Article

Evaluation of Speech Quality Through Recognition and Classification of Phonemes

Svetlana Pekarskikh ¹, Evgeny Kostyuchenko ^{1,*} and Lidiya Balatskaya ^{1,2}

¹ Tomsk State University of Control Systems and Radioelectronics, Tomsk 634050, Russia; svetlana.pekarskih@mail.ru (S.P.); balatskaya@oncology.tomsk.ru (L.B.)

² Tomsk Cancer Research Institute, Tomsk 634050, Russia

* Correspondence: key@keva.tusur.ru

Received: 24 August 2019; Accepted: 20 November 2019; Published: 25 November 2019

Abstract: This paper discusses an approach for assessing the quality of speech while undergoing speech rehabilitation. One of the main reasons for speech quality decrease during the surgical treatment of vocal tract diseases is the loss of the vocal tract's parts and the disruption of its symmetry. In particular, one of the most common oncological diseases of the oral cavity is cancer of the tongue. During surgical treatment, a glossectomy is performed, which leads to the need for speech rehabilitation to eliminate the occurring speech defects, leading to a decrease in speech intelligibility. In this paper, we present an automated approach for conducting the speech quality evaluation. The approach relies on a convolutional neural network (CNN). The main idea of the approach is to train an individual neural network for a patient before having an operation to recognize typical sounding of phonemes for their speech. The neural network will thereby be able to evaluate the similarity between the patient's speech before and after the surgery. The recognition based on the full phoneme set and the recognition by groups of phonemes were considered. The correspondence of assessments obtained through the autorecognition approach with those from the human-based approach is shown. The automated approach is principally applicable to defining boundaries between phonemes. The paper shows that iterative training of the neural network and continuous updating of the training dataset gradually improve the ability of the CNN to define boundaries between different phonemes.

Keywords: speech quality; syllable intelligibility; speech recognition; speech rehabilitation; cancer

1. Introduction

One of the most common types of tumors of the speech-forming tract organs is cancer of the tongue [1]. Surgical treatment, consisting of a glossectomy [2], leads to the loss of part of the tongue involved in the formation of a number of phonemes. In particular, even during reconstructive surgery, a disruption of the tongue's symmetry occurs. Ultimately, this leads to a disruption of the pronunciation, in particular, of pre-lingual consonants, which leads to a decrease in syllabic intelligibility and speech intelligibility in general.

In previous papers [3,4], we examined the application of automation in speech rehabilitation therapy for people who undergo a surgical intervention on speech organs. Currently, the evaluation of speech quality during rehabilitation is given by several experts (speech therapists) for providing more objective results. This procedure is time-consuming for the experts and requires a patient to come to a hospital. It is not always convenient and possible for a patient. Due to these problems, the idea of automating the rehabilitation process came up.

A speaker-dependent neural network is trained for each patient on audio records pronounced by a patient before a surgery. The neural network learns how a patient pronounces each type of

phoneme, thus aiming to recognize whether the patient pronounces phonemes in the same manner after the surgery. Speech recorded by a patient is a set of syllables which includes the most problematic phonemes, such as /k/, /s/, /t/, etc. (hereinafter referred to as problematic phonemes). The list of these phonemes was compiled at the first stage of the study [4]. We could potentially use the complete classical table of syllables from GOST R 50840-95 [5] (5 tables, 250 syllables according to the method of evaluating syllable intelligibility). However, recording 250 syllables per session is a tedious task for a patient. Therefore, in agreement with physicians engaged in speech rehabilitation, it was decided to limit the complete list to 90 syllables. The sample is oriented towards the main problematic phonemes (/k/, /s/, /t/ and their soft implementations). The most problematic phoneme /r/ was excluded from consideration, because the mechanism of producing this phoneme changes fundamentally after the operation. Consequently, its comparison with the standard is meaningless. The audio records obtained before a surgery were considered as a benchmark. Subsequently, the quality of the speech after the surgery was evaluated in comparison with this benchmark.

