УДК 004.04:004.942 https://doi.org/10.15407/pp2021.02.076

V. L. Shevchenko, Y. S. Lazorenko, O. M. Borovska

INTONATION EXPRESSIVENESS OF THE TEXT AT PROGRAM SOUNDING

As the amount of media content increases, there is a need for its automated sounding with the built-in means. The factors influencing the intonation were analyzed, the dependences of sound characteristics in accordance with the intonations were mathematically described. In the course of the work, the numerical analysis of sentences was improved using the moving average for smoothing audio, approximation lines for generalization of emotions as functions, and Fourier transform for volume control. The obtained dependences allow to synthesize intonations according to the punctuation, emotionally colored vocabulary and psycho-emotional mood of the speaker. Software for emotional sounding of texts was developed, which provides the perception of audio information easier and more comfortable based on the use of built-in processors of mobile devices.

Key words: text analysis, sound characteristics, intonation expressiveness.

Introduction

ost of the information comes to us in a graphical and audio representation. Therefore, the automation of sounding texts is an urgent problem. At the same time it is necessary to bring the intonation of the voice synthesized by the computer as close as possible to the human one. Usually monotonously read text is processed manually by a person.

The development of automated means of emotional sounding of the text is somewhat constrained due to the complexity of the task and, consequently, uncertainty about the success of its solution based on available mobile devices with built-in processors. At the same time, to solve similar problems in related fields, such as recognizing dangerous situations based on smartphone sensors, examples of successful solutions using machine learning exist [1].

In our case, to introduce emotion (intonation), in addition to machine learning methods, it would be desirable to identify and use mathematical patterns of texts. Therefore, the topic of the study of the formation of intonation for different emotions is relevant.

Analysis of the State of Research Issues

Theoretical research of intonation from the philological point of view was actively investigated by Bagmut A.Y. [2], [3]. Her works provide a very detailed analysis of intonations according to the syntactic and semantic features of sentences – some monographs contain an analysis of only a narrative sentence with equal intonation. But at the same time, the intonations given to speech depending on the psycho-emotional state of the speaker, i.e. the emotions themselves from the read text, have been little studied.

Minnigalimov R.T. [4] mathematically described the patterns of changes in the frequencies of the fundamental tone for narrative affirmative and negative, as well as interrogative sentences. However, his practical experiments sometimes contradict each other. The author explains this by the fact that different speakers have different reading styles, so it is necessary to increase the statistical base of speakers for further practical development.

American experts from AT&T Laboratories worked on the synthesis of voices for sound. They created a program that imitates the human voice after processing 10-40 hours of real recording. However, as noted by the developers themselves, the program is not yet able to fully reproduce the voices of real people, and the sound of synthesized recordings is quite technical and does not take into account the emotions of the speaker when sounding [5].

Last year, the British team Sonantic tried to add emotion to the computer voice. The project uses artificial intelligence, which analyzes large amounts of human records [6]. The synthesized voice is indeed similar to a natural sound, but so far the development is focused only on the negative emotion of despair and crying.

The main problem is that in practice everyone reads the same text in their own way, keeping only the basic intonation or mood. The aim of the work is to identify techniques and characteristics of voice change to increase the expressiveness of speech and formalize these features.

One of the available approaches to sounding texts is monotonous reading with constant pitch, volume, etc. And without pauses. This method of transmitting audio information requires constant focus and independent logical division of the listened text into syntactically integral fragments in content.

Another improved approach is to read in an even voice, but with punctuation pauses. So, after a comma there is a minimum pause, and, for example, after a dash or a point – more. This method facilitates the perception of the text due to the fact that syntactically whole units of speech (whole sentences or their parts) are perceived separately due to pauses [7]. The disadvantage is the lack of intonation difference between the fragments of the text. This approach is used in the built-in libraries of the Python programming language, in applications for viewing text files with the sound function, browsers, etc.

One of the best existing approaches is machine-based sounding. This method generates a number of sound effects for a certain set of emotions. However, the "assignment" of such an emotion to a particular sentence is done manually, i.e. the linguistic features of the text itself are not taken into account. This solution is used, for example, in the British startup Sonantic [6].

The contradiction between these approaches and practical needs is that the principles of intonation in the sound of texts are formulated rather vaguely and are based on human "sense of language", which is often explained by the skills of expressive reading in philological sources. However, if a person is able to unambiguously determine the mood of a sentence, then, probably, there are certain patterns in how it does it and that includes such a "sense of language." From this we can conclude that the found dependences can be generalized and formalized, to bring them closer to the concept of rule. This can be a good simplification and basis for the algorithms used in machine learning.

