International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:3, 2013

# The Main Principles of Text-to-Speech Synthesis System

K.R. Aida–Zade, C. Ardil and A.M. Sharifova

***Abstract*****—**In this paper, the main principles of text-to-speech synthesis system are presented. Associated problems which arise when developing speech synthesis system are described. Used approaches and their application in the speech synthesis systems for Azerbaijani language are shown.

***Keywords*****—**synthesis of Azerbaijani language, morphemes, phonemes, sounds, sentence, speech synthesizer, intonation, accent, pronunciation.

## I. INTRODUCTION

IN the XXI century widespread use of computers opened a new stage in information interchange between the user and the computer. Among other things, an opportunity to input the information to computer through speech, and to reproduce in voice text information stored in the computer have been made possible. The paper is dedicated to the second part of this issue, i.e. to the computer-aided text-to-speech synthesis, that is recognized as a very urgent problem. Nowadays the solution of this problem can be applied in various fields. First of all, it would be of great importance for people with weak eyesight. In the modern world, it is practically impossible to live without an information exchange. The people with weak eyesight face with big problems while receiving the information through reading. A lot of methods are used to solve this problem. For example, the sound version of some books is created. As a result, people with weak eyesight have an opportunity to receive the information by listening. But there can be a case when the sound version of the necessary book couldn't be found.

Therefore, the implementation of the speech technologies for information exchange for users with weak eyesight is of a crucial necessity. In fact, computer synthesis of speech opens a new direction for an information transfer through the computer. For today it is mainly possible through the monitor.

Synthesis of speech is the transformation of the text to speech. This transformation is converting the text to the synthetic speech that is as close to real speech as possible in compliance with the pronunciation norms of special language. TTS is intended to read electronic texts in the form of a book, and also to vocalize texts with the use of speech synthesis. When developing our system not only widely known modern methods but also a new approach of processing speech signal was used.

Kamil Aida-Zade is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.
Cemal Ardil is with the National Academy of Aviation, Baku, Azerbaijan.
Aida Sharifova is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.

Such systems can be used in communication systems, in information referral systems, it can be applied to help people who lost seeing and reading ability, in acoustic dialogue of users with computer and in others fields. In general, synthesis of speech can be necessary in all the cases when the addressee of the information is a person.

## II. PREVIOUS WORKS

The earliest efforts to produce synthetic speech date as far back as XVIII century. Despite the fact that the first attempts were in the form of mechanical machines, we can say today that these synthesizers were of a high quality. In 1779 in St. Petersburg, Russian Professor Christian Kratzenshtein explained physiological differences between five long vowels (/a/, /e/, /i/, /o/, and /u/) and made an apparatus to produce them artificially. In 1791 in Vienna, Wolfgang von Kempelen introduced his "Acoustic-Mechanical Speech Machine". In about mid 1800's Charles Wheatstone constructed his famous version of von Kempelen's speaking machine.

There have been three generations of speech synthesis systems [1]. During the first generation (1962-1977) formant synthesis of phonemes was the dominant technology. This technology made use of the rules based on phonetic decomposition of sentence to formant frequency contours.

The intelligibility and naturalness were poor in such synthesis. In the second generation of speech synthesis methods (from 1977 to 1992) the diphones were represented with the LPC parameters. It was shown that good intelligibility of synthetic speech could be reliably obtained from text input by concatenating the appropriate diphone units. The intelligibility improved over formant synthesis, but the naturalness of the synthetic speech remained low. The third generation of speech synthesis technology is the period from 1992 to the present day. This generation is marked by the method of "unit selection synthesis" which was introduced and perfected, by Sagisaka at ATR Labs. in Kyoto. The resulting synthetic speech of this period was close to human-generated speech in terms of intelligibility and naturalness.

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. "Infovox" [2] speech synthesizer family is perhaps one of the best known multilingual text-to-speech products available today. The first commercial version, Infovox SA-101, was developed in Sweden at the Royal Institute of Technology in 1982 and it is based on formant synthesis. The latest full commercial version, Infovox 230, is available for American and British English, Danish, Finnish, French, German, Icelandic, Italian, Norwegian, Spanish, Swedish, and Dutch.

