ISSN 0005-1179, Automation and Remote Control, 2013, Vol. 74, No. 7, pp. 1225–1232. © Pleiades Publishing, Ltd., 2013.
Original Russian Text © A.V. Savchenko, 2013, published in Sistemy Upravleniya i Informatsionnye Tekhnologii, 2013, No. 1, pp. 71–75.

CONTROL SYSTEMS AND INFORMATION TECHNOLOGIES

# Phonetic Words Decoding Software in the Problem of Russian Speech Recognition

**A. V. Savchenko**

*National Research University Higher School of Economics, Nizhni Novgorod, Russia*

Received July 19, 2012

**Abstract**—The prototype of the isolated words recognition software based on the phonetic decoding method with the Kullback-Leibler divergence is presented. The architecture and basic algorithms of the software are described. Finally, an example of application to the problem of isolated words recognition is provided.

**DOI:** 10.1134/S000511791307014X

## 1. INTRODUCTION

As is generally known [1, 2], nowadays there exist no high-performance software in the field of automatic speech recognition (ASR) of Russian language. The reason lies in exclusive linguistic features of Russian speech [2, 3], as well as in strict requirements to transmission and processing systems. One can claim that the major obstacle to wide dissemination of new speech technologies in Russia consists in two relevant problems, *viz.*, variability of the informal style of Russian speech and the problems of "large dictionaries." Even in cutting-edge systems of ASR (such as *Google Voice Search* [4], *Nuance Dragon NaturallySpeaking* [5] and others), this problem has not been completely solved. The minimal probability of recognition mistakes still maintains at the level of 15–20%. A similar situation has been established in automatic speech diarization (ASD) systems. Indeed, most successful developments [6] do not solve the existing problems of Russian language.

One requirement to ASR systems concerns *natural* speech recognition [7]. It leads to appreciable speech variability and reduces the reliability of ASR systems. Hypothetically, easing of this requirement would improve the recognition accuracy and, consequently, decrease the interaction time with a distant object (under voice control). For instance, investigations demonstrate that the best recognition belongs to stressed vocals, see [7, 8]. Therefore, in this paper we suggest the following constraint [8, 9]: a speaker must pronounce all commands with legible emphasis of each syllable. In this case, research in the field of syllabic phonetics [10] indicates that an input signal can be divided into an isolated sequence of syllables. Consequently, the problem of voice commands recognition is reduced to syllables recognition [11] based on the phonetic decoding method (PDM) [12]. For performance tests of the suggested approach to the problems of ASR and ASD, we have designed the prototype of phonetic words decoding software (PPWDS). This software implements the following methods:

- the PDM with the Kullback-Leibler divergence [13] for real-time ASR;
- speaker identification by a phonetic database (PDB) of all potential users (the basis of ASD) [14];
- automatic formation of a working dictionary from a text file by the topical principle [15];
- automatic formation of a PDB from a continuous talkspurt of a speaker [14].

The present paper considers the architecture and basic algorithms of the PPWDS. These algorithms are intended for a wide range of experts in automatic speech processing.

## 2. ARCHITECTURE AND USER INTERFACE

The PPWDS consists of four modules:

(a) *SpeechProcessing*, the library of speech signal processing;

(b) *SpeechRecognizer*, the executable file of the software;

(c) *MrcpSpeechRecog*, a plugin for UniMRCP server;

(d) the executable file of *mrcpspeechrecog-client*.

Modules (a), (b) have been developed using C# 4.0. They require preinstalled .NET Framework 4.0 platform. And modules (c), (d) have been developed using C++ and integrated with UniMRCP library (see www.unimrcp.org). It realizes MRCP (Media Resource Control Protocol).

The structure chart of the system (see [9]) is illustrated by Fig. 1.