Relationship between Intonation and Punctuation

The main lexical means of denoting emotions in the text is punctuation [8]. It gives instructions on how the sentence should begin and end, what interaction with the listener is envisaged, what feelings should be evoked in him, how long and frequent pauses should be endured. Emotions formed due to intonation do not depend on the speaker, his manner of narration and feelings, style of text and content, target audience, etc. That is, they will be dim, but always the same. If a comma and a dash indicate only the need for a pause and its length, then the key role in determining the intonation is played by punctuation at the end of the sentence.

According to the emotional color the sentence could be exclamatory and non-exclamatory. So, they have or do not have an exclamation mark at the end. They differ in how important the emphasis on their content is and determine how strongly the information will impress the listener [8].

Invocative sentences are pronounced in a calm voice, without extreme increase or decrease in pitch and volume, not oversaturated with accents on words and their meaning.

Exclamations are pronounced more sublimely, loudly, in a higher tone, the intonation may be less smooth, with tears and bright accents on keywords. Schematically, the differences are presented in fig. 1 – for volume and fig. 2 – for height.

In order to express a sentence, there are narrative, interrogative and motivating ones. Narrators report an event, fact, or phenomenon. At the end of such sentences a full stop is placed, sometimes – three full stops to indicate incompleteness of thought. If there is a full stop at the end, the intonation throughout the sentence remains equal, and at the end it drops, the voice subsides. The volume is kept evenly at the same level, decreasing fairly quickly at the end of the sentence. This expresses the completeness of thought and confidence in what is being said.

If the sentence ends with period, the decline of intonation is smoother, moderately fading, the voice subsides more, but gradually. The opinion is not complete, there is room for the listeners' own thoughts and their personal assessment of what has been said. Often it is to enhance this effect at the end of a sentence with ellipsis is a longer pause [8]. The exclamation mark in such a sentence often expresses anxiety associated with feelings of fear.

Since in most cases with increasing volume, the pitch of the sound also increases, because the accent and expression are provided by both means, in the future they are accepted as one set of sound characteristics. However, in this case the approximation rejects the jump of volume important for the exclamatory sentence.

Motivational sentences also have a full stop at the end. They differ from narrative content: express a request or demand, and - from purpose: motivate to action. Intonation have a strong emphasis on keywords [7], which, in fact, determine the motivation for action (ask, tell, bring, etc.). Usually such words appear at the beginning of a sentence, inversion is rare and is rather an atypical phenomenon for motivational sentences. Therefore, both the increase in voice and the increase in volume are characteristic of the beginning of a sentence.

Interrogative sentences have a pronounced logical emphasis on the most significant word and the strengthening of intonation at the end, which is illustrated in fig.3. If such a sentence contains interrogative words (how ?, where? etc.), then they are logically emphasized. Then at the end of the sentence the intonation decreases, as in narrative sentences. The degree of amplification of sound characteristics determines how important it is to focus on this issue [9].

Construction of Intonation according to Punctuation

For further work with intonation and program processing of sentences, we will enter some mathematical designations.

The word consists of syllables:

$w = \{n, \dots, n, k, n, \dots, n\}$, where

n – unstressed syllable,

k – stressed syllable.

