستاذ الدكتور هلال محمد يوسف
عميد كلية الرافدين الجامعة

الدكتور مؤيد احمد فاضل
معهد المعلوماتية للدراسات العليا
الجامعة التكنولوجية

يحيى مهدي هادي
كلية الرافدين الجامعة
طالب دكتوراة (مرحلة البحث)
في معهد المعلوماتية للدراسات العليا
الجامعة التكنولوجية

بغداد كانون اول 2004

المستخلص

تعتبر عملية انتاج وتوليف الكلام احد أهم الجوانب العلمية المهمة في حقل معالجة الاشارة الرقمية. يعني توليف الكلام أنتاج الصوت بواسطة نموذج الحاسب الآلي والذي يبنى اساسا على نظام انتاج الكلام عند الأنسان. من أهم الجوانب الاساسية في عملية توليد الكلام هو تحويل النص المكتوب الى أشارة صوتية . لقد شهد العقدين الاخيرين تطورا ملحوظا في هذا الجانب خصوصا في الدول المتقدمة، اما في العالم العربي فقد شهدت السنوات الاخيرة تطورا معتدلا ، خصوصا في مصر والجزائر. أما في العراق فقد اهتم الباحثين بصورة اكثر في موضوع تميز الكلام من موضوع توليف الكلام.

في عالم اليوم تحويل النص المكتوب طوعيا الى الى اشارة صوتية مفهومة بواسطة الحاسب الالي ، ذات أهمية عالية ؛ حيث يستخدم في مجالات عديدة مثل: خدمات الاتصالات، التربية والتعليم، ألصحافة والاعلام ، وفي مساعدة المعوقين.

بصورة عامة يمكن ان نقسم نظام تحويل النص الى كلام (TTS) الى جزئين أساسين : جزء معالجة اللغة الطبيعية (NLP) ، الذي يهتم في تجميع المعلومات ذات العلاقة في موضوع علم اللسانيات ، وفحص النص المكتوب وتحويله الى متغيرات ممكنة التحويل الى كلام مفهوم مستند الى قواعد اللغة المستخدمة. والجزء الآخر معالجة الاشارة الرقم ية (DSP) ووظيفتة تحويل النص المستلم من جزء ال(NLP) ألى أشارة كلام.

في هذا البحث سنعرض ما تم التوصل الية من تطور في موضوع أل(TTS) مع اجراء تطبيق لتحويل نص عربي الى أشارة صوتية بأستخدام أسلوب ترابط وتداخل ألاصوات لحروف النص المكتوب.

1. Introduction

Speech has evolved as a primary form of communication between humans. Speech can be considered as unique human signature, where speech not only varies widely from speaker to speaker, but also from time to time.[1]. Normally, there often occur conditions under which we measure and then transform the speech signal to another form in order to enhance our ability to communicate. [2]

Speech nowadays represents a big challenge of computer technology in both sides: the Hardware, and the Software.

New telecommunication services include the capability of a machine to speak with a human in a "natural way"; to this end , a lot of work must be done in order to improve the actual voice quality of text-to-speech and concept-to-speech systems.[3].

Speech synthesis technology plays an important role in many aspects of man-machine interaction, especially in telephony applications.

It should be pointed out that most commercial products that produce human sounding speech do not synthesize it, but merely play back a digitally recorded segment from a human speaker.

Computer generation and recognition of speech are faced problems, many approaches have been tried with only mild success. This is an active area of Digital Signal Processing (DSP) research, and will remain so for many years to come. [4].

This paper is intended to provide a background on text-to-speech synthesis with an emphasis on text-to-speech synthesis for Arabic TTS.

The paper begins with a brief overview of general text-to-speech systems and introduce a first practical step on Arabic TTS.

2. Human Speech Production Mechanism

All speech synthesis and recognition are base on the model of human speech production. The Human Speech organs are divided into three main groups: [2]

- The lungs, they act as power supply and provide airflow to the larynx.
- The larynx, it modulates airflow from the lungs and provides either a periodic puff-like, or a noisy airflow source to the third organ group, the vocal tract.
- The Vocal tract, it consists of: oral, nasal ,and pharynx cavities.

Figure (1), illustrate the view of the anatomy of speech production. [5].

Just as written language is a sequence of elementary alphabet, speech is a sequence of elementary acoustic symbols (known as phonemes) that convey the spoken form of language. Speech sounds are produced by air pressure vibrations generated by pushing air from the lungs through the vocal cords and vocal tract and out from the lips and nose airways.

