# Analysis of the Intelligibility of Phonemes at Different Mid-frequency Intervals

Anton Konev [a], Evgeny Kostyuchenko [a], Alexander Shelupanov [a], Evgeny Choynzonov [a,b] and Andrey Nikolenko[a]

[a] *Tomsk State University of Control Systems and Radioelectronics, 40 Lenina Prospect, Tomsk, 634050, Russia*
[b] *Cancer Research Institute of Tomsk National Research Medical Center of the Russian Academy of Sciences (Tomsk NRMC), 5 Kooperativny Street, Tomsk, 634009, Russia*

**Abstract**

The article analyzes the intelligibility of phonemes at various mid-frequency intervals. As part of the work, 6 vowel phonemes were considered, a table of their intelligibility was formed and analyzed. The correctness of the work was assessed by listening to the changed sound track by 7 persons. The results obtained allow us to identify parts in the spectrum that most influencing the intelligibility of phonemes. That parts can be used to assess the quality of speech and the intelligibility of phonemes during the rehabilitation of patients after surgical treatment of the organs of the speech-forming tract.

**Keywords** [1]

Speech intelligibility, speech quality, frequency ranges

## 1. Introduction

To date, automatic speech recognition systems have achieved significant results, investing and promoting commercial applications in this area is beneficial. So, according to BCC Research, by 2021 the world market for speech recognition technologies will be estimated at $ 184.9 billion [1].

Continuous speech recognition [2] and person identification by voice [3] are especially difficult problems for an open set of speakers. The main disadvantages of existing software products are: the need for long-term training of the system and insufficient quality of work with spontaneous speech. It is known that the recorded speech signal differs to one degree or another from the original one. This difference, first of all, is explained by the presence of interference and distortions in the composition of the speech signal recorded at the source.

Another important area where speech recognition can be used is the assessment of its quality, in particular, intelligibility. To assess the quality, standard approaches [4] can be used, provided that they replace the auditor with a recognition system [5, 6].

One of the important aspects in speech recognition is data preparation. Analysis methods can be resource-intensive, therefore, preparing an optimal data set in terms of volume is an urgent task. In terms of optimizing the size of the parameters, the potential reduction in the frequency analysis ranges is important. Filtering out areas that do not affect legibility can have a significant impact on system performance or reduce the amount of resources it consumes.

It is also important that the application of many existing methods for assessing intelligibility requires the participation of experts and, as a result, does not claim to be completely objective. However, such methods claim to be objective by increasing the number of experts, for example, up to five people [4]. The question arises: is there a need for signal preprocessing, the formation of a data

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia

EMAIL: kaa1@fb.tusur.ru (Anton Konev); key@fb.tusur.ru (Evgeny Kostyuchenko); saa@tusur.ru (Alexander Shelupanov); choynzonov@tnimc.ru (Evgeny Choinzonov); office@fb.tusur.ru (Andrey Nikolenko)
ORCID: 0000-0002-3222-9956 (Anton Konev); 0000-0001-8000-2716 (Evgeny Kostyuchenko); 0000-0003-2393-6701 (Alexander Shelupanov); 0000-0002-9084-683X (Evgeny Choinzonov), 0000-0002-7580-5853 (Andrey Nikolenko)

set for building and evaluating a speech recognition and intelligibility system, or in individual tasks (for example, related to assessing intelligibility during speech rehabilitation, when the flow of patients is relatively small, 1-2 per day, and the number of records does not exceed two hundred [4]), is it possible to use expert assessments without problems?

Based on these considerations, preliminary data analysis with the aim of forming the most informative features for further analysis using machine learning methods is an urgent task.

In order to increase the capabilities of continuous speech systems, it is necessary to consider sounds separately. In this work, the features of the behavior of phonemes at various intervals of mid-frequencies are considered and analyzed.

