AZERBAIJAN TEXT-TO-SPEECH SYNTHESIS SYSTEM

Kamil Aida-Zade[1], Aida Sharifova[2]

Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*kamil_aydazade@rambler.ru*, [2]*aid_tr@yahoo.com*

The idea of combination of methods concatenation and synthesis by rules lays at the heart of speech synthesis' system developed by us. The rough, primary basis of a formed acoustic signal is created on a basis of concatenation of the fragments of an acoustic signal taken from speech of the announcer – a "donor". Further this acoustic basis is exposed to updating by the rules, function of which consists of giving the necessary prosodies characteristics (frequency of the basic tone, duration and energy) to the "stuck together" fragments of an acoustic signal. As a rule, concatenate synthesis gives naturalness to sounding of the synthesized speech. Formant-synthesized speech can be reliably intelligible. Parameters, as own frequency and sounding vary after the lapse of time and are created the form of a signal of artificial speech [1].

The structure of the majority of systems of speech synthesis, as well as a structure of our system of automatic synthesis can be presented by the fig.1 [2]:

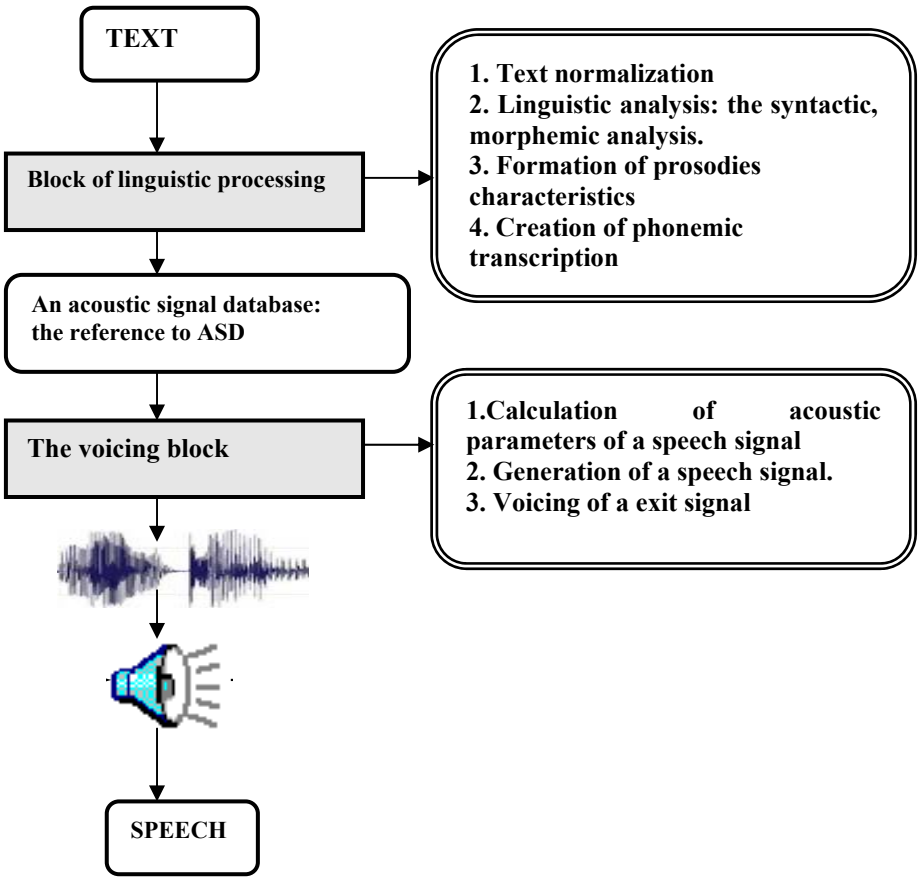


Figure 1 . Structure scheme of our TTS .

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 transcription 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.

Block of linguistic processing

Text normalization. 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.

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.