Digital Equipment Corporation [3] (DEC) talk system is originally descended from MITalk and Klattalk. The present version of the system is available for American English, German and Spanish and offers nine different voice personalities, four male, four female and one child. The present DECtalk system is based on digital formant synthesis.

AT&T Bell Laboratories [4] (Lucent Technologies) also has very long traditions with speech synthesis. The first full text-to-speech (TTS) system was demonstrated in Boston 1972 and released in 1973. It was based on articulatory model developed by Cecil Coker (Klatt 1987). The development process of the present concatenative synthesis system was started by Joseph Olive in mid 1970's (Bell Labs 1997). The current system is available for English, French, Spanish, Italian, German, Russian, Romanian, Chinese, and Japanese (Mcbius et al. 1996).

Currently, different research teams are working at this problem. Some of them are described in the Table 1.

TABLE I TTS SYSTEMS

| | | |
|---|---|---|
| German Synthesis | | |
| German Festival [5] | IMS Uni Stuttgart | diphone synthesis |
| Hadifix [6] | IKP Uni Bonn | mixed inventory |
| Waveform Synthesis[7] | TU Dresden | waveform synthesis |
| Multilingual TTS system [8] | TI Uni Duisburg | formant synthesis |
| English Synthesis | | |
| Laureate [9] | British Telecom | unit selection |
| YorkTalk [10] | University of York | non-segemental (formant) synthesis |
| SPRUCE [11] | Essex Speech Group | high level synthesis (using other synthesis backends) |
| French Synthesis | | |
| SpeechMill [12] | The University of Lausanne | diphone synthesis |
| ICP [13] | Grenoble | diphone synthesis |
| CNET [14] | Lannion | diphone synthesis |
| Spanish Synthesis | | |
| University of Madrid [15] | | concatenative and formant synthesis |
| LIMSI TTS System [16] | LIMSI-CNRS | diphone synthesis |
| Greek Synthesis | | |
| University of Patras [17] | | diphone synthesis |
| DEMOSTHeNES[18] | University of Athens | diphone synthesis |
| Arabic Synthesis | | |
| Sakhr Software [19] | Cairo, Egypt | concatenative |
| | | synthesis |
| MBROLA [20] | Le Mons, Belgium | diphone synthesis |
| Turkish Synthesis | | |
| TTTS [21] | Fatih University | syllable-based concatenative |

Despite existing various approaches, it is still difficult to tell which of these approaches is more suitable or more useful. In general, the problems faced during text-to-speech synthesis depend on many factors: diversity of languages, specificity of the pronunciation, accent, stress, intonation etc.

## III. SPECIFICATION OF LANGUAGE RESOURCES FOR SPEECH SYNTHESIS

Text to speech synthesis is converting the text to the synthetic speech that is as close to real speech as possible according to the pronunciation norms of special language. Such systems are called text to speech (TTS) systems. Input element of TTS system is a text, output element is synthetic speech. There are two possible cases. When it is necessary to pronounce the limited number of phrases (and their pronouncing linearly does not vary), the necessary speech material is simply recorded in advance. In this case, certain problems are originated. For example in this approach, it is not possible to sound the text, which is not known in advance. For this purpose the pronounced text has to be kept in computer memory. And it will lead to increase of the size of memory required for information content. This will bring to essential load of computer memory in case of much information and can create certain problems in operation. The main approach used in this paper is voicing of previously unknown text based on a specific algorithm.

It is necessary to note that the approach to solving problem of speech synthesis essentially depends on the language for which it will be used [22], and that the majority of currently available synthesizers basically were generated for UK-English, Spanish and Russian languages, and these synthesizers had not been applied to the Azerbaijani language yet. Azerbaijani language like as Turkish is an agglutinative language. In the view of the specificity of the Azerbaijani language, the special approach is required.

Every language has its own unique features. For example: there are certain contradictions between letters and certain sounds in English language. Thus, two different letters coming together, sound differently than when they are used separately. For example: letters (t), (h) separately do not sound the same as in chain (th).