In order to train the neural network to recognize phonemes in after-surgery records, it was necessary to know the start and the end point of each phoneme in before-surgery records. The phonemic composition of all syllables was known. However, the start and the end points of each phoneme were not defined. Consequently, this raised the question of phoneme alignment.

Initially, audio records were segmented at a phoneme level manually. The start and the end points of each phoneme in a syllable were defined acoustically and visually using a software tool that displayed spectrograms. For manual segmentation into phonemes and spectrograms, the software Praat [6] and Wavesurfer [7] were used. This approach was quite time intensive. As a result, it was necessary to move on to an automatic segmentation.

The aim of this work was a reduction in the time taken to align phonemes manually in before-surgery audio records.

2. Evaluation of Speech Quality Through Recognition and Classification of Phonemes

2.1. Methods of Syllable Recognition

The task of recognizing syllables and phonemes is part of a more general task of speech recognition. To solve it, when conducting hierarchical analysis (from the level of phonemes and syllables to the level of sentences and their groups), the following approaches can be used:

- hidden Markov models [8,9];
- hidden Markov models and Gaussian mixture models [10,11];
- deep learning neural networks [12,13];
- different hybrid models [14–16].

The second approach is end-to-end speech recognition. It differs from sequential hierarchical analysis in that it allows you to analyze the original signal and move to higher levels of analysis (for example, the level of words), bypassing lower levels [17,18].

In this paper, the recognition of phonemes within isolated syllables without context is of interest. This makes adjustments impossible at the upper levels of analysis. A statistical imbalance in the phonetic material increases the share of “problem” phonemes and contributes to the impossibility of using ready-made speech models. This makes it impossible to use ready-made solutions in the field of speech recognition, such as the Hidden Markov Model Toolkit (HTK) [19], Kaldi [20,21], Sphinx [22], and their local-language analogues.

In assessing the quality of syllable pronunciation, the following options may be identified:

- the direct recognition of pronounced phonemes within a syllable. The main disadvantage of this approach is a large number of classes (which is equal to the number of phonemes). As a result, we have a high level of error. On the other hand, the positive aspects include the straightforward result in the form of a phoneme sequence and the ease of error determination;

- the recognition of phonemes within a syllable as instances of classes formed of phonetic groups. The positive side is that the accuracy increases due to the reduced number of classes. However, it leads to a lack of direct interpretation, hence a lack of direct evaluation of the assessment according to GOST R 50840-95;
- the identification of boundaries between phonetic segments using a neural network for the follow-up application of parametric approaches in comparing these segments. In this case, the accuracy of determining transitions between phonemes is more important than the classification accuracy. A measure of difference between the selected segments is determined on the basis of previously developed parametric methods [3]. The disadvantage of this option is the unavailability of quantitative assessment such as the classic syllable intelligibility (as defined in Section 2.1) in the output of the system.

2.2. Direct Recognition of Phonemes

In the context of recognition, we used an approach for searching audio fragments similar to those in the training dataset.

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

1. Dependence on a speaker. A model for assessing the quality of speech was built for each new speaker. There was no task to improve the quality of speech in relation to the already established manner of pronouncing phonemes or to the presence of speech defects. The task in the rehabilitation process was to maximize the conformity of a patient's after-surgery speech with their speech before the operative treatment. This limitation significantly simplifies the task because there is no need to use a large database of records from many speakers for training.
2. Limited number of phonemes. We were primarily interested in the quality of pronouncing the phonemes that were most susceptible to change after the operation. For this reason, the table of syllables focuses specifically on those problematic phonemes.
3. The assessment speed of one syllable, pronounced by a patient while practicing at the rehabilitation stage, should be as fast as possible. Currently, the evaluation takes 3 seconds per syllable. The training time of the convolutional neural network (CNN) takes less than one hour. The Adam optimizer with a mini-batch size of 128 was used for training of the neural network [23]. However, the training time did not matter much since the period of time between a before-surgery session and the first rehabilitation session was approximately one week.
4. Within the framework of the paper, the term "syllable intelligibility" refers to the proportion of correctly recognized syllables among all of them pronounced by a patient in accordance with a predefined set of syllables. In the future, values of the output layer of the neural network will be used to assess the degree of similarity between a pronounced phoneme and the correct one in order to implement the biofeedback mechanism in the rehabilitation process. However, in this paper, the idea was to prove the applicability of this automated speech quality assessment approach to speech rehabilitation.
5. It was known in advance which syllable was pronounced. There was no need to interpret the sequence of recognized phonemes, transforming it into a syllable; it was only necessary to estimate the proportion of correctly pronounced phonemes in that sequence.