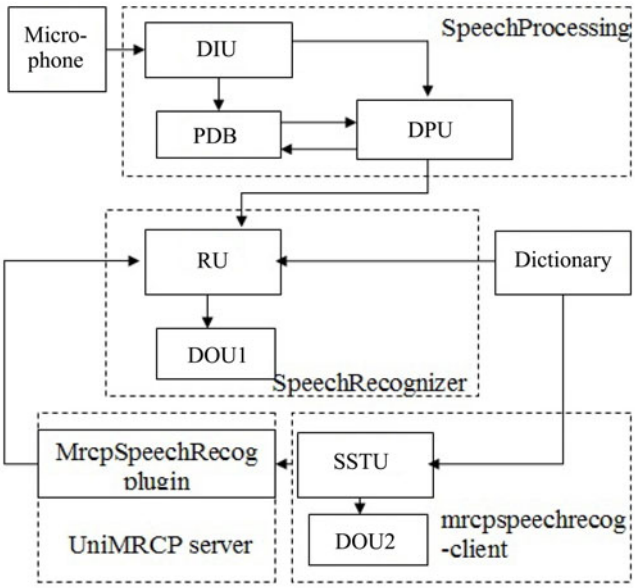


**Fig. 1.** The structure chart of the PPWDS: Microphone—a dedicated microphone, DIU—data input unit. DOU1, DOU2—data output units, DPU—data processing unit, RU—recognition unit, SSTU—speech signal transmission unit, Dictionary—a list of reference words/word combinations in text representation. The units in dashed-line rectangles are realized as corresponding modules and form the PPWDS.

The software supports two modes of operation as follows.

(1) *Main* mode, which employs just two modules, *SpeechProcessing* and *SpeechRecognizer*. Microphone supplies a speech signal to a PC for being recorded as a wav file (sampling rate of 8 or 16 kHz, 16 bits per sample). Next, DIU reads this file and performs its preliminary processing. Depending on data input mode, DIU saves the sound signal in a PDB (if the input mode serves for PDB creation) or sends the processed signal to DPU for recognition and speaker identification. DPU is intended for partitioning a speech signal into nonintersecting segments, computing the distances between segments, as well as for maintaining and filling of a PDB. The output signal of DPU enters RU for ASD and ASR using text representation of words in Dictionary. And the output signal of RU is displayed through DOU.

(2) *Client-server* mode, enabling any client application with MRCP support to address a server (more specifically, a UniMRCP server) for speech recognition from a given dictionary. A simple realization of MRCP client is *mrcpspeechrecog-client* application. As input data, it receives a wav file and a text file of a dictionary. In SSTU, these data are transferred to *MrcpSpeechRecog* plugin as a dictionary in speech recognition grammar specification (SRGS) and a SIP message (a sound

signal). By-turn, this plugin passes the data to *SpeechRecognizer* application via a named pipe of MS Windows. Consequently, recognition takes place according to the above scheme in the main mode. But the results are not displayed in DOU1; instead, they are transmitted as an NLSML file (Natural Language Semantics Markup Language) to the client, and the latter displays the results in DOU2.

## 3. ISOLATED WORDS RECOGNITION ALGORITHM

Consider a given set of $L > 1$ reference commands $\{X_l\}$, where $l = \overline{1, L}$ indicates the number of a reference word. Within the framework of the phonetic approach [1, 7], a reference command is divided into a sequence of phonemes (a transcription) $X_l = \{c_{l,1}, c_{l,2}, \ldots, c_{l,L_l}\}$. Here $L_l$ specifies the duration of the command (in phonemes), and $c_{l,j} \in \{1, \ldots, R\}$ are the numbers of phonemes from the phonetic alphabet $\{\mathbf{x}_r^*\}, r = \overline{1, R}$ ($R$ denotes the number of phonemes in the alphabet). The problem lies in the following. For an input speech signal $X$ with sampling rate of $F$ (in Hz), find the closest reference word.

To solve the problem, at stage 1 the signal $X$ is divided into nonintersecting segments $\{\mathbf{x}(t)\}$, $t = \overline{1, T}$ having the length of $\tau = 0.01$–$0.015$ s, where $T$ designates the total number of segments. Next, each partial signal $\mathbf{x}(t) = \|x_1(t) \ldots x_M(t)\|$ (here $M = \tau \times F$) is analyzed using a finite list of vowel phonemes $\{\mathbf{x}_r^*\}$ (a PDB contains vowels only). A partial signal is identified with a vowel phoneme corresponding to the minimum of a closeness measure of the signal $\mathbf{x}(t)$ and the reference $\mathbf{x}_r^*$ (such measure is selected by an investigator). To choose a closeness measure, we adopt the autoregressive model (AM) of a speech signal. A major benefit of the AM concerns normalization of speech signals by the variance of generating processes: $\sigma_0^2 = \sigma_x^2$, where $\sigma_x^2$ defines the sample estimate of the variance of the generating process $\mathbf{x}(t)$. In this case, the asymptotically optimal solution [12] is yielded by the criterion

$$\nu(t) = \arg \min_{r \in \{1, \ldots, R\}} \frac{1}{2} \left[ \frac{\sigma_r^2(\mathbf{x})}{\sigma_0^2} - 1 \right]. \tag{1}$$

