The development of speech technologies in Belarus

BM Lobanov

Introduction

Of all living things only a man has the gift of speech, by which he was able to to develop his intellectual abilities, and, according to many philosophers, become human. Something similar is happening before our eyes with the computer, which is intensely possessing a wide range of speech technologies: from coding telephone signals to the synthesis, recognition and speech understanding. We really hope that the further development of human-machine systems of verbal communication will steadily lead to a permanent improvement of the language "writing" of computers and, as a result, a continuous approximation of their intellectual abilities to human being. Here is the quotation of Academician NAS Viačaslava Usievaladaviča Ivanov from "Linguistics of the third millennium":

"We are the first in the history of the species who begin to make extensive use of technical devices to talk - in other words, not only to produce tools (by which the man differs from animals), **but also to teach them to our language** (as we begin to differ from all the people who lived before)**".**

However, we still need to overcome many difficulties towards the creation of virtually demanded systems that implement speech technologies. Here is a statement of Microsoft Vice-President G. Sinha: "... *Bill Gates is upset by how long speech recognition technology needs to" break through " its way. After all, the corporation still invests the project at least since 1991. ...* "(According to CNET News).

Modern speech technologies can be divided into the following types:

- Speech recognition (including commands, keywords, word strings, spontaneous speech).
- Text-to-speech synthesis(including vociferious, personalized, emphatic, multimodal synthesis).

- Voice biometrics (identification and verification of the person's voice and speech).Fig. 1 illustrates two spectral images of the w
- Diagnosis of personal characteristics and state of the speaker).
- Medical diagnostics (diagnosis of the voice and speech diseases).

Currently, the speech technology application is constantly growing and affects more and more spheres of human activity. For example, in 2012 the Sheremetyevo international airport in Moscow announced the launch of a voice information system on the status of the flight. Service is available at +7 (495) 956-46-66. This project induced natural interest at users: a convenient friendly interface distinguishes it from other automated services, which work in local companies.

1. The initial stage of speech research development.

The modern history of speech research in the USSR starts from the mid 60s of the last century, when All-Union Workshop on the automatic recognition of auditory images (ARUA) began their work. The staff counted nearly 300 participants. The initial stage of the speech research development in Belarus also belongs to this period. In the scientific laboratory of the radio receivers department of the Minsk Radio Engineering Institute a group headed by the author of this article was organized for research of speech signals in 1965. At that time they were M.P. Dziehciaroŭ, B.V. Pančanka, M. Faciejeŭ and others, who for long time, and even now have still been working in this direction.

The first studies were associated with the development of the general principles of the speech signals analysis and the allocation of informative features, which would introduce a continuous sequence of phonetic speech signal segments. The results of these studies were summarized in the thesis of the author "Some questions of speech signals analysis", defended in 1968 in Moscow National research Radio Institute. The most important results of this work were later published in reputable international journals [1-2]. On the basis of these studies a relatively simple voice recognition device "Sesame - 2" was developed in the Soviet Union. In 1968 this progect received a silver medal of

VDNKh USSR. The device composed two parts: a speech signal analyzer such as "Voice," "Noisy", "vowel", etc., and the counter number of signs in the speech command. There was a fairly high recognition reliability of 20 commands (including the numbers), regardless of the speaker's voice, the volume and speech tempo. In the same period, a specialized equipment for experimental phonetic speech research has been developed: the analyzer of dynamic spectra and intonograf. In the following years numerous studies were conducted with the help of these tools in the phonetic laboratories of the Institute of Linguistics NSA and the Minsk Institute of Foreign Languages.

The research of the speech dynamic spectra gave impetus to the development of nonlinear methods of matching the oral recognized words, including their standards. Speech spectral images, unlike ordinary visual images of objects may be subjected to uncontrolled distortion of an hour axis. Fig. 1 illustrates two spectral images of the word "car", uttered by the same speaker at different tempo. The figure shows that at accelerated tempo (lower spectrogram) with an overall reduction of the length by 30% the sounds duration of the stressed syllable " ШЫ " was virtually unchanged, while the sounds of the "T" and "H" in unstressed syllables were reduced by more than 2 times. This example demonstrates that it is not enough to have a simple scaling of spectral images to their reliable recognition. The situation, so to speak, like the one that

