ANALYSIS OF APPROACHES TO TEXT TO SPEECH SYNTHESIS AND THEIR APPLICATION TO AZERBAIJANI LANGUAGE

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

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

Two parameters - naturalness and intelligibility of speech- are applied to the description of quality of speech synthesis' system. The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. The ideal speech synthesizer should possess both characteristics: naturalness of sounding and legibility, and these two characteristics try to optimize various ways of synthesis of speech.

There were a different approaches to this problem since XVII century, but it hasn't found it's solution till now. A summary of the progress in speech synthesis, since 1962 year, is given in (fig.1) [1].

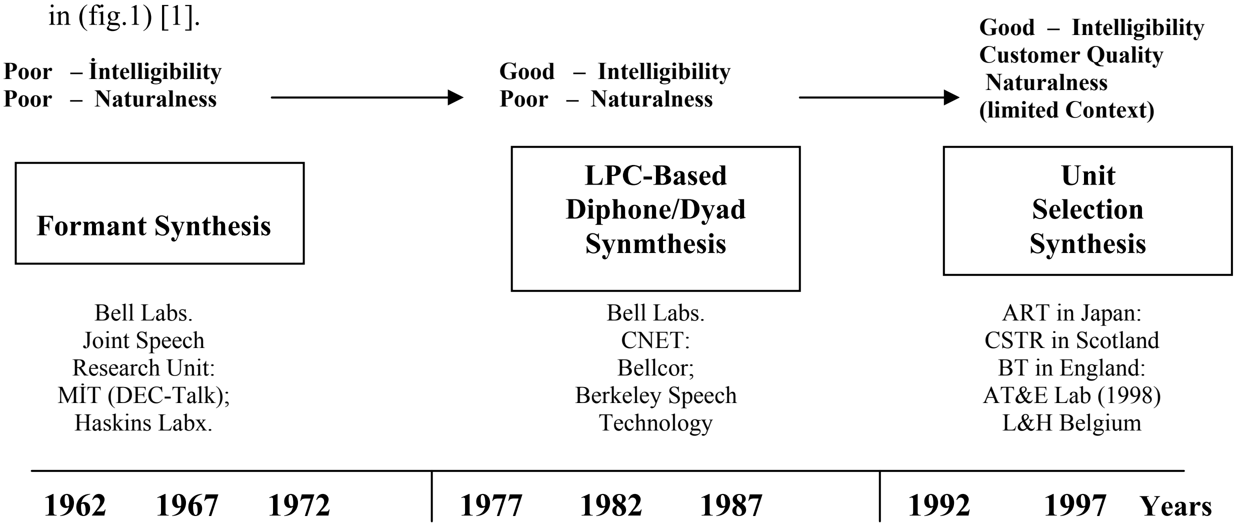


Fig. 1. Summary of the progress in speech synthesis, since 1962 year

This figure shows that there have been 3 generations of speech synthesis systems. During the first generation (between 1962 and 1977) formant synthesis of phonemes using a terminal analog synthesizer was the dominant technology using rules which related the phonetic decomposition of the sentence to formant frequency contours. The synthesis suffered from poor intelligibility and poor naturalness. The second generation of speech synthesis methods (from 1977 to 1992) was based primarily on an LPC representation of sub-word units such as diaphones (half phones). By carefully modeling and representing diaphones units via LPC parameters, it was shown that good intelligibility synthetic speech could be reliably obtained from text input by concatenating the appropriate diaphones units. Although the intelligibility improved dramatically over first generation formant synthesis, the naturalness of the synthetic speech remained low due to the inability of single diaphone units to represent all possible combinations of sound using that diaphone unit. The third generation of speech synthesis technology was the period from 1992 to the present, in which the method of "unit selection synthesis". The resulting synthetic speech from this third generation technology had good intelligibility and naturalness that approached that of human-generated speech.