The basic unit of the phonetic level of the language is the phoneme. The concept of a phoneme is associated with the development of understanding of language as an integral system. Professor of Kazan University I. A. Baudouin de Courtenay, who was the first to develop the concept of a phoneme, emphasized that the allocation of a phoneme is possible only when the entire system of phonemes of a given language is taken into account [7].

The phoneme is the minimal meaningful unit of the language, which does not independently have lexical or grammatical meaning, but serves to distinguish morphemes and words.

The phoneme as an abstract unit of language corresponds to the sound of speech as a concrete unit.

The spectrum of speech sound can be decomposed into tone (periodic) and noise (non-periodic) components. Tone sounds are formed with the participation of the vocal cords, noise sounds - by obstacles in the oral cavity. By the presence of these components, the first classification of speech sounds can be made:

- Vowels – tone
- Voiceless consonants – noise
- Sonorous consonants - tones with a slight admixture of noise
- Voiced consonants - noise with tone participation [8]

Differential features of phonemes are associated with the difference in acoustic features of sounds, which, in turn, is associated with differences in their articulation, that is, with a difference in the work of the speech organs. Voicedness - the presence of not only noise in the sound, but also the tone created by the work of the vocal cords; softness - a large pitch of sound caused by a change in the shape of the oral cavity as a result of additional articulation - the rise of the middle part of the back of the tongue to the hard palate.

But for the selection of phonemes, it is not the articulatory and acoustic aspects of these signs themselves that are important, but their opposition, their use to distinguish other linguistic units. The vowel sounds [a], [o], [i] can be pronounced in Russian with different durations (compare the extension of vowels in words when expressing surprise, doubt, indignation, etc.: [ta: m?], [ kn'iga?], [vo: n!] and under.), but the duration of pronunciation in Russian is not used to distinguish between words and forms of words, and therefore, the difference between sounds [o] and [o:], [i] and [i:] do not develop into phonemic differences.

Phonemes that differ in only one differential feature are called paired. Paired in Russian are the phonemes [b] and [b'], [b] and [p], [d] and [d'], [d] and [t] and so on. For example, the phonemes [ts] and [ch] are unpaired, since there is no phoneme that would differ from [ts] or from [ch] just one feature.

Different languages have different types of syllables. The types of syllables differ according to the ratio of the syllable (G) and non-syllable (S) element. When taking into account the end of a syllable, open - the syllable ends with a syllable element (SG) - and closed - the syllable ends with a non-syllable element (GS) - syllables. When taking into account the beginning of a syllable, they distinguish between covered (first sound of a non-syllable) and naked (first sound of a syllable) syllables. In the word [o\kno], both syllables are open, but the first is open and the second is covered, in the word [go\rod] both syllables are covered, but the first is open and the second is closed [9].

For the Russian language, open syllables are more characteristic, consisting of a consonant and a vowel, they make up more than half of all syllables found in speech. Based on this, we can conclude that the analysis of vowel phonemes is an important component of speech recognition and assessment of its quality.

## 2. Description of the data used

In this work, audio recordings are taken from the base of vowel phonemes of the male and female speaker, vowel phonemes are recorded, and also taken from the database of audio recordings of the syllables of the female and male speaker. Audio files in Russian were used for the analysis.

Number of syllables (recordings) from a speaker: 50.

Vowel phonemes: [a], [i], [o], [u], [ɨ], [e].

Before use, audio files were converted to wav format, 16 bit, mono.

The number of auditors is 7 people.

Total number of assessments received from auditors: 350.

Since one of the goals in the work is to demonstrate the problems in assessing intelligibility by expert methods even on small amounts of data, it was not the goal to form a large set intended solely for the application of automated analysis methods based on machine learning methods. It was necessary to identify and evaluate the problems associated precisely with the accuracy and objectivity of the obtained assessments of intelligibility when using standard expert assessments [4].

## 3. Research methods

The Butterworth filter [10] was applied to the recordings to highlight the frequency range of interest. Further, all the data were assessed in the form of a questionnaire, for this 7 persons were selected who had not previously listened to the audio recording data for a more accurate assessment. The results were compared with the baseline data to assess intelligibility.

