Speech Rehabilitation After Combined Treatment of Cancer and the Formation of a Set of Syllables for Assessing Speech Quality [*]

Evgeny Kostyuchenko
key@keva.tusur.ru

Dariya Novokhrestova
devijas@yandex.ru

Tomsk State University of Control Systems and Radioelectronics,
Tomsk, Russia

Abstract

The paper considers the organization of patient rehabilitation after surgical treatment of oncological diseases of the organs of the vocal tract. A review of the current state in the field of speech quality assessment is carried out. The emphasis is not on the assessment in the transmission through communication channels, but on the quality of pronunciation. The use of direct comparison in such a problem is difficult due to the high degree of variability of the pronunciation implementations of the same fragment. The current implementation of the software package for rehabilitation is considered. On the basis of its trial operation, a conclusion is drawn on the need to expand the list of analyzed phonemes, which previously contained only the most subject to change phonemes. An algorithm for generating the shortest list of phonemes that meets the requirements of a speech therapist is proposed. The integration of this approach into the existing software package has been carried out. The results of testing the proposed approach of speech rehabilitation are presented.

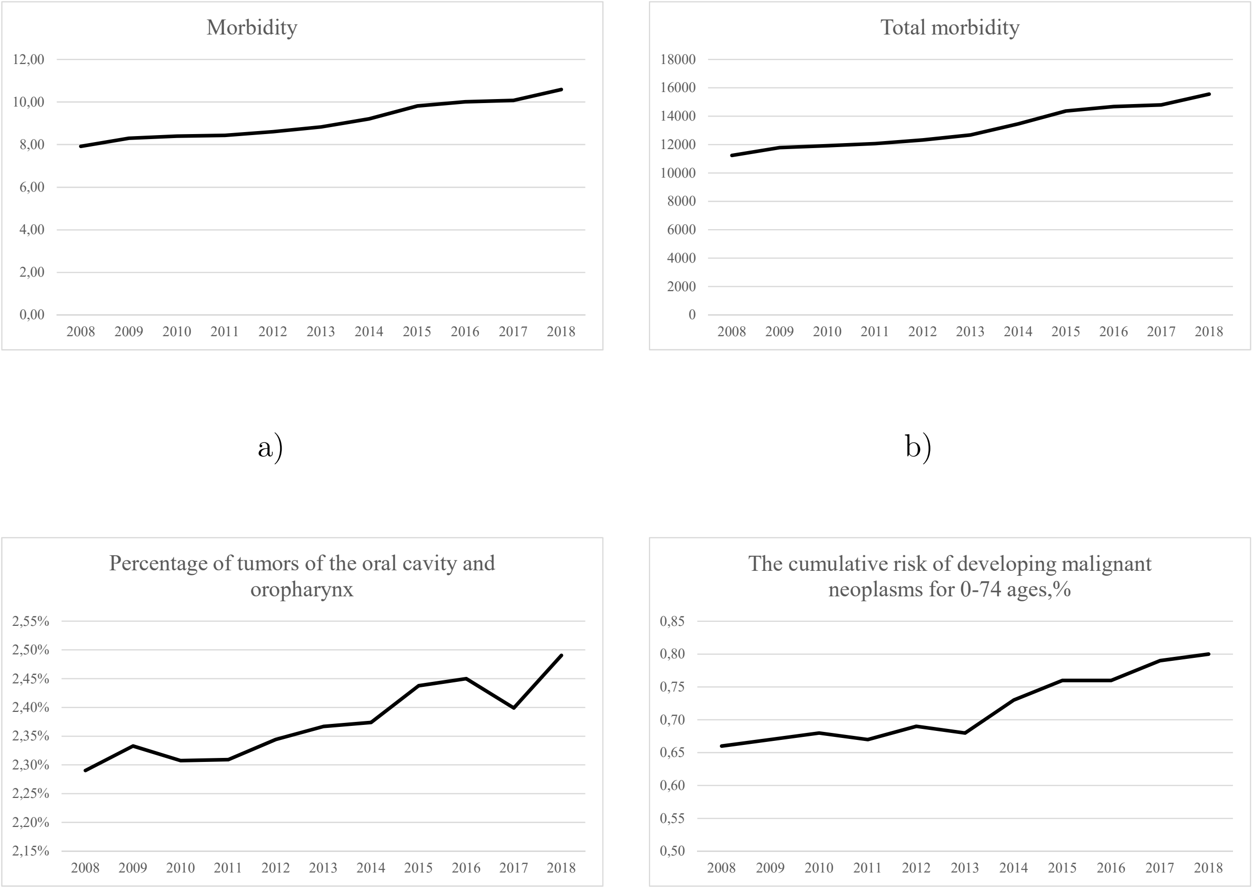
Keywords: *syllable segmentation, syllabic intelligibility, assessment of speech quality, speech rehabilitation*