Currently, different research teams are working on this problem: voice technology club of MDU and PROMPT-"Magic Goody" [2], Sakrament (Misnk city) [3], Microsoft speech SDK [4], TextAssist application based on DecTalk system, Monologue, used in ProVoice system, the product of Centrigram Communications Corporation (USA) for format synthesis - TruVoice. The methods used in initial approaches and their analysis are widely described in [5].

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.

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. But it is necessary to note that, research of generation of one-syllabic words from four letters (stol, dörd) still proceeds and consequently they inclusion into base entirely. The speech units used in creation ASD saved in WAV format. Process of creation of ASD itself consists of several stages:

The database is created in next levels:

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

Stage 2: The speech units from speaker's speech are processed before adding into database. It's done in 2 parts:

a) Speech signal were sampled 16 Hkz frequency and it helps to define the period by the T=10^{-4} precision. Later, the signal is smoothed by 100 Hz and 900 Hz filters for removing the noise of alternating current.

b) Written down speech units cleared of surrounding noise. The algorithm of division of a phrase realization on speech and pauses used for this purpose. It is supposed, that the first 10 shots do not contain a speech signal. On this site average value and a dispersion of each of sizes E_t, Z_t for definition of statistical characteristics of noise are calculated [7].

$$E_s(m) = \sum_{n=m-L+1}^{m} s_p^2(n) \qquad Z_s(m) = \frac{1}{L}\sum_{n=m-L+1}^{m} \frac{|\mathrm{sgn}(s_p(n)) - \mathrm{sgn}(s_p(n-1))|}{2}$$

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

L is the count of shots of speech signal

Then taking into account these characteristics and maximum on realization of a phrase of values E_t, Z_t thresholds T_E for energy limit of a signal and T_Z for number of zero of intensity are calculated. Following formulas have experimentally been chosen:

$$T_E = M(E,10) + 3\sqrt{D(E,10)} \leq k_1 \max_{1\leq t\leq L} E_t,$$

$$T_Z = M(Z,10) + 3\sqrt{D(Z,10)} \leq k_2 \max_{1\leq t\leq L} Z_t,$$

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

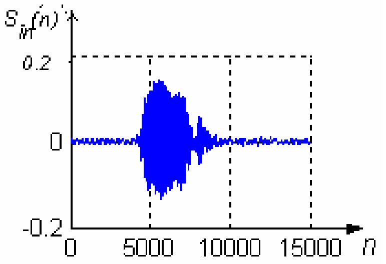
$$D(P,n) = \frac{1}{n-1}\sum_{t=1}^{n}(P_t - M(P,n))^2$$

We should equalize $X^{(t)}$ to binary sign b_t equal 1 if the shot contains speech, and 0 - otherwise. At first it is necessary to mark units shots, on which energy limit is $E_t \geq T_E$, and zero – the other shots. Signs b_t can accept only two values. Therefore the filtration is reduced to that consecutively for $t=h+1,...,L-h$ value b_t replaced with 1, if $\sum_{i=t-h}^{t+h} b_i > h$.

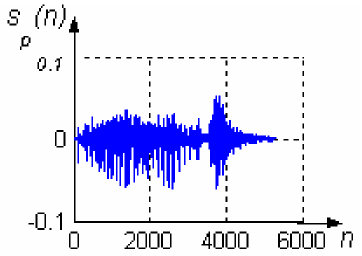
Otherwise value b_t replaced with a zero.

$$b_t = \begin{cases} 1, \ \sum_{i-t-h}^{t+h} b_i > h, & t = h+1 >,...L-h \\ 0, \ \sum_{i-t-h}^{t+h} b_i \leq h, & t = h+1 >,...L-h. \end{cases}$$

In the result the continuous sites containing speech allocated. Further each such site try to expand. For example, the site begins with a shot $X^{(N1)}$ and comes to an end on a shot $X^{(N2)}$. Move to the left from $X^{(N1)}$ (to the right from $X^{(N2)}$) and compare number of zeros of intensity Z_t to threshold T_Z. This moving should not exceed 20 shots to the left of $X^{(N1)}$ (to the right of $X^{(N2)}$). If Z_t has exceeded a threshold in three and more times the beginning of a speech site transferred to that place where Z_t time exceeds a threshold for the first time. Otherwise a shot $X^{(N1)}$ could be considers as the site beginning. The same arrive with $X^{(N2)}$. If two sites are blocked, they could be uniting into one. Thus, the continuous sites containing speech are allocated definitively. We will name such sites as realizations of words.