The following division of the total frequency range of 200-3000 Hz into sub-ranges was used, presented in tables 1-3.

**Table 1**

Splitting a range 200-1000 Hz

| 200-600 Hz | 600-1000 Hz |
|---|---|
| 200-400 Hz | 600-800 Hz |
| 300-500 Hz | 700-900 Hz |
| 400-600 Hz | 800-100 Hz |

**Table 2**

Splitting a range 1000-1800 Hz

| 1000-1400 Hz | 1400-1800 Hz |
|---|---|
| 1000-1200 Hz | 1400-1600 Hz |
| 1100-1300 Hz | 1500-1700 Hz |
| 1200-1400 Hz | 1600-1800 Hz |

**Table 3**

Splitting a range 1800-3000 Hz

| 1800-2200 Hz | 2200-2600 Hz | 2600-3000 Hz |
|---|---|---|
| 1800-2000 Hz | 2200-2400 Hz | |
| 1900-2100 Hz | 2300-2500 Hz | |
| 2000-2200 Hz | 2400-2600 Hz | |

This division is associated with the distribution of the resonant frequencies of the speech-forming tract (formant) for the first formant (table 1) and the second formant of various phonemes (tables 2 and 3) [11]. The lack of division of the last frequency range is due to the fact that even with its full use, the intelligibility turned out to be equal to 0.

For each of the subranges, the intelligibility was assessed as the proportion of correctly defined phonemes averaged over all speakers.

Further, on the basis of the obtained intelligibility values, the most important parts of the spectrum of vowel phonemes influencing the intelligibility of vowels were found.

## 4. Description of the experiment

Let's consider the order of research using the example of the phoneme "a".

1. Select the phonemes of interest
2. We carry out filtering to select the frequency range of interest
3. Save the resulting file
4. Submitting to the experts for assessing intelligibility

Signal spectra for the ranges 200-600 Hz, 200-400 Hz, 300-500 Hz and 400-600 Hz are shown in Figures 1 a-d.

The results of the assessment of intelligibility for the given fragments for the first auditor: 1, 1, 0, 0.

Similar assessments were carried out for all auditors, phonemes and spectrum regions. The final results for assessing intelligibility for isolated phonemes are presented in Tables 4-6.