The air is modulated and shaped by the vibrations of the glottal cords, the resonance of the vocal tract, the position of the tongue and the openings and closings of the mouth. Speech signals convey more than spoken words. The information contents in speech include: Acoustic phonetic symbols, Gender information, Age effects, Accent, speaker's identity which is determined by the vocal tract anatomy, Emotion and health, and Prosody.

We set the air coming up from the Lungs in motion using our vocal cords and the we can channel this air through the vocal tract using our tongue, lips, etc.

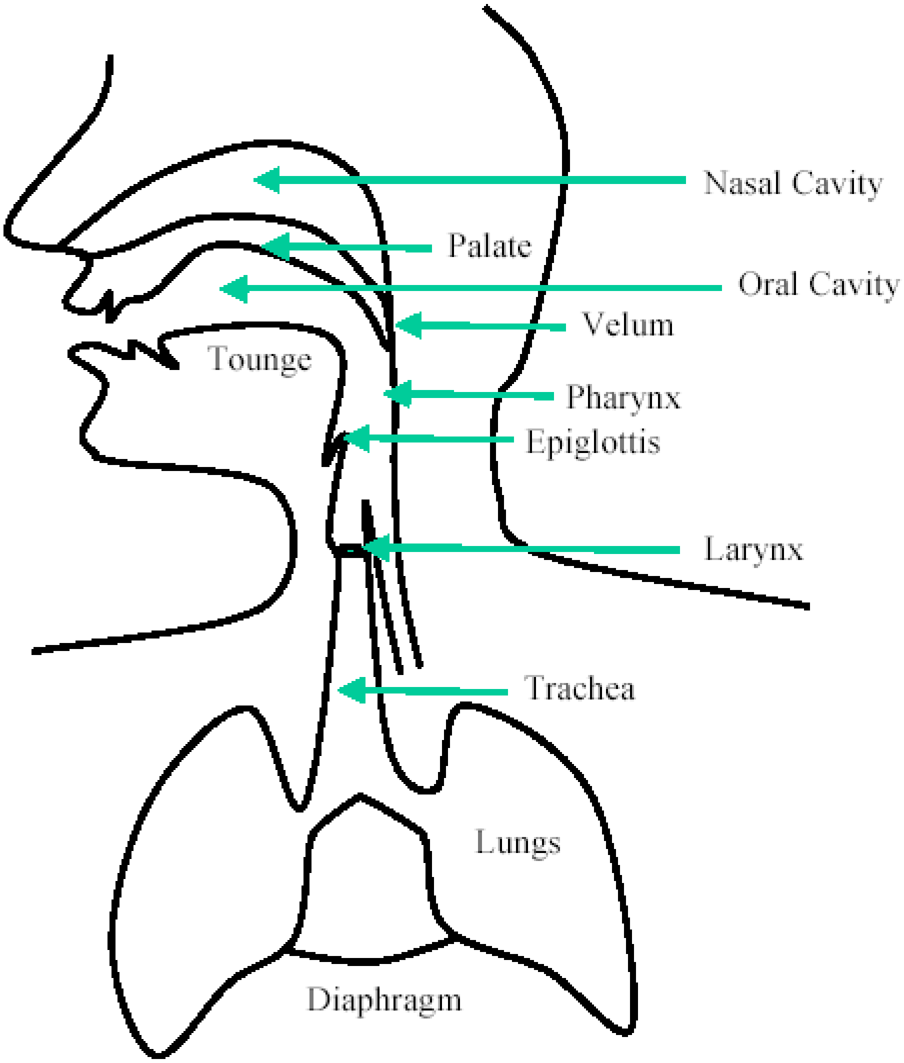


Figure (1) A view of the anatomy of speech production

There are three main subjects in Digital Signal Processing (DSP)
Area: Linguistics, Physiology, and Acoustics.

Linguistics is concern with how the language constructed, include the units of language, what are they? And the grammar of language.

Physiology relates to how the sounds are produced through neural and muscular activity. Acoustics, describes the generation and transmission of the sounds.

The relationships between these areas explained in the figure (2). [6].