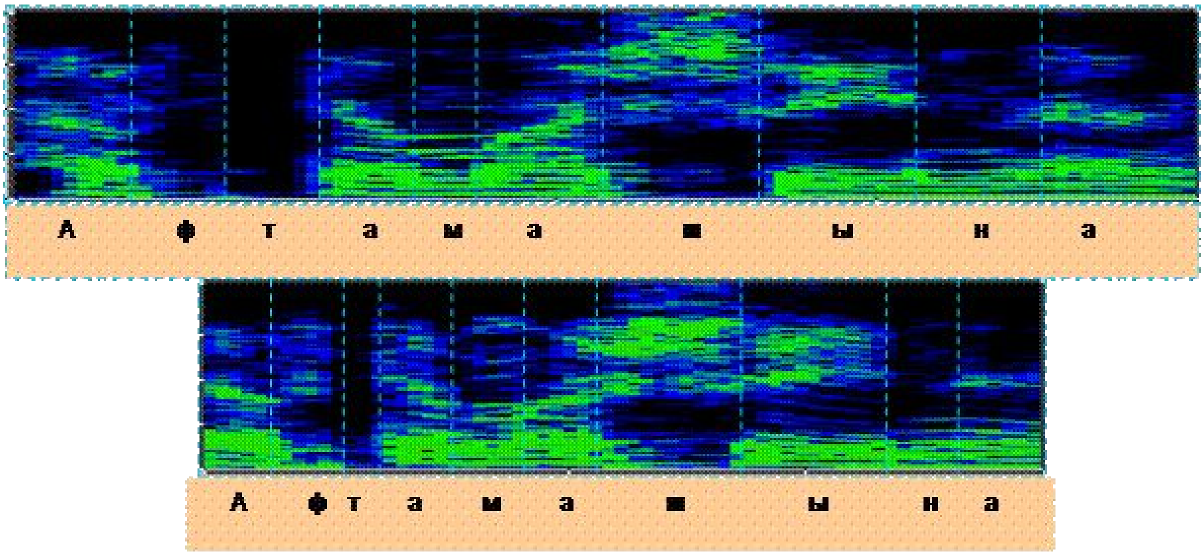


Figure 1 - Spectral Image speech

would be in a "distorting mirror" in recognition of visual images. The solution of the fundamental problems of speech recognition is related to the non-linear distortion of the time axis, suggested independently and almost simultaneously by H.S. Sluckier (the Moscow National research Radio Institute) and T. Vinciuk (the Institute of Cybernetics USSR) in the late 60s. The essence of the proposed solution was to find a method by dynamic programming (DP method) of optimal path on a matrix of distances between the local time reference vectors of distinctive and reference spectra. In 1969, the author, together with the staff of the Moscow National research Radio Institute published an article [3], which describes further development of DP-method for the extremely important practical case where the boundary distinguishing words are unknown, e.i. solutions to the problem of detection and identification of sound combinations in a continuous speech signal. DP method has been widely recognized by foreign researchers and, along with the method of hidden Markov Models (HMM), it is still used in modern speech recognition systems.

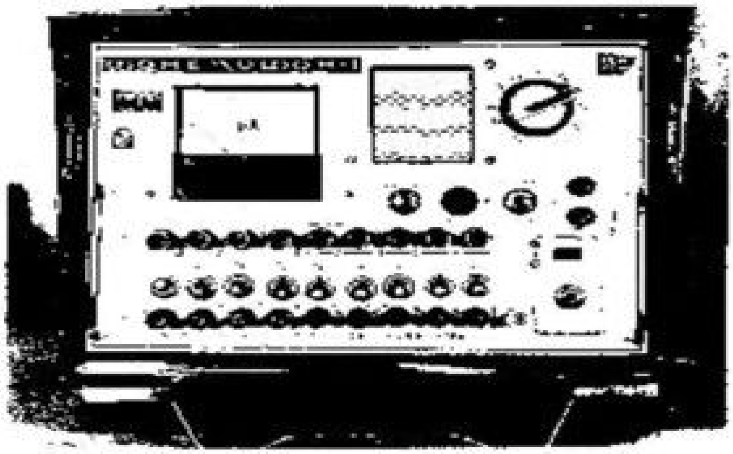


Fig. 2 - the Synthesizer "Fanemafon-1"

The works on creation of speech synthesizers also belong to the late 60s . The impetus of that was the recognition that the development and research of speech synthesis models are a direct path to a more detailed knowledge of education and properties of the speech signal. Thanks to this it became possible to build a more sophisticated analysis algorithms and speech recognition. An important role in the development of the world's technological level of speech synthesis was a scientific training of the author in the laboratory of Professor Lorenc (Edinburgh University) in 1970. At that time one of the first formant models of speech signals synthesis was developed. The scientists recieved first produced high-quality samples of Russian synthesized speech with the help of that synthesizer.

The first, but not completely perfect model of Russian text-to-speech synthesizer "FANEMAFON-1" (see. Figure 2)appeared in the early 70-ies. The success of its creation was primarily associated with the development of new methods for hardware implementation of speech signals formant synthesis. The principle of speech signals formant in synthesis is based on modeling the properties of excitation sources (voice and noise) and resonant (formant) characteristics of human speech organs. As a result of experimental studies a complete set of formant "portraits" of phonemes was created, which caused the first Russian speech synthesis of random text. Later, an improved version of the synthesizer - "Fanemafon-2" with the additional conversion unit "phoneme - alafon" was made.