This block foreseen carrying out of spelling check of the text, filtration of symbols that don't need voicing ("(", "-",), number expansion (exam: the number "1" in expressions "01.10.2009" and "1998" is read differently), punctuation analyze, non-standard words observation (abbreviations, Money, Mail address, Internet addresses), disambiguation is carried out. It should be noted that the punctuation symbols gives some information about intonation of the sentence, which one used for great information database to determine the intonation. The general structure of normalizer explained in **Figure 2**. Apparently from the figure this module has several stages [3]:

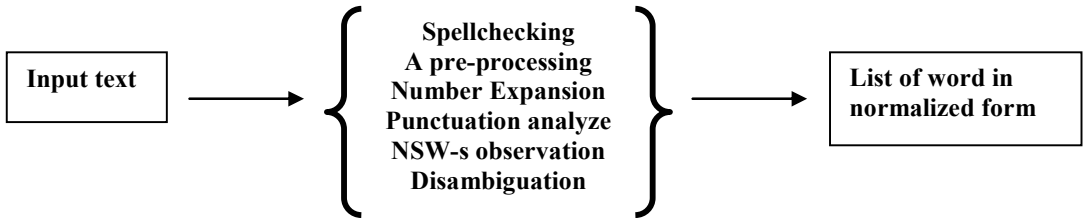


Figure 2. Text Normalization System

II. Linguistic analysis: the syntactic, morphemic analysis.

Linguistic analysis of the text takes place after the normalization process. By using morphological, syntactic characteristics of the Azerbaijan language the text is partitioned into sub-layers. The new text is divided into words and other subdivisions. The morphological and syntactic characteristics of the language used here are taken into consideration. As it was above mentioned the speech can not be only divided into words and word combination but also into phoneme subdivisions like morpheme and diphone. ([4 p. 15-17]).

"Phoneme", "Word" and "Sentence" are the least, middle and biggest units of the language, correspondingly. There are some additions in language to connect these units.

Text and speech signal have clearly defined hierarchical nature. Phoneme, morphem, word, "word-combination"sentence are the elements of the phonetical structure of the language. These units have an hierarchical interconnection – each unit is connected to the next unit. The least unit here is "speech unit" and the biggest one is "sentence". This kind of unit separation is called as the layers of language. Elements on the most top layer could be split into many bottom layer elements.

Example:

Sentence ---- word + word and word-combination.....

Word-combination----- word + word ...

Word --- morpheme + morpheme ...

Morpheme--- phoneme + phoneme ...

phoneme--- ...

The unit of bottom layer is the construction element of the next upper layer. That's why top units may be split into bottom layer units. There units are the 5 main parts of language.

III. Formation of prosodies characteristics – The voice-frequency, accent and rhythmic characteristics are 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 operating information for the subsequent generation of an acoustic signal.

VI. Creation of phonemic transcription

In this stage, it is necessary to attribute data on its pronunciation to each word of the text (each word form), i.e. transform into a chain of phonemes or, in other words, to create it phonemic transcription. In many languages, as well as in Azerbaijani, there are sufficiently regular reading rules – rules of conformity between letters and phonemes (sounds). The rules of reading are very

irregular in English language, and therefore the task of given block for English synthesis becomes more complicated. In any case, there are serious problems at definition of a pronunciation of proper names, the loanwords, new words, acronyms and abbreviations. Due to great volume of the vocabulary and contextual changes of a pronunciation of the same word in a phrase, it is not possible to simply store a transcription for all words of the language.

There are following phonematic content in Azerbaijan language: «i:», «i», «ü», «e», «ö», «ə:», «ə», «a», «a:», «o», «u», «ı», «öü», «ou», «close vowel», «open vowelt», «n», «b», «m», «f», «v», «t», «d», «n», «s», «z», «ş», «ç», «c», «l», «r», «k», «g», «y», «q», «n», «ğ», «x», «h», «mixed consonant». [5, p.16-17].

In general, the phonematic structure of the Azerbaijan language consists of 43 phonemes, including 18 vowels and 25 consonants. In the structure of the given the phoneme two vowels ("close vowel" (ı,i,u,ü) and "open vowel" (a,ə)), and also one vowel ("mixed consonant" phoneme (g,k,ğ)) need a separate explanation

Addressing to acoustic signal database (ASB)

As it was above mentioned an ADB is created without initial login. The core of any speech synthesizer based on concatenative method is formed by acoustic database (ASB) that consists of fragments of real acoustic signal – concatenation elements (CE). The issue of optimal choice of speech units that voicing module will use in creation of acoustic speech base is raised. It is known that the different measures of speech unit exist: allophone, diphone, syllable, phoneme etc [6]. Depending on selected synthesis method these elements can be allophone, diphone, syllable, word, word combination, etc [2, p 227-231] .

