Article

Assessment of speech quality during speech rehabilitation based on the solution of the classification problem

Evgeny Kostyuchenko [1,*], Ivan Rakhmanenko [1], Alexander Shelupanov [1], Lidiya Balatskaya [1,2] and Ivan Sidorov [3]

[1] Tomsk State University of Control Systems and Radioelectronics, 40, Lenina str., 634050 Tomsk, Russia; key@fb.tusur.ru
[2] Tomsk Cancer Research Institute, 5, Kooperativniy av., 634050 Tomsk, Russia; nii@oncology.tomsk.ru
[3] Irkutsk Supercomputer Center of SB RAS, 134, Lermontova, Irkutsk, 664033, Russia; ivan.sidorov@icc.ru
* Correspondence: key@fb.tusur.ru; Tel.: +7-3822-70-15-29

Abstract: The article considers an approach to the problem of assessing the quality of speech during speech rehabilitation as a classification problem. For this, a classifier is built on the basis of an LSTM neural network for dividing speech signals into two classes: before the operation and immediately after. At the same time, speech before the operation is the standard to which it is necessary to approach in the process of rehabilitation. The metric of belonging of the evaluated signal to the reference class acts as an assessment of speech. An experimental assessment of rehabilitation sessions and a comparison of the resulting assessments with expert assessments of phrasal intelligibility were carried out.

Keywords: Speech Rehabilitation, Speech Quality Assessment, LSTM

1. Introduction

1.1. Relevance of work

The problem of oncological diseases of the organs of the speech-forming tract is urgent. According to statistical studies [1], for the period from 2009 to 2019, there has been a steady increase in such indicators for the localization of the lip, oral cavity, pharynx, as the incidence per 100,000 people, the overall incidence, and the cumulative risk of this type of disease in the age category 0 -74 years old. At the same time, the proportion of tumors of the organs of the speech-forming tract in the total number of oncological diseases remains practically unchanged due to a decrease in the proportion of diseases of the lips. These trends are graphically presented in Figure 1. These quantitative values determine the relevance of research related to oncological diseases of the organs of the vocal tract.

Another feature of this localization of diseases is the influence of its treatment on the quality of life. Surgical treatment requires relearning to speak. This requires a speech rehabilitation procedure. The particular importance of this procedure is due to the fact that the bulk of the sick are of working age, and the lack of speech function significantly reduces the quality of life, preventing most of the communicative functions from being performed both at work and at home. Based on these facts, it can be concluded that developments in the field of increasing the effectiveness of speech rehabilitation are relevant. One of the subspecies of such studies is obtaining objective quantitative assessments of speech quality, which this work is devoted to.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2021 doi:10.20944/preprints202112.0196.v1

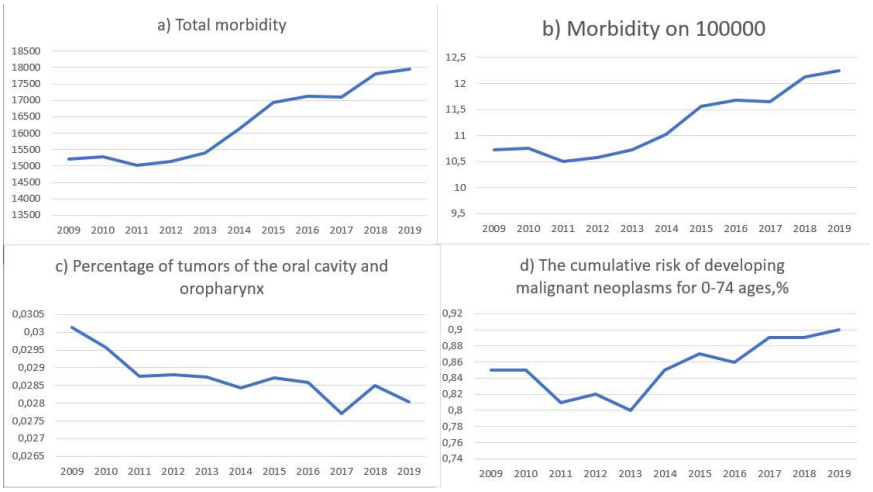

Figure 1. Dynamics of the incidence of oncological diseases of the organs of the speech-forming tract (lips, oral cavity, pharynx). a) morbidity per 100,000 people, b) newly detected cases of diseases, c) the proportion of diseases in the studied localization, d) The cumulative risk of developing malignant neoplasms for 0-74 ages, %