2. The history of the "average" years (70 - 80 years)

A decisive role in the orientation of speech technology usage has played the establishment of the Laboratory of processing speech signals in the central department of the Minsk Research Institute of Communications. In 1976 the All-Union seminar ARSV-9, held in Minsk, firstly demonstrated a prototype of an automatic telephone inquiry service with a synthesized voice response. Since the early 80s for a long time automatic telephone calls debtors worked for long-distance calls in Minsk (the authors of this development also recieved calls). By the mid 80s, this system has been implemented in many cities of the USSR - from Brest to Petropavlovsk-Kamchatsky.

Figure 3 - "Fonematon-3" at the exhibition in Geneva

The successful implementation of speech synthesis was preceded by a long work both for the improvement of the synthesized speech indicators' quality and technology implementation as a new class of computer peripherals. The main disadvantages of the first "Fonemafon-1, 2" models were low legibility and quality of the synthesized speech, which caused, first of all, a very simplified models of the sound interaction during speech production (effects of coarticulation and reduction) and insufficiently modified models of speech intonation in the text. The next model - "Fonemafon-3" introduced additional blocks of articulation and speechintonation modelling processes, which significantly increased the quality parameters of the synthesized speech.

In 1979, "FONEMAFON-3" was shown at the World Exhibition "Telecom-79" in Geneva (see. Figure 3). Renowned science fiction Arthur C. Clarke, visitting the USSR pavilion and getting acquianted with the speech synthesizer, wrote in the guestbook: *"You have surpassed my imagination of the movie" Space Odyssey - 2001 ". The Swiss newspaper" Observer "published an article:" Now the Russian study foreign languages with the help of PC speaker "*.

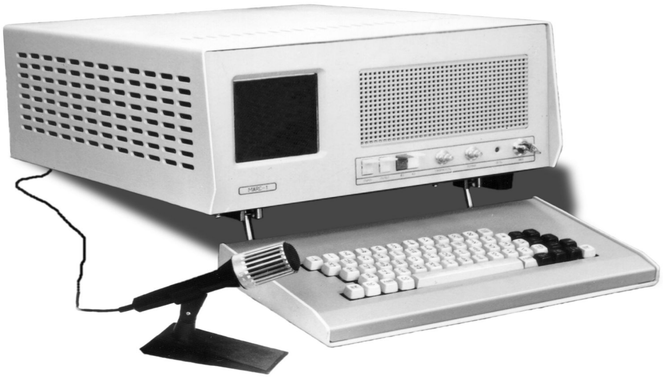
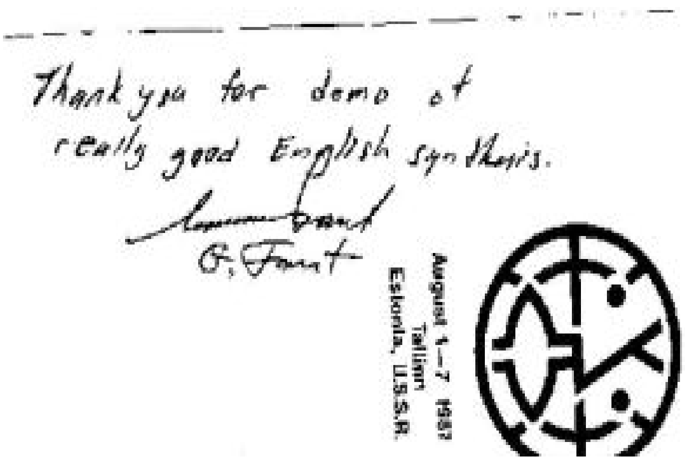


Figure 4 - The voice terminal "MARS"

An important role in the creation of the industrial development of speech synthesis played a fully digital model of a speech synthesizer "Fonemafon-4", a speech recognition system "Sesame" and the voice terminal - "Mars" (Fig. 4). For the first time their serial production in the USSR (1983) has been set at "Quartz" in Kaliningrad due to the enthusiasm of the design department employees, which was headed by Valery Afanasiev. The laboratory staff of M.Dziehciarov and IV.Šaternik played the key role in the creation of that models. The voice terminal "MARS" firstly integrated features of speech recognition and synthesis. DP-verbal method of decision-making on the basis of a formant speech attributes of speech signal became the frame of the speech recognition algorithms. The prototypes of systems "Sesame" and "Mars" were made on the basis of the microprocessor. The application parameters did not come short of the best foreign analogues at that time. The originality of the technical solutions used in the creation of "Fonemafon", "Sesame" and "Mars" systems was protected by numerous copyright certificates of the USSR for inventions.

The final formulation, theoretical and experimental development of a common linguistic and acoustic approach to the problem of speech synthesis on the text, its implementation in the form of technical systems and practical implementation in the automated systems of control and communication referred to the beginning of 1984. The results of these studies were summarized in the doctoral thesis of the author "Methods of utomatic synthesis of Russian speech on the text", defended in 1984 at the Institute of Electronics and Computer Science, Academy of Sciences of the Latvian SSR. Later, the results have been adapted for speech synthesis systems in other European languages. In particular, by 1987 thanks to the collaboration with Professor of the Minsk Institute of Foreign Languages Je. Karnieŭskaja a Russian version of the synthesizer was developed [4], which was shown at the World Congress of Phonetic Sciences, and has been praised by the English-speaking professionals. That was a facsimile response of one of the most prominent researchers of speech in the world Gunnar Fanta (see. Fig. 5).