1 Introduction

The problem of cancer incidence is relevant both in the world and in Russia. So, in Russia alone in 2018, more than 624,000 [Kap18] cases of the first time in the life of established diagnoses of malignant neoplasms were registered. At the same time, the number of registered cases is growing from year to year.

A tumor of the organs of the vocal tract, in particular the oral cavity and oropharynx is a one of the actual localizations. The dynamics of the incidence rate of cancer with this localization per 100,000 population is negative, demonstrating an increase of more than 30%

over the past 10 years. The general dynamics of changes in the incidence of these types of tumors per 100,000 over the past 10 years is presented in Figure 1, a. The dynamics of the total number of identified diseases is also negative. As can be seen from Figure 1, b, the total share of localization presented (Figure 1, c) and the cumulative risk of developing malignant neoplasms in the range of 0-74 ages (Figure 1, d) are growing.



a)

b)

c)

d)

Figure 1: The dynamics of the incidence of tumors of the oral cavity and oropharynx

None of the statistics presented allows us to speak about the positive dynamics of the incidence of tumors of the mouth and oropharynx over the past 10 years.

In addition, for the localization under consideration, the average age of disease registration is 60 years, which is one of the lowest in comparison with other localizations. This suggests that a significant proportion of the population is affected to disease by the working-age. The surgical treatment procedure leads to the need to learn to speech, that reducing the quality of life and the possibility of working.

The above arguments indicate the high relevance of the study of speech rehabilitation after surgical treatment of tumors of the oral cavity and oropharynx.

2 Existing approaches to assessing the quality of pronunciation of speech units

One of the fundamental obstacles (at the time the research began) was the lack of objective tools for measuring the pronunciation quality of individual phonemes and their groups. This did not allow obtaining an objective assessment of the quality of speech and tracking its dynamics. Consider the classification of methods for assessing the quality of speech signals in general.

Initially, all methods can be divided into two categories: subjective and objective assessments. It may seem that obtaining subjective assessments during implementation is easier. However, if you need high-quality assessments, then this is not so. The main disadvantages of subjective assessments are the need to attract experts, as a result, poor automation, dependence on expert opinions, the need to formulate rigorous assessment methods and scales, the time and effort required to obtain estimates.

As subjective assessment methods, we can distinguish:

1. Standard GOST 50840-95 Voice over paths of communication (1995) Methods for Assessing the Quality, Legibility and Recognition [GOST] and its adaptation to work with patients during rehabilitation [Spec13]. Of greatest value is the opportunity to obtain syllabic and phrasal intelligibility with minimal expert influence due to the rigorous formalization of they actions and an unambiguous binary scale;

2. Mean opinion score (MOS) - described in ITU Recommendation R.800. It evaluates the absence or presence of an echo, voice distortion, delay from end to end and overall assessment of the quality of speech, as a subjective assessment of experts. This assessment is formed as an arithmetic mean, where the main evaluation parameters are: intelligibility, the naturalness of the sound of the voice and the level of effort of the listener [MOSPh]. Application of the same MOS technique, but already on IP networks, is described in [MOSIp].

Objective methods can be divided into the following two two categories:

1. Methods designed to compare the physically same signal before and after exposure (for example, transmission over communication channels);

2. Methods designed to compare different signal implementations, for example, the same phrase pronounced before and after surgical treatment.