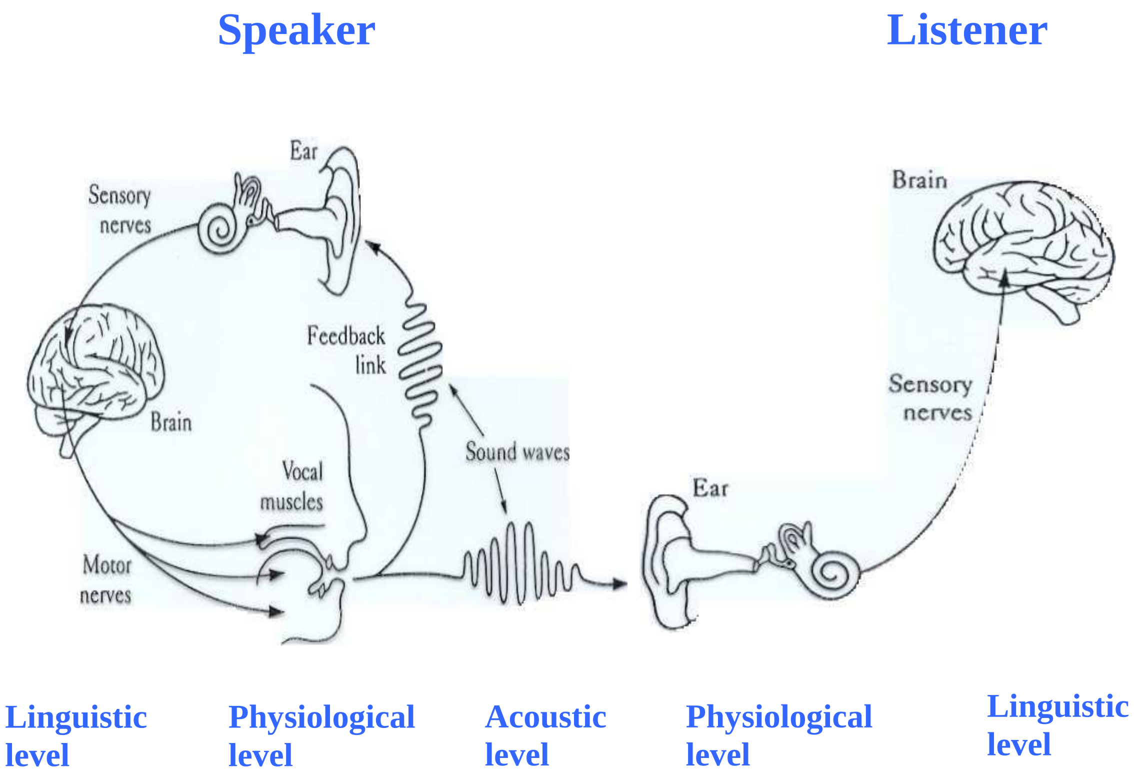


Figure (2) speech chain

The elementary linguistic unit of spoken speech is called a *phoneme* and its acoustic realization is called a *phone*. There are between 60 to 80 phonemes in spoken English, the exact number of phonemes depends on the dialect.

For the purpose of automatic speech processing the number of phonemes is clustered and reduced to between 40 to 60 phonemes depending on the dialect. Syllables are also sub-word units, but they are larger than phonemes. A word may be composed of one or more syllables and a syllable may be composed of one or more phonemes. Words are the most commonly understood speech units , from the word we can construct the sentences.

Every communication system has a set of elementary symbols (or alphabet) from which larger units such as words and sentences are constructed. For example in digital communication the basic alphabet are “1” and “0”, and in written English the basic units are *A* to *Z*.[6]. In Arabic language the basic units are Alf to Yaa. **(** أ- ي).

For speech recognition context-dependent triphone units are used. Assuming that there are about 40 phonemes, in English language theoretically there will be about 40×40=1600 context dependent variations of each phone, and hence a total of 40×1600=64000 triphones.

For example, the word ‘*imagination*’ can be deconstructed into the following sub-word units:

Word *imagination*

Phonetic transcription *iy m ae g iy n ay sh e n*

Triphone transcription *iy+m iy-m+ae m-ae+g ae-g+iy g-iy+n iy-n+ay n-ay+sh ay-sh+ay sh-e+sh e-n+sh e-n*

Syllable transcription *iyma giy nay shen*. [5].

3. An Overview of Text-to-Speech (TTS) Synthesis

A Text-To-Speech (TTS) synthesizer as a computer-based system that should be able to read text aloud, regardless whether the text is introduced by computer input stream or a scanned input that is submitted to an optical character recognition (OCR) engine. This TTS synthesis should be intelligent enough to read “new” words/sentences and the speech it produces should be “natural” like human. Thus, a formal definition of text-to-speech is “the production of speech by machines, by way of the automatic phonetization of the sentences to utter”. [4].