Before application of algorithm



after application algorithm

Apparently from drawings the continuous sites containing speech are definitively allocated after algorithm application.

Stage 3: 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
-

Stage 4: At a following stage another corresponding variant is created on the basis of everyone EC. In spite of the fact that this increases the quantity of ASD elements but at the same time helps to reduce quantity of modules of generation of a target signal.

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 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.2) [8]:

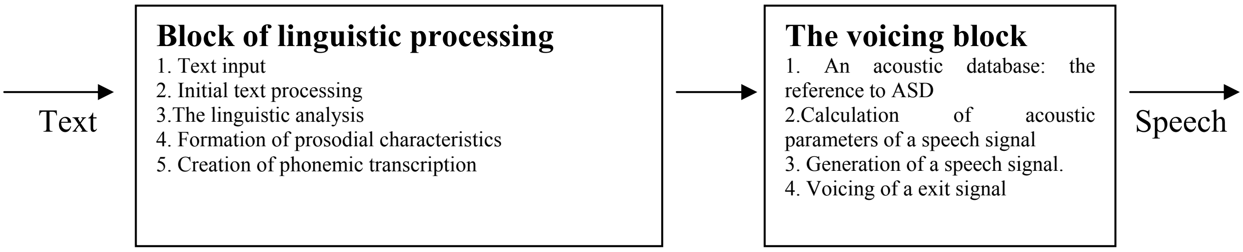

Text
Block of linguistic processing
1. Text input
2. Initial text processing
3.The linguistic analysis
4. Formation of prosodial characteristics
5. Creation of phonemic transcription
The voicing block
1. An acoustic database: the reference to ASD
2.Calculation of acoustic parameters of a speech signal
3. Generation of a speech signal.
4. Voicing of a exit signal
Speech

Fig. 2. Structure scheme of our system

Apparently from the scheme two blocks lay on a basis of system: the block of linguistic processing and the voicing module. It is possible to allocate two main blocks in our synthesizer: the block of linguistic text processing and the block of voicing or actually formation of a speech signal.

The sounded text can be entered in any form. The size or font type here does not matter. The main requirement is that the text must in Azerbaijan language. After text normalization, the linguistic analysis is done, where text is split into words and other parts. Phonemic transcription builds to the entrance text the sound transcription corresponding to standard rules operating the Azerbaijan reading language.

In system developed by us voicing of words of any text in the Azerbaijan language is carried out on the basis of the aforesaid with the limited base. It is necessary to note that there are some unsolved questions for instance work on intonation is not finished because segmentation was made manually and thus there are appreciable hindrances in scoring. Further it is planned will apply independent segmentation and will improve quality of synthesis.

References

1. Lawrence R. Rabiner və Ronald W. Schafer "Introduction to Digital Speech Processing "
2. www.prompt.ru
3. www.sakrament.com/products/tts
4. www.alantts.com/accueil.html
5. K.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)
6. «Vocal structure of the Azerbaijan language» // Bakı, 1989. (in Azerbaijani)
7. Sagisaka Y. Spoken Output Technologies. Overview//Survey of the state of the art in human language technology. Cambridge, 1997.
8. Sharifova A.M The Computer Synhtesis of the Azerbaijan Speech / (Azerbaijani). Application of information-communication technologies in science and education. International conference. Baku, 2007. Volume II, p. 47-52.