The first methods include:

1. Perceptual Evaluation of Speech Quality (PESQ) [PESQ] - defined in ITU-T Rec. R.862. It is an objective methodology for determining the quality of voice communication in telephone systems, which predicts the results of a subjective assessment of the quality of this type of communication by expert listeners. To determine the quality of voice transmission, PESQ provides a comparison of the input or reference signal with its distorted version at the output of the communication system;

2. E-model - The primary output of the model is a scalar rating of transmission quality. A major feature of this model is the use of transmission impairment factors that reflect the effects of modern signal processing devices [E-model];

3. Methods in which the signal-to-noise ratio (SNR) and the segmented signal-to-noise ratio (segSNR) are evaluated. The SNR method is also called the total signal to noise ratio criterion. It takes into account the overall ratio of signal power and noise over the entire signal duration. However, at a low intensity of the useful signal at any interval, it can be masked by another part of the signal with a higher intensity of the useful signal, which ultimately distorts the estimate. segSNR is an evolution of the signal to noise ratio method. In this case, the signal-to-noise ratio is estimated at intervals of 15 to 20 ms, which makes it possible to obtain a more accurate estimate as a whole due to the fact that the uneven signal intensity does not distort the whole picture [SNR];

4. Enhanced Modified Bark Spectral Distortion (EMBSD) - objective speech quality measure based on audible distortion and cognition model [EMBSD]. Based on experiments with Time Division Multiple Access (TDMA) data containing distortions encountered in real network applications, the performance of the MBSD has been further enhanced by modifying some procedures and adding a new cognition model. The Enhanced MBSD (EMBSD) shows significant improvement over the MBSD for TDMA data. Also, the performance of the EMBSD is better than that of the ITU-T Recommendation P.861 for TDMA data. The performance of the EMBSD was compared to various other objective speech quality measures with the speech data including a wide range of distortion conditions. The EMBSD showed clear improvement over the MBSD and had the correlation coefficient of 0.89 for the conditions of MNRUs, codecs, tandem cases, bit errors, and frame erasures.

A comparison of these and similar methods can be found in [SPIIRAS2018]

The second group includes methods:

1. Estimating the quality of speech based on methods of temporal normalization of signals and subsequent comparison using metrics from the previous group, or similar [Spec2017]

2. methods based on the use of speech recognition. Such methods can be represented as methods of expert assessment based on listening, in which the speech recognition algorithm acts as an expert [Niko2002], [Spec2019].

The classification is shown in Figure 2.

Based on the analysis of existing methods for evaluating the quality of speech, the most interesting for solving the problem of evaluating the quality of pronunciation during speech rehabilitation are objective methods of assessment based on a comparison of different realizations of pronunciation of speech units. Records before the operation (despite the presence of a tumor, the legibility of such records is close to unity and they can act as a reference) and after the operation (during rehabilitation) can be used for comparing. The degree of proximity to the first record is an assessment of the quality of the current pronunciation and rehabilitation in general. However, the lack of ready-to-use methods of this class talks about the need for their development to solve the problem of speech rehabilitation after surgical treatment of oncological diseases of the organs of the vocal tract.