The concept of high quality TTS synthesis appeared in the mid-eighties, as a result of important developments in speech synthesis and natural language processing techniques, mostly due to the emergence of new technologies like Digital Signal and Logical Inference Processors [7].

Text-to-speech synthesis can be used in many areas, such as telecommunications services, language education, vocal monitoring and, multimedia applications as well as an aid to handicapped people.

Furthermore, the potential applications for such technology
include teaching aids, text reading, and talking books/toys [7]. However, most TTS systems today only focus on a limited domain of applications, e.g. travel planning, weather services, and baggage lost-and-found [2].

TTS systems and synthesis technology for Arabic languages have been developed in the last decade [9]. The main difference between general purpose TTS systems and Arabic TTS systems ,the later still under the research.

Figure (4) is a simple functional diagram of a general TTS synthesizer.

A TTS system is composed of two main parts, the Natural Language Processing (NLP) module and the Digital Signal Processing (DSP) module.[4].

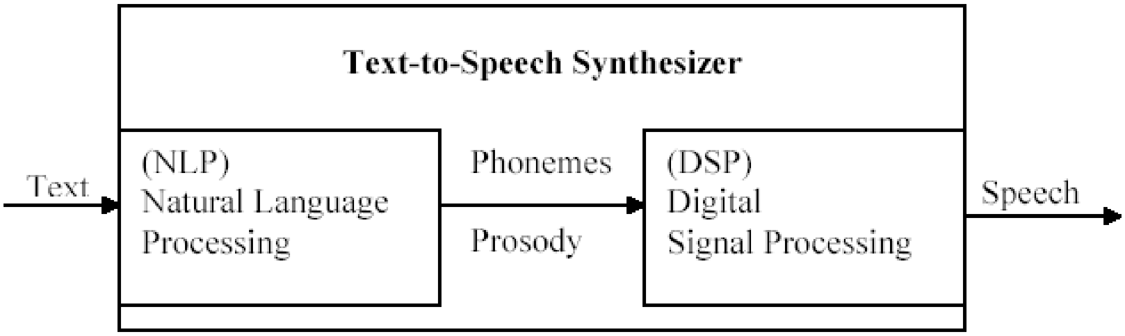


Figure (4) . A General TTS Synthesizer

The NLP module takes a series of text input and produces a phonetic transcription together with the desired intonation and prosody (rhythm) that is ready to pass on the DSP module. There are three major components within the NLP module, the letter-to-sound component, the prosody generation component, and the morpho-syntactic analyzer component .The DSP module takes the phonemes and prosody that were generated by the NLP module and transforms them into speech. There are two main approaches used by DSP module: rule-based-synthesis approach and concatenative-synthesis approach . Many researchers, refer to the NLP module as text-to-phoneme module and the DSP module as the phoneme-to-speech module. [4].

4. Natural Language Processing Module

Figure (5) introduces the functional view of a NLP module of a general Text-to-Speech Conversion system. The NLP module is composed of three major components: text-analyzer, letter-to-sound (LTS), and prosody generator.

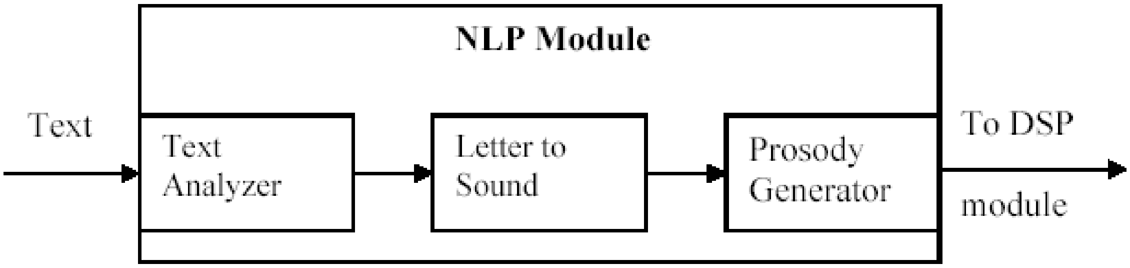


Figure (5) A Simple NLP Module

5. Arabic Language Phonology

The variety of Arabic dialects reflect the ethnic and social diversity of its speakers. There are two main clases of Arabic dialects: The Easterian dialects (Egypt, Sudan, and the Middle East), and the Western dialects of North Africa. These dialects classes are distinguished by the reduction of the vowel sounds[8]. Standard Arabic language has twenty-eight consonants and six vowels. The six vowels are divided into three long vowels (Alf, Yaa, and Wau), and three short vowels (fatha, kasra, and thoma). The long vowels have similar spectral properties like their short vowels version with longer durations than the short version. [3].

