Evaluation of syllable intelligibility through recognition in speech rehabilitation of cancer patients

Evgeny Y. Kostyuchenko
Security Dept.
Russia, Tomsk, Lenina 40
key@keva.tusur.ru

Dariya I. Novokhrestova
Security Dept.
Russia, Tomsk, Lenina 40
devijas@yandex.ru

Lidiya N. Balatskaya
Research Dept.
Russia, Tomsk, Kooperativniy 5
balatskaya@oncology.tomsk.ru

Abstract

The approach to the evaluation of speech quality through recognition of spoken syllables in speech rehabilitation of cancer patients after combined treatment of tongue cancer is applied in the work. A neural network of in-depth training is used to assess the pronunciation of syllables during speech rehabilitation. The structure of the neural network has been selected. Estimations of the recognition quality for normal speech (syllables) and the speech after operative intervention during the process of speech rehabilitation are obtained. A conclusion about the applicability of this approach is made and specific recommendations on the choice of the neural network parameters, taking into account the limited volume of records during its training and obvious dependence from speaker were obtained.

1 Introduction

Today the problem of treating patients with oncological diseases of the oral cavity and oropharynx is actual. In accordance with the latest statistics, only in Russia for 2016 about 25 thousand new cases of diseases revealed, while the total number of patients with this tumor localization exceeds 100 thousand people [Kap18, Kap17]. At the same time, an important feature of such disease is the fact that after passing through a combined treatment involving surgical intervention, significant changes occur in the organs of the speech-forming tract. This leads to the need to undergo a procedure of speech rehabilitation by the patient. In fact, the patient is re-learning to pronounce the individual phonemes group considering the existing organ changes. Within the framework of the project, the task of automating the process of assessing the quality of phonemes for the development of an automated system of speech rehabilitation using the biofeedback mechanism is considered. The approaches used at the time of starting work were subjective [Kor15] or required the personal long-term participation of several speech therapists-experts to obtain a quantitative evaluation having an objective character [Kos14]. This

In: Marco Schaerf, Massimo Mecella, Drozdova Viktoria Igorevna, Kalmykov Igor Anatolievich (eds.): Proceedings of REMS 2018 – Russian Federation & Europe Multidisciplinary Symposium on Computer Science and ICT, Stavropol – Dombay, Russia, 15–20 October 2018, published at http://ceur-ws.org

work presents one of the options for obtaining objective quantitative characteristics based on the replacement of auditors by a neural network of deep learning, operating both in the simple recognition mode (in fact, a direct automated analog of the algorithm based on GOST R 50840-95 Voice over paths of communication. Methods for assessing the quality, legibility and recognition [GOST95]), or focusing on specific values outputs for quality assessment.

2 Current state of the problem

The technique from GOST R 50840-95 was considered in more detail, since in fact this work is its direct automation. In comparison with GOST R 51061-97 [GOST97], which also uses tables from GOST R 50840-95, the standard allows the speech therapist to use more ”understandable” estimates. In the framework of the study of speech quality assessment, methods for evaluating syllabic and phrase intelligibility were selected and implemented. As such estimates, the proportion of correctly heard syllables and, accordingly, phrases chosen from special tables were taken. These tables are formed in such a way as to cover all possible combinations of phonemes that arise in real speech. This technique can also be used to assess the quality of a speech source if the influence of the communication channel is absent or negligible. To exclude the technical communication channel, a direct evaluation of the heard syllables by the speech therapist conducting the lesson is used. As part of the assessment, syllable intelligibility is chosen as criterion of pronunciation quality, because it characterizes the quality of phoneme pronunciation without dependence on the context [Kos14].

This work is aimed at replacing the auditor with a deep neural network in this technique [Kip16]. Based on the reference records made before the operation, its training is conducted. Trained neural network is used to recognize syllables in the process of rehabilitation. Due to this, the effect of dependence on the speaker appears, which in this case has a positive effect. This is because all the features of the utterance characteristic of a particular speaker and the parameters of the speech-forming tract are preserved.

In previous works [Kos17] considered the work with the time form of the speech signal to extract parameters. This work implements a fundamentally different approach to quality assessment, proposed, in particular, in [Nik02]. The aggregation of these approaches to obtain characteristics that is most sensitive to changes in the speech quality is assumed in the future.