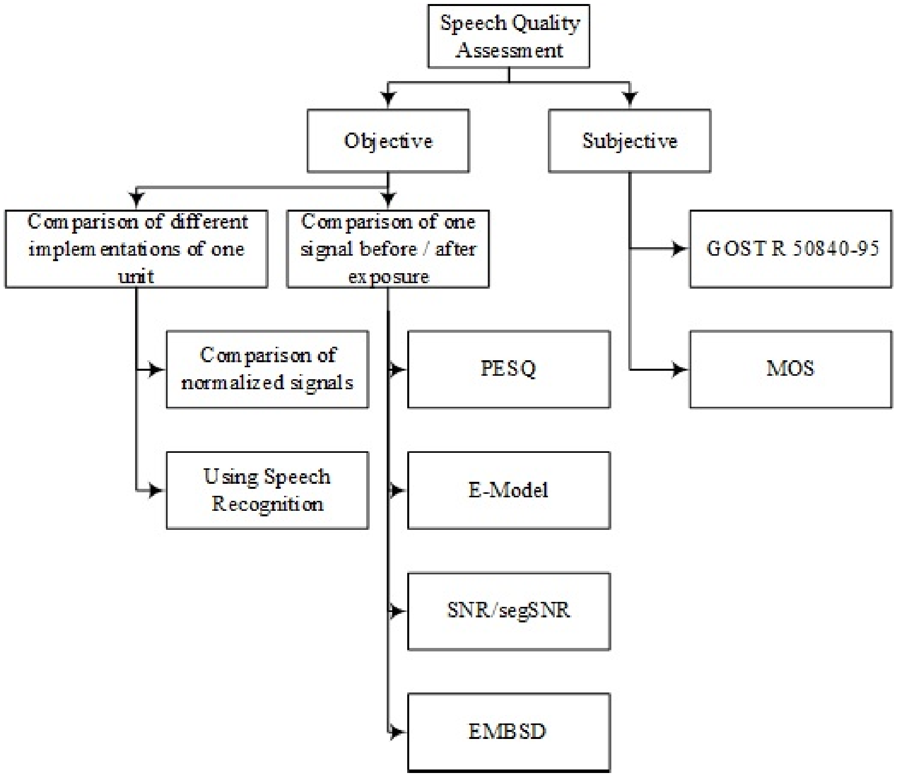

Speech Quality Assessment
Objective
Subjective
Comparison of different implementations of one unit
Comparison of one signal before / after exposure
GOST R 50840-95
MOS
Comparison of normalized signals
Using Speech Recognition
PESQ
E-Model
SNR/segSNR
EMBSD

Figure 2: Classification of algorithms for assessing the quality of speech signals

3 Methods for comparing syllables in assessing syllabic intelligibility and a software package for speech rehabilitation

When assessing the quality of pronouncing phonemes and syllables, the metric is the distance of the estimated syllable from the standard one made before the operation. However, to carry out the comparison, it is necessary that the signals have the same duration. To do this, before the assessment, it is necessary to conduct a time alignment procedure. In the study, two methods were tested:

1. Alignment using linear transformation. Used in the early stages of the study. To obtain the result, the oversampling mechanism of the signal for assessment was used to bring it into line with the reference one. The method is quite crude, since the dependencies and transients inside the syllable during pronunciation are not linear. When comparing individual phonemes, this approach is more acceptable. However, there is a problem of segmentation and allocation of individual phonemes.

2. Alignment using the dynamic time warping algorithm [DTW]. It is more flexible due to non-linearity and, in fact, compares the extrema of the smoothed signal and carries out their alignment in time.

After a time alignment, the question arises of finding a metric for comparing the received signals and identifying the relationship between them. The following metrics can be used:

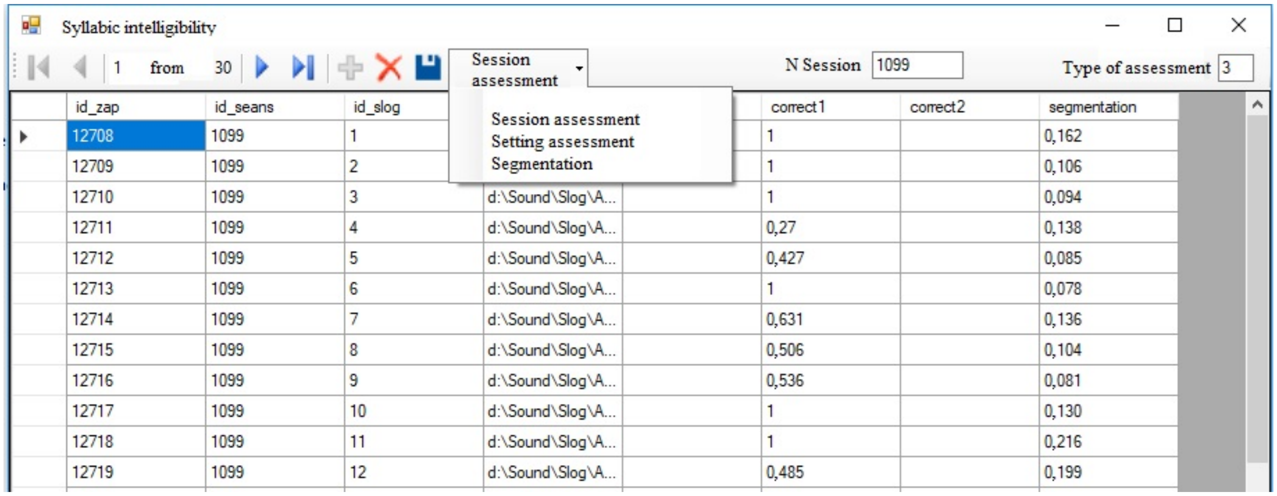
1. The correlation distance obtained after applying the dynamic time warping. It is obtained as an assessment of the leveling quality after it has been carried out. It has not been investigated if time alignment is applied to other methods, since in this situation it combines two different alignment method.