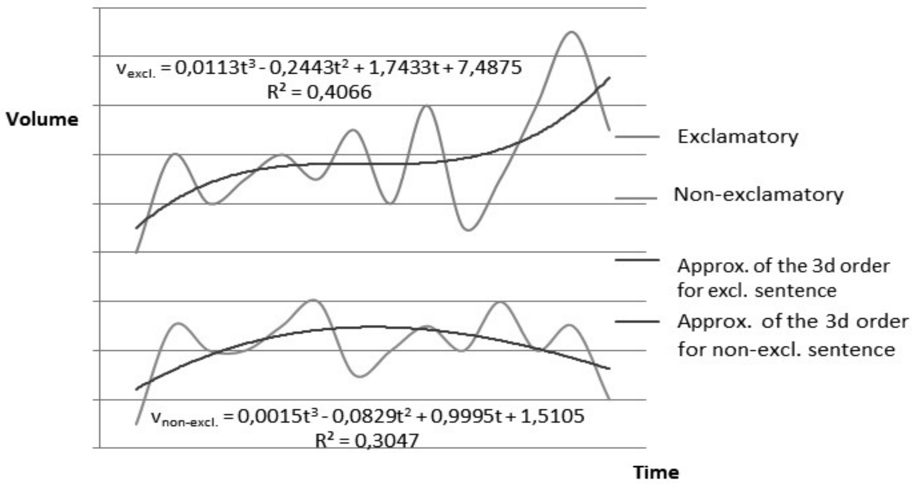

Fig. 1. Changing the volume of exclamatory and non-exclamatory sentences

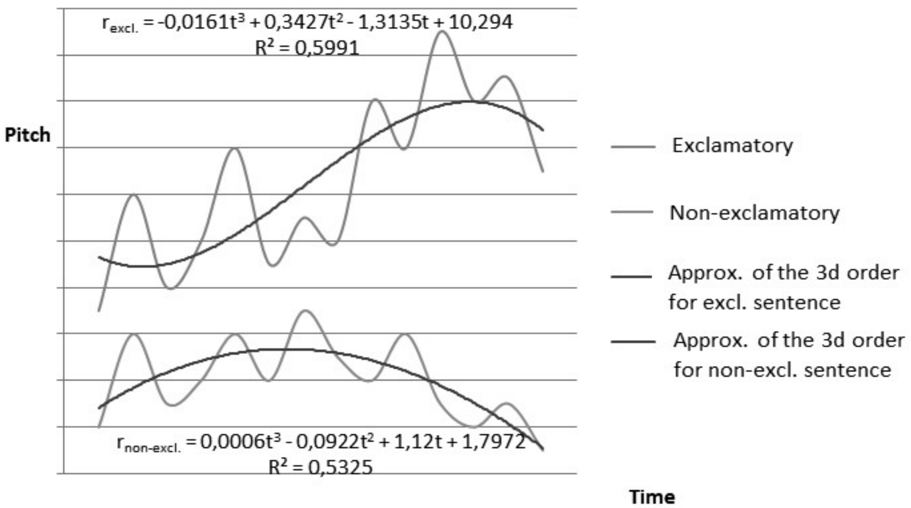

Fig. 2. Changing the pitch of exclamatory and non-exclamatory sentences

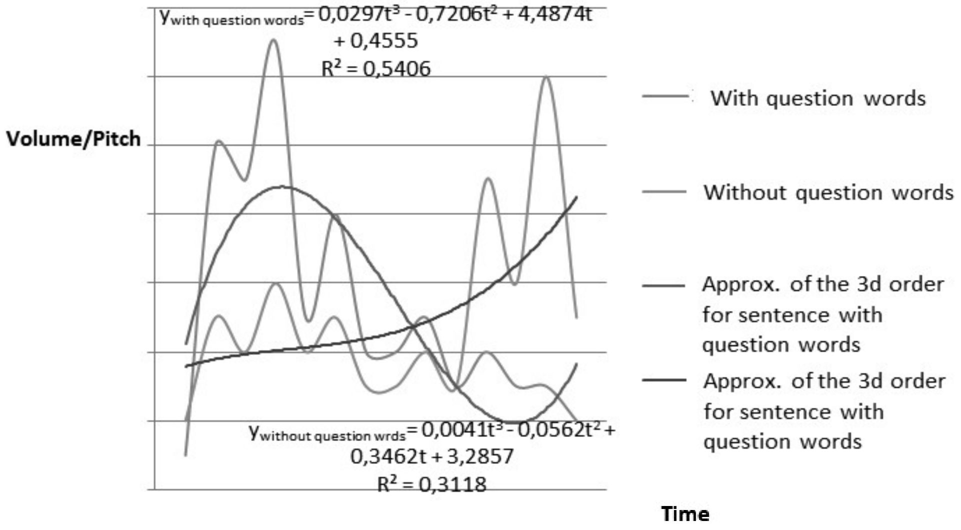

Fig. 3. Changing the sound characteristics of interrogative sentences

The sentence consists of the words:

$s = \{q, \dots, q, l, q, \dots, q\}$, where

q – logically unstressed word,

l – logically emphasized, the most significant word.

But a sentence can also be represented as a sequence of syllables, then you can write it like this:

$s = \{n, \dots, n, k, n, \dots, n\}$, where

n – syllable of a logically unstressed word or any unstressed,

k – stressed syllable of the most significant word.