3 Description of the proposed approach

3.1 The application of deep neural networks for assessing the quality of speech

In this work, for the process of assessing the quality of speech through recognition, an approach based on the application of deep neural networks for image analysis applied to the spectrogram is applied. The recognition procedure uses a Fourier spectrogram in log-Bark scale (40 bands), taking into account the features of perception. For its construction, a speech signal with a sampling frequency of 16,000 Hz is used, the size of the analysis window cut out from the spectrogram, is 10 ms (120 samples). Since in this work the speed of the algorithm is fundamentally not interesting, but the applicability of the approach is important, the step between the windows is chosen to be 1 sample. Further, the resulting matrix of 120×40 size was fed to the input of the neural network for learning and subsequent recognition.

3.2 The principal features of training and limitations under the problem to be solved

The approach implemented within the framework of this task has several limitations, some of which are introduced artificially.

1. Dependence on the speaker. The model for assessing the quality of speech is built every time for a specific speaker. There is no task to improve the quality of speech in relation to the already established manner of pronouncing phonemes and the presence of speech defects. The task in the rehabilitation process is to maximize the speech to the existing standard, the corresponding speech of this particular patient before the procedure of operative treatment of the disease. This limitation significantly simplifies the task from the point of view of speech recognition, there is no need to use a large database of records from a lot of speakers for training.

2. Limited amount of phonemes. We primarily are interested in the quality of pronouncing the phonemes that are most susceptible to change after the operation. By this reason the table of syllables oriented specifically to these problematic phonemes was chosen. The list of these phonemes was compiled at the first stage of the

study [Kos16]. It would be possible to use the complete classical table of syllables from GOST R 50840-95 (5 tables, 250 syllables according to the method of evaluating syllabic intelligibility), however, recording 250 syllables per session is quite problematic for the patient, therefore, in agreement with physicians engaged in speech rehabilitation, it was decided to limit to 90 syllables, but focused specifically on the main problematic phonemes ([k], [s], [t] and their soft implementations). The most problematic phoneme [r] is excluded from consideration, because the mechanism of its utterance changes in principle and direct comparison with the standard is meaningless.

3. The orientation on obtaining a quality assessment as quickly that does not create a problems for patient in the process of training. Now the quality evaluation takes 3 seconds per syllable, the learning time is not important, but takes lesser that one hour.

4. Within the framework of this work, syllable intelligibility refers to the proportion of correctly recognized syllables. In the future, the values of the output layer of the neural network will be used to assess the degree of proximity to the correct phoneme for implementing the biofeedback mechanism in the rehabilitation process. However, in this paper, it was precisely the applicability of the approach to evaluate the quality of phoneme pronunciation in process of speech rehabilitation.

5. It is known in advance which syllable is pronounced. There is no need to interpret the sequence of recognized phonemes, transforming it into a syllable, it is only necessary to estimate the proportion of correct phonemes in this sequence.

3.3 The current state of the database for learning and assessing pronunciation quality

To assess the applicability of this approach, two databases are used. The first is the database of healthy speakers, who pronounce syllables with and without the use of tongue. In this database, there are records of 3 speakers participating in 3 recording sessions (2 sessions with using the tongue for assessing the variation in pronunciation, and 1 session without using the tongue). A small number of speakers in the database relates to the verification of the applicability of the approach. After that, the test was carried out on real patients. The number of patients with records before and after the operation was 79 people.

4 Construction of a deep training neural network and its training

To implement the deep neural network for recognition of syllables in the framework of assessing the quality of their pronunciation, computing environment MATLAB 2018a [Mat18] containing a package Neural Network Toolbox was used, which allows to design flexible deep neural networks without deepening their low-level design. The internal architecture of the neural network (30 layers) was chosen based on the recommendations of the Matlab test pattern for command recognition and looks like this: the input layer, 2×(Convolutional Layer, Batch Normalization Layer, Rectified Linear Unit Layer, Max Pooling Layer), 2×(Dropout Layer, Convolutional Layer, Batch Normalization Layer, Rectified Linear Unit Layer), Max Pooling Layer, 2×(Dropout Layer, Convolutional Layer, Batch Normalization Layer, Rectified Linear Unit Layer) , Max Pooling Layer, Fully Connected Layer, Softmax Layer and Weighted Cross EntityLayer.

The outputs have the following structure: vocalization output, softness output, 21 classes for phoneme identification - total 23 outputs. Input layer contains 4800 neurons.

The total volume was more than 28480 sets (according to an example, the question of selecting the best structure and sampling will be considered in the future), 25000 sets were selected for the training sample.

The final accuracy of training for the 25 epochs and 4875 iterations was 95.75%. Accuracy was calculated as a ratio of problematic phonemes that was correctly detected by neural network on the validation part of dataset.