2. Correlation coefficients between signals. The application of the correlation coefficients of Pearson, Spearman and Kendall is investigated. The last two are ranked.

3. The use of Euclidean, Mahalanobis, Manhattan and generalized Minkowski metrics. Based on the results of the study, recommendations were made on choosing the Minkowski metric parameter from the range [1.6; 3.1] to ensure maximum consistency between records before surgery.

According to the results of the analysis, all the investigated options were implemented in the software package to evaluate the pronunciation quality, however, by default, alignment is proceeded using dynamic time warping and the correlation coefficient as a proximity metric. In this form, theoretically, the degree of proximity is in the range [-1; 1]. However, in reality, after analysis of the sessions of 21 patients, no examples of negative correlation were found.

The appearance of the module of the software package, which is responsible for obtaining assessments of the quality of pronunciation of syllables and phonemes, as well as their accounting within the database is presented in Figure 3.



1 from 30

N Session 1099

Type of assessment 3

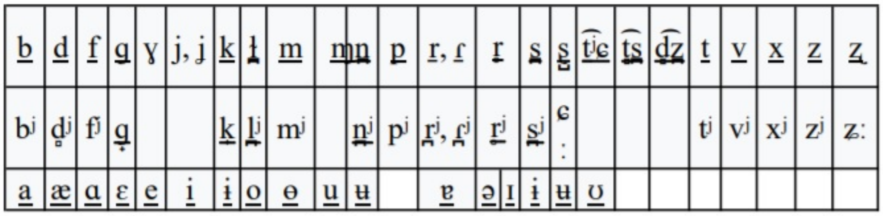
id_zap	id_seans	id_slog			correct1	correct2	segmentation
12708	1099	1			1		0,162
12709	1099	2			1		0,106
12710	1099	3	d:\Sound\Slog\A...		1		0,094
12711	1099	4	d:\Sound\Slog\A...		0,27		0,138
12712	1099	5	d:\Sound\Slog\A...		0,427		0,085
12713	1099	6	d:\Sound\Slog\A...		1		0,078
12714	1099	7	d:\Sound\Slog\A...		0,631		0,136
12715	1099	8	d:\Sound\Slog\A...		0,506		0,104
12716	1099	9	d:\Sound\Slog\A...		0,536		0,081
12717	1099	10	d:\Sound\Slog\A...		1		0,130
12718	1099	11	d:\Sound\Slog\A...		1		0,216
12719	1099	12	d:\Sound\Slog\A...		0,485		0,199

Figure 3: The module of the software complex, responsible for obtaining assessments of the quality of pronunciation of syllables and phonemes

4 An algorithm for generating a set of syllables for evaluating pronunciation quality

The algorithm for generating a set of syllables is intended to form such a set of syllables for which their total number would be minimal while providing the requested number of different phonemes. The phoneme realizations are divided according to such signs as voicedness / deafness, softness / hardness, stressedness / shocklessness. Based on this approach in the

Russian language, 59 different phonemes can be distinguished. The general set of phonemes allocated in the notation of the international phonetic alphabet [IPA] is presented in Figure 4 [Cub02].



b	d	f	g	ɣ	j, ʝ	k	ɫ	m	m	n	p	r, ɾ	r	s	ʂ	t͡ʲɕ	t͡s	d͡z	t	v	x	z	ʐ
bʲ	d̥ʲ	fʲ	g̟			k	lʲ	mʲ		nʲ	pʲ	rʲ, ɾʲ	rʲ	sʲ	ɕ :				tʲ	vʲ	xʲ	zʲ	ʑː
a	æ	ɑ	ɛ	e	i	ɨ	o	ɵ	u	ʉ		ɐ	ə ɪ	ɨ	ʉ	ʊ							

Figure 4: Sounds of the Russian language in the notation of the international phonetic alphabet

The general ideas used in constructing the algorithm are as follows:

1. The value of a syllable is estimated as the sum of the values for each phoneme within the framework of the resulting set.

2. The value of each phoneme is estimated as the number of phonemes that need to be added to the set to meet the requirements. With this construction, the algorithm allows the presentation of requirements for the count in the set for each of the phonemes individually. However, based on the experience of using the software complex in practice, in the framework of the previous operation of the complex, the requirements for occurrence in the set were the same for all phonemes of interest.

