Voice Control of Robots and Mobile Machinery

Ruslan Sergeevich Shokhirev
Institute of Technical Cybernetics and Informatics,
Kazan State Technical University,
Kazan, Russia
ruslan.shohirev@gmail.com

***Abstract*—I develop a system of Russian voice commands recognition. Wavelet transformation is used to analyze the signal key characteristics. Kohonen neural network is used to recognize spoken sound based on these characteristics. Besides, I'll give a brief overview of the current state of the problem of speech recognition**

Keywords—speech recognition; voice control; wavelet; neural network

I. INTRODUCTION

One of the ways to improve the human-machine interaction is using of voice control interface. This approach allows to control activities of the technical devices in situations where the operator's hands are busy another work, as well as people with disabilities. In addition, this approach can be used to improve ease of use device.

There are many approaches to solving the problem of voice control at the present moment. There are many speech recognition systems in Russia and in the world. The main problems of modern Russian language recognition systems include the following:

1) Phonetics and semantics of the Russian language be formalized much worse compared with the English language.
2) There has been a little research and produced a few works on the subject of speech recognition in Russia since the USSR. This complicates the task of creating systems of recognition, because there is no well documented theoretical basis and description of modern approaches to solving this problem[1].
3) Existing systems that recognize the Russian language are often built on the principle of client-server, which makes them dependent on availability and quality of communication from global network of Internet. In addition it often puts the user in relation to corporations that own these servers. This is not always possible from the point of view of safety.

The most popular speech recognition systems today can be called the client-server solutions from the corporations Google and Apple: Google Voice Search and Apple Siri. These systems are similar in their work and are based on distributed cloud computing made in corporate date-centers. Systems have extensive vocabularies in different languages, including Russian. The number of recognizable words by Google is hundreds of billion[2]. The main application of these systems is mobile devices and gadgets. Disadvantages are the dependence on the Internet and corporate data centers.

Both foreign and Russian companies are currently engaged in a number of studies related to speech recognition. However, to date there is no public system of Russian speech recognition.

II. STATEMENT OF THE PROBLEM

One of the applications of speech recognition systems is the control of mobile machines. At present, manual data input from the keyboard, and specialized controllers – joystick are widely used to interaction with mobile machinery. However, there are situations where it is impossible or inefficient to use these interfaces for control. Operator's hands may be busy doing other work. For example, voice commands can be used to control external video cameras during outdoor work on the space station, while the operator's hands are operating the manipulators. Just such systems can be used to control various household devices by people with limited physical abilities. In such systems the reliability speech recognition and independence of the system from external factors plays an important role, even at the expense of the number of recognizable words. On the other side the recognition of spontaneous speech does not required for these systems. They are used to enter predefined control commands in most cases.

Thus had the following research objectives:

1) The developed system must be autonomous and independent. I.e. all calculations related to the speech recognition must be made directly on the device, or on the local server.
2) The developed system should have a limited vocabulary of recognizable words. The system must be universal, namely: adding and removing commands must be performed as quickly as possible.

III. COMPOSITION OF SPEECH RECOGNITION SYSTEMS

A. General Scheme

In the general case speech recognition system (SRS) can be represented by scheme in figure 1[3]. But some units may be missing or combined into one in real SRS. SRS that used to control some devices requires a limited set of commands, and we can use more simple scheme (figure 2) for our system.