Arabic syllable must start with only one consonant and the syllabic structure prevent three consonants or two vowels to appear adjacently, letter to sound conversion for Arabic usually has simple one to one mapping between orthography and Phonetic transcription for given correct diacritics.

Syllabification for Arabic language has only six syllable types (CV,CVC,CVVC,CVCC,and,CVVCC). [1]

The number of vowels and the number of syllable in our Arabic phrase must be equal**,** any stream be accurately parsed according to these rules. Unfortunately, most modern standard written Arabic omits the diacritics, using partially is explicitly written but other vowels are left out.

It is therefore essential to recover the diacritics before applying grapheme to allophone rules. Furthermore Arabic TTS needs an intensive study of the morphological, syntactic, semantic, and pragmatic aspects to support the phoneme to allophone module and to derive prosody models requires for natural speech synthesis [5] .

6. Synthesis By Concatenation Approach

This approach use a real recorded speech as the synthesis units such as : phoneme, syllable, or word, and concatenate the units together to produce speech. Most researchers believed that the concatenate speech synthesis is the simplest and the most effective approach. In addition, they indicate that this approach is adapted by most of the TTS systems today. Thus, by using concatenate approach unit selection becomes critical for producing high-quality speech. Speech units have to be chosen to minimize future concatenation problems such as disjoint speech. Usually, speech units that are chosen affect the size of the database.

In the past, phonemes have been adopted as the basic synthesis units. Using phonemes as the synthesis unit requires a small storage, but it causes a lot of discontinuity between adjacent units. As the result, the researcher suggests that other synthesis units, such as dip hones and trip hones are often chosen as speech units because they are involved in the most articulation while requiring affordable an amount of memory.[4]

The models employed in concatenate synthesis are mostly based on signal processing tools and the most representative tools are Linear Prediction Coding (LPC) synthesizers, Harmonic/Stochastic (H/S), and Time-Domain Pitch-Synchronous-OverLap-Add (TD-PSOLA).,Wavelet,Multi-wavelet,and Hidden Marcov Model (HMM). Each one of these models has advantages and disadvantages, that related requirements of memory and computation time. Figure (6), represent the Human speech model.

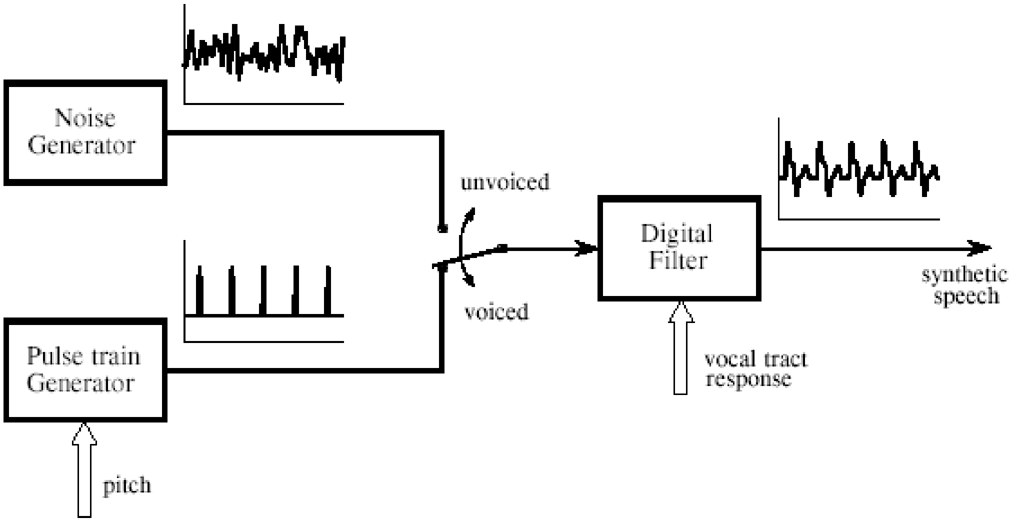


Figure (6) Human Speech Model

7. Synthesis An Arabic word based on All-pole Modeling

Our approach is to synthesize the Arabic waveform from model parameter estimated using linear prediction analysis. The synthesized signal is given by the following equation:

$$s[n]= \sum_{k=1}^{p} aks[n-k]+Ah[n]$$

Where h[n] is while noise, and the parameters (ak) and A are estimated by way of the normal equations and specific gain requirements. We use the auto correction method because it gives a stable and efficient solution. Typically we select the window to be fixed and equal to 23 ms to give a satisfactory time-frequency tradeoff.