Along with the tuple of syllables, we will write a tuple that contains commands for the computer. They contain information about the change of the main technical parameters of the sound. Enter the appropriate symbols for them:

- volume – v;
- length – t;
- frequency (tone of sound) – r;
- pause (before/after the syllable) – p.

Then the phrase (sentence or word) for the computer will look, for example, like this: $s = \{t, v, r, q, \dots, v, r, q, \dots, v, r, p, q\}$ – the duration of the whole record is set, and then before each syllable indicates the required volume and frequency, if necessary, a pause, followed by the text of the syllable.

Consider the basic emotions that cover most human sensations. We introduce their corresponding notations, some of them are opposite to others:

joy – j;

sadness – j';

aggression (attacker's reaction) – a;

confusion / anxiety (protective reaction of the victim) – a';

calm – c;

irritation / dissatisfaction – c'.

Each emotion has its own characteristic pattern in time, which is represented by a set of functions of the main technical parameters of sound, using the previous notation:

$v = j(t); r = j(t)$ – for joyful intonation;

$v = j'(t); r = j'(t)$ – for sad;

$v = a(t); r = a(t)$ – for angry, aggressive;

$v = a'(t); r = a'(t)$ – for anxious;

$v = c(t); r = c(t)$ – for calm;

$v = c'(t); r = c'(t)$ – for irritated.

To work with audio recording, we turn the emotions of sentences into functions for processing arrays, because the sound signal during processing is usually given as a set of values like an array.

For program work with the text we will read it from a file. In the beginning it is enough to use ready means of sounding and to receive monotonous reading of the text. It is inconvenient to work with an integral sound file – change of any characteristic will be superimposed on all sound series. Therefore, it is necessary to divide the record into a number of identical fragments and give them a numerical representation. Therefore, for further sound processing, the conversion of the sound series into an array of volume values with a certain frequency is used. The wave format is used as standard for such purposes.

During sounding it is necessary to work simultaneously with the text (to process sound according to punctuation), and directly with its audio representation. The text from the file is generally a string. So, first we divide the text into an array of sentences, the following punctuation marks will serve as delimiters: «.», «!», «?», «…». Then in a cycle we process each of them. To determine the final punctuation mark, we analyze the sentence character by character from the end.

The main intonation emphasis is not given to the whole combination of words, but focuses on one syllable of the keyword in the phrase. The increase in intonation may be more or less smooth, but not sudden. Therefore, the syllable (or several syllables) before the most intonationally outlined and after it will also be somewhat pronounced. This will help smooth out the difference in volume and pitch and make the sound more natural and pleasant.

Multiplication of values by a certain factor should not be used to amplify the sound characteristics of a piece of audio recording. This approach can give an unexpectedly strong or too small result. In addition, if the recording has a large amplitude of volumes or pitches, the highest of them can be extremely amplified, which will lead to unpleasant intermittent sound and poor sound quality due to the appearance of noticeable noise. To adjust the sound characteristics, it is better to add a certain number to their values. You need to consider how quiet the original recording was and what the initial pitch was, so that the adjustment is not too sharp when adding a large number or, conversely, the number is not too small and the changes are not noticeable. If the original recording already had inhomogeneous volume and pitch, it should be generalized.

The arithmetic mean of the values of one of the characteristics does not always give the desired result: for example, if there are values that are several times greater than the bulk [10]. Therefore, you must first take the middle range, discarding too large and small values. This can be done by averaging or other smoothing methods. You can also approximate an array of values. For such purposes, linear is enough, because the nature of the function itself does not interest us. The goal is to select the range in which most values are concentrated. Schematically, the principle of this method is illustrated in fig. 4. The middle (vertical) range used for further calculations, the upper and lower ranges contain discarded values.

Another way to discard redundant values, more convenient for software implementation – setting the lower and upper limits of acceptable values. After that, you can take the average value or mode as the basic initial value of a characteristic. The result of setting such thresholds is constructed by software tools and the principle is shown in fig. 5, where the value of the initial recording frequencies is indicated in the upper and lower ranges, and the selected range is indicated in the middle (vertical) range.

