http://dx.doi.org/10.35596/1729-7648-2020-18-1-67-73

Оригинальная статья
Original paper

UDC [004.934+004.056.5]:811.411.21

INTELLIGIBILITY OF THE KAZAKH SPEECH WHEN IT'S PROTECTED WITH COMBINED MASKING SIGNALS

[1]YERZHAN N. SEITKULOV, [1]SEILKHAN N. BORANBAYEV
[2]HENADZI V. DAVYDAU, [2] ALEKSANDR V. PATAPOVICH

[1]L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
[2]Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

Submitted 3 December 2019

Abstract. The article is devoted to assessing the intelligibility of the Kazakh speech when it's masked by combined signals, including «white» noise and speech-like signals. The phonetics features of the Kazakh language have been considered taking into account the law of syngarmonism and the spectrum differences of speech in the Kazakh language and speech in the Russian language. A technique for assessing the intelligibility of the Kazakh speech when it's masked by «white» noise and speech-like signals is proposed. The aim of the work is to analyze well-known methods for speech intelligibility assessing and applying these methods to assess speech intelligibility in the Kazakh language, taking into account masking by its combined signals. Due to the fact that the use of the articulation method of assessing intelligibility for the Kazakh speech requires a dependence of intelligibility on the articulation index for this particular language (the application for the Kazakh language has not been experimentally tested), the use of the formant approach to speech intelligibility assessing will be examined in more detail. The curried out experimental studies of the spectral density of speech in the Kazakh language made it possible to obtain it's approximate dependence on the frequency and take into account the phonetic features of the Kazakh speech when assessing the security of the speech information using the formant method.

Keywords: combined masking signal; security of voice information; «white» noise; speech-like signal.

Conflict of interests. The authors declare no conflict of interests.

Gratitude. This work was supported by grant funding from the Ministry of Education and Science of the Republic of Kazakhstan, No. AP05130293.

For citation. Seitkulov Y.N., Boranbayev S.N., Davydau H.V., Patapovich A.V. Ittelligibility of the Kazakh speech when it is protected with combined masking signals. Doklady BGUIR. 2020; 18(1): 67-73.

Introduction

A sufficient number of methods for assessing speech intelligibility on the background of noise or masking signals have been developed. There is an International Standard. All such methods might be divided into two classes: methods based on the formant approach (formant method) [1, 2] and methods based on the using of the articulation index (articulation method) [3–5]. The formant method has been developed for the Russian language. It was primarily aimed at ensuring the quality

of the speech transmission over communication channels. The articulation method and the articulation index were developed in Bell's laboratory to ensure the quality of communication in aviation technology and was oriented to the English language. Both formant and articulation methods for assessing speech intelligibility have been developed for areas of speech intelligibility above 50 % and after significant improvements they found application for areas of speech intelligibility of several percent for solving problems of protection of the speech information.

The essence of the formant method for speech intelligibility assessing is to find the sum of speech intelligibility in each of the bands of the speech frequency range. Such an assumption is possible if the signal and the masking noise at these frequencies are independent. Formant speech intelligibility is calculated from the expression [1, 5].

$$A = \sum_{k=1}^{k} p_k \cdot w(E_k), \quad (1)$$

where k is frequency band number for which formant speech intelligibility is calculated; p_k is the probability of the formants location in the k^{th} frequency band; $w(E_k)$ is speech perception coefficient as a function of signal-masking noise ratio in the k^{th} frequency band; E_k is the ratio of the level of the speech signal in the k^{th} band to the level of the masking signal in this frequency band.

The probability of formants finding in the k^{th} band is calculated by the distribution function of formants in the speech frequency range and is determined from the expression [1]

$$p_k = F(f_{hk}) - F(f_{lk}), \quad (2)$$

where $F(f_{hk})$ and $F(f_{lk})$ are frequency distribution functions of formants at the highest and lowest frequencies of the k^{th} band accordingly.

The articulation method for assessing speech intelligibility is based on the calculation of the articulation index for a given frequency range.