The model order of p=14 poles.LPC analysis is performed on 256 point frames, and the frames are Hamming-weighted before analysis.[2].

In the LPC synthesis process each phonetic unit is created by synthesis from the LPC parameters stored in the database and is concatenated with other to produce the Arabic synthetic messages. With the LPC of each unit we can intervene in its parameters in order to improve the quality of the produced synthetic speech.[10].

In the overlap-add synthesis using an all-pole model, the wave is generated frame by frame by convolution synthesis and the fitter output on each frame is over-lapped and added with adjacent frame output,
as in the figure (7).

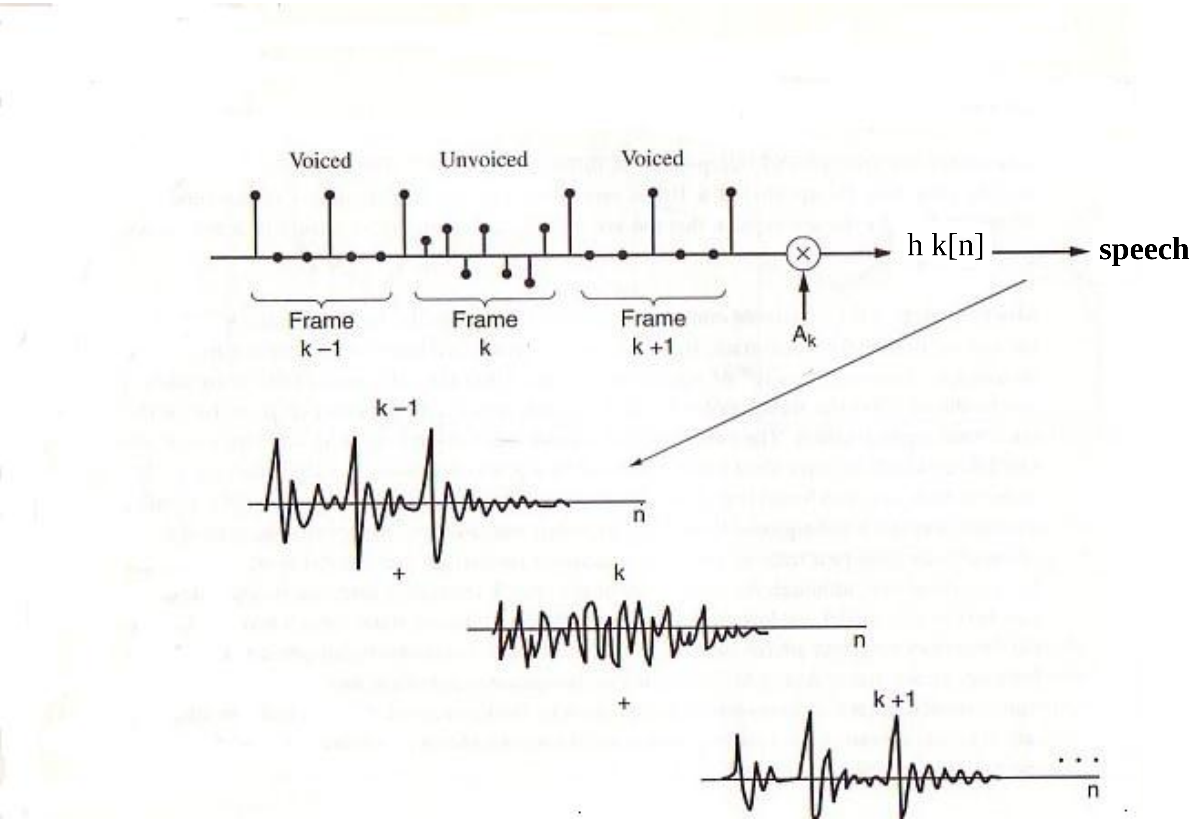


Figure (7) Overlap-add synthesis using an all-pole model.

8. The implementation of the model.

The computer system configuration used for the synthesis practical work has the following specification:

- PC Pentium 4 with RAM 512 GB, and CPU 1.7 GHz.
- XP Window Operating system.
- Recording system (Head phone type).
- Sound Forge program (for speech recording).
- Math Lab programming language version 6.5 release 13.

The practical work is done according to the following algorithm:

8.1 The first work step is recording the Arabic spoken phonemes with the Following data recording properties:

- The input file type is wave file.