To implement a deep neural network for the recognition of syllables in the framework of assessing the quality of their pronunciation, the computing environment MATLAB 2018a (MathWorks, Natick, MA, USA) [24] containing the Neural Network Toolbox package was used. The internal architecture of the neural network (30 layers) was chosen based on recommendations of the MATLAB test pattern for command recognition. The set of the CNN was as follows: 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, $2 \times$ (Dropout Layer, Convolutional Layer, Batch Normalization Layer, Rectified Linear Unit Layer), Max Pooling Layer, Fully Connected Layer, Softmax Layer, and Weighted Cross Entity Layer. The outputs had the following structure: vocalization output, softness output, and 21 classes for the phoneme identification, for a total of 23 outputs. The input layer contained 4800 neurons.

2.3. Algorithm of Automatic Time Alignment at Phoneme Level

First of all, the speaker-independent neural network was trained on a Russian language audio dataset containing 7 sentences recorded by 10 speakers of both genders. Every audio file was complemented by a transcript describing its phonemic composition and time periods.

The algorithm to align an unsegmented audio file included the following steps:

1. Training a neural network (initially on data from an audio-aligned corpus).
2. Recognition of a phonemic composition as phonetic sequences for each syllable.
3. Determining time-aligned phonemic transcriptions of syllables.
4. Adjustment of recognized phonemic compositions and selection of syllables with a correctly defined phonemic composition.
5. Forming additional data from transcriptions of correctly recognized syllables.
6. Going to step 1 and retraining the neural network on an updated dataset (including new data formed at step 5).

The steps were repeated until the required level of quality was reached.

The main sequence of actions in IDEF0 notation [25] is shown in Figure 1.

Figure 1. The main sequence of actions.

2.3.1. Phoneme Recognition and Time Alignment

At the data preprocessing stage, WAV format audio files at a sample rate of 16 kHz were sliced into overlapping 20 ms frames with a frame step of 1 ms. Mel-frequency cepstral coefficients (MFCCs) were extracted for each frame.

One audio file of a syllable record lasts about 1000 ms (1 s). Thus, we obtained 981 frames of 20 ms length each. The arg max rule was used to compute the label for each time step. We referred to the independent labelling of each time step, or frame. Figure 2 depicts the best path decoding example [26,27] for a 1 s audio file.

Frame position \ Phoneme	0	...	100	101	102	103	...	771	772	773	774	...	980
ch + sh	0.00	...	0.00	0.00	0.00	0.00	...	0.09	0.00	0.00	0.00	...	0.00
d + d'	0.00	...	0.21	0.10	0.00	0.00	...	0.00	0.00	0.00	0.00	...	0.00
g + g'	0.00	...	0.65	0.88	1.00	1.00	...	0.00	0.00	0.00	0.00	...	0.00
k + k'	0.00	...	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	...	0.00
pause	1.00	...	0.00	0.00	0.00	0.00	...	0.13	0.01	0.01	0.00	...	1.00
s + s'	0.00	...	0.00	0.00	0.00	0.00	...	0.78	0.99	0.99	1.00	...	0.00
t + t'	0.00	...	0.00	0.00	0.00	0.00	...	0.00	0.00	0.00	0.00	...	0.00
vow	0.00	...	0.14	0.02	0.00	0.00	...	0.00	0.00	0.00	0.00	...	0.00

Figure 2. Example of a convolutional neural network (CNN) output. The cells highlighted in bold show the highest probable classes. X is the number of frames. Y is the number of outputs for eight target classes.

As a result, we obtained a sequence of labels. Many labels were certainly repetitive because a phoneme lasted a few milliseconds and frames were overlapped. The algorithm of removing the duplicates and calculating the phoneme's start and end points was considered further.