In the works [4,] it has been proposed to evaluate the security of the speech information using indicators of intelligibility, audibility and cadence (rhythm). It has been proposed to calculate of the speech intelligibility through the SNR or SPI indicator, using signal-to-noise ratios for 16 1/3 octave frequency bands [4, 5, 7, 8]. SNR is determined from the expression

$$SNR = \sum_{f=160}^{5000} \left[L_{ts}(f) - L_n(f) \right] / 16, \text{ dB} \quad (3)$$

where $L_{ts}(f)$ is the level of the speech transmitted to the position of the disturber; $L_n(f)$ is the level of the external noises on the position of the disturber.

The sum (3) is defined for each of 1/3 octave bands with an average frequency f.

The value in square brackets of the expression (3) couldn't be less than –32 dB.

If the signal-to-noise ratio in a particular frequency band is less than –32 dB, then this value is significantly lower than the auditory threshold and such extremely low values will inappropriately exaggerate the speech confidentiality degree. Therefore, it is necessary to limit the values of the difference in signal-to-noise levels in each 1/3 octave frequency band with a value of at least –32 dB.

In this case, the transition from the values of the SNR parameter to the indicator of speech intelligibility is performed using the dependence presented in graphical form. However, this dependence is characteristic of the English speech and the application of this dependence for other languages, including Kazakh, characterized by own phonetic specificity, is very problematic.

Assessing the security of speech information by the parameter speech intelligibility for languages other than Russian and English using the considered methods could introduce a system error due to differences in the speech spectrums of for different languages. In addition, different distributions of languages phonemes on the frequency range will also impact to the speech intelligibility. The differences in the speech spectrums for 12 languages have been studied in [7]. The significant differences both at low and high frequencies have been shown.

Research and comparison of formant properties of the Ukrainian and Russian speech have been performed in [8, 9]. It was found that when signal-to-noise ratios are small and levels of sound pressure of the speech signal are high, intelligibilities of the Russian and Ukrainian speech are almost the same, but when signal-to-noise ratios are large, intelligibility of the Ukrainian speech is noticeably

lower. In this case, the formant method for speech intelligibility assessing was used. When signal-to-noise ratio is small, the same intelligibility for the Ukrainian and Russian speech is due to the influence of the factor that the speech apparatus of the speakers is formed in the conditions of bilingualism and they equally easily know each language.

If the phonetic structure and intelligibility indicators of the Ukrainian and Russian speech are close, then it is necessary to take into account the phonetic features of the Kazakh language, when assessing intelligibility of the speech in this language. In addition, the existing methods for speech intelligibility assessing, discussed above and used in voice information protection systems, are focused on the use of the masking signal – «white», «pink» or another type of noise. The use of combined masking signals in modern voice information protection systems imposes its own characteristics on the speech intelligibility assessment as an indicator of the security of speech information [10–14] and was not reflected in the publications.

Combined masking signals

Combined masking signals used for the speech information protection from leakage via technical channels usually contain a noise component in the form of «white» noise and speech-like signals formed based on the structural units of speech taking into account the probability distribution of their appearance in a given language [11–13].

Quite often, it is recommended to use «pink» noise as a noise component – it's a noise whose spectral density decreases with increasing frequency according to dependence f_0 / f_c, where f_0 is low frequency value of the noise; f_c is current frequency value [15, 16].

An important requirement for masking signals is the requirement that they are generated randomly, i.e. that «white» noise is generated due to thermal noise of semiconductor devices or other nature of physical noise. This requirement is due to the need to exclude any possibility of cleaning the noise from intercepted acoustic signals.

Speech-like signals intended for masking speech information are similar in their formal properties to continuous speech, however, there are temporary sections where the speech-like signal is absent (as well as in natural speech). These sections should be filled with a noise signal to exclude gaps and cases when information signal has got empty temporary sections which are not filled with noise.

Such approach to the formation of speech masking combined signals provides higher levels of security of the speech information, as the difficulty of isolation and processing of any signal increases when interference (combined masking signals, including speech-like signals) becomes closer to the protected signal in shape and frequency. Therefore, one of the promising options for the formation of masking speech-like signals is their formation on the base of the structural units of the speech of the speakers, whose speech signals require an increased degree of protection. Moreover, the formants of the protected speech signals and the formants of the masking speech-like signals are difficult to distinguish.