- Sampling Rate Fs = 11025 Hs.
- Sample size = 16-bits.
- Channel =Mono.
- Frame length= 256 samples (23 ms).
- Overlap shift = 50% of frame size.

8.2 Linear Predictive coding analysis is performed with the 14th order for each frame of the phoneme using Durbin autocorrelation method with hamming window.

8.3 Record all Arabic phonemes (34 phonemes: 6 vowel sounds and 28 Consonants sounds), by using the Sound forge program, i.e create 34 Wave files , one for each phoneme.

8.4 Remove all silence period from each phoneme signal using visualization interactive manual process. (Sound Forge editor facilities).

8.5 for f=1 to 34 do steps 8.6 to 8.9

8.6 Read each file using MathLab Wavread function to find the size of each file and calculate number of frames for each file= file size/ frame size. The result will be rounded up to get the nearest integer frame number.

8.7 For k=1 to number of frames do the steps 8.7.1 to 8.7.3

8.7.1 Calculate the LPC Coefficients for frame k.

8.7.2 Pass the LPC vector through Filter, which is provided by MathLab as a function with name FILTER.

8.7.3 Calculate the estimated coefficient value

8.8 Concatenate frame k coefficient with frame k-1 coefficients to form LPC coefficient array.

8.9 Store Coefficient array in phoneme matrix indexed by phoneme letter value.

8.10 Read input text s.

8.11 Take each letter inturn and convert it to sound using the LPC Coefficient array and equation s[n]= $\sum_{k=1}^{p} aks[n-k] + Ah[n]$.

8.12 Concatenate each letter sound with previous letter using overlap 50% to eliminate discontinuity between phonemes.

8.13 Use wave play mathlab command to hear the synthesized wave.

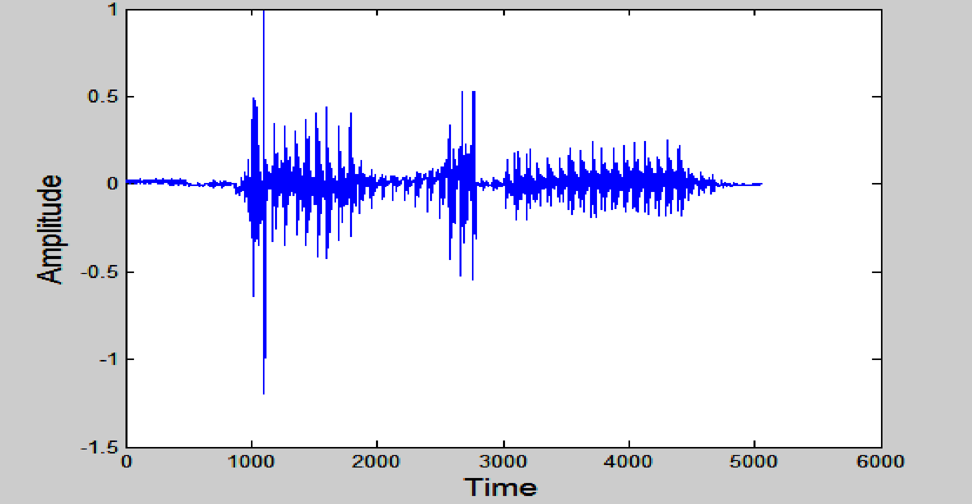
9.The results.

We test the proposed model through record a signal for Arabic utterance word (رَباب), and consider this signal as an original signal. The signal is shown in the figure (8).

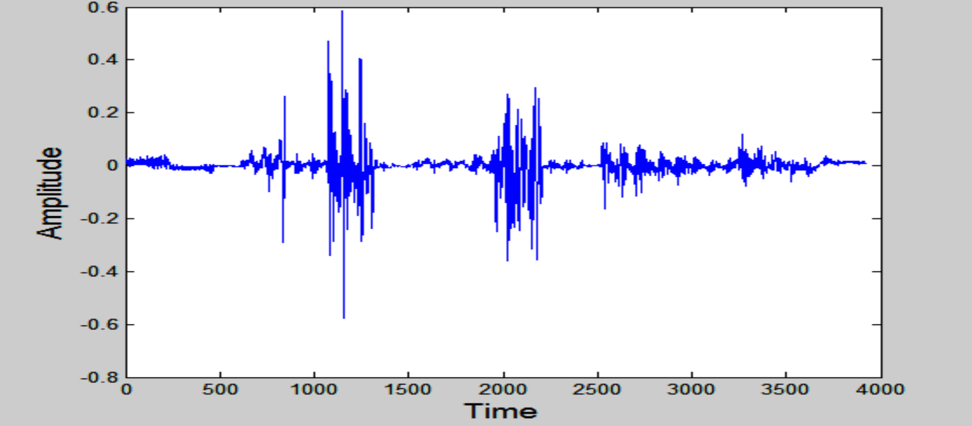
The test signal for the input text (رَباب), is composed of linguistics as “CVCVVC” and passed to the TTS model using the phoneme database to generate the sound. The output test signal is shown in figure (9).