In the formula above, $\sigma_r^2(\mathbf{x})$ is the sample estimate for the variance of response of the whitening filter $\|y_{r;1}(t) \ldots y_{r;M-p}(t)\|$ with index $r$, $p$ corresponds to the order of the AM, and

$$y_{r;j}(t) = x_{j+p}(t) - \sum_{m=1}^{p} a_{r,m} x_{j+p-m}(t). \tag{2}$$

Each reference $\mathbf{x}_r^*$ from the PDB is determined by its vector of autoregressive coefficients $\{a_{r,m}\}$, $m = \overline{1, p}$ (obtained by the Berg algorithm or the Levinson–Durbin recursion [16]).

Suppose that each input word $X$ is divided into $N$ syllables, and the boundaries of syllable $n$ ($n = \overline{1, N}$) are defined under the accuracy within the number of quasi-stationary segment $\left(t_n^{(1)}, t_n^{(2)}\right)$. Next, for each syllable $n$, recognition takes place [8, 9, 15] only among vowels. In other words, the phonetic alphabet $\{\mathbf{x}_r^*\}, r = \overline{1, R}$ (i.e., the PDB) consists of the references of vowel phonemes. For this, based on all $\nu(t), t = \overline{t_n^{(1)}, t_n^{(2)}}$, syllable $n$ is associated with a sequence of frequencies $\mu_n(r)$, $r = \overline{1, R}$, where

$$\mu_n(r) = \frac{1}{t_n^{(2)} - t_n^{(1)} + 1} \sum_{t=t_n^{(1)}}^{t_n^{(2)}} \delta(\nu(t) - r), \tag{3}$$

with $\delta(x)$ indicating the discrete delta function. Subsequently, for each reference command $X_l$, estimate its correlation to the recognized speech signal

$$\mu_l = \begin{cases} \sum\limits_{n=1}^{N} \mu_n\,(c_{l,n}), & L_l = N \\ 0, & L_l \neq N. \end{cases} \tag{4}$$

And the ASR problem is solved by choosing the word $X^*$ by the following criterion:

$$X^* = \arg \max_{X_l, l=\overline{1,L}} \mu_l. \tag{5}$$

The expressions (1)–(5) describe the proposed approach to voice command recognition. The flowchart of the recognition algorithm [14] is shown in Fig. 2.