The method of synthesis of speech-like signals in the Russian and Belarusian languages should be used as a base for the formation of speech-like sequences in the Kazakh language. The features of the formation of speech-like sequences in the Kazakh language are associated with the law of syngarmonism (harmony of vowels and harmony of consonants). Only hard or only soft vowels can be combined in a word of the Kazakh language. Words of foreign origin (Arabic, Persian and Russian) could contain both soft and hard vowels. A restriction on the use in one word (in the root of the word and derivative bases) of either soft or hard vowels was applied when forming speech-like sequences of the Kazakh language. Consonant assimilation is used for voicing and deafness. If the last sound of the root of the word is deaf or ends in voiced *b, v, g, d,* then the initial consonant sound of the affix is deaf. If the last sound of the root of the word is dull consonant *q, k, p* and the initial affix sound is vowel, then the deaf *q, k, p* go into *ġ, g, b*. Sound *a* is not used in words with vowels *á, e, i, ı, ú,* as well as with soft consonants *g, k.*

Sound *g* is not used at the beginning and at the end of the Kazakh language words and in combination with vowels *a, o, u, y*. It is not used in words and in combination with strong consonants *ġ, ķ*. *L* is written at the beginning of the Kazakh language words, but not pronounced,

therefore, it is not used at the beginning of the words of the speech-like sequences. The sound *o* is not used at the end of the Kazakh language words. These features of the Kazakh language have been when forming of the speech-like sequences, which were converted into acoustic speech-like signals to mask speech. Speech-like signals also could be formed as a dialogue of negotiators. At the same time, the ratio of the speech-like signal to the masking «white» noise should be –6 dB. It provide the case when the level of consonant sounds in speech-like signals exceeds the vowels sound in speech information signal.

It is difficult to evaluate analytically the intelligibility of speech masked by combined signals, as the ratio of the speech signal to the combined noise will change over time within small limits and it is necessary to take into account the probability of coincidence of the formants of the speech signal and the formants of the speech-like interference. Therefore, the most acceptable solution is to use the limit state method in calculations and to take into account the specific features of the phonetics of the language.

Method for assessing the intelligibility of Kazakh speech and experimental results

The phonetics features of the Kazakh language, which may affect the intelligibility of the speech in this language, compared to the Russian and other languages, are as follows.

The law of syngarmonism of the Kazakh language is that the vowels of the Kazakh language can be hard or soft. In one word, the vowels can be either hard or soft. Solid vowels are *a, o, u, y*. Soft vowels are *e, á, ó, ú, i*. Experimental studies have shown that the sound pressure level of the words spoken with soft vowels are lower on 1 dB than words the sound pressure level of the with hard vowels. Moreover, the number of words with hard vowels in the Kazakh language is 59 %, and the number of words with soft vowels is 41 % (data obtained from the texts analysis).

The feature of the speech signals is that they have a formant character. Formant is the area of the frequency band in which the main energy is concentrated when pronouncing a certain vowel phoneme. For each vowel phoneme, the number of formants could range from 3 to 5. If consonant sounds have an energy distribution over a frequency band, then vowel sounds are characterized by a concentration of energy in certain areas of the frequency band.

Experimental studies of the energy characteristics of vowels and consonants, deaf and voiced, hard and soft sounds has showed that the energy performance of vowels is about 70–78 dB with a rms sound pressure of 70 dB. In this case, stressed vowels are pronounced at a sound pressure of 73–78 dB. For hissing and whistling sound without clearly expressed formants in the spectrum are characteristic pressure values are 58–63 dB. Speech intelligibility is determined by the relationship between informational speech signals and the level of masking noise with speech-like signals.

The combination of vowels and consonants pronounced with an increased level of sound pressure strongly influents on the speech intelligibility. Vowel sounds are formed on the base of vibrations of the vocal cords and have a greater power than consonants, which are formed by modulating the air stream. The intelligibility of consonants is not the same. The intelligibility of sonor consonants is higher than hissing ones, and the intelligibility of solids consonants is higher than soft ones. To account for the above mentioned phonetic features of the Kazakh language, experimental speech amplitude spectra were carried out in the range from 100 to 8000 Hz, performed for a sample of 14 people. Fig. 1 shows the averaged amplitude spectra of speech in the Kazakh language (lighter dependence is for a female voice, darker dependence is for a male voice).

