Comparison of recognition using Google and Kaldi to solve the problem of assessing intelligibility

E Y Kostyuchenko¹, L N Balatskaya¹, S S Kharchenko¹, M A Lapina²

¹Biomedical Research Laboratory Tomsk State University of Control Systems and Radioelectronics Tomsk, Russia

² Institute of Mathematics and Information Technologies named after Professor N.I. Chervyakova North Caucasus Federal University Stavropol, Russia

E-mail key@fb.tusur.ru

Abstract. The paper is devoted to improving the functionality of the speech intelligibility assessment solution. A previously implemented approach based on the Google Cloud Speech-to-Text solution has shown controversial quality indicators due to its attachment to the global news context. The paper proposes and evaluates the use of the Kaldi system for recognition. This solution not only made it possible to assess the quality of speech without using the Internet, but also to improve the quality of recognition in isolated phrases by excluding the global news context.

Keywords: speech intelligibility, speech recognition, verbal intelligibility, Kaldi

1. Introduction

The problem of obtaining automatic objective assessments of speech quality through intelligibility is highly relevant due to its potential use in several critical areas. The first such direction is the use of intelligibility when assessing the possibility of information leakage through acoustic channels [1]. This task is relevant both for assessing the current acoustic environment in a room and for assessing the effectiveness of using noise to protect information circulating in the room [2].

Another important area is the use of the obtained assessments of the speech quality during speech rehabilitation of patients after the treatment of oncological diseases of the organs of the speech-forming tract [3].

The problem of the incidence of cancer is relevant both in the world and in Russia. So, in Russia alone in 2018, more than 624,000 [4] cases were registered for the first time in their life, diagnosed with malignant neoplasms. At the same time, the number of registered cases is growing from year to year.

One of the localizations encountered are tumors of the organs of the speech-forming tract, in particular the oral cavity and oropharynx.

Another negative side of the statistics given in the above work is the susceptibility to diseases of the population of working age. Loss of working capacity and getting a disability is a colossal problem from both an economic and psychological point of view.
One of the ways to improve the effectiveness of rehabilitation is to form objective assessments of the quality of pronunciation without the participation of experts, which will allow individual training sessions to be carried out independently. This will increase the intensity of the rehabilitation process through independent training and work with a speech therapist within the framework of staged exercises and checkpoints.

Based on this, the task of automating the receipt of objective assessments of speech quality is urgent. Within the framework of this work, we compare the use of two speech recognition systems: Google Cloud Speech [5] and Kaldi [6] in solving this problem, and the possibility of obtaining such estimates without using access to the Internet is carried out.

2. Review of related works

Consider the classification of existing methods for assessing speech quality, shown in Figure 1 [7].

![Figure 1. Relationships and relationship Collections](image)

It should be borne in mind that this classification is rather arbitrary. So, the methods referred to it as subjective (GOST R 50840-95 [8], Mean opinion score [9], Modified Rhyme Test [10]) because they use direct expert judgment. Within the framework of the methods themselves, such a method of increasing objectivity as using several experts and increasing objectivity by increasing the number of experts is used. However, the number of such experts varies, as a rule, from 3 to 5, it is impossible to exclude them, therefore the category is designated as Subjective, not only and not so much from the point of view of the subjectivity of the estimates obtained, but also from the point of view of the impossibility of directly excluding the participation of auditors-subjects.

Objective methods are divided into 2 categories: comparison of signals before / after transmission / processing (Perceptual evaluation of speech quality [11], E-model [12], Signal-to-noise ratio [13], Enhanced Modified Bark Spectral Distortion [14]). This approach is popular when assessing the quality of communication channels, however, its use for assessing the intelligibility of a single speaker in its pure form seems to be poorly possible, since even if by analogy we compare the signals before / after the operation, then from the point of view of the methods of this category, these are two fundamentally different signals and their comparison using such methods will not be correct.
The second category is for comparing differing signals. There are two approaches to assessment:

1. Use various methods of signal normalization in order to bring them to such a form that it would be possible to apply methods from the previous category or use difference metrics for signals of identical duration.

2. Use an assessment approach using speech recognition methods, in which subjective expert methods can be used, in which the expert is replaced by a recognition system.

Within the framework of this work, it is assumed that the latter category of methods will be used, which is used to obtain estimates of intelligibility according to GOST R 50840-95 [8] and compares the application of Google Cloud Speech and Kaldi speech recognition systems.