We defined two parameters for this algorithm:

- The minimum length of a phoneme (the minimum number of consecutive frames labeled as the same phoneme) as *min_seq_len*.
- The maximum length of deviations from a consecutive sequence of the same phoneme labels as *max_dev_len*.

These two parameters might vary depending on the type of phoneme. For example, phonemes /g/ and /d/ usually lasted less than phoneme /t/ or /s/.

Considering a simplified sequence shown in Figure 3, which contains only 30 labels, the result of applying the algorithm described further (below Figure 3) is a time-aligned phonemic transcription as presented in Table 1.

1	2	3	4	5	6	7	8	9	10
pause	pause	pause	pause	s + s'	pause	g + g'	g + g'	g + g'	g + g'
11	12	13	14	15	16	17	18	19	20
g + g'	g + g'	d + d'	d + d'	d + d'	vow	vow	vow	vow	vow
21	22	23	24	25	26	27	28	29	30
s + s'	s + s'	s + s'	s + s'	s + s'	pause	pause	pause	pause	pause

Figure 3. Example of the output label sequence at each time step.

If we suppose *max_dev_len* = 1 and *min_seq_len* = 5 for all the phonemes, then:

1. **SEQUENCE_1.** 1:pause/ 2:pause/ 3:pause/ 4:pause/ 5:s + s' (*dev_len* = 1)/ 6:pause/ 7:g + g' (*dev_len* = 2 > *max_dev_len*, *seq_len* = 6 ≥ *min_seq_len*) ≥ "pause 1–6" is appended to the result

2. **SEQUENCE_2.** 7:g + g'/ 8:g + g'/ 9:g + g'/ 10:g + g'/ 11:g + g'/ 12:g + g'/ 13:d + d' (dev_len = 1)/ 14:d+d' (dev_len = 2 > max_dev_len, seq_len = 6 ≥ min_seq_len) ≥ “g + g' 7–12” is appended to the result
3. **SEQUENCE_3.** 13:d + d'/ 14:d + d'/ 15:d + d'/ 16:vow (dev_len = 1)/ 17:vow (dev_len = 2 > max_dev_len, seq_len = 3 < min_seq_len) ≥ nothing is appended to the result
4. **SEQUENCE_4.** 16:vow/ 17:vow/ 18:vow/ 19:vow/ 20:vow/ 21:s + s' (dev_len = 1)/ 22:s + s' (dev_len = 2 > max_dev_len, seq_len = 5 ≥ min_seq_len) ≥ “vow 16–20” is appended to the result
5. **SEQUENCE_5.** is formed in the same way as SEQUENCE_4. “s + s' 21–25” is appended to the result
6. **SEQUENCE_6.** 26:pause/ 27:pause/ 28:pause/ 29:pause/ 30:pause (last frame reached, seq_len = 5 ≥ min_seq_len) ≥ “pause 26–30” is appended to the result

Table 1. Time-aligned phonemic transcription of the syllable [g'is].

No.	Phoneme	Start	End
1	pause	1	6
2	g + g'	7	12
3	vow	16	20
4	s + s'	21	25
5	pause	26	30

2.3.2. Adjustment of Recognized Phonemic Composition

Since time-aligned phonemic transcriptions were defined for all the syllables, it was possible to apply some operations to adjust partly incorrect phonemic compositions of some syllables. We proposed three operations for the adjustment:

- Substitution
- Shift
- Removal

These operations are simple and intuitive, and applied as shown in Figure 4.

Operation	Substitution	Shift	Removal
Real	[k+k'], [vow], [s+s']	[d+d'], [vow], [k+k'], [t+t']	[sch], [vow], [s+s']
Recognized	[t+t'], [vow], [s+s']	[d+d'], [vow], [t+t'], [k+k']	[sch], [vow], [sch], [s+s']
Applying operation	[k+k'], [vow], [s+s'] [t+t']	[d+d'], [vow], , [k+k']	[sch], [vow], [sch] , [s+s']
Result	[t+t'], [vow], [s+s']	[d+d'], [vow], [k+k'], [t+t']	[sch], [vow], [s+s']

Figure 4. Operations for adjusting phonemic transcriptions.