3. When adding a syllable to the final set, the values of phonemes included in the added syllable are reduced by 1.

4. In the process, an array of syllables containing phonemes of interest is analyzed. If there are no such phonemes in the syllable, the syllable its value becomes equal to 0 and it is excluded from consideration. The initial set of syllables - a complete list of syllables contained in the table for evaluating syllabic intelligibility [GOST]

5. If 0 is the value of all the syllables being evaluated, a conclusion is made about the readiness of the requested set or the impossibility of its formation.

The value of functions and variables in the framework of the algorithms:

Phonemes - a list of phonemes for selection in the generated set;

Phcount - the number of phonemes to select. Two index-linked arrays;

Syllist - a list of syllables being considered for addition at the current iteration;

OutData - formed list;

Range - ranks Syllist in descending order of the presence of the required phonemes (CurPhcount) from the Phonemes list. The output is an ordered list of syllables and utility for each syllable;

CurPhcount - The utility of each syllable. The array is indexed to Syllist;

Corr - removes syllables for which CurPhcount = 0;

Add - adds the most useful syllable to the OutData final set;

Sub - a procedure that corrects after adding CurPhcount - the usefulness of each syllable, a list of syllables and selected phonemes Syllist and Phonemes when achieving zero utility, based on the results of adding the most useful syllable.

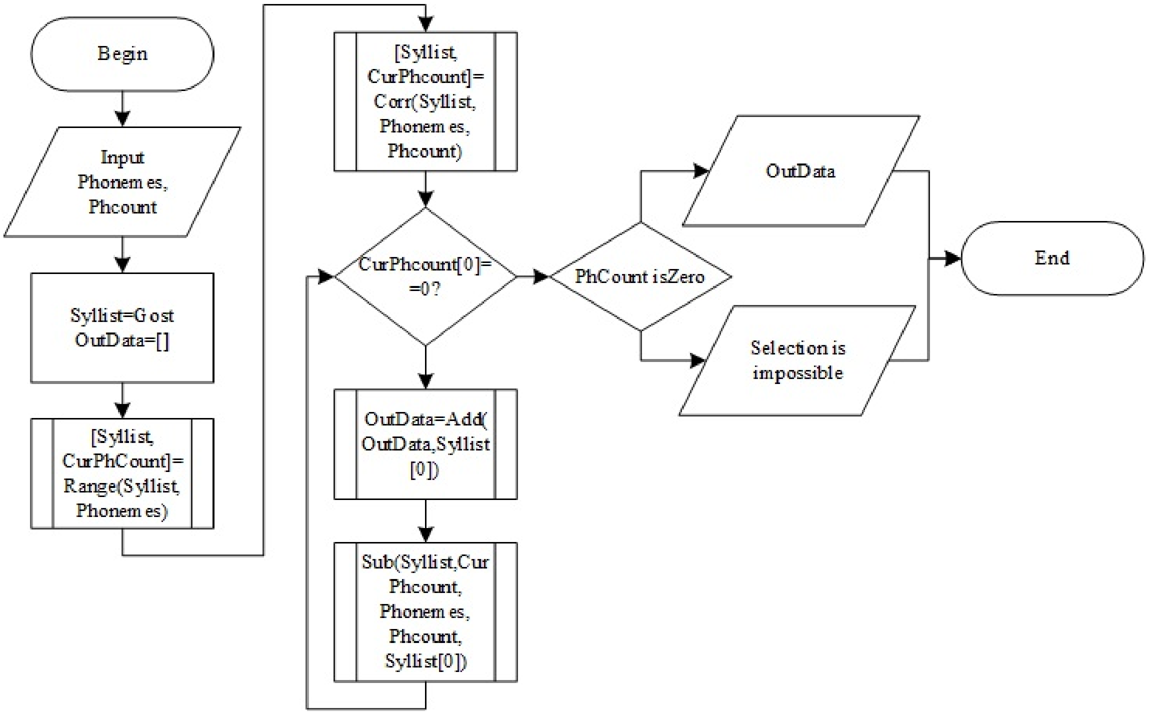

Begin
Input Phonem es, Phcount
Syllist=Gost OutData=[]
[Syllist, CurPhCount]= Range(Syllist, Phonem es)
[Syllist, CurPhcount]= Corr(Syllist, Phonem es, Phcount)
CurPhcount[0]= =0?
OutData=Add(OutData,Syllist [0])
Sub(Syllist,Cur Phcount, Phonem es, Phcount, Syllist[0])
PhCount isZero
OutData
Selection is impossible
End

Figure 5: An algorithm for forming a set of syllables for assessing speech quality and speech rehabilitation

The above description of the algorithm is presented in Figure 5.