This is only one of problems faced in English language. In other words, the place of the letters affect on how they should be or should not be pronounced. Thus, according to the phonetic rules of English language the first letter (k) of the word (know) is not pronounced.

As well, Russian language has certain pronunciation features. First of all, it should be noted that the letter (o) does not always pronounced like sound (o). There are some features

based on phonetic rules of Russian language. For example: the first letter (o) in the word (korova) is pronounced like (a). Moreover, if we take into consideration that the letter (ь) is not pronounced at all and only gives softness to the pronounced word.

From the foresaid it is clear that it the synthesizer programs developed especially for one language cannot be used in a different language, because the specificity of one language is not presumably typical for the others. Each program is based on algorithms corresponding to the phonetic rules of certain language. Till now there are no any programs of a synthesizer type that take into consideration the specificity of Azerbaijani language.

The Azerbaijani language has its specific features [23]. Some words aren't pronounced as its written form in Azeri. For example, the Azeri word "ailə" is pronounced like [ayilə], "Aidə" like [Ayidə], "müəllim" like [mə:lim]. As it is shown the sound "y" is added to the first and second words, the sounds "ü" and "l" aren't pronounced in the second word. Here is another example: the word "toqqa" is pronounced like [tokqa], here the first sound "q" is changed into "k".

## IV. USED APPROACH

Two parameters, naturalness of sounding and intelligibility of speech, are applied for the assessment of the quality of synthesis system. One can say that naturalness of sounding of a speech synthesizer depends on how many generated sounds are close to natural human speech. By a intelligibility (ease for understanding) of a speech synthesizer is meant the easiness of artificial speech understanding. The ideal speech synthesizer should possess both characteristics: naturalness of sounding and intelligibility. Existing and being developed systems for speech synthesis are aimed at improvement of these two characteristics.

The idea of combination of concatenation methods and of formant synthesis is the fundament of the system we developed. The rough, primary basis of a formed acoustic signal is created on the basis of concatenation of the fragments of an acoustic signal taken from speech of the speaker, i.e., a "donor". Then, this acoustic database is changed by the rules. The purpose of these rules is to give the necessary prosodies characteristics (frequency of the basic tone, duration and energy) to the "stuck together" fragments of an acoustic signal.

The method of concatenation together with an adequate set of base elements of compilation provides for qualitative reproduction of spectral characteristics of a speech signal, and the set of rules provides for the possibility of generating natural intonation-prosodial mode of pronouncements.

Formant synthesis does not use any samples of human speech. On the contrary, the speech message of the synthesized speech is created by means of acoustic model. Such parameters as own frequency, sounding and noise levels vary in order to generate natural form of a signal of artificial speech.

In systems of concatenate synthesis (earlier it was called compilation), synthesis is carried out by sticking together necessary units from available acoustic units. Concatenation of segments of written speech lays in the basis of concatenate synthesis. As a rule, concatenate synthesis gives naturalness to sounding of the synthesized speech. Nevertheless, the natural fluctuations in speeches and the automated technologies of segmentation of speech signals create noise in the generated fragment and this decreases the naturalness of sounding.

An acoustic signal database (ASD), which consists of fragments of a real acoustic signal, i.e. the elements of concatenation (EC), is the basis of any system of synthesis of the speech based on concatenation method. Dimension of these elements can be various depending on a concrete way of synthesis of speech, it can be phonemes, allophones, syllables, diaphones, words etc [24].

In the system developed by us the elements of concatenation are diaphones and various combinations of vowels. But it is necessary to note that the study of generation of one-syllabic words consisting of four letters (stol, dörd) is still underway and that is why this words are included into the base as indivisible units. The speech units used in creation ASD are saved in WAV format. The process of creation of ASD consists of the following stages:

Stage 1: In the initial stage, the speech database is created on the basis of the main speech units of the donor speaker.