However, using such dependencies when performing calculations is not entirely convenient. So in [17] it's proposed to approximate the spectral density of the Russian speech in the frequency range from 200 to 5000 Hz by the dependence

$$S_\xi(\omega) = \frac{\rho \cdot \sigma_\xi^2}{\pi}\left[\frac{1}{\rho^2 + (\omega_0 - \omega)^2} + \frac{1}{\rho^2 + (\omega_0 + \omega)^2}\right], \quad (4)$$

where $\rho = 1.14 \cdot 10^3$ s^{-1}, $\omega_0 = 2.98$ s^{-1} ($f_0 = 210$ Hz), σ_ζ^2 is the speech dispersion.

In this case, $\omega_0 = 2.98 \cdot 10^3$ s^{-1} characterizes the frequency near which the maximum of the spectral density of the sound pressure of the speaker's speech is located, and the first term, taking into account the coefficient $\rho = 1.14 \cdot 10^3$ s^{-1}, is the degree of the maximum manifestation.

The second term characterizes the decrease of the spectral density of the sound pressure of speech with increasing frequency. The choice of specific approximations is determined by the nature of the dependencies, which are close in appearance to the class of widely used functions. On the other hand, the approximation coefficients should have their own physical interpretation.

Comparison of this dependences with the speech spectrum presented in Fig. 1 showed that they cannot be used for the Kazakh language. In this regard, the amplitude spectra of the Kazakh language speech were averaged for 14 speakers. The averaged dependence is shown in Fig. 2.

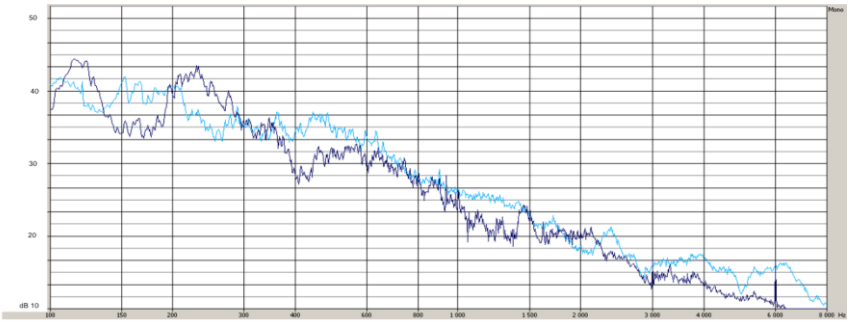

Fig. 1. Speech spectrum of Kazakh speakers when reading a text in Kazakh

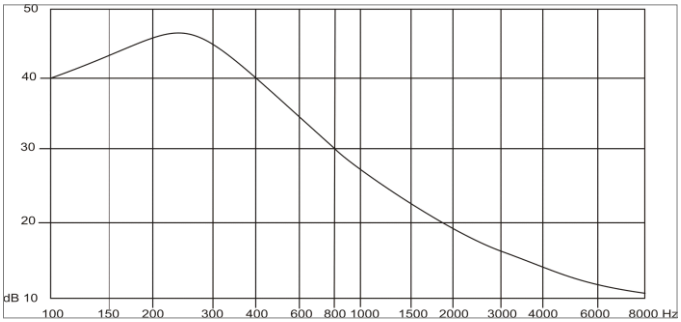

Fig. 2. Averaged amplitude spectrum of the Kazakh language speakers' speech

According to the results of the experimental studies performed for a sample of 14 people, the amplitude spectrum of the Kazakh speech, presented in Figure 2, could be approximated for the frequency range from 100 to 8000 Hz and for the integral sound pressure level from $6.3 \cdot 10^{-3}$ Pa to 0, 36 Pa (from 50 to 85 dB) in this frequency range. The approximation is realized by the expression

$$S(f) = \rho \cdot P \cdot K \left[\frac{1}{\rho + |(f_0 - f)|} + \frac{1}{\rho + (f_0 + f)} \right], \quad (5)$$

where ρ = 100 Hz, P is the sound pressure level of speech in the frequency band from 100 to 8000 Hz expressed in Pa, f_0 = 225 Hz, K is proportionality coefficient.

In this case, f_0 = 225 Hz characterizes the frequency near which the maximum of the spectral density of the sound pressure of the speaker's speech. Coefficient ρ is close to the value of the frequency of the fundamental tone. The proportionality coefficient has a dimension of $s^{-1/2}$ and is equaled to 0,0585 $s^{-1/2}$. The first term, taking into account the coefficient ρ = 80 Hz, characterizes the severity of the maximum in the speech spectrum. The second term characterizes the decrease in the spectral density of the sound pressure of speech with increasing frequency.