The proposed algorithm allows you to add the most useful syllables (from the point of view of the necessary phonemes) to the set, which allows you to create an data set for speech rehabilitation. The formation of a completely optimal set is a solution to the problem of several backpacks, so the possibility of solving it by methods other than exhaustive seems impossible [Mart1990]. Since the addition of redundant phonemes based on the results of the algorithm work is not critical and there is no need to ensure an exact match of the number of phonemes required (an excess over their number is allowed), the proposed algorithm looks workable. Additionally, its performance is confirmed by the sets of syllables formed on its basis.

5 The results of applying an objective approach to assessing the quality of speech and methods of speech rehabilitation based on them

The developed software package allows you to evaluate the effectiveness of traditionally used in the rehabilitation of respiratory and articulatory gymnastics. The conclusion can be made on the basis of the dynamics of changes in the quality of pronouncing phonemes, syllables and phrases when passing control measurements after passing the stages of speech rehabilitation.

Upon receipt of the patient before the operation, the first reference record is used, which serves as a reference for this patient.

Recovery therapy is started in the early postoperative period 10-12 days after surgery, after suturing and removing the nasophageal probe.

In the first lesson, a second recording of the patient's speech is made for an objective assessment that arose as a result of surgical treatment of speech disorders. The speech therapy classes are planned by the results. Individually for each patient, a plan of rehabilitation measures is made, depending on the general condition, the volume of surgical intervention, age, psychological state, profession, and labor orientation. Observe the basic principles of rehabilitation: early initiation of rehabilitation therapy, continuity, continuity, staging, complex nature, the transition from simple to complex.

Articulation exercises for the tongue are carried out 3-5 times a day for 7-10 minutes. Subsequently, when the postoperative wound is cleansed and swelling of the stump of the tongue is reduced, the time is increased to 10-12 minutes and classes are more intense, but the general condition of the patient is taken into account. Each exercise is performed 5-7 times in a row.

After an improvement in the mobility of the articulation organs is noted, patients proceed to the stage of correcting sound pronunciation.

Every 5-7 days, a voice recording is performed to evaluate changes in speech function and to correct tactics for further speech training. Voice training is carried out for problem phonemes.

Speech training with a speech therapist was conducted daily 2 times a day with an interval of 2 hours. When testing the developed solution for one of the patients, the following results were obtained.

Assessment of speech quality was performed using the Speech Quality software package for an objective assessment of speech quality: before the start of combined treatment on July 17, 2018, after the surgical stage of combined treatment at the beginning of speech rehabilitation on August 01, 2018; after the end of speech rehabilitation on August 18, 2018.

Before treatment, the indicator of impaired sound pronunciation was 0.9, this is due to the fact that the tumor occupied a part of the tongue, with metastases of the lymph nodes of the neck. Before the beginning of speech rehabilitation after the surgical stage of the combined treatment in the volume of 1/2 of the tongue and the FFIC of the neck on the right, a violation of speech function was noted and pronunciation rating was equal to 0.466. After completing the restoration of sonorous speech, this indicator was 0.78 (with a maximum of 1.0). When evaluating individual phonemes [k], [t], [s] and their soft variants, the following parameters were obtained, presented in Table 1:

Table 1: The results of assessing the quality of pronunciation of problem phonemes before and after rehabilitation

	01/08/2018	18/08/2018
K	0,404	0,711
K'	0,383	0,606
S	0,342	0,506
S'	0,266	0,371
T	0,662	0,808
T'	0,458	0,708

The proposed method was carried out restoration of speech function in 21 patients.

6 Conclusion

In this paper, we consider an approach to obtaining objective assessments of speech quality. This result combines the previous results obtained in earlier works and offers an algorithm for improving them by expanding the spectrum of the analyzed phonemes. This allows focusing not only on those phonemes that are most susceptible to change following the results of surgery. There is an opportunity to form a set of phonemes based on the wishes of a speech therapist, taking into account the included phonemes and their count. Summarizing the results of studies and practical testing, we can conclude that the development of a software package for speech rehabilitation allows to reduce its terms. According to the results of the developed solution, a patent "A method of restoring speech function in patients with cancer of the oral cavity and oropharynx after organ-preserving operations" for an invention was obtained [Pat19].