If a phonemic composition of a syllable became completely correct after the adjustment, the time-aligned phonemic transcription of this syllable was used to create new data for further neural network training.

3. Results of Experiments

3.1. The Direct Recognition of Problem Syllables

At this stage, a phoneme was considered to be correctly recognized if more than 50% of the correct samples were present. Results of the assessment of syllable intelligibility by five experts and using the proposed approach are presented in Table 2. The recognition of the whole syllable with problematic phonemes was considered. Estimation with its standard deviation by five different experts and individual neural networks for every person/patient is presented. The group of patients and the group of healthy speakers included three people in each group. The healthy speakers spoke with and without the use of their tongues in order to imitate the pronunciation of patients before and after surgery accordingly. The patients made their audio records before and after undergoing the surgery. “Person No.” are healthy speakers and “Patient No.” are patients who began the rehabilitation. “Normal” is a standard speech for healthy speakers. “Before-surgery” is a standard speech before operation for patients. “Without tongue” is speech without the use of a tongue for healthy speakers. “After-surgery” is speech after operation for patients. Records contain syllables with problematic phonemes (/t/, /k/, /s/, /tʰ/, /kʰ/, /sʰ/ [3]). The list of audio records contains 90 syllables.

Table 2. The results of the assessment of syllable intelligibility by experts and using the proposed approach based on CNN for healthy speakers with and without the use of a tongue and for patients before and after surgery.

Person No.	Normal by the Experts	Without Tongue by the Experts	Normal by the CNN	Without Tongue by the CNN
Person1	1.00	0.43	0.64	0.24
Person2	1.00	0.43	0.43	0.20
Person3	1.00	0.57	0.50	0.29
Patient No.	Before-Surgery by the Experts	After-Surgery by the Experts	Before-Surgery by the CNN	After-Surgery by the CNN
Patient1	1.00	0.44	0.53	0.21
Patient2	0.99	0.68	0.47	0.31
Patient3	1.00	0.57	0.56	0.20

Table 3 contains the same information, but for the calculation of the scores only problematic phonemes were used instead of the whole phoneme composition of syllables. This nuance is not substantial for experts, but has an important influence on the neural network. The main reason for this fact is the larger number of problematic phonemes in the training dataset in comparison with other phonemes. As a result, the neural network has a much smaller number of errors with respect to problematic phonemes.

Table 3. The results of the assessment of syllable intelligibility by experts and using the proposed approach based on CNN for healthy speakers with and without the use of a tongue and for patients before and after surgery. The recognition of only problematic phonemes from syllables is considered.

Person No.	Normal by the Experts	Without Tongue by the Experts	Normal by the CNN	Without Tongue by the CNN
Person1	1.00	0.43	0.95	0.60
Person2	1.00	0.43	0.97	0.51
Person3	1.00	0.57	0.97	0.65

Patient No.	Before-Surgery by the Experts	After-Surgery by the Experts	Before-Surgery by the CNN	After-Surgery by the CNN
Patient1	1.00	0.44	0.99	0.56
Patient2	0.99	0.68	0.97	0.69
Patient3	1.00	0.57	0.99	0.59

After considering Tables 2 and 3, the following conclusions were drawn.

1. In Table 2, even for a healthy speaker, the syllable intelligibility calculated by the CNN did not reach 100%, thus diverging from the opinion of the experts. However, mistakes mostly arose from “non-problematic” phonemes, which is explained by their small share in the syllable set 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 3, the difference is statistically insignificant when using the Student’s *t*-test with the 0.95 significance level. 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 qualitative assessment of syllable intelligibility given by the CNN corresponds to the experts’ ones. This fact allows a discussion about the applicability of the proposed approach for solving the problem of speech quality assessment during speech rehabilitation. It also confirms the consistency at the level of ranking positions between the classical expert method for estimating syllable intelligibility and the proposed method using neural networks.