1.2. Existing approaches to assessing speech quality

If we consider the general structure of methods for assessing the quality of speech (Figure 2), then they can be divided into 2 categories: objective and subjective. Subjective assessment methods are based on research and assessment of pronounced units by experts. At the same time, the units themselves can differ significantly: individual phonemes, syllables, phrases. The most striking example of this category is the assessment based on GOST R 50840-95 [2]. For rehabilitation tasks, this standard allows one to obtain estimates of syllable and phrasal intelligibility [3].

Objective assessment methods, in turn, can be divided into 2 classes: they work on the basis of comparison of the same signal before and after transmission and use different signal realizations for assessment. At the same time, the use of the former for the tasks of speech rehabilitation is extremely problematic, since the recordings of the patients' speech before and after the operation are not the same signal before / after exposure. In the second category, many assessment methods have been developed with input from our team. It is possible to distinguish assessment approaches based on the normalization of signals and their subsequent comparison [4, 5] and the use of recognition tools for assessment as a substitute for an expert in the GOST method [6, 7].

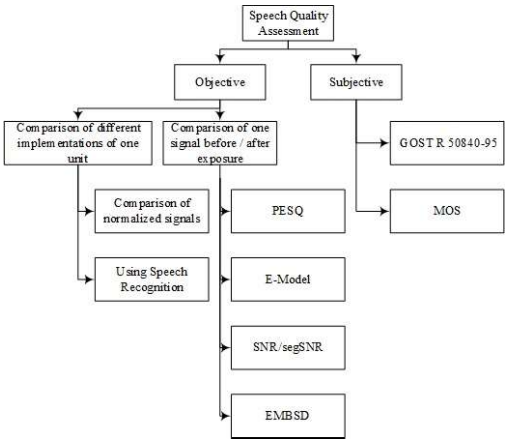

Figure 2. Classification of methods for assessing speech quality

 doi:10.20944/preprints202112.0196.v1

However, the limited number of assessment methods and their incomplete coincidence with the GOST reference method in terms of the accuracy of the esti-mates obtained and their interpretability, suggests the relevance of the search for new approaches to obtaining such estimates.

In this paper, we propose to consider a new class of such methods - based on the application of machine learning methods to obtain an estimate of the speech quality as a result of solving the classification problem.

2. Materials and Method

3.1. Dataset description

During the experiment, we used a previously collected set of recorded phrases from GOST. Made 25 records of phrases in one session. The number of patients with two sessions (before and after surgery) is 24, with three sessions - 18, with four sessions - 7. The total number of records is 3250. The sampling rate is 12000 Hz. The number of pairs of sessions suitable for constructing the classifier was 49. To construct the classifier, 80% of the sets were selected into the training set, the remaining 20% into the test set.

During processing, each signal was converted into a spectral form using the Fourier transform, block length 64 ms, 50% overlap. After that, the obtained spectrograms were transferred to the input of the classifier. This approach to con-structing inputs is basic [8] and is suitable as the first iteration in constructing a classifier.

3.2. Speech quality assessment based on the classification problem

The main idea of the proposed approach is easy to understand. At the time of the visit to the clinic, the patient, despite the presence of the disease, practically does not disturb the intelligibility of speech. The resulting grades of phrasal intelligibility are almost always equal to 1, and the grades of syllabic intelligibility are close to 1 (differences may arise more due to incorrect reading of syllables than due to their incorrect pronunciation). This fact allows us to speak about the possibility of using the notes before the operation as a standard of speech for a particular patient. This approach allows us to take into account the presence of speech features and individual defects in the patient, because further comparison will go exactly with the speech of a particular patient.

After the operation, speech intelligibility is significantly reduced. The final value depends on the volume and localization of the surgical intervention, however, syllabic intelligibility in some cases may fall below 0.1.