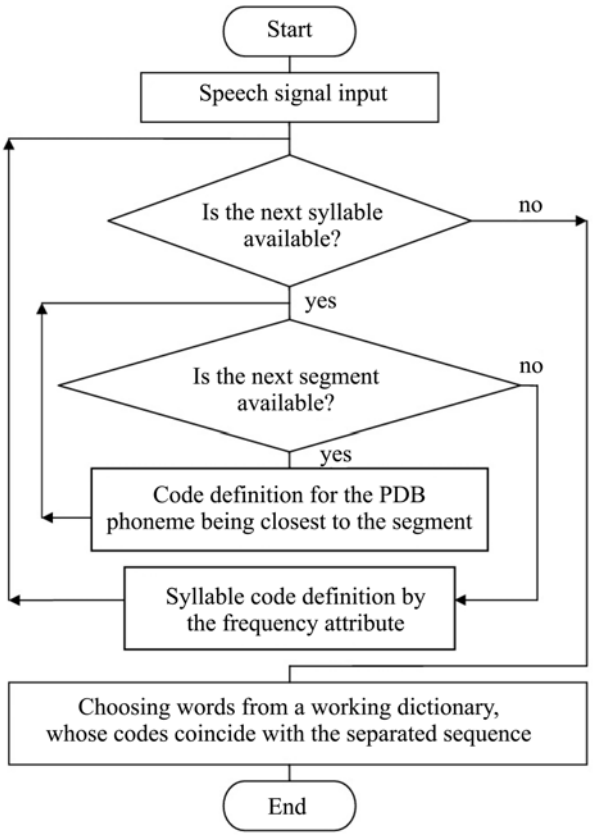


**Fig. 2.** Automation recognition algorithm for Russian speech by the PDM.

A speech signal supplied by the microphone serves for automatic extraction of syllables using a simple amplitude limiter (or complicated algorithms [17] based on preliminary processing of speech signals). Each syllable is partitioned into nonintersecting segments having the duration of $\tau$ s. For each segment, one evaluates the closest phoneme (1) in the sense of the chosen divergence (2), (3). Afterwards, the sequence of frequencies $\mu_n\,(r)$ is determined for the whole syllable according to (4). Next, for each reference word, define its closeness $\mu_l$ to the input word (4) and choose all closest reference commands (5) (with maximal values of $\mu_l$). They are displayed as the results of ASR.

## 4. BASIC ALGORITHMS OF THE PPWDS

The recognition procedure (see Fig. 2) possesses the following advantages. First, speaker identification is possible (under saved PDBs of all potential speakers). For each segment of a speech signal, phonetic code construction involves references from united PDBs of all speakers. For a syllable, the identification procedure chooses a speaker, whose PDB contains sounds being closest to segments of this syllable (according to the frequency principle—by analogy with (4)–(6)). Figure 3 presents the flowchart of the diarization algorithm for Russian speech [14] used in speaker identification.

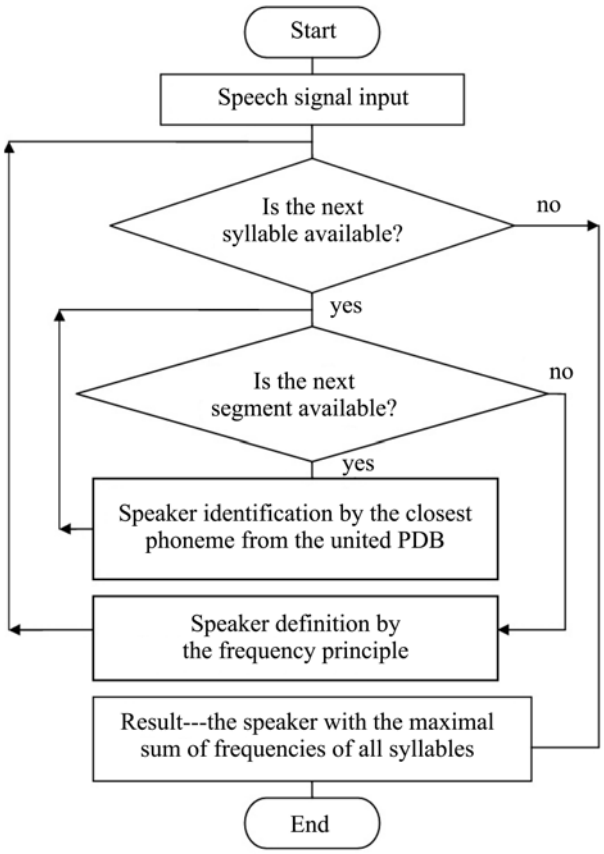


**Fig. 3.** Automatic speaker identification algorithm by a fragment of Russian speech.

Second, the procedure of assigning a sequence of vowel phoneme codes $\{c_{l,1}, c_{l,2}, \ldots, c_{l,L_l}\}$ to each reference command can be also performed automatically (by text representation of a command using phonetic speech synthesis algorithms [18]). Consequently, the adjustment procedure with respect to a new dictionary is completely automatized [15].

Finally, adjustment with respect to a new speaker takes place by recording his pronunciation of all vowel phonemes. Generally, their number is rather small ($R = 6$–$20$). Thus, such adjustment procedure does not consume much time. Moreover, this procedure reveals one more benefit of the AM of an acoustic pipe, *viz.*, verification of reference sounds. Imagine that a discrete delta pulse with the fundamental tone frequency (100–150 Hz) is supplied to the input of the autoregressive process with the coefficients estimated by the reference signal from a speaker. In this case, the output autoregressive signal can be passed to an audio device (for "insonification"). If this signal is not perceived by a speaker as his vowel syllable, the AM should be readjusted.