Thank you for demo of
really good English synthesis.
G. Fant

August 1–7 1987
Tallinn
Estonia, U.S.S.R.

Figure 5 - The facsimile response H. Fanta

3. Recent History

In 1988, in CTI NSA the speech synthesis and reognition laboratory was established.The writer of the article was invited to head the laboratory. As many people still remember, the end of the 1980s was marked by the appearance of the first personal computers, so it was natural that all the plans of the laboratory were related to the PC hardware system of voice input-output. Formant method, playing a key role in speech synthesis systems on text for a long time, became unsuitable because of the need for a large amount of calculations in real time, which was not available for PC at the time. At the end of the 80s a new microwave (MX) method was offered for the synthesis of speech signals [5], in which pre-prepared set of microwaves of natural speech was used instead of computing formant vibrations (sound waves). The set consisted of the microwave signal segments equal to the period duration, and their number was required to generate any sound of speech up to several hundred. MX-method implemented in the synthesizer "FONEMAFON-5" by Aliaksandr Ivanoŭ was in the form of specialized software synthesizer based on the work with the internal sound card or with self-contained device that was connected to the RS-232 port (see. Fig. 6). Strange to many, the compactness of its software (total 64K bytes)allowed to provide first IBM PC-XT and even the domestic PC ES1840 with speech. The Speech Synthesizer was in demand in many practical applications, but particularly well it was used by blind PC users (more than one hundred sets of specialized hardware and software for the blind have been created and distributed by the laboratory of Hieorh Losik in Russia, Ukraine and Belarus in the first half of the 90s). Its sound, quite legible, can be heard on the Internet or in CD ROM "Talking Mouse." Later, based on the MX-method the versions of the Czech and Polish languages were developed (approximately 3 months of staying in the country at the invitation of investigators), as well as a self-combained monoboard module of speech synthesis, the Ukrainian-language version of which has long been working on a line of the Kiev Metro.

The difficult economic situation in the country in the mid-90s forced to seek sources of funding abroad, primarily in the form of joint international projects. The first international project **"The bilingual speech synthesis - German / Russian"** (1995-1996) was implemented in cooperation with the Dresden University of Technology and funded by the German Science Foundation FTU Karlsruhe.

The next project was **"The Analysis of natural language and speech"** (1996-1997), which was carried out jointly with Saarbrukensk University (Germany), Manchester University (UK) and the Institute for Information Transmission Problems (Russia), and funded by the European Foundation INTAS. The participation in this project was associated with the further development of models for speech synthesis by integrating them into the system of natural language processing techniques of computational linguistics.

An important role in the integration into European community by the Belarusian researchers in the field of linguistics and speech made an international project **"The European network development of linguistics and speech towards the east"** (1997-1998), funded by the European Fund COPERNICUS. Since 1998 the speech synthesis and recognition laboratory UIIP is the focal point of the network in Belarus.

In addition to European research organizations interested in cooperation with the laboratory we have found some commercial organizations. In 1996, the French firm "Sextant Avionics" proposed to implement research project **"The recognition of voice commands in terms of a noise in the cockpit".** The project was financed by the French Ministry of Defence. In spite of the exceptional complexity of the task, the project was successfully completed in 1997 and accepted by the customer. Main scientific results of the project are presented in [6]. Another commercial project was the development of a **"Smart phone attendant"**, which was carried out in 1997 under a contract with NovCom NV (USA). The essence of the project was to solve the problem of the person's name recognition, spoken on the phone, and other service information so that the system can perform the functions of a telephone attendant. The project was completed in 1999 and the main research results are published in refference [7]. A key role in the implementation of these international projects has played aboratory staff: Tatiana Levkovskaya l, Alexander Ivanov and Andrei Kubashyn. The work on these projects gave an impetus to the revival and the further development of speech recognition algorithms proposed in 1969 [3].

4. Modern researches and developments

Since 2000, the research interests of the Laboratory move again from recognition to the speech synthesis problems. A new multi-wave model of text-to-speech synthesis was tdeveloped, which provides high quality and personalization of the synthesized speech. The employees of the laboratory, Vitaĺ Kisialioŭ i Zmicier Žadziniec made significant contribution to the creation of the speech synthesizer "Multyfon", based on this model. At the same time theoretically new scientific direction in speech research - cloning of computer speech - was substantiated and developed. It was nesessary for high-quality of text-to-speech synthesis to approximate the voice and manner of a particular person. At the same time the technology and special software cloning of acoustic, phonetic and intonation characteristics of human speech were produced. The laboratory staff Andrej Davydaŭ i Lilija Cyruĺnik brought significant contribution to the creation of software cloning speech projects "Fonoklonatar" and "Intoklonatar" .