For software work with text and sound, a function was created that first sounds a given text from a file, and then converts it into an array of values and amplifies its corresponding fragments according to the desired condition. This is the main function, which is then referenced by oth-

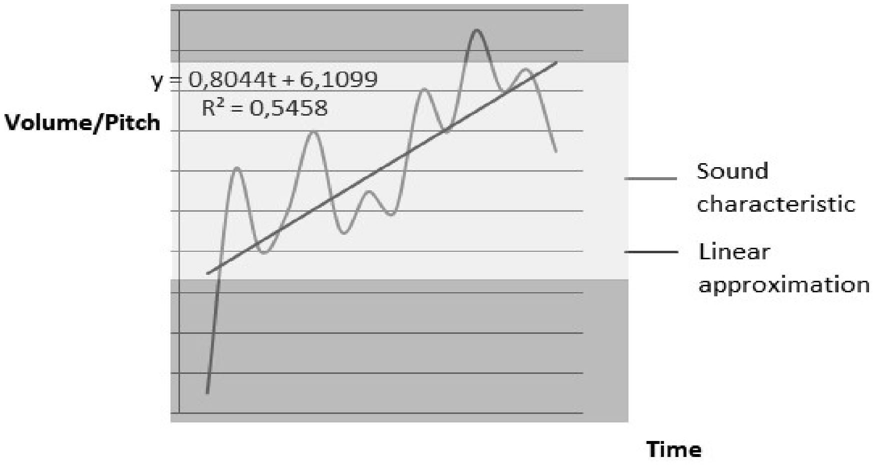

Fig. 4. Selection of the average range of values by means of linear approximation

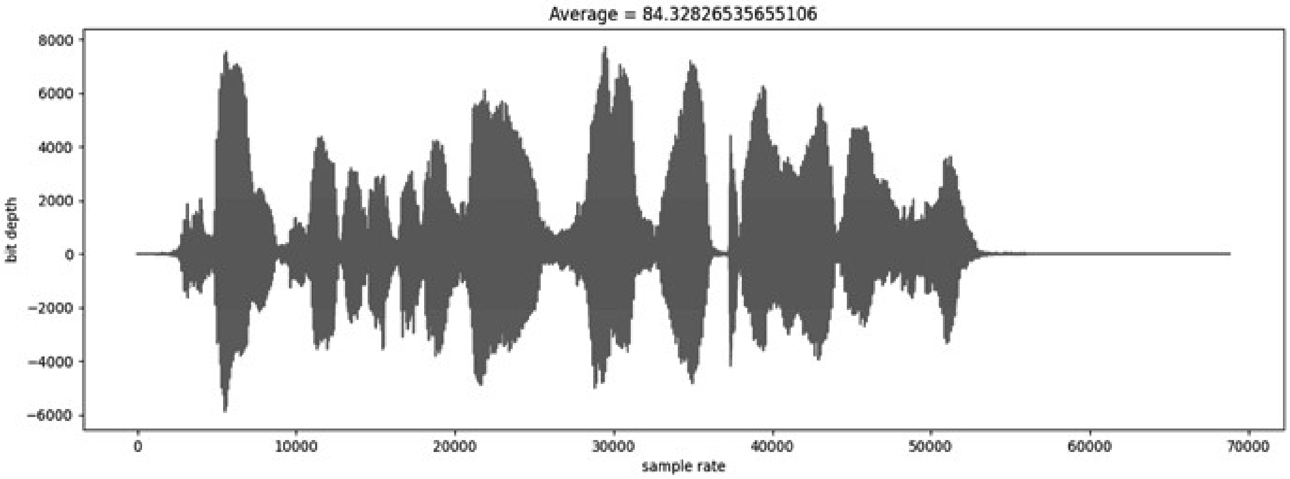

Fig. 5. Select a range of values for averaging

ers. It also controls the imposition of several gain conditions on the same fragment. For example, if the interrogative word «when» is defined as an adverb, it should be strengthened twice – as a question word and as an emotionally colored word (because it is adverbs and adjectives that most often indicate the shade in the meaning) [10]. This can have the undesirable effect of a sharp drop in values and an increase in volume or tone too much. Therefore it is necessary to check up, whether the word was strengthened already on other sign, and in that case not to strengthen it repeatedly (or to strengthen much less).

In addition to placing emphasis on words that belong to certain parts of speech, in natural language a person intonationally expresses a part of a phrase that contains objections. The usual marker of such words in the text is the negative part «no». Thus, the words after it are also emotionally amplified in the software implementation. To find such words in a sentence, a function is created that determines the ordinal numbers of words with a negative shade of meaning.