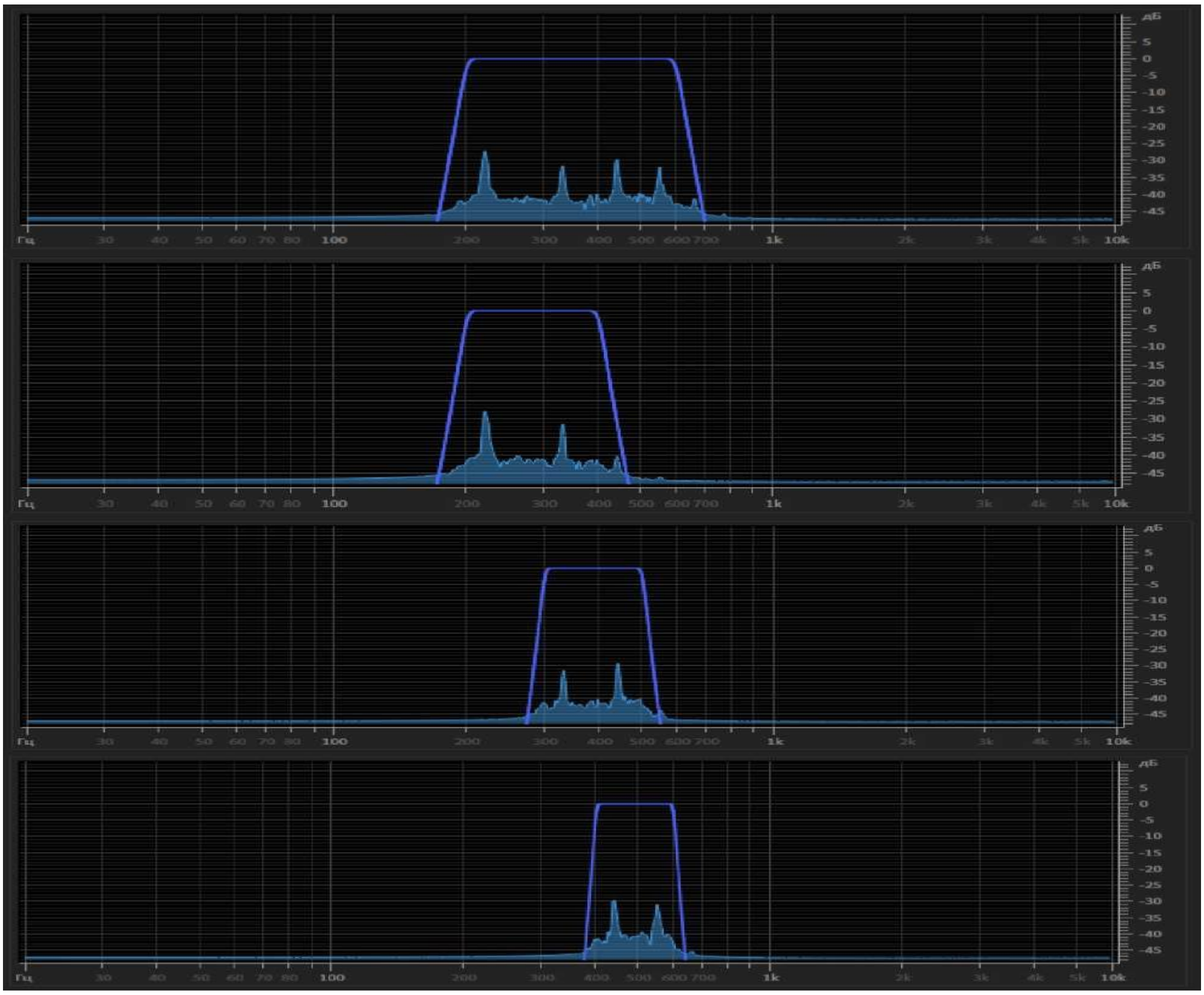

**Figure 1**: Cutting out spectra for signal synthesis in various frequency ranges (200-600 Hz, 200-400 Hz, 300-500 Hz, 400-600 Hz)

**Table 4**
Intelligibility for the range 200-1000 Hz

| | [a] | [i] | [o] | [u] | [ɨ] | [e] |
|---|---|---|---|---|---|---|
| 200-1000 Hz | 1 | 1 | 1 | 1 | 1 | 1 |
| 200-600 Hz | 1 | 1 | 0,148 | 1 | 1 | 1 |
| 600-1000 Hz | 1 | 1 | 1 | 1 | 1 | 1 |
| 200-400 Hz | 0 | 0 | 0 | 0 | 1 | 1 |
| 300-500 Hz | 0 | 0 | 0 | 0,149 | 0 | 0 |
| 400-600 Hz | 0,148 | 0 | 0,148 | 1 | 0 | 0 |
| 600-800 Hz | 0,444 | 0 | 0,147 | 0 | 0 | 0 |
| 700-900 Hz | 0 | 0,147 | 0 | 0 | 0 | 0 |
| 800-1000 Hz | 0 | 1 | 0 | 0 | 0 | 0 |