The obtained in the early 2000s theoretical and experimental results have allowed to carry out a number of joint application projects.

In 2005 - 2007 an international INTAS-project **"The Development of vociferous and multilingual system synthesis and speech recognition" (Belarusian, Polish, Russian languages)** was hold. Among the Participants of the project were Belarus, Germany, Poland, Russia. The project resulted in the foundation for the creation of a multilingual system of automatic speech translation for three Slavic languages.

In 2006 - 2007 on request of the LLC "Invaservis" (Minsk, Belarus) a system **"Reader"** was developed, focused on the creation and editing of "speaking" textbooks for blind students of special Belarussian schools.

In 2007 - 2009 BRFFR and RFFR made a project **"The Development of multimedia system of audio-visual speech synthesis"** - the so-called "personalized talking head" (see. fig. 7). The system significantly improved the quality of speech perception in number of practical applications (particularly in remote training).

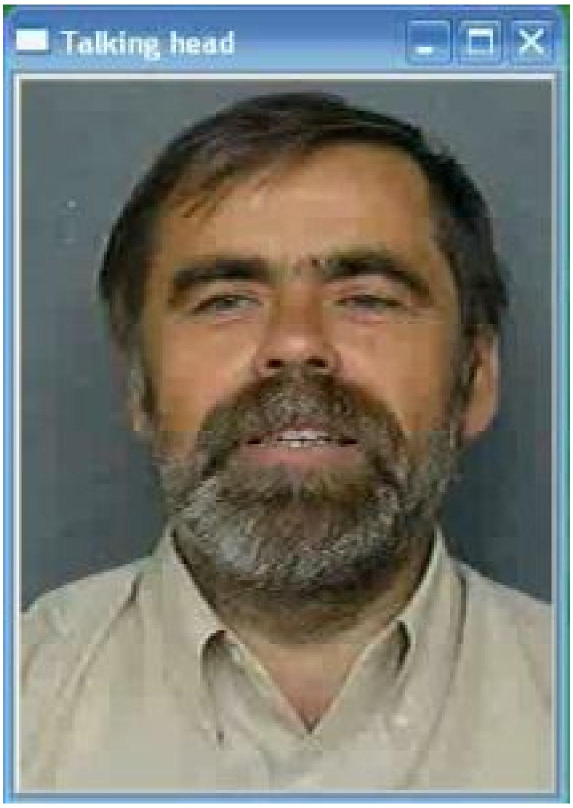


Figure 7 - "Talking Head"

In 2009 - 2011 BRFFR and RFFR made a project **"The emphatic speech synthesis on the basis of deep text parsing ".** The project resulted in the first bilingual synthesis system of emphatic speech in Russian and Belarusian, which was shown at a number of scientific and technical exhibitions (see. Fig. 8).

Figure 8 - Belarussian Prime Minister Mikhail Myasnikovich and

vice president of the Russian Academy of Sciences Žores ALFIORAŬ

listen report to a young scientist Juraś HIECEVIČ

about the synthesis development of the Belarusian and Russian speech. ("Republic" 16.11.2011)

At the same time a large amount of researches aimed at the development of Belarusian speech synthesis model were worked out. The main role in the creation of the Belarusian speech synthesizer played the works of YS Getsevich, an acting head of the of speech synthesis and recognition laboratory.

In 2011 - 2012 by order of the Yakut State University the world's first Yakut speech synthesis system was developed . The Yakut speech synthesizer is used in computer training to the Yakut language, as well as a sound system for the blind.

You can see two examples of modern laboratory developments below.

4.1. Computer cloning of personal voice and diction

Long-term studies of the twentieth century helped to create synthesizers, which provide the speech quality and emphasis, quite suitable for a wide range of practical applications. However, despite all efforts, the synthesized speech was still far from the quality of the natural with machine recognizable accent. The reason for this was not so much the level of our knowledge about the processes of speech generation and phonetics as a lack of computer resources of that time. Now we are not limited with disk and internal memory, volume demand computing, and it has become a prerequisite to the problem solution of the speech synthesis on the text as close as possible to the voice and manner of a particular person.

This formulation of the task resembles the well known problem of biological cloning. In this case, in contrast to the classical problem of cloning it was an attempt to create a close copy, not the biological but computer, not the human being as a whole, but only one of its intellectual functions: for example reading any orphografic text. The objective is to maximize the preservation of personal acoustic voice characteristics, pronunciation and phonetic peculiarities of personal accent and prosodic speech (melody, rhythm, dynamics).

General structural diagram of the personalized speech synthesizer on the basis of which the cloning processes is shown in fig.9.