To speed up writing, we use a function that discards single values from an array with a certain step. At not very small step distortions of a sound are almost not appreciable and are corrected in the subsequent smoothing. A similar function to slow down the recording, on the contrary, adds elements. That is, it also duplicates single values with a certain step.

Fourier transform was used to shift the tone of the sound in the work. This allows you to make the voice both lower and higher while almost completely preserving the original recording time. The degree of deterioration of sound quality is proportional to the length of the audio fragments with which it is processed – the smaller the pieces of the recording undergo changes, the better the sound.

Construction of Intonation that Expresses the Feeling of the Speaker

In general, all emotions can be divided into 3 groups: positive, negative and neutral.

The first group expresses high spirits and satisfaction. Such emotions are high in tone of voice, with normal or slightly increased volume, slight rhythm, but with smooth differences, ascending intonation of phrases. At the end of the sentence, the voice subsides smoothly, but not stretched. Polynomial approximation is used in the work to determine the general nature of the decrease or increase of graphs of sound parameters and their comparison with other emotions. It most accurately reflects intonation changes. In this case, the approximation of the 2nd order is not enough - the differences in volume or height are not taken into account, although they are important for building emotions. The approximation of the 4th and higher orders approaches the initial graph too accurately, reflecting even minor fluctuations in the voice. The best option is the 3rd order: sufficient smoothing of graphs is provided, the largest differences of values remain. Graphic generalization of positive emotions is shown in fig.6.

So you can mathematically generalize the functions of positive emotions. Since the best result of the approximation is achieved in the 3rd order, the function is cubic. The coefficients of the approximation lines can be rounded, because the goal is an approximate form of the function. Then for volume and height of positive intonations (we will take a joyful voice as a basis) accordingly it turns out:

$$v = 0.01t^3 - 0.17t^2 + 1.34t + 1.63,$$
$$r = 0.01t^3 - 0.16t^2 + 1.11t + 0.95,$$

where t is the time.

Similar generalized functions are built for all the emotions considered below.

The second group, on the contrary, means unpleasant feelings, dissatisfaction with something and denial. If positive emotions, in general, are very similar in nature and sound, then negative ones give a much wider palette of such patterns.

To further improve our approach, we will analyze the features and differences of emotions in the text from a philological point of view in more detail.

The main difference in sound from the positive is the lack of smoothness, gradation of the signal. Also, most (not all) negative emotions are characterized by a decrease in pitch. But the volume can be both reduced and increased, depending on the severity of the emotion and the purpose of its expression. Let's analyze these shades in more detail.

As a rule, when a person feels irritated or dissatisfied, his voice becomes lower. But the feeling of fear and anxiety is accompanied by the opposite phenomenon: the voice becomes louder, very inhomogeneous, smooth and longer. Often vowel sounds are lengthened and amplified, while consonants are lost and replaced by short pauses in live speech caused by minor sudden breaths.

An increase in volume indicates an «attacking» mood. Such emotions arise under critical psychological stress, develop rapidly and grow intonationally. This is evidenced by the sharp and confident ending of sentences expressing similar emotions in speech. Acceleration or a gradual increase in the speed of sound is sometimes used for additional expression.

The opposite tool (decrease in volume and stretching of phrases) is a protective reaction, excitement, confusion and helplessness. Such experiences depress a person, worsen mood, well-being, reduce productivity, prudence. The range of such soft emotions is extremely wide: sadness, grief, confusion, despair, guilt and many others. They actually differ in the root cause, i.e. in the text - in content. Intonations are almost identical, therefore, the program requires a single implementation.

Neutral emotions characterize a calm, balanced state of the speaker. In such emotions there are no jumps of sound characteristics, the voice is smooth. Usually correspond to narrative unpronounceable sentences. An example of such emotions is interest or indifference.

Programmatically, the main difference from the construction of intonation on the basis of punctuation of the sentence is that changes should be applied to the whole sentence, and not only to its individual parts (words, syllables). This will ensure a smooth transition of emotion, sound quality and proximity to natural human language.