In fact, we can say that we have 2 classes of records: before and after surgery. Within the framework of the proposed approach, it is proposed to build a ma-chine learning system that solves the problem of determining whether the presented record is a record before or after the operation. If you train such a system to solve the described problem, then there is an opportunity to present it with the notes made during the rehabilitation process and use the metric of belonging to the reference class as an assessment of the quality of pronouncing the phrase.

3.3. Speech quality assessment based on the classification problem

A neural network was chosen as a machine learning method for constructing the classifier. The use of such networks is typical for solving a variety of speech analysis problems, such as speech recognition [9, 10], authentication [11, 12], sentiment determination [13], and others. For this reason, it was decided to use this particular type of city when constructing the classifier.

Considering the small amount of data and examples of using these networks for speech analysis tasks [14], a neural network based on LSTM was chosen [15].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2021 doi:10.20944/preprints202112.0196.v1

To combat overfitting, regularization, dropout and batch normalization were applied.

The architecture of this neural network is shown in Figure 3.

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 16, 128)	328704
batch_normalization_1 (Batch	(None, 16, 128)	64
lstm_2 (LSTM)	(None, 64)	49408
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 64)	4160
activation_1 (Activation)	(None, 64)	0
dense_2 (Dense)	(None, 16)	1040
activation_2 (Activation)	(None, 16)	0
dense_3 (Dense)	(None, 1)	17
activation_3 (Activation)	(None, 1)	0

Total params: 383,393
Trainable params: 383,361
Non-trainable params: 32

Figure 3. Neural network architecture

The next section describes the experimental study of the proposed approach and the establishment of its applicability.

3. Results

3.1. All-user training and personalized training

Training was carried out according to two methods: for all users and a separate one only for the user of interest. The second training is based on a limited set of data, but the output is a classifier designed to work with a specific patient. A system trained on all users is more capable of generalizing data, however, due to the lack of focus on working with an individual user, it is likely to show less accurate results in the final assessment. Graphs of changes in accuracy and loss-function depending on the iteration number are presented in Figures 4 and 5.

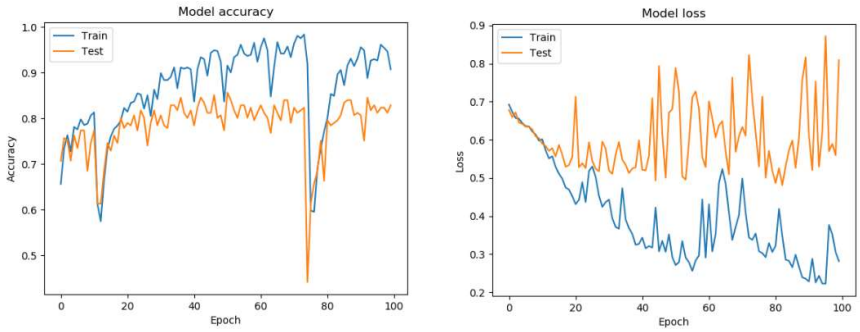

Figure 4. Neural network architecture

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2021 doi:10.20944/preprints202112.0196.v1

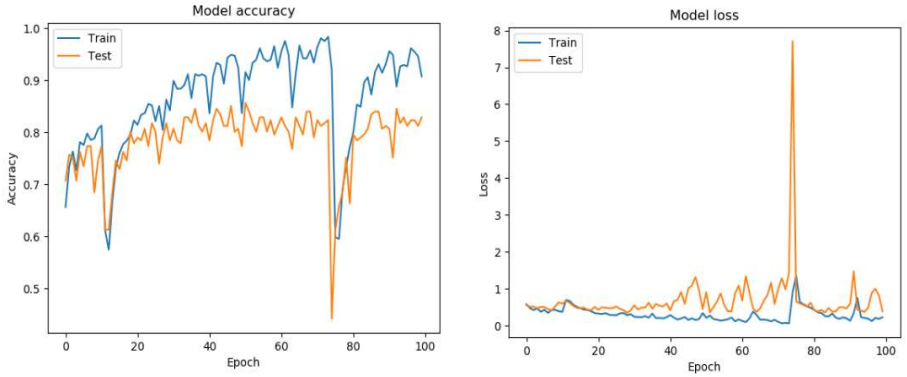

Figure 5. Neural network architecture

It can be seen that it is possible to train the neural network without retraining for one user. The final accuracy for the case without separation of users was 0.8, and there are signs of overfitting.