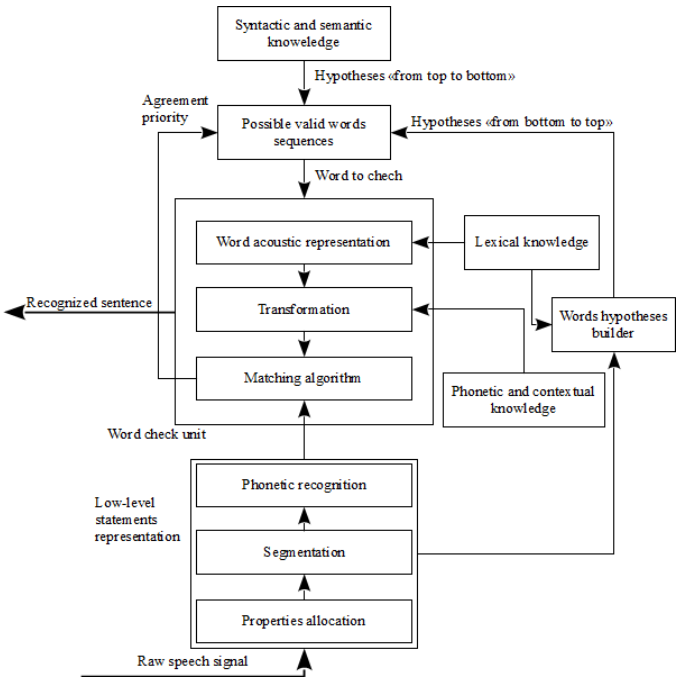

Fig. 1. SRS Common scheme

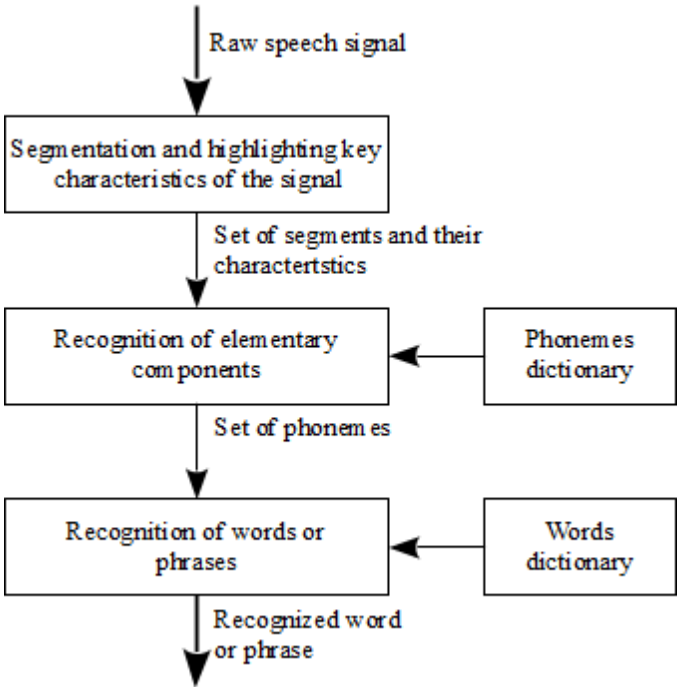

Fig. 2. Command recognition system scheme

Auxiliary algorithms for pre-filtering and system learning are also used in addition to these basic steps.

The second part is often skipped in voice commands recognition and whole words are recognized at once. Advantages of this approach is reducing the number of calculations. But retraining such system for recognition new commands will take more time than retraining system which recognizes phonemes. Because in the second case phonemes of which consists the new command is already in the system database, and we only need to train it to identify a new order of phonemes.

B. Selection of Signal Characteristics

Frequency spectrum changing in time is the natural characteristic of speech signal. The human brain recognizes speech exactly based on its time-frequency characteristics. Correct identification of the signal characteristics is extremely necessary for successful speech recognition. There are many approaches to solve this problem:

- Fourier spectral analysis.
- Linear prediction coefficients.
- Cepstral analysis.
- Wavelet analysis.
- And other.

Wavelet is a mathematical function that analyzes different frequency components of data. Graph of the function looks like a wavy oscillations with amplitude decreases to zero away from the origin. However, this is particular definition. Generally, the signal analysis is performed in the plane of the wavelet coefficients. Wavelet-coefficients are determined by the integral signal transformation. The resulting wavelet spectrogram clearly ties spectrum of various characteristics of the signals to the time[4]. This way they are fundamentally different from the usual Fourier spectra. This difference gives the advantage of wavelet transformation in the analysis of speech signals that non-stationary in time.

The wavelet transformation of the signal (DWT) consistently selects more and more high-frequency parts, thus breaking the signal into several levels of wavelet coefficients. The coefficients on the first levels are the lowest frequency signal. These coefficients give a good frequency resolution and low time resolution. The coefficients on the last levels of decomposition are the highest frequency of the signal. They give good time resolution and low frequency resolution.

Thus, selection of the signal characteristics using wavelet analysis is transformation of signal into wavelet-coefficients and calculation of average values of these coefficients at each level of the wavelet decomposition.

Segmentation of the signal on phonemes is performed at this stage. A phoneme is the minimal unit of the sound structure of language. DWT can solve this problem. The signal is changing on many decomposition-levels at once in transition between phonemes. Thus, the determination of the phonemes boundaries can be reduced to finding the moments of the wavelet-coefficients changing in most of the decomposition-levels[5].

First signal is divided into overlapping regions (frames), each of witch applies DWT. We can calculate energy for each frame *i* and decomposition-level *n*:

$$E_n(i)=\sum_{i=1}^{2^n-1} d^2_{n,j+2^{n-1},} \quad (1)$$

The signal energy (1) rapidly changes from frame to frame for each level. This is due to unavoidable noise during speech signal recording. We define E'_n to smooth energy changes. For this we replace value of E_n in window of 3 – 5 frames on the maximum value of E_{max} in this window. We calculate derivative R to determine the rate of energy change. The transition between phonemes are characterized by small and rapid changes of energy level at one or more decomposition-levels. Thus, criterion of the phonemes boundary finding is fast change of the derivative at a low energy level[6].

C. Recognition of Phonemes

The recognition result depends on the correct identification of the detected phonemes in many respects. However, the solution of this task is not trivial. Person never pronounces sounds the same. Pronunciation depends on physical health of speaker and his emotional state. Therefore it is impossible to identify phoneme simply comparing its characteristics with the characteristics of the standard phoneme. However, all versions of pronouncing the same phoneme will somehow resemble the standard pronunciation. In other words, they will be around in the signal characteristics domain. Identification of the pronounced phoneme can be reduced to solving the problem of clustering.