## 5. AN EXAMPLE OF PRACTICAL APPLICATION

We provide the following example of practical application of the PPWDS. Consider the problem of medications name recognition in a drugstore located in the city of Nizhnii Novgorod (the size of dictionary is $L = 1913$). The sampling rate of a speech signal in ADC equals 8 kHz, the order of the AM makes up $p = 20$, and the length of one segment is $M = 120$ samples (equivalently, $\tau = 0.015$ s). A built-in microphone serves for speech input. The PDB varies from one speaker to another and, in each case, comprises all $R = 10$ vowel syllables of a corresponding speaker. The basic requirement to speakers lies in legible emphasis of all open syllables while pronouncing each word. During recognition, syllables are automatically separated by the amplitude detector of silence not smaller than 70 ms. System operation in recognition mode is illustrated by Fig. 4.

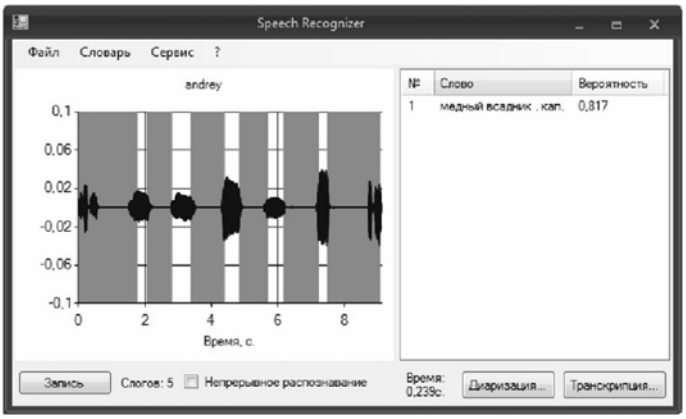


**Fig. 4.** The main application window of the PPWDS in recognition mode.

For continuous speech of a user **(“dobryi den’! mednyi vsadnik kap. Do svidaniya!”)**, the program separates fragments pronounced syllable by syllable **(“mednyi vsadnik kap”)**. A stationary long ($> 100$ ms) vowel phoneme is preliminary separated in each emphasized syllable. Commands are differentiated from continuous speech by comparing with a certain threshold of the ratio of homogeneous segment duration (separated in a signal) to syllable duration. Indeed, common continuous speech admits no pauses between syllables (sometimes, between words); thus, the stated ratio appears small. Figure 4 shows that the phrase pronounced syllable by syllable is actually recognized (continuous speech is ignored).

## 6. CONCLUSION

Let us dwell upon the advantages of the ASR approach involved in the PPWDS.

(1) Computations acceleration is a central problem in the field of ASR [1, 14]. When a dictionary contains several thousand words, real-time realization of most famous algorithms (based on words segmentation into phonemes and their leveling by speech pace) requires considerable computational capabilities unavailable to a modern laptop or mobile phone. No wonder that precise implementation of the classic approach appears possible only within the framework of large corporations such as Microsoft, Google, Apple, IBM and Nuance Communications. The PPWDS employs a fundamentally different approach proceeding from the PDM; its major advantage concerns appreciable (by one order of magnitude or even higher) reduction of computational costs owing to the procedure of dynamic leveling of words and calculation of the closeness measure based on the adaptive algorithm (1), (2).

(2) Automation problems for the formation procedures of PDBs [12] and working dictionaries [15] in ASR systems provoke a heightened interest of experts in speech technologies. The developed

system ensures complete automation for the formation processes of a PDB and working dictionary; furthermore, the time required for their formation is significantly decreased.

(3) Due to the automatic procedure of working dictionary formation [15], there is no need in labor-intensive preliminary formation of large universal speech databases (several hundred thousand words). As a result, one can perform rapid adjustment of the system with respect to a specific subject domain. Consequently, the size of a working dictionary is cut down, whereas the performance and accuracy of ASR gets improved.

Therefore, the PPWDS can serve for designing a commercial software product for voice control systems with Russian language such as

• industrial control systems and autonomous communication networks (with an additional option of voice verification of a manager);

• automatic management systems "Intelligent House" (with an additional option of individual adjustment of commands for occupants);

• telephone ordering systems (with automatic processing of voice orders, door-to-door delivery of goods and services, prompt changes in product/service assortment, prices, etc.).