3. Research methods and data used

3.1. Research methods

As part of the work, the quality of recognition was assessed using three variations of recognition systems:

1. Google Cloud Speech systems in free implementation using the web interface on the site;
2. a commercial version of the Google Cloud Speech system integrated into the software for assessing speech quality using the Google Cloud Speech API;
3. Kaldi speech recognition libraries.

The fundamental difference between Google Cloud Speech and Kaldi systems is that the former uses Internet resources to transmit speech information for processing to Google servers. This approach has the obvious advantage of being able to use Google's vast computing power to recognize our information. This allows using more time-consuming but accurate recognition algorithms or simply increasing the search depth, improving the quality of the selection of the final recognized phrase from the many analyzed variants. However, the use of external resources has three drawbacks, the first of which, the most obvious, is the need for a continuous connection to the Internet for the system to work. In addition, you need to pay for access to Google Cloud Speech resources when implementing recognition of many phrases from an application for assessing speech quality. The third drawback was revealed at the stage of regular operation of the speech quality assessment system, an effect was revealed, the influence of which was imperceptible at the stage of choosing a recognition system within the framework of work[]. The operation of the system has shown that the use of Internet resources and the computing power of Google has a drawback in the form of attachment of the used recognition model to global news content. As part of the operation of the system, this was expressed in the regular appearance in the recognized phrase of words that dare to be directly related to what is happening in the world (terrorist attacks, elections, the coronavirus pandemic), but have nothing to do with the words actually used. This is due to the fact that, based on operating experience, the recognition model built by Google depends on the frequency of words used in search queries. This approach is relevant when recognizing search queries, especially related to news. But for our task related to the recognition of a fixed set of phrases, this is a disadvantage. In addition, it is not obvious whether information about the address of the requestor and his previous recognition requests is used in making a decision. If so, this can have a positive effect on the quality of recognition, but negatively on the quality of the estimates obtained (we recognize the same sets of phrases). During the preliminary selection, these effects were not noticed, in this regard, when building the research plan, 2 questions were formulated:

1. Is there a difference between the recognition accuracy on a fixed set of phrases for the Google Cloud Speech system in the free implementation using the web interface on the site and the commercial version of the Google Cloud Speech system? Is the discovery of the above effects related to the fact that the free version does not have access to the additional world context or for some other reason, or was this effect simply missed during preliminary testing?

2. When using the offline version of the Kaldi recognition system, which of the effects will be greater: lack of access to more powerful computing resources, which reduces the quality of recognition, or a
reliable lack of access to the global context and previous recognized phrases, which in our case is an advantage.

To answer these questions, it was decided to conduct an experiment to assess phrasal and verbal intelligibility as part of a comparison of approaches 1-3, as well as to compare the recognition results between the implemented versions of the 2-3 system in terms of recognition of individual characters using the Levenshtein distance.

Verbal and phrasal intelligibility is understood as the proportion of correctly recognized words or phrases of the reference signal in the recognized [8].

The Levenshtein distance is defined as the minimum number of one-character operations (namely, insert, delete, replace) required to convert one sequence of characters into another [15].

The measurements are taken on the records made before the operation, before the beginning of rehabilitation and during its process. At the same time, it is not the absolute values of the indicators that are of decisive importance, but the ratio between them (after the surgical treatment, the speech quality decreases, but then increases in the process of rehabilitation).

Thus, during the study, the aforementioned indicators were measured in at least three marked control points both for individual users (the assessment is speaker-dependent, since the comparison occurs with the speech of a particular speaker before the operation), and by the averaging method after normalization according to the indicator before the operation.

3.2. Description of the used speech base

For the experiment, we used patient records made at the Oncology Research Institute of the Tomsk National Research Medical Center during treatment and speech rehabilitation procedures. For the study, records of patients with at least 3 sessions were selected. The number of such patients is 24.