Stage 2: The speech units from speaker's speech are processed before being added into database. It's done in the following steps:

a) Speech signals were sampled at 16 kHz and it makes possible to define the period T with a precision of $10^{-4}$.

b) Removal of surrounding noise from the recorded speech units. For this purpose we use the algorithm of division of a phrase realization into speech and pauses. It is supposed, that the first 10 frames do not contain a speech signal. For this part of signal we calculate mean value and dispersion of $E_t$ and $Z_t$ and obtain statistical characteristics of noise [25].

$$E_s(m) = \sum_{n=m-L+1}^{m} s_p^2(n) \tag{1}$$

$$Z_S(m) = \frac{1}{L} \sum_{n=m-L+1}^{m} \frac{\left|\operatorname{sgn}(s_p(n)) - \operatorname{sgn}(s_p(n-1))\right|}{2} \tag{2}$$

where

$$\operatorname{sgn}(s_p(n)) = \begin{cases} 1, & s_p(n) \geq 0, \\ -1, & s_p(n) < 0. \end{cases} \tag{3}$$

where L is the number of frames of speech signal.

Then taking into account these characteristics and maximal values of $E_t$, $Z_t$ for realization of a given phrase we calculate the threshold $T_E$ for short-time energy of a signal and the threshold $T_Z$ for number of zeros of signal intensity. The following formulas have been chosen experimentally:

$$T_E = M(E,10) + 3\sqrt{D(E,10)} \le k1 \max_{1\le t\le L} E_t \quad (4)$$

$$T_Z = M(Z,10) + 3\sqrt{D(Z,10)} \le k2 \max_{1\le t\le L} Z_t \quad (5)$$

where

$$M(P,n) = \frac{1}{n}\sum_{t=1}^{n} P_t \quad (6)$$

$$D(P,n) = \frac{1}{n-1}\sum_{t=1}^{n} (P_t - M(P,n))^2 \quad (7)$$

If a frame $X^{(t)}$ contains speech, then we assign binary variable $b_t$ to 1, otherwise we assign $b_t$ to 0. At first, it is necessary to assign the value 1 to the frames with short-time energy $E_t \ge T_E$, and to assign the value 0 to the other frames. Variables $b_t$ can accept only two values. Therefore the filtration is reduced to the following procedure. Consecutively for t=h+1,...,L-h the values of $b_t$ are replaced with 1, if $\sum_{i=t-h}^{t+h} bi > h$ .

Otherwise the values of $b_t$ are replaced with zero.

$$b_t = \begin{cases} 1, & \sum_{i=t-h}^{t+h} b_i > h, \quad t = h+1 >,\ldots L-h \\ 0, & \sum_{i=t-h}^{t+h} b_i \le h, \quad t = h+1 >,\ldots L-h. \end{cases} \quad (8)$$

As a result, the continuous parts containing speech are determined. Then, we try to expand each part of this type. For example, the part begins at the frame $X^{(N1)}$ and comes to an end at the frame $X^{(N2)}$. Moving to the left from $X^{(N1)}$ (or to the right from $X^{(N2)}$) the algorithm compares the number of zeros of intensity $Z_t$ with the threshold $T_Z$. This moving should not exceed 20 frames to the left of $X^{(N1)}$. If $Z_t$ has exceeded the threshold three and more times, then the beginning of a speech part is transferred to the place where $Z_t$ exceeds the threshold for the first time. Otherwise the frame $X^{(N1)}$ could be considered as the beginning of the speech part. The same is valid for $X^{(N2)}$. If two parts are overlapped, they can be combined into one part. Thus, the continuous parts containing speech are finally determined. Such parts will be called realizations of words.

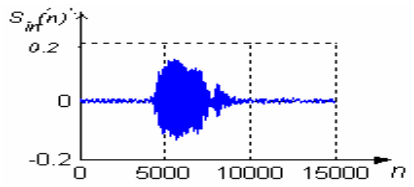


Fig. 1. Before application of the algorithm

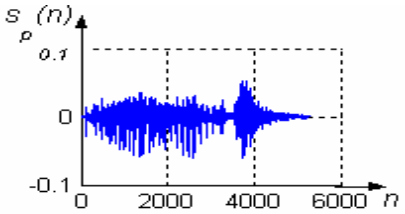


Fig. 2. After application of the algorithm

As may be seen from the figures (Figure 1, 2), the continuous parts containing speech are finally determined after the algorithm application.