## ACKNOWLEDGMENTS

This study was carried out within "The National Research University Higher School of Economics' Academic Fund Program in 2013–2014," research grant no. 12-01-0003, and the Federal Grant-in-Aid Program "Research and Development on Priority Directions of Scientific-technological Complex of Russia for 2007–2013" (Governmental Contract no. 07.514.11.4137).

## REFERENCES

1. Sorokin, V.N., Fundamental Study of Speech and Applied Problems of Speech Technologies, *Rechevye Tekhnol.*, 2008, no. 1, pp. 18–48.
2. Babin, D.N., Mazurenko, I.L., and Kholodenko, A.B., On Prospects of Designing an Automatic Recognition System for Continuous Russian Speech, *Intellektual. Sist.*, 2004, vol. 8 (1-4), 45–70.
3. Ronzhin, A.L. and Li, I.V., Automatic Recognition of Russian Speech, *Vestn. Ross. Akad. Nauk*, 2007, vol. 77, no. 2, pp. 133–138.
4. Schuster, M., Speech Recognition for Mobile Devices at Google, in *Lecture Notes in Computer Science*, 2010, vol. 6230, pp. 8–10.
5. Grant, R. and McGregor, P.E., Speech Recognition System and Method, US Patent 8 175 883 B3, 2012.
6. Huijbregts, M. and Wooters, C., The Blame Game: Performance Analysis of Speaker Diarization System Components, in *Proc. Interspeech*, 2007, pp. 27–31.
7. Anusuya, M.A. and Katti, S.K., Speech Recognition by Machine: A Review, *Int. J. Computer Sci. Inf. Security*, 2009, vol. 6, no. 3.
8. Savchenko, A.V., Automatic Speech Transcription Based on Minimum Information Discrimination Principle, *Vestn. Komp. Inform. Tekhnol.*, 2012, no. 8, pp. 14–19.
9. Savchenko, A.V., Savchenko, V.V., and Akat'ev, D.Yu., A Device for Phonetic Analysis and Recognition of Speech, *Offits. Byull. Federal. Sluzh. Intel. Sobstv., Patent. Tovarnym Znakam*, 2011, registr. number 2011125526/08.
10. Belyavskii, V.M. and Svetozarova, N.D., Syllable Phonetics and Three Phonetics of L.V. Shcherba, in *Teoriya yazyka, metody ego issledovaniya i prepodavaniya* (The Theory of Language, Methods of Its Analysis and Teaching), Leningrad: Nauka, 1981, pp. 36–40.
11. Sirigos, J., Fakotakis, N., and Kokkinakis, G., A Hybrid Syllable Recognition System Based on Vowel Spotting, *Speech Commun.*, 2002, vol. 38, pp. 427–440.

12. Savchenko, A.V., Words Phonetic Decoding Method in a Problem of Speech Automatic Recognition on the Basis of Information Mismatch Minimum Principle, *Izv. Vyssh. Uchebn. Zaved. Ross., Radioelektron.*, 2009, no. 5, pp. 31–41.

13. Kullback, S., *Information Theory and Statistics*, New York: Dover, 1997.

14. Savchenko, V.V., The Development of Phonetic Algorithms of Speech Recognition and Diarization with Automatically Reconfigurable Dictionary, *Sist. Upravlen. Inform. Tekhnol.*, 2012, no. 3(49), pp. 99–104.

15. Savchenko, V.V. and Savchenko, A.V., The Procedure of Working Dictionary Formation in Automatic Speech Recognition Systems by a Topical Text File, *Sist. Upravlen. Inform. Tekhnol.*, 2012, no. 2.2(48), pp. 284–289.

16. Marple, S.L., *Digital Spectral Analysis with Applications*, Englewood Cliffs: Prentice Hall, 1987. Translated under the title *Tsifrovoi spektral'nyi analiz i ego prilozheniya*, Moscow: Mir, 1990.

17. Janakiraman, R., Kumar, J.C., and Murthy, H.A., Robust Syllable Segmentation and Its Application to Syllable-centric Continuous Speech Recognition, in *Proc. Natl. Conf. Commun.*, 2010, pp. 1–5.

18. Kipyatkova, I.S. and Karpov, A.A., Development and Assessment of a Transcribing Module for Recognition and Synthesis of Russian Speech, *Iskusstv. Intellekt*, 2009, no. 3, pp. 178–185.