**Table 5**
Intelligibility for the range 1000-1800 Hz

| | [a] | [i] | [o] | [u] | [ɨ] | [e] |
|---|---|---|---|---|---|---|
| 1000-1800 Hz | 0,297 | 1 | 0 | 0 | 0 | 0 |
| 1000-1400 Hz | 0,296 | 1 | 0 | 0 | 0 | 0 |
| 1400-1800 Hz | 0 | 0,444 | 0 | 0 | 0 | 0 |
| 1000-1200 Hz | 0 | 1 | 0 | 0 | 0 | 0 |
| 1100-1300 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1200-1400 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1400-1600 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1500-1700 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1600-1800 Hz | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 6**
Intelligibility for the range 1800-3000 Hz

| | [a] | [i] | [o] | [u] | [ɨ] | [e] |
|---|---|---|---|---|---|---|
| 1800-2600 Hz | 0 | 1 | 0 | 0 | 0 | 0 |
| 1800-2200 Hz | 0 | 0,149 | 0 | 0 | 0 | 0 |
| 2200-2600 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 2600-3000 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1800-2000 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 1900-2100 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 2000-2200 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 2200-2400 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 2300-2500 Hz | 0 | 0 | 0 | 0 | 0 | 0 |
| 2400-2600 Hz | 0 | 0 | 0 | 0 | 0 | 0 |

Similar characteristics for male and female voices are presented in Table 7, the total is averaged over all vowels.

**Table 7**
Intelligibility for the range 1800-3000 Hz

| | m | w | | m | w |
|---|---|---|---|---|---|
| 200-1000 Hz | 1 | 1 | 1200-1400 Hz | 0 | 0 |
| 200-600 Hz | 1 | 1 | 1400-1600 Hz | 0 | 0 |
| 600-1000 Hz | 1 | 1 | 1500-1700 Hz | 0 | 0 |
| 200-400 Hz | 0 | 0 | 1600-1800 Hz | 0 | 0 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 300-500 Hz | 0 | 0 | 1800-2600 Hz | 1 | 1 |
| 400-600 Hz | 0 | 0 | 1800-2200 Hz | 0 | 1 |
| 600-800 Hz | 0 | 0 | 2200-2600 Hz | 0 | 0 |
| 700-900 Hz | 0 | 0 | 2600-3000 Hz | 0 | 0 |
| 800-1000 Hz | 0 | 0 | 1800-2000 Hz | 0 | 0 |
| 1000-1800 Hz | 1 | 1 | 1900-2100 Hz | 0 | 0 |
| 1000-1400 Hz | 1 | 1 | 2000-2200 Hz | 0 | 0 |
| 1400-1800 Hz | 1 | 1 | 2200-2400 Hz | 0 | 0 |
| 1000-1200 Hz | 0 | 0 | 2300-2500 Hz | 0 | 0 |
| 1100-1300 Hz | 0 | 0 | 2400-2600 Hz | 0 | 0 |

Once the intelligibility scores have been obtained, you can analyze and compare them.

## 5. Analysis of results

It can be noted that in the interval from 1800 to 3000 Hz, intelligibility is not preserved for all vowel phonemes. As an exception, the phoneme [a] and [i] can be distinguished, their ranges are close to high frequencies, but to maintain intelligibility, a wide band is required at intervals of 400-800 Hz. You can also see that the intelligibility of the male voice is lower in the frequency range than the female, which is most likely due to the lower value of the pitch frequency.

It can also be seen that some of the listeners noted the intelligibility of the phoneme [a] in the range from 400 to 600 Hz and in the band from 600 to 800 Hz, therefore, you need to look at the frequency range from 500 to 700 Hz. Additional analysis of this range confirmed the intelligibility within its limits equal to 1.

Phoneme [a] is partially legible at 1000-1800 Hz and 1000-1400 Hz, which means that for legibility it may be necessary to partially expand the range towards 200-1000 Hz. As a result, with the extended range, it turned out that intelligibility was preserved in the range 800-1400 Hz, while the frequencies in the range significance for 400- 600 Hz and 600-800 Hz cannot be denied.

Phoneme [i], the frequency range from 700 to 900 Hz, 14% of listeners noted that intelligibility was preserved, since in the range from 800 to 1000 Hz, intelligibility is preserved, it is possible to distinguish that the range partially coincides, but the main band lies higher, and this assessment is related to individual differences of the listener.