Stage 3: As described above, the ASD plays the main role in speech synthesis. The information stored in AED is used in different modules of synthesis. In our system, CE is stored in .wav format, with 16 kHz frequency. Each wav file includes the following elements:

0. the description of CE
1. the count of speech signal parts – N
2. energy of speech signal – E
3. amplitude of CE - A
4. the frequency of crossing zero– Z

Stage 4: At the following stage another corresponding variants of each CE are created. In spite of the fact that it increases the quantity of ASD elements, but at the same time it makes possible to reduce the quantity of modules for generation of a target signal.

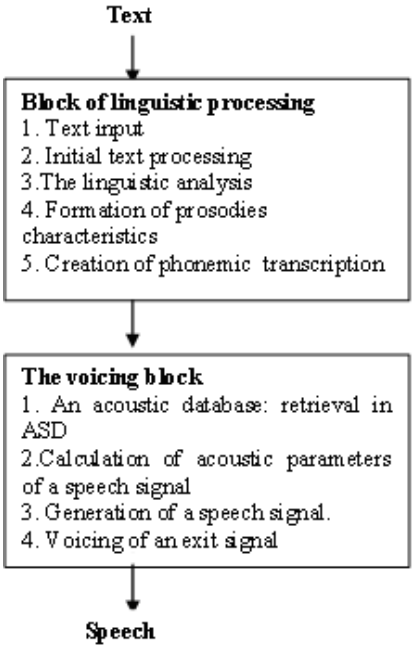


Fig. 3. The structure of typical system

These stages are used only in the beginning of process of creating CE database for ASD. In the subsequent stages we do not turn to them any more.

The structure of the majority of systems of speech synthesis, as well as the structure of our system of automatic synthesis can be presented by the following flow chart (Fig. 3) [26].

As may be seen from the shown diagram, there are two blocks in our system: the block of linguistic processing and the voicing module. At first, the input text is processed in the Linguistic block and the obtained phonemic transcriptor is passed to the second block, i.e., to the Voicing block of system. In the Voicing block after certain stages the obtained speech signal is sounded.

*4.1 Block of linguistic processing*

*4.1.1 Text input*

The sounded text can be entered in any form. The size or font type is of no importance. The main requirement is that the text must be in Azerbaijani language.

*4.1.2 Initial text processing*

For forming of transcriptional record, the input text should be shown as sequence of accentuated spelling words separated by space and allowed punctuation marks. Such text can conditionally be named as "normalized". Text normalization is a very important issue in TTS systems. The general structure of normalizer is explained in Figure 4. This module has several stages as it is shown in the figure.

Stage 1: Spell-checking of the text

The spell-checkers are used in some cases (modules of correction of spelling and punctuation errors). The module helps to correct spelling errors in the text thereby to avoid voicing of these errors.

Stage 2: A pre-processing module

A pre-processing module organizes the input sentences into manageable lists of words. First, text normalization isolates words in the text. For the most part this is as trivial as looking for a sequence of alphabetic characters, allowing for an occasional apostrophe and hyphen.

It identifies numbers, abbreviations, acronyms, idiomatics and transforms them into full text when needed.

Stage 3: Number Expansion

Text normalization then searches for numbers, times, dates, and other symbolic representations [27]. These are analyzed and converted to words. (Example: "$54.32" is converted to "fifty four dollars and thirty two cents, "$200" appears in a document it may be spoken as "two hundred dollars". Similarly, "1/2" may be spoken as "half", "January second", "February first", "one of two") Someone needs to code up the rules for the conversion of these symbols into words, since they differ depending upon the language and context.

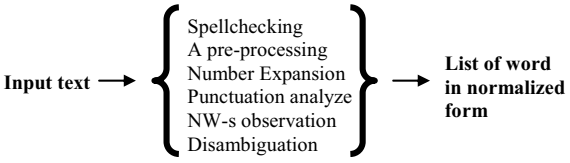


Fig. 4. Text Normalization System

The following non-standard representations are described in the Table 2.