The spectral density of the speech signal at the places of speech intelligibility assessing outside the room is determined from the expression

$$S_r(f) = \rho \cdot P \cdot K \left[\frac{1}{\rho + |(f_0 - f)|} + \frac{1}{\rho + (f_0 + f)} \right] \cdot K_r(f), \quad (6)$$

where $K_r(f)$ is speech transmission coefficient as a function of frequency (room soundproofing).

The spectral density of the combined masking signals is determined from the expression

$$S_{ms}(f) = \rho \cdot P_{sl} \cdot K\left[\frac{1}{\rho + \left|\left(f_0 - f\right)\right|} + \frac{1}{\rho + \left(f_0 + f\right)}\right] + S_{wn}, \quad (7)$$

where S_{wn} is spectral density of «white» noise in the speech frequency range (this value is constant); P_{sl} is sound pressure level of speech-like signals in the frequency band from 100 to 8000 Hz. The affiliation indexes of the spectral density of speech and speech-like signals are not used in expressions (6) and (7), as they are the same in nature depending on the frequency, but differ in amplitude due to different levels of the speech signal and the speech-like signal P and P_{sl}. The spectral density of masking noise is constant over time for the white noise component at all control points. The spectral component of the speech-like signals at the control points changes over time and is redistributed in frequency in accordance with expression (5). In this case, the amplitude of the spectral components of speech-like signals significantly exceed the amplitude of the spectral components of «white» noise (by 6–12 dB in the frequency range up to 500 Hz), but they are short-term at a given frequency. It should be noted that if the speech-like signals are formed on the base of the allophone of the speaker whose speech is necessary to protect, then the probability of overlapping of the frequency components of the information signal by the frequency components of the speech-like signal is much higher, as the formants of a certain phoneme of the information signal, for example, phonemes *a*, exactly coincide with the formants masking speech-like phoneme signal *a*, because the first and the second phonemes belong to the same speaker.

Speech intelligibility assessment for information security systems should be performed according to the limit states. It is indicated in [18] that the intelligibility limit is –18.5 dB, and to ensure complete security of speech information, the signal-to-noise ratio should be –27 dB (taking into account the burst nature of speech).

The averaged spectral components of the speech-like masking signal and the protected speech signal have approximately the same value if the sound insulation is uniform in frequency and the generated noise is close to dependence (5).

Formal speech intelligibility is determined from the expression

$$A = \sum_{k=1}^{k} p_k \cdot w\left(S_r / S_{ms}\right), \quad (8)$$

where p_k is the probability of finding formants in the k^{th} frequency band in case of band-frequency analysis; $w(S_r/S_{ms})$ is speech perception coefficient and the ratio of the speech signal and the masking signal in a given frequency band.

Verbal speech intelligibility could be determined by formant speech intelligibility using the expressions presented in [2]. In addition, as Bradley points out, when protecting voice information, it's necessary to take into account intelligibility, recognition (coding) and audibility. Recognition is the case when speech intelligibility is absent but can be determined by the speaker's timbre if the auditor is familiar with the recordings of this speaker. The auditor will hear what the preset speaker is saying, but it is not clear whether it is an information signal or a speech-like masking noise. S.J. Bradley has showed in [4] that the probability of increasing the signal-to-noise ratio depends on the level of background noise at different times of the day. Since speech and noise level vary from moment to moment, therefore, the actual intelligibility of speech will similarly change over time.

Conclusion

The curried out experimental studies of the spectral density of speech in the Kazakh language made it possible to obtain it's approximate dependence on the frequency and take into account the phonetic features of the Kazakh speech when assessing the security of the speech information using the formant method.

References

1. Pokrovsky N.B. [*Calculation and measurement of speech intelligibility*]. M.: Svyazizdat; 1962. (In Russ.)
2. Zheleznyak V.K., Makarov Yu.K., Horev A.A. [Some methodological approaches to assessing the effectiveness of voice information protection]. *Special Technique.* 2000;4:39-45. (InRuss.)
3. French N. Factors Governing the Intelligibility of Speech Sounds. *Acoust. Soc. Am.*;1947;19:90-119.