Clustering of phonemes in the developed system uses a network of vector quantization based on Kohonen neural network[7][8]. The advantage of neural network over k-means algorithm is that it less sensitive to outliers as it uses universal approximator – neural network.

Kohonen neural networks is a class of neural networks, their main element is the Kohonen layer. Kohonen layer consists of adaptive linear combiners. Typically, the output signals of Kohonen layer are processed by the rule "winner takes all": the largest signal is converted into one, others in zeros. Problem of vector quantization with k code vectors W_j for a given set of input vectors S is formulated as a problem of minimizing the distortion in encoding. The basic version Kohonen network uses the method of least squares and distortions D is given by:

$$D=\sum_{j=1}^{k}\sum_{x\in K_j}\left\|x-W_j\right\|^2$$

where K_j is consists of those points of $x\in S$, which are closer to W_j than to other W_l $(l\neq j)$. In other words, K_j consists of those points $x\in S$, which are encoded code vector W_j. Set S is not known when the network not configured to the speaker. In this case online method is used to learn network. Input vectors x are processed one by one. The nearest code vector (a "winner" who "takes all") $W_j(x)$ is sought for each of them. After that, this code vector is recalculated as follows:

$$W_{j(x)}^{new}=W_{j(x)}^{old}(1-\theta)+x\theta$$

where $\theta\in(0,1)$ is learning step. The rest of the code vectors do not change in this step. The online method with fading rate of learning is used to ensure stability: if T is the number of steps of training, then we put $\theta=\theta(T)$. Function of $\theta(T)>0$ is chosen so that the $\theta(T)\to 0$ monotonically as $T\to\infty$ and the series $\sum_{T=1}^{\infty}\theta(T)$ diverges such, $\theta(T)=\theta_0/T$.

D. Recognition of Words

After receiving the sequence of phonemes from the original signal we must map this sequence to voice command in the system database or indicate that the spoken word is not recognized. However, this problem is also a non-trivial. Differences in the pronunciation of sounds can be so significant that the same sound pronounced by a person will be identified by the system as two entirely different phonemes. Thus, only based on comparison the sequence of spoken phonemes to the standard sequence of phonemes of command, we can not say that this or that command was pronounced. One of solutions this problem is using of algorithm for finding the shortest distance between spoken word and standard system commands.

In the developed system Levenshtein distance (edit distance) is used as a measure of distance between the words. The Levenshtein distance is a string metric for measuring the difference between two sequences[9]. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other. Mathematically, the Levenshtein distance between two strings a, b is given by $lev(|a|,|b|)$ where

$$lev(i,j)=\begin{cases}max(i,j) & ,min(i,j)=0\\ min\begin{cases}lev(i-1,j)+1\\ lev(i,j-1)+1\\ lev(i-1,j-1)+[a_i\neq b_j]\end{cases} & ,else\end{cases}$$

In this case, the characters is a phoneme, the source string is pronounced sequence of phonemes, and the resulting string is a sequence of phonemes in the standard command

IV. Conclusion and Future Works

At the moment, I realize algorithms described above in Matlab environment. The most immediate problem is study of selected algorithms efficiency and taking action to improve it. Here are some possible directions for improving the system:

- Using of pre-filtering algorithms.
- Experiments on the choice of the most suitable wavelet for speech processing.
- Check the efficiency of the wavelet packet analysis instead of the usual.
- Check the efficiency of the Kohonen neural in comparison with the other clustering algorithms.
- Check of efficiency of other algorithms to determine the distance between the spoken word and standards.
- Assessing the impact of size and composition of the commands dictionary on the system performance.

Later, algorithms tested in Matlab environment will allow us to develop software system in the C++ language. After that I

will be able to make field testing of the system in controlling educational mobile robot.

REFERENCES

[1] Hitrov M. “Распознавание русской речи: состояние и перспективы” in “Речевые технологии”, vol.1, 2008, pp. 83-87.

[2] M. Pinola “Speech Recognition Through the Decades: How We Ended Up Siri” article on PCWorld web-site, 2011. URL: http://www.pcworld.com

[3] Li U. “Методы автоматического распознавания речи”, vol.1, vol.2, Moscow, “Наука”, 1983.

[4] Daubechies I. “Ten Lectures on Wavelets”, SIAM, 1 edition, 1992.

[5] Ermolenko T., Shevchuk V. “Алгоритмы сегментации с применением быстрого вейвлет-преобразования” Papers accepted for publication on the website of the international conference “Диалог”, 2003. URL: http://www.dialog-21.ru

[6] Vishnjakova O., Lavrov D. “Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования” in “Математические структуры и моделирование” vol. 23, 2011, pp. 43-48

[7] Tan Keng Yan, Colin “Speaker Adaptive Phoneme Recognition Using Time Delay Neural Networks” National University of Singapore, 2000

[8] Hecht-Nielsen R., “Neurocomputing”, Reading, MA: Addison-Wesley, 1990

[9] Levenshtein V. “Двоичные коды с исправлением выпадений, вставок и замещений символов”. Доклады Академий Наук СССР 163 (4), pp. 845–8, 1965.