TABLE II POSSIBLE TOKEN TYPE IN TEXT

| TYPE | TEXT | SPEECH |
|---|---|---|
| *Decimal Numbers* | 1,2 | One and two tenth |
| *Ratios* | 1/2 | One second |
| *Ordinal numbers* | 1-st | first |
| *Roman Numerals* | VI, X | sixth, tenth |
| *Alphanumeric strings* | $10^a$ | Ten a power of a |
| *Phone Number:* | (+9941 2)5813 433 | Plus double nine, four, one, two, five, eight, one, three, four, double, three |
| *Count:* | 25 | Twenty five |
| *Date :* | 01.11.1999 | First of November nineteen ninety-nine |
| *Year* | 1989 | Nineteen eighty nine |
| *Time* | 10:30 pm | Half past ten post meridiem |
| *Mathematical:* | 2+1=3 | Two plus one is equal to three |

Stage 4: Punctuation analyze

Whatever remains is punctuation. The normalizer will have rules dictating if the punctuation causes a word to be spoken or if it is silent. (Example: Periods at the end of sentences are not spoken, but a period in an Internet address is spoken as "dot.")

In normal writing, sentence boundaries are often signaled by terminal punctuation from the set: full stop, exclamation mark, question mark or comma {. ! ? ,} followed by white spaces. In reading a long sentence, speakers will normally break up the sentence into several phrases, each of which can be said to stand alone as an intonation unit. If punctuation is used liberally so that there are relatively few words between the commas, semicolons or periods, then a reasonable guess at an appropriate phrasing would be simply to break the sentence at the punctuation marks though this is not always appropriate. Hence, determining the sentence break and naming the type of sentence has to be done so as to apply the prosodic rules.

In natural speech, speakers normally and naturally give pauses between sentences. The average duration of pauses in a natural speech has been observed and a lookup table (Table 3.) is generated. Finally, the lookup table is used to insert pauses between sentences that improve naturalness.

TABLE III OBSERVATION NATURAL SPEECH

| Sentence Type | Duration in seconds |
|---|---|
| Affirmative (.) | 1 |
| Exclamatory (!) | 0.9 |
| Interrogative(?) | 0.8 |
| Comma (,) | 0.5 |

This block creates information base for the creation of phonemic transcription.

Stage 5: NSW-s observation

The aim of this stage is converting None Standard Words (NSWs) into their standard word pronunciations. In Table 4. some of NSWs are explained.

TABLE IV NSW-S OBSERVATION

| NSW | Their explained. |
|---|---|
| *Abbreviations* | |
| ASOA | Azerbaijan State Oil Academy |
| AR | Azerbaijan Republic |
| BSU | Baku State University |
| Internet addresses | |
| http://www.Microsoft.com | w w w dot Microsoft dot com |
| Mail address | |
| ibrahim@mail | ibrahim at mail dot ru |
| Money | |
| $ | Dollar |
| AZN | Manat |

Once the text has been normalized and simplified into a series of words, it is passed onto the next module, homograph disambiguation.

Stage 6: Disambiguation

In some system disambiguation module is generally handled by hand-crafted context-dependent rules [28]. However, such hand-crafted rules are very difficult to write, maintain, and adapt to new domains.

The simple cases are the ones that can be disambiguated within the word. In that case, the pronunciation can be annotated in the dictionary, and so long as the word parsing is correct, the right pronunciation will be chosen.

Currently we have only implemented the contextual disambiguation. We will continue to implement other cases.

By the end of this step the text to be spoken has been converted completely into tokens.

*4.1.3 Linguistic analysis: the syntactic, morphemic analysis*

Linguistic analysis of the text takes place after the normalization process. By using morphological, syntactic characteristics of the Azerbaijani language the text is partitioned into sub-layers.

Text and speech signal have clearly defined hierarchical nature. In view of hierarchical representation, we can conclude that for the qualitative construction of systems of speech synthesis it is necessary to develop a model of mechanism of speech formation. In the system, initially we should define the flow of the information which should proceed according to the scheme presented in Fig. 5.