4. Bradley J.S. Designing and Assessing the Architectural Speech Security of Meeting Rooms and Offices. *IRC Research Report*. 2006. DOI: 10.4224/20377425
5. Didkovsky V.S., Prodeus A.N. [Comparison of formant properties of Ukrainian and Russian Speech Electronics and communications]. *Thematic «Electronics and Nanotechnology»*. 2009:2:88-94. (In Russ.)
6. Bradley S.J. [Speech Levels in Meeting Rooms and the Probability of Speech Privacy Problems]. *Acoust. Soc. Am*. 2010;127 (2):815-822. DOI: 10.1121/1.3277220.
7. Byrne D. [An international comparison of long-term average speech spectra]. *Acoust. Soc. Am.* 1994;96(4):2108-2120. DOI: 10.1121/1.410152.
8. Gavrilenko O.V., Didkovsky S.V., Prodeus A.N. [Calculation and measurement of speech intelligibility at small signal-to-noise ratios]. *Electronics and Communications, Thematic issue «Problems of Electronics»*. 2007:137-141. (In Russ.)
9. Gavrilenko O.V., Didkovsky S.V., Prodeus A.N. [Calculation and measurement of speech intelligibility at small signal-to-noise ratios]. Electronics and Communications, Thematic issue «Problems of Electronics». 2007;142-147. (In Russ.)
10. Davydau H.V. [Method for protecting speech information]. *Doklady BGUIR=Doklady BSUIR*. 2015;94:107-110. (In Russ.)
11. Seitkulov Y.N., Davydov G.V., Potapovich A.V. Substantiation of the Method of Forming Combined Speech Masking Signals. *Bulletin of KazNTU*. 2014;102.
12. Seitkulov Y.N., Davydov G. V., Potapovich A. V. [The base of speech structural units of Kasakh language for the synthesis of speech-like signals]. *Proceeding of the IEEE 12th International Conference on Application of Information and Communication Technologies*. 2018. DOI: 10.1109/ICAICT.2018.8747120.
13. Seitkulov Y.N., Davydov G.V., Potapovich A.V. [Algoritym of forming speech base units using the method of dynamic programming]. *Theoretical and Applied Information Technology*. 2018;96:7928-7941. DOI: 10.1109/ICAICT.2018.8747120.
14. Davydov G.V. [Synthesis of speech-like signals in the Belarusian language]. *Doklady BGUIR=Doklady BSUIR*. 2015;90:27-32. (In Russ.)
15. Zaitsev A.P. [*Technical means and methods of information protection*]. M.*:* Publishing House Engineering, LLC; 2009. (In Russ.)
16. Khorev A.A. [Technical information protection. In 3 parts. P. 1. Technical channels of information leakage]. M.: SPC «Analytics»; 2008. (In Russ.)
17. Velichkin A.I. [Amplitude limitation of speech]. *Acoustic Journal*. 1962;8:168-174. (In Russ.)
18. Bradley S.J. [Developing a new measure for assessing architectural speech security]. *Canadian Acoustics*. 2003;31:50-51.

Authors contribution

Seitkulov Y.N. realized the selection and recording of speakers in the Kazakh language.

Boranbayev S.N. realized the selection of auditors in the Kazakh language, assessed speech intelligibility

Davydau H.V. set tasks that need to be solved in the course of the study, interpreted the results.

Patapovich A.V. realized the results processing and calculation.

Information about the authors

Seitkulov Y.N., PhD, Director of the Institute of information security and cryptology of L.N. Gumilyov Eurasian National University.

Boranbayev S.N., PhD, Professor of L.N. Gumilyov Eurasian National University

Davydau H.V., PhD, Researcher of SRL 5.3 of R&D Department of Belarusian State University of Informatics and Radioelectronics.

Patapovich A.V., Researcher of SRL 5.3 of R&D Department of Belarusian State University of Informatics and Radioelectronics.

Address for correspondence

20013, Republic of Belarus,
Minsk, P. Brovka str., 6,
Belarusian State University
of Informatics and Radioelectronics
tel. +375-29-670-30-40;
e-mail: nil53@bsuir.edu.by
Patapovich Aleksandr Vladimirovich