The comparison degree between the original signal and the test one is (82%) Accuracy. We found that, this result is accepted at this stage , and we hope , it can improved in next research progress by applying more intelligent techniques.

Figure (8). The Orginal Signal Wave



Figure(9). The test Signal wave



10. Conclusion

This paper gives a general over view for the speech synthesis, and some details for the Text-To-Speech. The research still has a long way to go before delivering natural speech output for any input text with any intended emotions. Arabic TTS needs an intensive study of the morphological, syntactic, semantic, and pragmatic aspects to support the phoneme to allophone module and to derive prosody models required for natural speech synthesis.

11. References

1. *Yasser Hifny, Shady Quarany, Salah Hamid. 2003*)**.”ARABICTALK, An Implementation For Arabic Text To Speech System”.** *http://www.rdi-eg.com/*.

2. *Thomase F. Quatiieri (2002)*. **“Discrete-Time Speech Signal Processing Principles And Practice”.** *Prentice Hall PTR*.

3. *Nedhal Abdul Majied Abdul Saiyd (2000). “***Real – Time Concatenative Arabic Text - To- Speech System”.** *Department of Computer Science and Information Systems At The University of Technology / Baghdad*.

4. *Yuan Yuan Li , Stenven Case.(2003*) **“Text-To-Speech Synthesis For Mandrin Chinese”.** *Steven.case@mnsu.edu*

5. *Furui S. (1989),* **Digital Speech Processing, Synthesis and Recognition,Marcel Dekker.**

6. *RABINER L.R. and JUANG B.H. (1993)* **Fundamentals of Speech Recognition.***Prentice-Hall, Englewood Cliffs, NJ.*

7. *Paual A. Lynn , WolfGrang Fuerst. (1998*) **“Introductory Digital Signal Processing With Computer Applications”.** *John Wiley & Sons.*
8. *Laura MayfieldTomokiyo, Alan W Black,and Keven A. Lenzo.(2003).”* **Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic”.** *laura@ cepstral. Com.*

9. *Sana'a Wafa Tawfiq Al-Sayegh. (2002)."* **Arabic Phoneme Recognozer Based On Neural Networks***". Ph.D Thesis.Iraqi Commission for Computers and Informatics,Informatics for Postgraduate Studies.*

10. *P. Stathopoulou-Zois.(1997).* **"The UOP Text-To-Speech system for Greek Speech Synthesis".***Pstath@ee.upatras.gr***.**

11. *Lawrence R. Rabiner , Ronald W.Schafer.(1978*) **" Digital Processing Of Speech Signals".***By Bell Laboratories,Incorporated***.**

12. *Emmanuel C. Ifeachor, Barrie W. Jervis. (1998*)**" Digital Signal Processing A practical Approach".** *Addison_Wesley***.**

13. *Gyorgy Balogh, Ervin Dobler.(2002*)**." Flexvoice: A Prarametric Approach To High-Quality Speech Synthesis".** *Email:Grobler@mindmaker.hu*

14. *Allan Ramsay , Hanady Mansour.(2003)*"**Text to Speech for Modren Standard Arabic".** *Allan@co.umist.ac.uk*

15. *Manal Hasson Mouhammed.(1992) "***Stochastic Modeling And Quantization Techniques Applied For Arabic Speech Processing".** ***Ph.D Thesis.****The College Of Enginerring /University Of Bagdad* **.**

16. *Saad Najim Bashik Al-Saad. (2001)."***Recognition of Arabic Phonemes Using Vector Quantization***". Ph.D Thesis .Natinal Computer Center/ Institute of Higher Studies in Computer and Information/Baghdad.*

17. *Denes A. Pinson. (1995)."* **The Scientist and Engineer's Guide to Digital Signal Processing**"**.**

18.*Thomase W. Parsons.(1995*) **"Voice And Speech Processing".** *McGraw_Hill Book Company.*