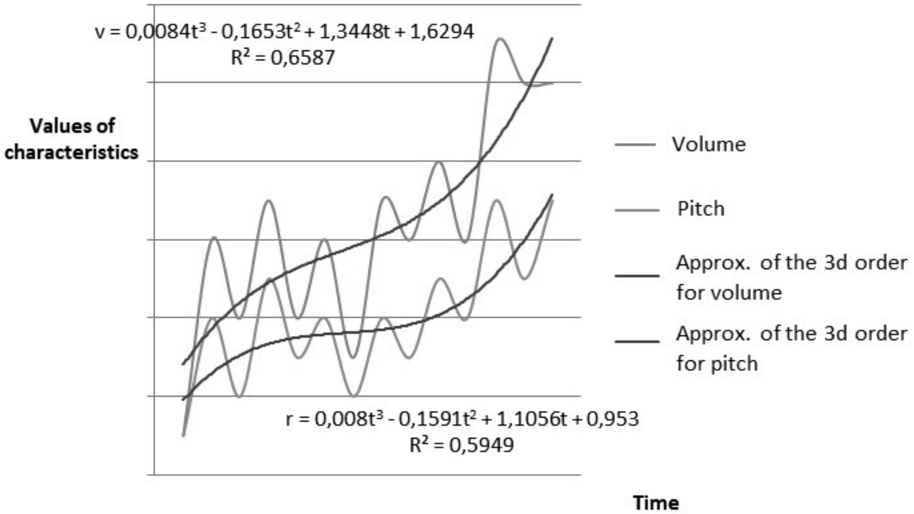

Fig. 6. Changing sound characteristics over time for positive emotions

Since it is not necessary to analyze the text itself to give the phrase a certain emotion based on the speaker's feelings, it was possible to process the recording without the initial textual representation. The main goal is to divide the sound series into fairly small phonetic units. Due to the fact that accents (both strong and weak) in pronunciation fall on vowel sounds, it is advisable to divide the recording into syllables, because one syllable contains exactly one vowel sound.

In the model, the soundtrack was divided into different fragments over time, for example, 0.5s. But there is a problem: the basic phonetic unit, which may contain an accent, is the syllable, but the syllables differ in duration of sound. Accordingly, after processing the recording, the accents may be misaligned, which will affect the sound quality.

Therefore, another method was used in the work. If we compare the graphical representation of the volume array of a record and the text version of the phrase, we see that the volume fluctuations occur in each syllable. This is illustrated in fig.7.

Thus, the volume increases at the vowel sound, and at the consonant level it decreases and almost completely subsides at the hissing and whistling. From this follows the conclusion that it is possible to break the audio recording into fragments-compositions, i.e. particles of amplification-attenuation, if you set the threshold of "silence". We will assume that the values that are higher than this threshold correspond to vowel sounds - the key component of the syllable (sound), and those that are lower - consonant (silence). Then we will process the sound separately in parts (audio fragments of phrases of text or syllables), and then combine them back into one file.

It is these fragments of silence that serve as dividers when splitting a record into syllables. By default, it is assumed that the reading occurs at a speed of 100% (the value can be both lower and higher) and a volume of 1 (values from 0 to 1). The optimal "silence" interval for determining the composition limit is 50 ms, and the silence threshold is approximately minus 30 dB (dBFS). Then to adjust the silence time, reduce it by the same percentage as increase the speed by one percent (for 150% of the speed, the silence time will be approximately 40ms). To adjust the silence threshold, reduce the volume by the same percentage and reduce the threshold by the same percentage, i.e. increase the modulus of the threshold value (module, because this value remains negative). Thus, for a volume of 0.8 we obtain a threshold of silence minus 36 (approximately minus 40).

Conclusions

1. The speech techniques and voice characteristics that give emotional color to the read text were studied in the work; the regularities of change of sound characteristics for transfer of various intonations are formalized and programmatically realized.

2. The paper proposes a method of formalizing emotions when breaking sentences into syllables and mathematical formalization of patterns of change of sound characteristics of the synthesized voice in accordance with the intonation of sentences; proposed formulaic dependences for different types of sentences