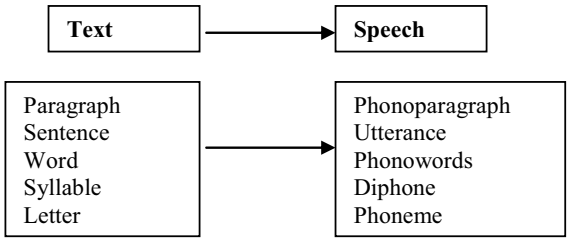


Fig. 5. Difference between text and speech

*4.1.4 Formation of prosodies characteristics*

The voice-frequency, accent and rhythmic characteristics belong to the prosodies characteristics of utterance. The frequency of the basic tone, energy and duration are their physical analogues. These characteristics are informative for formation of the control information for the subsequent generation of an acoustic signal.

*4.1.5 Creation of phonemic transcription (PT)*

Phonemic transcription forms the sound transcription relevant to the inputted text, based on standard rules of reading in the Azerbaijani language.

At this stage, it is necessary to assign each word of the text (each word form) the data on its pronunciation, i.e. to transform each word into a chain of phonemes or, in other words, to create its phonemic transcription. In many languages, as well as in Azerbaijani, there are sufficiently regular reading rules, i.e., the rules of conformity between letters and phonemes (sounds). The rules of reading are very irregular in English language, and therefore the task of this block for English synthesis becomes more complicated. In any case, there are serious problems at definition of pronunciation of proper names, the loanwords, new words, acronyms and abbreviations.

It is not possible to simply store a transcription for all words of the language because of great volume of the vocabulary and because of contextual changes of a pronunciation of the same word in a phrase. In addition, it is necessary to consider the cases of graphic homonymy correctly: the same sequence of alphabetic symbols in various contexts can represent two different words/word forms and can be read differently. Often, this problem can be solved by grammatical analysis; however, sometimes only wider use of semantic information helps.

*4.2 The voicing block*

*4.2.1 An acoustic database: the retrieval in ASD*

As it was already mentioned processing the EC and creation of ASD takes place at the initial stage. But at the first stage of the voicing block the retrieval is only made in ASD. The phonemic chain is created by using generated phonemic transcription from EC in ASD.

*4.2.2 Calculation of acoustic parameters of a speech signal*

This block generates the created phonemic chain on the basis of prosodic characteristics.

The purpose of the rules of this block includes definition of energy, time and voice-frequency characteristics that should be assigned to the sound units forming a phonetic transcription of the synthesized phrase.

*4.2.3 Generation of a speech signal.*

The joining of speech units occurs independently of the size of EC. Thus, there can be rather sensitive distortions of a speech signal for hearing. To prevent this effect, local smoothing of left (i) and right (j) joined waves is carried out by the following algorithm [29]:

1. From the last (zero) reading of left ($i^{th}$) joined wave we count the $3^{rd}$ reading for which new average value Si3m is calculated from values of $i^{th}$ and $j^{th}$ waves by the formula:

$$Si3m = 1/9 * (Si7+ \dots +Si3 \dots + Si0+Sj0) \quad (9)$$

2. Then, we reiterate the process according to the following recurrent scheme until we receive the last new value for zero reading of the $i^{th}$ wave:

$$\begin{aligned} Si2m &= 1/9 * (Si6+ \dots +Si2 \dots + Sj0+Sj1) \\ Si1m &= 1/9 * (Si5+ \dots +Si1 \dots + Sj1+Sj2) \\ Si0m &= 1/9 * (Si4+ \dots +Si0 \dots + Sj2+Sj3) \end{aligned} \quad (10)$$

3. Then, the new values of the $j^{th}$ wave are calculated:

$$\begin{aligned} Sj0m &= 1/9 * (Si3+ \dots Sj0m \dots + Sj3+Sj4) \\ Sj1m &= 1/9 * (Si2+ \dots Sj1m \dots + Sj4+Sj5) \\ Sj2m &= 1/9 * (Si1+ \dots Sj2m \dots + Sj5+Sj6) \end{aligned} \quad (11)$$

4. The process ends after reception of new value for the $4^{th}$ reading of the $j^{th}$ wave:

$$Sj3m = 1/9 * (Si0 + Sj0+ \dots Sj3m \dots + Sj6+Sj7) \quad (12)$$

*4.2.4 Voicing of an output signal*

Using available EC from the received sequence of speech units is sounded.