5 The results obtained for the evaluation of syllable intelligibility

At testing at this stage, the phoneme is correctly recognized if more than 50% of the correct samples were present. The results of testing for assessing intelligibility using experts and the proposed approach for healthy speakers with and without the use of tongue in pronunciation and patients before and after surgery are presented in Table 1. The table was compiled for 3 speakers and 3 patients. PersonN are healthy speakers, PatientN are patients that proceed rehabilitation. "Normal" - standard speech for healthy speaker and speech before operation for patients. "Without tongue" - speech without using of tongue to pronunciation for healthy speaker and speech

Table 1: The results of testing for assessing intelligibility using experts and the proposed approach for healthy speakers with and without the use of tongue in pronunciation and patients before and after surgery. Whole syllable with problematic phonemes. Estimation with its standard deviation for five different experts or neural networks from every person.

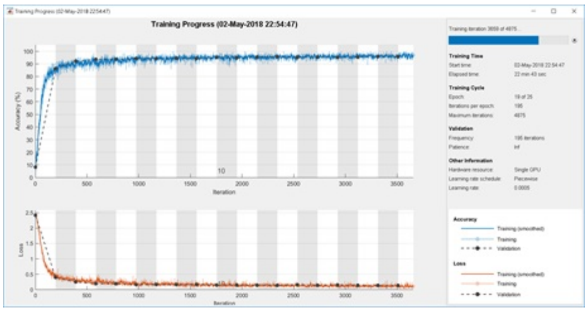
PersonID	Normal/std Expert	Without tongue/std Expert	Normal/std Network	Without tongue/std Network
Person1	1.00/0.000	0.43/0.011	0.64/0.045	0.24/0.057
Person2	1.00/0.000	0.43/0.011	0.43/0.039	0.20/0.059
Person3	1.00/0.000	0.57/0.006	0.50/0.042	0.29/0.061
Patient1	1.00/0.000	0.44/0.011	0.53/0.046	0.21/0.043
Patient2	0.99/0.006	0.68/0.017	0.47/0.017	0.31/0.064
Patient3	1.00/0.000	0.57/0.011	0.56/0.040	0.20/0.064

Table 2: The results of testing for assessing intelligibility using experts and the proposed approach for healthy speakers with and without the use of tongue in pronunciation and patients before and after surgery. Just problematic phonemes from syllable.

PersonID	Normal/std Expert	Without tongue/std Expert	Normal/std Network	Without tongue/std Network
Person1	1.00/0.000	0.43/0.011	0.95/0.017	0.60/0.029
Person2	1.00/0.000	0.43/0.011	0.97/0.029	0.51/0.039
Person3	1.00/0.000	0.57/0.006	0.97/0.033	0.65/0.038
Patient1	1.00/0.000	0.44/0.011	0.99/0.033	0.56/0.039
Patient2	0.99/0.006	0.68/0.017	0.97/0.021	0.69/0.042
Patient3	1.00/0.000	0.57/0.011	0.99/0.040	0.59/0.029

after operation for patients. Records contains syllables with problematic phonemes ([t], [k], [s], [t'], [k'], [s'] [Kos16]). List of records contains 90 syllables.

Expert score is the ratio of the number of correctly recognized syllables (phonemes) to the total number of pronounced syllables (phonemes). Network score are the same, but for neural network instead of expert. The diagram of the neural network training is shown in Figure 1. The figure represents typical increasing of accuracy (and decreasing of loss) depending from the time.



Training iteration 3659 of 4875...

Training Time
Start time: 02-May-2018 22:54:47
Elapsed time: 22 min 43 sec

Training Cycle
Epoch: 19 of 25
Iterations per epoch: 195
Maximum iterations: 4875

Validation
Frequency: 195 iterations
Patience: Inf

Other Information
Hardware resource: Single GPU
Learning rate schedule: Piecewise
Learning rate: 0.0005

Figure 1: The diagram of the neural network training.

Table 2 contains the same information, but for calculation of score were used just problematic phonemes instead of whole syllables. This is not principle for expert, but this have very big influence for neural network. The main reason of this fact is bigger count of problematic phonemes in teaching dataset in comparing with other phonemes. As result, network have lesser count of mistake on problematic phonemes.

Count of patients was selected just three for comparing with the model speaker with/without tongue using.

From the consideration of this table the following conclusions are drawn:

1. Even for a healthy speaker, the intelligibility did not significantly reach 100%, thus spreading with the opinion of experts. However, mistakes mostly arose in "non-problematic" phonemes, which is explained by their small share in the syllable table, in particular, some of the phoneme implementations in the table are missing, since recognition was not the ultimate goal of the system.