The recordings were made in mono with a sampling rate of 11025 Hz. The number of recordings within one session is 25. The spoken phrases and their transcription are presented in Table 1.

| № | No. | Phrase and transcription |
|---|-----|--------------------------|
| 1 | A white veil lay in the fields | bělaja pělēna ležala na polax |
| 2 | Heroes of the front came to school | v školu přijezžali čereži franta |
| 3 | White steam spreads over puddles | běâj par rasstilajetsja nad kužami |
| 4 | The tank crew understood the task | ekipaz tanka pional zadatjju |
| 5 | This unit works well | ēst blisk rabotaet xarsgo |
| 6 | Steppe fires begin | natsjajutsja stepnije požari |
| 7 | Students water the garden | uštjenikili polivajut voyrod |
| 8 | Heavy pouring is over | tjaželij poljem zakončjilsa |
| 9 | The path ran into a clay ledge | trpînka upejlasj v ýlînîstj ustup |
| 10 | The sun rose over the forest | slântse podnâlajse nal lesam |
| 11 | There were nurses at the entrance | v podjezdje stjajali sanitarki |
| 12 | The meeting place became known | stâb izvîstno mjesto vstreťj |
| 13 | The site is under surveillance | na utjastke vëdješja nabludjenije |
| 14 | The teacher entered the class | v klas vajgel prepodavatelj |
| 15 | The log split in two | polenoko raskîlbsja nadvojse |
| 16 | Need to load the gun | nadž zaradît ružîje |
17. The cart began to creak
direktor sravnit’ doxad s rasxadom

18. Mother took the child to the garden
visokaja raž kõõxalaši

19. The guys were sitting on the shore
švëti prestrelë v dolinë

20. The store sells apples
šudnij zapax ësa osivëžajt

21. The director compared income with expense
den’ biul udívátõlõ xòrõš

The total number of recording sessions is 79. The total number of records analyzed in the evaluation is 1975. The distribution of the number of users depending on the number of sessions is presented in Table 2.

Table 2. Distribution of the number of users depending on the number of sessions

| Number of sessions | Number of users |
|--------------------|-----------------|
| 3                  | 17              |
| 4                  | 7               |

4. Description of experiment results

Table 3 shows examples of the original phrase and the result of its recognition by all three considered methods, as well as the quality metrics used for them.

Table 3. Examples of the original phrase and the result of its recognition by the three considered methods, as well as the quality metrics used for them (1 - phrasal intelligibility, 2 - verbal, 3 - Levenshtein distance normalized by the length of the phrase, subtracted from 1)

| Phrase recognizer type | Number of users | 1    | 2    | 3    |
|------------------------|-----------------|------|------|------|
| The original           | The site is being monitored | 1    | 1    | 1    |
| Google web             | Well, let’s dream plot. | 0    | 0,25 | 0,55 |
| Google API             | Kirov observation | 0    | 0,25 | 0,45 |
| Kaldi                  | Observed frequently | 0    | 0,5  | 0,72 |

Table 4 shows the average values for the recognition quality of the three considered methods, as well as the quality metrics used for them (1 - phrasal intelligibility, 2 - verbal, 3 - Levenshtein distance normalized by the length of the phrase, subtracted from 1). The study for the Google web interface using Levenshtein distance was not conducted due to the lack of the ability to automate the receipt of estimates.
Table 4. Average values for the recognition quality of the three considered methods, as well as the quality metrics used for them (1 - phrasal intelligibility, 2 - verbal, 3 - Levenshtein distance normalized by the length of the phrase, subtracted from 1). Results before surgery, after and after speech rehabilitation sessions

| Phrase Recognizer Type | 1   | 2   | 3   |
|------------------------|-----|-----|-----|
| Google web             |     |     |     |
| Until about.           | 0,53| 0,79| -   |
| After about.           | 0,22| 0,39| -   |
| After p.               | 0,38| 0,61| -   |
| Google API             |     |     |     |
| Until about.           | 0,47| 0,70| 0,78|
| After about.           | 0,15| 0,31| 0,68|
| After p.               | 0,30| 0,52| 0,76|
| Kaldi                  |     |     |     |
| Until about.           | 0,51| 0,73| 0,92|
| After about.           | 0,27| 0,51| 0,76|
| After p.               | 0,44| 0,61| 0,92|

5. Analysis of the obtained results
Consider the results obtained.

1. Speech quality scores obtained using the API are lower than those obtained using the web interface. This fact is significant and within the framework of the conduct can be explained by the influence of two factors.

   - Manual estimation using the web interface was spread over time (due to manual labor costs) and because of this could be obtained taking into account the changing temporal context. However, this assessment is obtained in a manual processing mode and its effectiveness is achieved through the use of methods that do not correspond to the initial class of assessment methods. This is due to the need to perform manual actions, which contradicts the initial principle of automating the assessment.

   - there is no information about the same Google recognition algorithms for data received from the web interface and API. In the conditions of the negative impact of accounting for the global information context, additional capabilities implemented in the API interface can be negative for solving the recognition problem on a limited set of phrases.