## V. CONCLUSION

On the abovementioned grounds, the voicing of words of any text in Azerbaijani language is carried out with the help of a limited base of EC.

In this study the framework of a TTS system for Azerbaijani language is built. Although the system uses simple techniques it provides promising results for Azerbaijani language, since the selected approach, namely the concatenative method, is very well suited for Azerbaijani language. The system can be improved by improving the quality of the speech files recorded.

In particular, the work on intonation is not finished because segmentation was made manually and there is noticeable noise in voicing. It is planned to apply independent segmentation and to improve the quality of synthesis in the future.

The punctuations are removed in the preprocessing step just to eliminate some inconsistencies and obtain the core system. In the future versions of the TTS, the text can be synthesized in accordance with the punctuations for considering the emotions and intonations as partially achieved in some of the researches [30]. The synthesis of a sentence ending with a question mark can have an interrogative intonation and synthesis of a sentence ending with an exclamation mark can be an amazing intonation. In addition to these, other punctuations can be helpful for approximating the synthesized speech to its human speech form such as pausing at the end of the sentences ending with full stop and also pausing after the punctuation comma.

## REFERENCES

[1] Lawrence R. Rabiner and Ronald W. Schafer , *Introduction to Digital Speech Processing* , Now Publishers ,USA, 2007.
[2] http://www.infovox.se
[3] http://www.digital.com/
[4] http://www.bell-labs.com/project/tts
[5] http://www.ims.uni-stuttgart.de/phonetik/synthesis
[6] http://www.ikp.uni-bonn.de/~tpo/Hadifix.html
[7] http://www.et.tu-dresden.de/ita/ita.html
[8] http://www.fb9-ti.uni-duisburg.de/demos/speech.html
[9] http://www.labs.bt.com/innovate/speech/ laureate/index.htm
[10] http://www.york.ac.uk/~lang4/Yorktalk.html
[11] http://www.essex.ac.uk/speech/research/spruce/demo-1/demo-1.html
[12] http://www.unil.ch/imm/docs/LAIP/LAIPTTS.html
[13] http://www.icp.grenet.fr/ICP/index.uk.html
[14] http://www.cnet.fr/cnet/lannion.html
[15] http://lorien.die.upm.es/research/synthesis/synthesis.html
[16] http://www.limsi.fr/Recherche/TLP/theme1.html
[17] http://www.clab.ee.upatras.gr/
[18] http://www.di.uoa.gr/speech/synthesis/demosthenes
[19] http://demo.sakhr.com/tts/tts.asp
[20] http://tcts.fpms.ac.be/synthesis/
[21] http://fatih.edu.tr
[22] R.Ayda-zade, A.M.Sharifova. "The analysis of approaches of computer synthesis Azerbaijani speech". Transactions of Azerbaijan National Academy of sciences. "Informatics and control problems". Volume XXVI, №2. Baku, 2006, p.227-231. (in Azerbaijani)
[23] Mammadov N. The theoretical principles of Azerbaijan linguistics. Baki: Maarif, 1971, 366 p. (in Azerbaijani)
[24] Akhundov A. The phonetics of Azerbaijani language. Baki: Maarif, 1984, 392 p. (in Azerbaijani)
[25] Sagisaka Y. Spoken Output Technologies. Overview// Survey of the state of the art in human language technology. Cambridge, 1997.
[26] Sharifova A.M The Computer Synhtesis of the Azerbaijani Speech / (Azerbaijani). Application of information-communication technologies in science and education. International conference. Baku, 2007. Volume II, p. 47-52.
[27] http://www.clsp.jhu.edu/ws99/projects/normal/
[28] Yarowsky D., "Homograph Disambiguation in Text-to-Speech Synthesis" 2nd ESCA/IEEE Workshop on Speech Synthesis, New Paltz, NY, pp. 244-247.
[29] Lobanov B.M. Retrospective review of researches and workings out of Laboratory of recognition and speech synthesis. "Automatic recognition and speech synthesis", ITC NAS Belarus, Minsk, 2000.-S.6-23. (In Russian)
[30] Kurematsu, M., Hakura, J., Fujita, H., The Framework of the Speech Communication System with Emotion Processing, Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19, 2007, 46-52