3.2. Obtaining final speech quality scores

After constructing a ready-made classifier, the signals were processed during the rehabilitation process, their quality was assessed and the resulting estimate was compared with the estimate obtained by an expert. Thus, values were obtained for 32 sessions. The obtained values are presented in table 1.

Table 1. Speech quality estimates obtained using the basic expert method and the proposed classification approach.

No. seans	Expert	Class all	Class one	No. seans	Expert	Class all	Class one
1	,84	,42	,73	17	,84	,46	,71
2	,68	,87	,52	18	,56	,12	,27
3	,88	,68	,66	19	,92	,90	,68
4	,96	,24	,94	20	,68	,84	,55
5	,78	,46	,50	21	,84	,46	,60
6	1,00	,32	,75	22	,92	,12	,88
7	,96	,92	,84	23	,84	,76	,84
8	1,00	,95	,91	24	,92	,46	,90
9	,96	,54	,68	25	,84	,14	,75
10	,96	,42	,79	26	,76	,86	,59
11	,56	,64	,45	27	,84	,42	,69
12	,96	,96	,90	28	,92	,18	,74
13	,70	,28	,63	29	1,00	,28	,83
14	,96	,18	,74	30	,96	,92	,96
15	,96	,94	,95	31	,92	,84	,63
16	,88	,82	,75	32	,84	,50	,55

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2021 doi:10.20944/preprints202112.0196.v1

After the expected receipt of quality assessments, you can proceed to the analysis of the results obtained.

4. Discussion

To assess the results obtained, we will find the correlation coefficient between them and check its statistical significance. The calculation will be carried out using Spearman's rank correlation coefficients. The calculation was carried out in the SPSS program.

The obtained values and the level of their significance are presented in Table 2.

Table 2. Assessment of the significance of the correlation coefficient. **. Correlation is significant at the 0.01 level (two-tailed).

			Expert	ClassAll	ClassOne
Rho Spearman	Expert	Correlation coefficient	1,000	,115	,772**
		Mean. (double-sided)	.	,532	,000
		N	32	32	32
	ClassAll	Correlation coefficient	,115	1,000	,116
		Mean. (double-sided)	,532	.	,526
		N	32	32	32
	ClassOne	Correlation coefficient	,772**	,116	1,000
		Mean. (double-sided)	,000	,526	.
		N	32	32	32

The results show that the results obtained for one user are consistent, which allows us to speak about the absence of obvious contradictions between the considered assessment method. For estimates based on all users, more significant discrepancies are visible and the correlation coefficient turns out to be statistically insignificant. This is due to the fact that during the training all users were united in one class, regardless of the volume and localization of the surgical intervention. Thus, the previous assumption about the best applicability of the method when working with one user is experimentally confirmed.

5. Conclusions

The experiment carried out has shown the potential applicability of the proposed approach based on the application of the classification. The efficiency in solving the problem of dividing speech into classes before / after the operation is shown. The applicability of this approach is shown when constructing a classifier for a specific patient. Spearman's correlation coefficient for estimates obtained by this method and estimates obtained by an expert way is 0.772 and is statistically significant. In the future, it is planned to analyze the applicability of the proposed approach when grouping patients into groups (gender, location and volume of surgery).

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: Conceptualization, E.K., L.B. and I.R.; methodology, E.K., L.B. and I.R.; software, E.K. and I.R.; validation, E.K. and I.R.; formal analysis, E.K. and I.R.; investigation, E.K. and I.R.; resources, E.K., I.S. and I.R.; data curation, E.K., L.B., I.S. and I.R.; writing—original draft preparation, E.K.; writing—review and editing, E.K.; visualization, E.K. and I.S.; supervision, A.S.; project administration, E.K.; funding acquisition, E.K. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Education and Science of the Russian Federation within the framework of scientific projects carried out by teams of research laboratories of educational institutions of higher education subordinate to the Ministry of Science and Higher Education of the Russian Federation, project number FEWM-2020-0042 (AAAA-A20-120111190016-9).