2. The tendency for the assessment to fall after surgery and its recovery during rehabilitation is confirmed regardless of the assessment method. Taking into account the verification of this approach for recognition using the web interface, it can be assumed that the results obtained for using an alternative software version based on the Kaldi library do not contradict the results obtained on the basis of expert assessments.

3. Estimates obtained on the basis of more small units of the series (phrase-word-sign) are less variable, which is confirmed by the possibility of accepting a larger number of intermediate values of the interval [0; 1] with a larger number of individual values, and the results of the experiment. An important consequence of this is the great promise of evaluations focused on individual signs than on words and,
especially, phrases. This result has been incorporated into speech quality assessment software as a separate metric based on normalized Levenshtein distance.

6. Conclusion
As a result of the study, answers were obtained to two main questions posed before its conduct:
1. The results obtained using the Google Cloud Speech web interface look more preferable than the API version. This fact can be explained:
   - lack of use of the global context,
   - various processing algorithms aimed at recognizing arbitrary spontaneous speech, but not individual fixed phrases taken out of context. An additional factor is the atypical nature of these phrases from GOST in comparison with standard search engine queries;
   - time-spaced testing of the web interface.
2. Benefits of using the offline version of the Kaldi library showed better results on fixed phrases. This allows us to talk about the modification of the existing program for assessing the quality of speech by replacing the recognition module.
3. In addition to the initial tasks of the study, the transition to elementary recognition units (from phrases / words to signs) makes it possible to increase the sensitivity of the resulting estimates of the quality of pronunciation, regardless of the difference between the concepts of letter-sound. The results obtained are implemented in the third version of the speech quality assessment based on intelligibility by replacing the recognition module with the Kaldi module and changing the main quality metric to the Levenshtein distance.

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

References
[1] Likov Y, Morozova G and Kukush V 2017 Influence of features of information leakage channels on intelligibility of eavesdropped voice messages Technology audit and production reserves 1 4–8
[2] Seikutlov Y N, Boranbayev S N, Tashatov N N, Davydau H V and Patapovich A V 2005 Speech information security assessing in case of combined masking signals Journal of Theoretical and Applied Information Technology 12 3270-3281
[3] Kostuchenko E, Novokhrestova D, Tirskaya M, Shelupanov A, Nemirovich-Danchenko M, Choyznzonov E and Balatskaya L 2019 The Evaluation Process Automation of Phrase and Word Intelligibility Using Speech Recognition Systems Speech and Computer Lecture Notes in Computer Science ed A A Salah, A Karpov and R Potapova (Cham: Springer International Publishing) 11658 pp 237–246
[4] Zaridze D G, Kaprin A D and Stilidi I S 2018 Dynamics of morbidity and mortality from malignant tumors in Russia Voprosy Onkologii 64 578–591
[5] Speech-to-Text: Automatic Speech Recognition Google Cloud https://cloud.google.com/speech-to-text
[6] Kaldi ASR https://kaldi-asr.org/
[7] Kostyuchenko E and Novokhrestova D Speech Rehabilitation After Combined Treatment of Cancer and the Formation of a Set of Syllables for Assessing Speech Quality CEUR-WS 2552 11
[8] GOST R 50840-95 Speech transmission over varies communication channels. Techniques for measurements of speech quality, intelligibility and voice identification 1995
[9] Daengsi T, Wutiwiwatchai C, Preechayasboon A and Sukparungsee S 2014 IP Telephony: Comparison of Subjective Assessment Methods for Voice Quality Evaluation Walailak Journal of Science and Technology (WJST) 11 87–92
[10] Brungart D S, Barrett M E, Cohen J I, Fodor C, Yancey C M and Gordon-Salant S 2020 Objective Assessment of Speech Intelligibility in Crowded Public Spaces Ear and Hearing 41 68S
[11] Rix A W, Beerends J G, Hollier M P and Hekstra A P 2001 Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (Cat. No.01CH37221) 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221) 2 pp 749–752

[12] G.107 : The E-model: a computational model for use in transmission planning 2015

[13] Quackenbush S R, Barnwell T P and Clements M A 1988 Objective Measures Of Speech Quality (Englewood Cliffs, N.J: Prentice Hall)

[14] Yang W and Yantorno R 1999 Enhanced modified bark spectral distortion (embsd): an objective speech quality measure based on audible distortion and cognition model 163

[15] Wagner R A and Fischer M J 1974 The String-to-String Correction Problem J. ACM 21 168–173