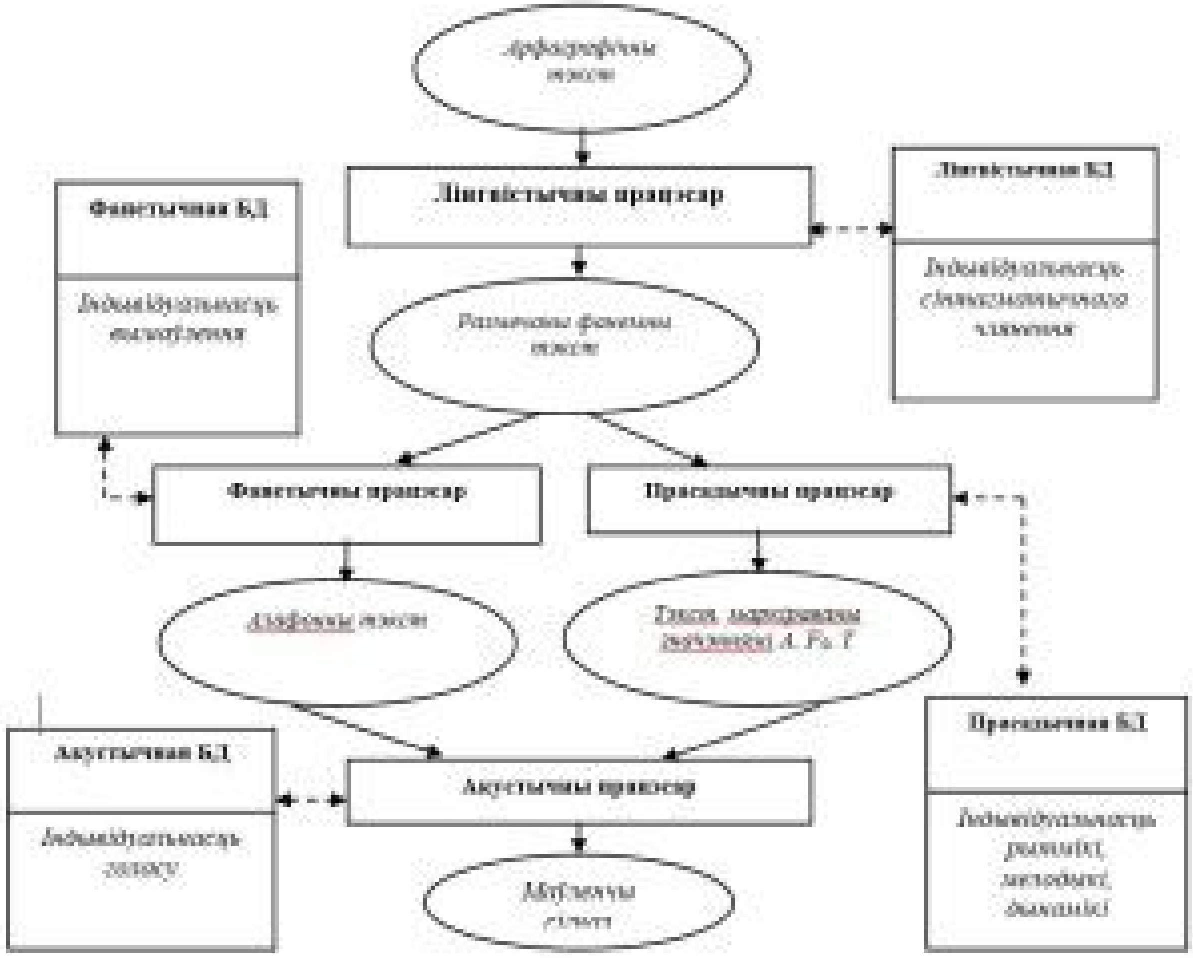


Figure 9 - The diagram of personalized speech synthesizer

The Synthesizer consists of four processors: linguistic, prosodic, phonetic and acoustic. Each processor uses a specialized data base (DB) for convertion. The general rules of language (linguistic, prosodic, phonetic, acoustic) and the rules relating to the individual characteristics of the speaker's voice and speech are laid down in these databases.

Cloning of the voice acoustic characteristics. Personal acoustic characteristics of the human voice are caused by many factors such as the anatomical features of the structure and speech apparatus (larynx, vocal cords, pharynx, oral cavity and others), the interaction of of the vocal cords vibrations and cavities of the vocal apparatus and much more. As you know, the attempts to simulate the characteristics of the personal voice in systems "text - speech" on the basis of physiological and acoustic speech production modeling because of their extreme complexity has not yet led to significant results. In this regard, the most reasonable to use segments of natural speech wave as a minimum "genetic material" for cloning voice. The preferable segment is allophone as the most studied phonetic substance. A limited set of allophones is capable of oral speech generation of any text. In this case the sound wave carries all the essential personal voice features which a particular allophone contains.

Cloning of personal phonetic pronunciation features. Unlike personal acoustic characteristics of the voice mainly due to static parameters of the vocal apparatus, the phonetic pronunciation is conditioned mainly due to the dynamics of articulation movements performed during the speech. Specific to the individual tempo of articulatory movements, particular characteristics of sound articulation (such as / R /), regional or foreign accent are responsible for the emergence of unique positional and combinatorial shades of phonemes and creation of allophone system. Thus, the successful cloning of the personal characteristics of the phonetic pronunciation can be achieved by simulating the characteristics of phoneme-allophone conversion, distinctive for any person during the speech.