Acknowledgments: The authors would like to thank Irkutsk Supercomputer Center of SB RAS for providing the access to HPC-cluster «Akademik V.M. Matrosov» [16].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaprin, A.; Starinskiy, A.; Petrova, G. Malignant Neoplasm in Russia in 2019 (Morbidity and Mortality).; Hertsen Moscow Oncology Research Center - branch of FSBI NMRRCof the Ministry of Helth of Russia: Moscow, 2020;
2. Standard GOST R 50840-95 Voice over Paths of Communication. Methods for as-Sessing the Quality, Legibility and Recognition; Publishing Standards: Moscow, 1995;
3. Balatskaya, L.N.; Choinzonov, E.L.; Chizevskaya, S.Yu.; Kostyuchenko, E.U.; Meshcheryakov, R.V. Software for Assessing Voice Quality in Rehabilitation of Patients after Surgical Treatment of Cancer of Oral Cavity, Oropharynx and Upper Jaw. In Proceedings of the Speech and Computer; Železný, M., Habernal, I., Ronzhin, A., Eds.; Springer International Publishing: Cham, 2013; pp. 294–301.
4. Kostyuchenko, E.; Meshcheryakov, R.; Ignatieva, D.; Pyatkov, A.; Choynzonov, E.; Balatskaya, L. Correlation Normalization of Syllables and Comparative Evaluation of Pronunciation Quality in Speech Rehabilitation. In Proceedings of the Speech and Computer; Karpov, A., Potapova, R., Mporas, I., Eds.; Springer International Publishing: Cham, 2017; pp. 262–271.
5. Meschryakov, R.V.; Kostyuchenko, E.Y.; Ignatieva, D.I.; Pyatkov, A.V.; Choynzonov, E.L.; Balatskaya, L.N. Speech Quality Measurement Automation for Patients with Cancer of the Oral Cavity and Oropharynx. In Proceedings of the 2016 International Siberian Conference on Control and Communications (SIBCON); May 2016; pp. 1–5.
6. Nikolaev, A.N. 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, Ekaterinburg, 2002.
7. Kostuchenko, E.; Novokhrestova, D.; Tirskaya, M.; Shelupanov, A.; Nemirovich-Danchenko, M.; Choynzonov, E.; Balatskaya, L. The Evaluation Process Automation of Phrase and Word Intelligibility Using Speech Recognition Systems. In Proceedings of the Speech and Computer; Salah, A.A., Karpov, A., Potapova, R., Eds.; Springer International Publishing: Cham, 2019; pp. 237–246.
8. Rippel, O.; Snoek, J.; Adams, R.P. Spectral Representations for Convolutional Neural Networks. arXiv:1506.03767 [cs, stat] 2015.
9. Kipyatkova, I.S.; Karpov, A.A. Variants of Deep Artificial Neural Networks for Speech Recognition Systems. Pr. SPIIRAS 2016, 6, 80, doi:10.15622/sp.49.5.
10. Graves, A.; Mohamed, A.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 2013; pp. 6645–6649.
11. Lim, C.P.; Woo, S.C.; Loh, A.S.; Osman, R. Speech Recognition Using Artificial Neural Networks. In Proceedings of the Proceedings of the First International Conference on Web Information Systems Engineering; June 2000; Vol. 1, pp. 419–423 vol.1.
12. Shukla, A.; Tiwari, R. A Novel Approach of Speaker Authentication by Fusion of Speech and Image Features Using Artificial Neural Networks. International Journal of Information and Communication Technology 2008, 1, 159–170, doi:10.1504/IJICT.2008.0191.
13. Kaya, H.; Karpov, A.A. Efficient and Effective Strategies for Cross-Corpus Acoustic Emotion Recognition. Neurocomputing 2018, 275, 1028–1034, doi:10.1016/j.neucom.2017.09.049.
14. Graves, A.; Jaitly, N.; Mohamed, A. Hybrid Speech Recognition with Deep Bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding; December 2013; pp. 273–278.
15. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to Forget: Continual Prediction with LSTM. Neural Computation 2000, 12, 2451–2471, doi:10.1162/089976600300015015.
16. Irkutsk Supercomputer Center SB RAS Available online: http://hpc.icc.ru/en/ (accessed on 16 January 2021).