2. On the other hand, for problematic phonemes, the results of which are presented in Table 2, the difference is statistically insignificant when using the Student's test and the significance level of 0.95. In the future, it is possible to increase this value due to the variation in the structure of the neural network used and its adaptation to the problem being solved.

3. The intelligibility in comparing qualitative assessments at the level "more" - "less" corresponds to expert estimates, which allows to talk about the applicability of the proposed approach for solving the problem of assessing the quality of speech in the process of speech rehabilitation. This fact confirms the consistency at the ranks level of the previously used classical expert method for estimating syllabic intelligibility and the proposed method using neural networks.

6 Conclusion

In this paper, the application of speech recognition based on a deep neural network for the problem of estimating syllabic intelligibility according to the method of GOST R 50840-95 Voice over paths of communication. Methods for assessing the quality, legibility and recognition is considered. In the framework of this method, the final deep neural network can act as an auditor and issue an appropriate quantitative estimate at the output. The received values allows to speak about absence of obvious contradictions between received results and the estimations received by experts. In addition for correct estimates obtaining it is necessary to have the opinion of 5 experts. This significantly reduces the practical applicability of the method with direct experts participation. The use of a neural network instead of experts solves this problem. It is also possible to formulate several points for a more accurate study of the proposed approach for improving the results obtained, additional confirmation of their reliability and implementation within the version of the speech quality assessment complex in the process of speech rehabilitation.

1. Verification of the operation of the system using several trained neural networks that can act as separate auditors (in accordance with GOST R 50840-95, it is planned to use 5 neural networks).

2. The use of a fraction of correctly recognized phonemes on a time interval, as well as the use of quantitative outputs of a neural network to increase the flexibility of the values obtained, currently at the level of a correctly / incorrectly recognized syllable.

3. Verification of the obtained approach on the full extent of available data for the process of rehabilitation of real patients.

6.1 Acknowledgements

Supported by a grant from the Russian Science Foundation (project 16-15-00038)

References

[Kap18] A. Kaprin, V. Starinskiy, G. Petrova /em Status of cancer care the population of Russia in 2016. P. A. Hertsen Moscow Oncology Research Center - branch of FSBI NMRRC of the Ministry of Health of Russia, Moscow, 2018

[Kap17] A. Kaprin, V. Starinskiy, G. Petrova *Malignancies in Russia in 2014 (Morbidity and mortality).* P. A. Hertsen Moscow Oncology Research Center - branch of FSBI NMRRC of the Ministry of Health of Russia, Moscow, 2017

[Kor15] N. Korotkikh, N. Mitin, D. Mishin, E. Ponomarev The Speech Rehabilitation of Patients After Surgical Operations. *Modern problems of science and education.* 1(1), 2015.

[GOST95] *Standard GOST R 50840-95 Voice over paths of communication. Methods for assessing the quality, legibility and recognition.* Publishing Standards, Moscow, 1995

[GOST97] *Standard GOST R 51061-97 Low biterate speech transmission systems. Speech quality characteristics and their evaluation.* Publishing Standards, Moscow, 1997

[Kos14] E. Kostyuchenko, R. Meshcheryakov, L. Balatskaya, E. Choinzonov Structure and database of software for speech quality and intelligibility assessment in the process of rehabilitation after surgery in the treatment of cancers of the oral cavity and oropharynx, maxillofacial area. *SPIIRAS Proceedings* 1(32): 116–124, 2014

[Kip16] I. Kipyatkova, A. Karpov, Variants of Deep Artificial Neural Networks for Speech Recognition Systems. *SPIIRAS Proceedings.* 6(49): 80–103, 2016

[Kos17] 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. *19th Interna-tional Conference on Speech and Computer (SPECOM 2017), LNCS* 10458: 262–271, 2017

[Nik02] A. Nikolaev *Mathematical models and a complex of programs for automatic evaluation of the quality of a speech signal* The thesis of a Cand.Tech.Sci .: 05.13.18., Ekaterinburg,2002

[Kos16] E. Kostyuchenko, D. Ignatieva, R. Meshcheryakov, A. Pyatkov, E. Choynzonov, L. Balatskaya. Model of system quality assessment pronouncing phonemes. *2016 Dynamics of Systems, Mechanisms and Machines*, Omsk, 2016

[Mat18] *Mathworks Homepage* https://www.mathworks.com/. Last accessed 30 Apr 2018