Cloning of personal prosodic speech characteristics. Complex prosodic (intonation) characteristics of speech, which include the melody, rhythm and energy, reproduce regularities of changes in time of the pitch frequenc, sounds' duration and amplitude of the vocal expression. The nature of these changes is determined not only by the specific text, but also the personal style of its reading. The solution of of prosodic speech characteristics cloning is a fairly complete set of personal intonation "portraits" of speech.

Cloning technology and its applications. For the successful cloning of the personal voice characteristics and diction, we must create a fairly complete set of allophone sound waves and intonation speech "portraits". If the speaker, who was physically cloned, is available, he reads a specially designed compact sound array of words and text passages in the studio or in normal conditions. Otherwise, developers use the existing record of his voice on the radio, television and others. The first results of cloning (for example, personal voice and diction of the author of this article) have been received in 2000. In 2001, a clone of a female voice was obtained and by, the end of 2005, a set of clones of 3 male and 2 female voices were already made. Currently established computer technology of cloning phonetic-acoustic and prosodic features of speech would greatly automate and speed up the process of creating speech cloning. The main scientific and practical results of the computer speech synthesis and cloning are reflected in the monograph [8].

Now we will note some potential commercial and practical aspects of computer cloning. Surely there is a large number of computer users who want their PC to speak with his own voice, or, for example, a voice close to him, or a favorite actor. An animated voice of great people, who have long passed away, can also be an interesting project. It may be reproduced by their gramophone or studio recordings. The development of voice cloning may present drastic means to combat so-called terrorism by telephone, providing identification of the individual voice by automatically comparing the operational voice with voice database filled by clones of potential offenders.

4.2. A computer model of speech virtual interlocutor

The computer model of oral speech virtual interlocutor (SYSTEM "REVIRS") is a new development of speech synthesis and recognition laboratory, which integrates the original scientific and technical solutions obtained by laboratory staff in recent years. The system allows you to create scripts of dialogues for a variety of applications and implement them by means of oral speech of man-machine communication. REVIRS is unique because it realizes:

- reliable recognition of the query keyword in a *continuous stream of speech*;
- multiannouncer recognition of keywords *in terms of acoustic noise and distortion*;
- *vociferous* speech synthesis of arbitrary text;

- the possibility of *"cloning" the voice*s during speech synthesis;
- *duplex mode in real time* (possibility of interrupting voice response).