3.2. The Speaker-Dependent Neural Network for Class Segmentation

Retraining the neural network on an updated dataset makes it more and more speaker-dependent as well as more precise for a certain speaker-patient.

Figure 5 depicts the graphical representation of gradual changes in a phonemic composition of syllable [g’is] recognized by the neural network.

Figure 5. Example of an output label sequence at each time step. “Iteration 0” represents the output of the speaker-independent neural network. Starting from the first iteration, the neural network becomes speaker-dependent. “Iteration 3” and “Iteration 6” allow us to see the progress of retraining.

Among the characteristic features, the decrease in phonemes belonging to the class $[d + d']$ in the classification results can be distinguished. The count of windows with this class decreases with an increase in the number of iterations. This fact can be explained by the relatively small representation of the given class in the general sample. As a result, with additional training, fewer values fall into the training set, which leads to a decrease in the detection of this class. If we compare this result with the set of sounds $[g + g']$ and $[k + k']$ when solving the ultimate problem of setting boundaries with the known phonetic composition of the syllable and an a priori representation of the class, the following dependence is visible. The significance of the class that is absent in the identified syllable decreases with an increase in the number of iterations.

The experiment of syllable segmentation at a phoneme level was conducted with following characteristics:

- 1 initial speaker-independent neural network trained on 7 sentences recorded by 10 speakers;
- 6 Russian-speaking patients (3 female and 3 male voices);
- 8 classes of phonemes were considered: $[ch + sh]$, $[d + d']$, $[g + g']$, $[k + k']$, $[pause]$, $[s + s']$, $[t + t']$, $[vow]$;
- a set of 15 syllables (see Table 4).

Table 4. Syllable set.

No.	Serial Number	Syllable (in accordance with the International Phonetics Association (IPA) [28])	Phonemic Composition (in accordance with 8 classes)
1	265	[k'as ^j]	[k + k', vow, s + s']
2	271	[d'okt]	[d + d', vow, k + k', t + t']
3	282	[k'æs ^j ɐj]	[k + k', vow, s + s', t + t']
4	283	[k'əs]	[k + k', vow, s + s']
5	289	[sk ^j et]	[ch + sh, k + k', vow, t + t']
6	295	[s'os ^j]	[s + s', vow, s + s']
7	298	[s'it ^ε]	[s + s', vow, ch + sh]
8	305	[d'is]	[d + d', vow, s + s']
9	306	[g'is]	[g + g', vow, s + s']
10	310	[s'it ^ε]	[s + s', vow, ch + sh]
11	311	[s'ies ^j]	[s + s', vow, s + s']
12	316	[ks'et]	[k + k', s + s', vow, t + t']
13	323	[s'is ^j]	[ch + sh, vow, s + s']
14	332	[st'it ^ε]	[s + s', t + t', vow, ch + sh]
15	351	[t ^ε 'æt ^j]	[ch + sh, vow, t + t']

Table 5 presents the results obtained for each patient. The neural network was retrained as many times as it needed to recognize an accurate phonemic composition for at least 12 out of 15 syllables.

Table 5. Summary of applying the algorithm of automatic syllable segmentation at a phoneme level.

Patient No.	Gender	Fully Correct Recognized Syllables	Problematic Syllables	Number of Iterations
1	female	13/15	[d'okt], [st'it ^ε]	6
2	female	12/15	[k'as ^j], [d'okt], [sk ^j et]	6
3	female	12/15	[sk ^j et], [g'is], [st'it ^ε]	5
4	male	12/15	[d'okt], [ks'et], [st'it ^ε]	2
5	male	12/15	[d'okt], [g'is], [s'it ^ε]	4
6	male	13/15	[ks'et], [st'it ^ε]	5

After considering the results in Table 5, the following conclusions were drawn.

- The proposed algorithm gives the expected result and may be applied for automatic identification of a phonemic composition of syllables as well as for determining the start and the end time points for each phoneme.
- The syllables [d'okt] and [st'it^ε] turned out to be the most problematic syllables. This may be due to two consecutive voiceless phonemes, such as /k/, /t/ for [d'okt] or /s/, /t/ for [st'it^ε].
- The efficiency of the algorithm may be improved by fine-tuning *min_seq_len* and *max_dev_len* parameters based on the specificities of phonemes.