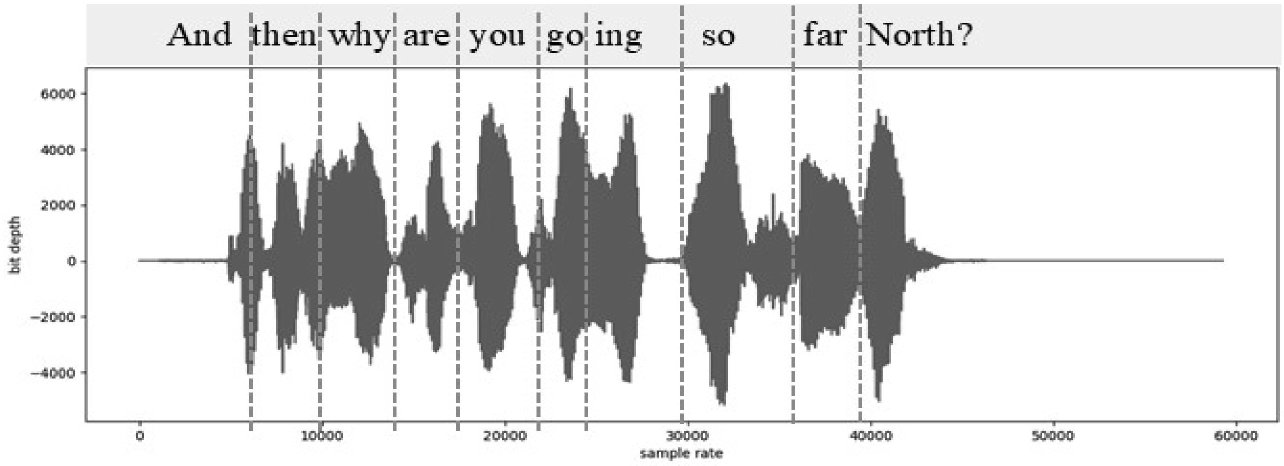

Fig. 7. Graphic correspondence of text and voice

by syntactic and semantic features; improvement of the method of numerical analysis of sentences using the method of moving average, approximation lines and Fourier transform; improving the method of sentence synthesis taking into account the given emotions with the help of lexical and syntactic analysis of sentences.

3. Special software has been developed in the Python algorithmic language, which allows you to voice text with appropriate intonations based on the use of built-in mobile processors.

4. In further research it is planned to investigate how the use of pauses - both short in words and larger in a phrase - affects the change of intonation; keep in mind that the final punctuation marks can be several: «?!», «!!!», «? ..», etc., so that the intonation may have different shades.

References

1. Shevchenko V. Dynamic Objects Emergency State Monitoring by Means of Smartphone Dynamic Data / Shevchenko A., Bychkov O., Shevchenko V. // 2017 14-th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM). Proceeding. - Polyana, February 21-25, 2017. - p.292-294. http://ieeexplore.ieee.org/document/7937138/ DOI: 10.1109/CADSM.2017.7916138
2. Bahmut A.Y. Semantics and intonation in the Ukrainian language. – 1991
3. Bahmut A.Y. Intonation structure of a simple narrative sentence in Slavic languages. – 1970
4. Minnihalimov R.T. Analysis and synthesis of Ukrainian speech / Minnihalimov R.T., Kyiv, 2015. - 90 p.
5. Official site AT&T Laboratories: https://about.att.com/sites/labs_research
6. Official site Sonantic: https://www.sonantic.io/
7. Blyznychenko L. A. Intonation of speech. – 1968
8. Peshkovskiy A. M. Punctuation marks and scientific grammar. – 1918
9. Peshkovskiy A. M. Intonation and grammar. – 1928
10. Zagumennov A. P. Computer sound processing [Electronic resource] / Zagumennov A. P. - Moscow: DMK Press, 2006.-- 384 p .: ill. - ISBN. - Text: electronic. - URL: https://znanium.com/catalog/product/407267

Received: 18.05.2021

About the authors:

Viktor L. Shevchenko,
Dr.Sc., Prof., Professor of Software systems and technologies Department of Taras Shevchenko National University of Kyiv.
Publications - more than 300.
Publications in foreign scientometric publications – 17.
H=3.
https://orcid.org/0000-0002-9457-7454.

Yana S. Lazorenko,
Bachelor student
Publications – 2.
https://orcid.org/0000-0002-3987-2338.

Olena M. Borovska , Chief designer
at the Institute of
Software Systems of the National Academy of Sciences of Ukraine.
Kiev, 03187, Acad. Hlushkov avenue, 40 building 5
Publications - 4

Affiliations:

Program system and technologies
Department
Taras Shevchenko National University
of Kyiv,
Bohdan Hawrylyshyn str. 24
UA-04116, Kyiv, Ukraine
E-mail: gii2014@ukr.net,
yana_lazorenko@knu.ua

Department of Automated Organizational Management Systems (№23)
Institute of Software Systems of the National Academy of Sciences of Ukraine.
Academician Hlushkov Avenue, 40, building 5, Kyiv, Ukraine, 03187.
E-mail: e.borovskaya@nas.gov.ua