Block diagram of the system is shown in Fig. 10

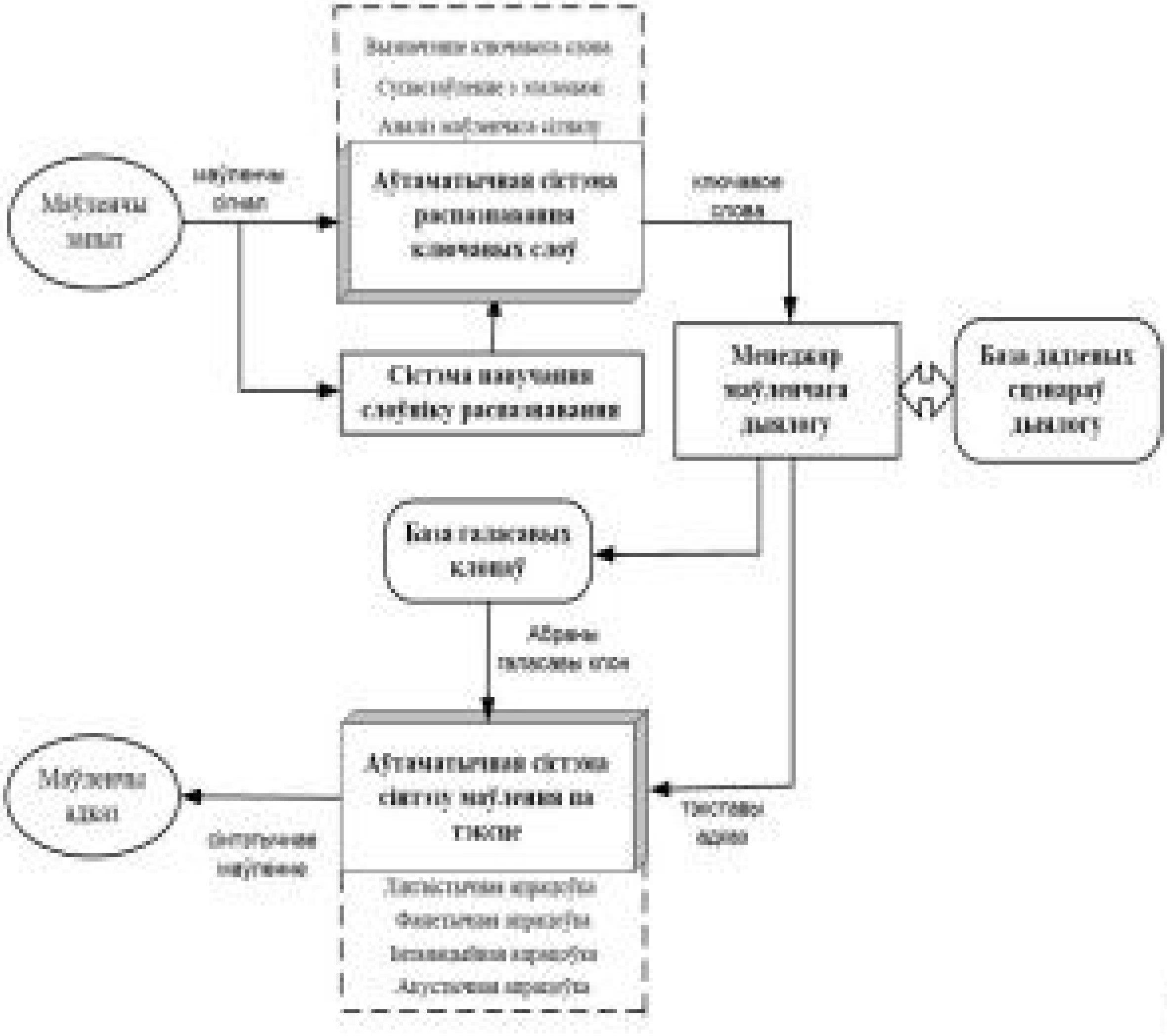

Вылучэнне ключавога слова
Супастаўленне з эталонам
Аналіз маўленчага сігналу
Маўленчы запыт
маўленчы сігнал
Аўтаматычная сістэма распазнавання ключавых слоў
ключавое слова
Сістэма навучання слоўніку распазнавання
Менеджар маўленчага дыялогу
База дадзеных сцэнарыяў дыялогу
База галасавых клонаў
Абраны галасавы клон
Маўленчы адказ
Аўтаматычная сістэма сінтэзу маўлення па тэксце
тэкставы адказ
Сінтэзаванае маўленне
Лінгвістычная апрацоўка
Фанетычная апрацоўка
Інтанацыйная апрацоўка
Акустычная апрацоўка

Figure 10 - Block diagram of the system REVIRS

The speech signal is input to a speech recognition system, which analyzes the signal informative features, their comparison with the standards of keywords, identification and makes decision on the spoken word. If recognizing unit have found a keyword, it is transmited to the manager of the speech dialogue, which forms a text response. The manager also selects the voice clone for synthesis of voice response. The selected clone and answer text are input to speech synthesis system, which provides linguistic, intonatsion, phonetic and acoustic processing, as a result the text is converted to a pronounced, wich perfoms the features of voice clone.

Conclusion

In this section we set out a brief outline of the history and development of speech technologies in Belarus over 40 years. Of course, it is not the whole list of all scientific and technical investigations. It is even impossible to mention the work of all the authors who have contributed to the development of this branch of knowledge in Belarus (in total they have published more than 300 scientific papers, including 5 monographs).

Despite the fact that we have covered a long way acquiring sensitive scientific and practical results, I believe that the history of speech research in Belarus is not completed. Like 40 years ago, the explanation of human speech is captivating for young researchers, and possible applications of speech technology are always relevant.

References

1. Lobanov B.M. More About Speech Signal and the Main Principles of its Analysis // ieee Transactions on Audio and Electroacoustics.- 1970, N 3. – P. 316-318
2. Lobanov B.M. Classification of Russian Vowels Spoken by Different Speakers // The Journal of the Acoustical Society of America.- 1971, N 2 (2). – P. 521 - 524
3. Лобанов Б.М., Слуцкер Г.С., Тизик А.П. Автоматическое распознавание звукосочетаний в текущем речевом потоке // Труды НИИР, Москва, 1969.- С. 67 - 75
4. Lobanov B.M. The Phonemophon Text-to-Speech System // Proceedings of the XI-th International Congress of Phonetic Sciences, Tallin, 1987.- 100 - 104
5. Lobanov B.M, Karnevskaya E.B. MW - Speech Synthesis from Text // Proceedings of the XII International Congress of Phonetic Sciences.- Aix-en-Provense, Franse 1991.-P. 387 - 391

6. Lobanov B.M., Levkovskaya T.V. Continuous Speech Recognizer for Aircraft Application // Proceedings of the 2nd International Workshop “Speech and Computer” – SPECOM’97.-Cluj-Napoca, 1997.-P. 97-102.
7. Lobanov B.M. et al. An Intelligent Answering System Using Speech Recognition // Proceedings of the 5th European Conference on Speech Communication and Technology – EUROSPEECH’97. V. 4.-Rhodes-Greece, 1997.- P 1803-1806.
8. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи // Минск: Белорусская Наука, 2008. – 342 с.