4. Conclusions

In this paper, we proposed an approach to speech quality assessment based on a convolutional neural network. In the context of this work, speech recognition was applied for estimating syllable intelligibility according to the method presented in GOST R 50840-95 "Speech transmission over various communication channels". Methods for assessing the quality of the recognition were considered. In this approach's framework, the final deep neural network can act as an auditor and issue an appropriate quantitative estimate at the output. The received values allow a discussion about the absence of obvious contradictions between the results from the CNN and the estimations provided by experts. In addition, for objective assessments made by humans, it was necessary to have the opinions

of five experts. This significantly reduces the practical applicability of the method with experts' direct 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 in order to improve the obtained results, ensure additional confirmation of their reliability, and confirm the approach's implementation within the version of the speech quality assessment system used during speech rehabilitation.

1. Verification of the system's functioning using several trained neural networks that can act as separate auditors [29] (in accordance with GOST R 50840-95, the plan is to use five neural networks).
2. Use of a fraction of correctly recognized phonemes in a time interval, as well as 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 the available data for the rehabilitation process with real patients.

For the task of classification and segmentation:

1. Recurrent training with the addition of data allows researchers to increase the accuracy of speech information analysis.
2. The total classification accuracy without reference to the speaker exceeds 82%.
3. The number of training iterations, at which an increase in the resulting accuracy is recorded, does not exceed six cycles of additional training.

According to the results, the approach for obtaining parametric estimates after segmentation on the basis of the classification of the phonemes of syllables is further proposed.

Author Contributions: Conceptualization, L.B.; Data curation, S.P.; Methodology, E.K.; Project administration, E.K.; Software, S.P.; Supervision, E.K.; Validation, E.K. and L.B.; Writing—original draft, S.P.; Writing—review & editing, L.B. and E.K.

Funding: This research was funded by a grant from the Russian Science Foundation (project 16-15-00038) (The application of neural networks and the subject area for the experiment). The results were obtained as part of the implementation of the basic part of the state task of the Ministry of Education and Science of the Russian Federation, project 8.9628.2017/8.9 (The algorithm of segmentation by phonemes).

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

References

1. del Carmen Migueláñez-Medrán, B.; Pozo-Kreiling, J.-J.; Cebrián-Carretero, J.-L.; Martínez-García, M.-Á.; López-Sánchez, A.-F. Oral squamous cell carcinoma of tongue: Histological risk assessment. A pilot study. *Med. Oral. Patol. Oral Cir. Bucal.* **2019**, *24*, e603–e609.
2. Kishimoto, Y.; Suzuki, S.; Shoji, K.; Kawata, Y.; Shinichi, T. Partial glossectomy for early tongue cancer. *Pract. Oto-Rhino-Laryngologica* **2004**, *97*, 719–723. [[CrossRef](#)]
3. Kostyuchenko, E.; Meshcheryakov, R.; Balatskaya, L.; Choinzonov, E. 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. *SPIRAS Proc.* **2014**, *32*, 116–124.
4. Kostyuchenko, E.; Ignatieva, D.; Meshcheryakov, R.; Pyatkov, A.; Choinzonov, E.; Balatskaya, L. Model of system quality assessment pronouncing phonemes. In Proceedings of the 2016 Dynamics of Systems, Mechanisms and Machines, Omsk, Russia, 15–17 November 2016. [[CrossRef](#)]
5. GOST R 50840-95. *Speech Transmission over Various Communication Channels. Techniques for Measurements of Speech Quality, Intelligibility and Voice Identification*; National Institute of Standards: Moscow, Russia, 1995.
6. Boersma, P. Praat, a system for doing phonetics by computer. *Glott Int.* **2002**, *5*, 341–345.