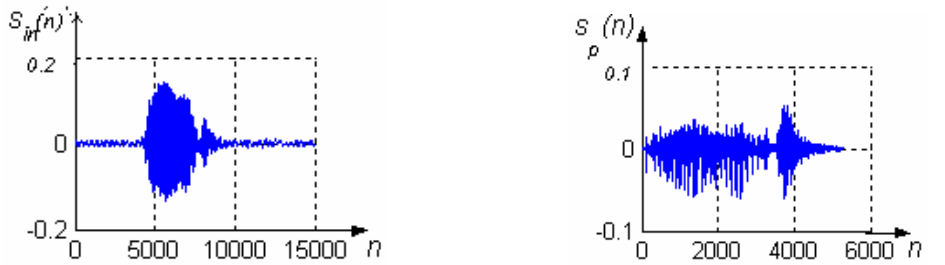
An acoustic signal database (ASD), which consists of fragments of a real acoustic signal - elements of concatenation (EC) is the basis of any system of synthesis of the speech based on concatenation a method. Dimension of these elements can be various depending on a concrete way of synthesis of speech, it can be phonemes, allophones, syllables, diaphones, words and et cetera ([6]).

The elements of concatenation in system developed by us are diaphones and various combination of vowels.

Process of creation of ASD itself consists of several stages.

Stage 1: In the initial stage, the speech database is created based on the basic speech units of the donor speaker. The speech units used in creation ASD saved in WAV format.

Stage 2: Written down speech units cleared of surrounding noise. The algorithm of division of a phrase realization on speech and pauses used for this purpose ([6, p. 37-40]).



Before application of algorithm after application algorithm

Figure 3. The clearing of speech units from surrounding noise

Apparently from **Figure 3** the continuous sites containing speech are definitively allocated after algorithm application.

Stage 3: The amplitudes of speech units are normalized on the following stage.

Stage 4: As described above, AVB plays a main role in speech synthesis. The stored information is used in different modules of synthesis. In our system, KE is stored in .wav format, with 16 kHz frequency. Each wav file includes the next elements of annotations:

- the description of KE
- the count of speech signal parts – N
- energy of speech signal – E
- amplitude of KE - A

- the frequency of crossing zero – Z

These stages are used only in the beginning of process for creation EC for ASD. In the subsequent stages we do not address to them any more.

The voicing block This block consists under stages:

I. Calculation of acoustic parameters of a speech signal

The modification of selected CE takes place based on given prosodic characteristics. 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.

II. 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 [7]:

1. From the last (zero) reading of left (ith) joined wave we count the 3rd reading for which new average value Si3m is calculated from values of ith and jth waves by the formula:

Si3m=1/9*(Si7+...+Si3+...+Si0+Sj0)

2. Then, we reiterate the process according to the following recurrent scheme until we receive the last new value for zero reading of the ith wave:

Si2m=1/9*(Si6+...+Si2+...+Sj0+Sj1)
Si1m=1/9*(Si5+...+Si1+...+Sj1+Sj2)
Si0m=1/9*(Si4+...+Si0+...+Sj2+Sj3)

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

Sj0m=1/9*(Si3+...+Sj0m+...+Sj3+Sj4)
Sj1m=1/9*(Si2+...+Sj1m+...+Sj4+Sj5)
Sj2m=1/9*(Si1+...+Sj2m+...+Sj5+Sj6)

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

Sj3m=1/9*(Si0+...+Sj0...+Sj3m+...+Sj6+Sj7)

III. Voicing of an output signal

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

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 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 voicing is taking place only by one speaker and only for declarative sentence.

References

1. Paul Taylor. Text –to- Speech Synthesis // Cambridge ,2007, 627 p.
2. 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, pp. 47-52.
3. Sharifova A.M, Dadalov V.A, Ibrahimov I.E Text normalization System for Azerbaijan TTS // "International Symposium on Innovations in Intelligent Systems and Applications" INISTA 2009, June 29-July 1, 2009, Trabzon, Turkey, pp. 67-71
4. Djalilov F.A. Morphology of Azerbaijani language // Baku, 1988, 285 p. (in Azerbaijani)
5. Akhundov A. The phonetics of Azerbaijani language. Baku: Maarif, 1984, 392 p. (in Azerbaijani)
6. Sagisaka Y. Spoken Output Technologies. Overview // Survey of the state of the art in human language technology. Cambridge, 1997.
7. Lawrence R. Rabiner and Ronald W. Schafer Introduction to Digital Speech Processing, //California, 2007, 194 p.