Acknowledgements

This research was funded by a grant from the Russian Science Foundation (project 16-15-00038)

References

[Kap18] A.D. Kaprin, V.V. Starinskiy, G.V. Petrova Malignancies in Russia in 2018 (Morbidity and Mortality). *MNIOI name of P.A. Herzen, Moscow*, 2018.

[GOST] *Standard GOST 50840-95. Voice over paths of communication (1995) Methods for Assessing the Quality, Legibility and Recognition.* Publishing Standards, Moscow January 01, 1997, p. 234.

[Spec13] L.N. Balatskaya, E.L. Choinzonov, S.Y. Chizevskaya, E.Y. Kostyuchenko, R.V. Meshcheryakov Software for assessing voice quality in rehabilitation of patients after surgical treatment of cancer of oral cavity, oropharynx and upper jaw *15th International Conference on Speech and Computer, SPECOM 2013, Pilsen, Czech Republic*, 2013, Volume 8113 LNAI, 2013, pp. 294-301.

[MOSPh] V.P. Poltorak, O.M. Morgal, Y.A. Zaika Assessment of voice quality in IP-telephony *Young scientist*, 2014, Volume 4, pp. 121-123.

[MOSPh] G.G. Yanovsky Assessment of voice quality in IP networks *Newsletter*, 2018, Volume 2, pp. 91-94.

[PESQ] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001, Proceedings, Volume 2, pp. 749-752.

[E-model] *ITU-T Recommendation G.107: The E-model: a computational model for use in transmission planning.* 2011.

[SNR] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements Objective Measures of Speech Quality *Prentice Hall, Englewood Cliffs*, 1988.

[EMBSD] W. Yang. Enhanced Modified Bark Spectral Distortion (EMBSD): an Objective Speech Quality Measrure Based on Audible Distortion and Cognition Model. *PhD thesis, Temple University Graduate Board*, May 1999.

[Spec17] E. Kostyuchenko, R. Meshcheryakov, D. Ignatieva, A. Pyatkov, E. Choynzonov, L. Balatskaya Correlation normalization of syllables and comparative evaluation of pronunciation quality in speech rehabilitation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 10458, LNAI, 2017, pp. 262-271.

[SPIIRAS2018] S. Kirillov, V. Dmitriev CA complex algorithm for objective evaluation of the decoded speech signal quality under the action of acoustic interference. *SPIIRAS Proceedings*, Issue 1(56), 2018, pp. 34-55.

[Niko2002] A.N. Nikolaev Mathematical models and a set of programs for automatic assessment of the quality of a speech signal. *The dissertation for the degree of candidate of technical sciences, specialty 05.13.18 - Mathematical modeling, numerical methods and program complexes*, 2002, Ekaterinburg

[Spec19] E. Kostyuchenko, D. Novokhrestova, M. Tirskaya, M. Nemirovich-Danchenko, E. Choynzonov, L. Balatskaya, A. Shelupanov The evaluation process automation of phrase and word intelligibility using speech recognition systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 11658, LNAI, 2019, pp. 237-246.

[DTW] L.R. Rabiner, A.E. Rosenberg, S.E. Levinson Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume 26, Issue 6, 1978, pp. 575-582.

[IPA] M. Duckworth, G. Allen, M.J. Ball Extensions to the International Phonetic Alphabet for the transcription of atypical speech. *Clinical Linguistics and Phonetics : journal*, Volume 4 Issue 4, 1990, pp. 273—280.

[Cub02] P. Cubberley The phonology of Modern Russian. *Russian: A Linguistic Introduction*, 2002, Cambridge University Press.

[Mart1990] S. Martelo, P. Toth Knapsack problems. *Great Britain: Wiley*, 1990, 306 p.

[Pat19] E.A. Krasavina, E.L. Choynzonov, D.I. Novokhrestova, E.Y. Kostyuchenko, S.Y. Chizhevskaya, L.N. Balatskaya A method of restoring speech function in patients with cancer of the oral cavity and oropharynx after organ-preserving operations. *Patent of the Russian Federation for an invention*, 2019.