7. Sjölander, K.; Beskow, J. Wavesurfer—An open source speech tool. In Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000; Volume 4, pp. 464–467.
8. Song, J.; Chen, B.; Jiang, K.; Yang, M.; Xiao, X. The Software System Implementation of Speech Command Recognizer under Intensive Background Noise. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *563*, 052090. [\[CrossRef\]](#)
9. Betkowska, A.; Shinoda, K.; Furui, S. Robust speech recognition using factorial HMMs for home environments. *Eurasip J. Adv. Signal Process.* **2007**, 20593. [\[CrossRef\]](#)
10. Mahadevaswamy; Ravi, D.J. Performance of isolated and continuous digit recognition system using Kaldi toolkit. *Int. J. Recent Technol. Eng.* **2019**, *8*, 264–271.
11. Thimmaraja Yadava, G.; Jayanna, H.S. Creation and comparison of language and acoustic models using Kaldi for noisy and enhanced speech data. *Int. J. Intell. Syst. Appl.* **2018**, *10*, 22–32.
12. Shewalkar, A.; Nyavanandi, D.; Ludwig, S.A. Performance Evaluation of Deep neural networks Applied to Speech Recognition: RNN, LSTM and GRU. *J. Artif. Intell. Soft Comput. Res.* **2019**, *9*, 235–245. [\[CrossRef\]](#)
13. Tóth, L. Phone recognition with hierarchical convolutional deep maxout networks. *Eurasip J. Audio Speech Music Process.* **2015**, *25*, 1–13. [\[CrossRef\]](#)
14. Mendiratta, S.; Turk, N.; Bansal, D. ASR system for isolated words using ANN with back propagation and fuzzy based DWT. *Int. J. Eng. Adv. Technol.* **2019**, *8*, 4813–4819.
15. James, P.E.; Mun, H.K.; Vaithilingam, C.A. A hybrid spoken language processing system for smart device troubleshooting. *Electronics* **2019**, *8*, 681. [\[CrossRef\]](#)
16. Novoa, J.; Wuth, J.; Escudero, J.P.; Fredes, J.; Mahu, R.; Yoma, N.B. DNN-HMM based Automatic Speech Recognition for HRI Scenarios. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, 5–8 March 2018; pp. 150–159.
17. Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End 2000 Automatic Speech Recognition. *Symmetry* **2019**, *11*, 1018. [\[CrossRef\]](#)
18. Wang, D.; Wang, X.; Lv, S. End-to-End Mandarin Speech Recognition Combining CNN and BLSTM. *Symmetry* **2019**, *11*, 644. [\[CrossRef\]](#)
19. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; et al. *The HTK Book (v3.4)*; Engineering Department, Cambridge University: Cambridge, UK, 2009.
20. Trmal, J.; Wiesner, M.; Peddinti, V.; Zhang, X.; Ghahremani, P.; Wang, Y.; Manohar, V.; Xu, H.; Povey, D.; Khudanpur, S. The Kaldi OpenKWS System: Improving low resource keyword search. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, 20–24 August 2017; pp. 3597–3601.
21. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. Available online: <https://infoscience.epfl.ch/record/192584> (accessed on 14 October 2019).
22. Lee, K.-F.; Hon, H.-W.; Reddy, R. An Overview of the SPHINX Speech Recognition System. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 35–45. [\[CrossRef\]](#)
23. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; p. 149801.
24. Mathworks Homepage. Available online: <https://www.mathworks.com/> (accessed on 14 October 2019).
25. IDEF: Integrated Computer-Aided Manufacturing (ICAM) Architecture, Part II (1981) Volume VI—Function Modeling Manual, 3rd ed.; USAF Report Number AFWAL-TR-81-4023; Wright-Patterson AFB: Dayton, OH, USA, June 1981.
26. Graves, A.; Fernandez, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
27. Scheidl, H.; Fiel, S.; Sablatnig, R. Word Beam Search: A Connectionist Temporal Classification Decoding Algorithm. In Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR 2018), Niagara Falls, NY, USA, 5–8 August 2018; pp. 253–258.

28. International Phonetic Association. Available online: <https://www.internationalphoneticassociation.org> (accessed on 30 April 2019).
29. Karpov, A.A. An automatic multimodal speech recognition system with audio and video information. *Autom. Remote Control* **2014**, *75*, 2190–2200. [[CrossRef](#)]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).