Next, consider the phoneme [i], as with the phoneme [a], we will shift the frequency range and evaluate the results. intelligibility appeared in the 1400 to 2200 Hz range, but as the range decreases, the intelligibility begins to fade, but these ranges significance cannot be denied.

For phoneme [o], the range was extended in the same way as for phoneme [a]. It can be concluded that a wider range is required for intelligibility and the data is due to the individual listener, but this range significance cannot be denied.

In the phoneme [u] in the range 300-500Hz, 14% of the listeners emphasized intelligibility, this is due to the individual differences of the listener and this range significance cannot be denied.

After analyzing the results in syllables, a clear difference can be distinguished in that a wider band is required to preserve audibility, and audibility is also preserved not for a single phoneme, but for the syllable as a whole. As a feature, the intelligibility of the syllables stands out in the aggregate, during the listening, the falling out syllables did not stand out, as a result, the intelligibility or absence is preserved for the entire audio file as a whole.

An intelligibility table was compiled for all midrange syllables. The data can then be used to develop a module for assessing the quality and intelligibility of speech.

## 6. Conclusion

In the course of the study, it can be concluded that the frequency ranges corresponding to the most informative set of features can be compared to the frequencies of the formants of sounds and the frequency of the main tone of the speaker. Indirectly, this dependence depends on the gender of the speaker through its influence on the frequency of the main tone [12]. The data obtained can be used to identify the most informative areas of the phoneme spectrum when solving speech recognition problems and assessing the quality of pronouncing phonemes.

In addition, the studies carried out have clearly confirmed that the use of existing expert methods for assessing intelligibility can introduce significant contradictions due to differences in the perception of messages by five (the recommended number [4]) experts. These results clearly substantiate the need to form a dataset for creating a system based on machine learning for assessing syllabic, verbal and phrasal speech intelligibility when solving, in particular, assessing the quality of speech in speech rehabilitation problems using machine learning for recognition [5].

## 7. Acknowledgements

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

## 8. References

[1] Understand me. "Promobot" engaged in speech recognition technologies”, 2019. URL: https://www.kommersant.ru/doc/3960716

[2] Kipyatkova I.S., Karpov A.A. Analytical review of Russian speech recognition systems with a large dictionary Trudy SPIIRAN – SPIIRAS Proceedings, 2010, vol. 12, no. 1, pp. 7–20

[3] Rakhmanenko, I.A., Shelupanov, A.A., Kostyuchenko, E.Y. “Automatic text-independent speaker verification using convolutional deep belief network”. Computer Optics. 2020.

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

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

[6] I.S. Kipyatkova, A.A. Karpov Variants of Deep Artificial Neural Networks for Speech Recognition Systems. Trudy SPIIRAN – SPIIRAS Proceedings, 2016, vol. 6, no. 49, pp. 80–103

[7] I.A. Boduen d’Kurtene Experience of phonetic alternations, 1895.

[8] L.V. Bondarko Phonetic system of the modern Russian language. Moscow: Prosveshcheniye, 1977.

[9] L.V. Zlatoustova, R.K. Potapova, V.V. Potapov, V.N. Trunin-Donskoi General and applied phonetics. M.: Izdatel’stvo Moskovskogo universiteta, 1997.

[10] L.R. Rabiner, R.W. Schafer. Digital Processing of Speech Signals. — Paramus, NJ: Prentice-Hall, 1978. — ISBN 0-13-213603-1.

[11] L.V. Bondarko, L.A. Verbitskaya, M.V. Gordina Fundamentals of general phonetics. - 4th ed., St. Petersburg: Academy, 2004, 160 p.

[12] H. Kaya, A.A. Salah, A. Karpov, O. Frolova, A. Grigorev, A., E. Lyakso Emotion, age, and gender classification in children's speech by humans and machines Computer Speech and Language Volume 46, November 2017, Pages 268-283