UDC (004.42: 004.93)
https://doi.org/10.31891/csit-2022-4-5

YANA BIELOZOROVA, KATERYNA YATSKO
National Aviation University, Kyiv, Ukraine

FEATURES OF THE IMPLEMENTATION OF THE SPEAKER IDENTIFICATION SOFTWARE SYSTEM

The proposed architecture of the identification software system in the form of class and sequence diagrams. The main criteria for assessing the accuracy of speaker identification were studied and possible sources of loss of speaker identification accuracy were identified, which can be used when building a speaker identification system. A software system based on the proposed architecture and previously developed identification algorithms and methods was created.

The following conclusions can be drawn on the basis of the performed research: approaches to the construction of existing announcer identification systems are considered; the main criteria for assessing the accuracy of announcer identification were investigated and the main sources of loss of accuracy during announcer identification were identified; the structural construction of the announcer identification system is considered, taking into account the identified sources of loss of accuracy during announcer identification; the proposed architecture of the speaker identification system in the UML language in the form of class and sequence diagrams; a software system was built that implements the functions of speech signal identification according to the methods and algorithm proposed in previous works.

The software system uses a ranking method based on three different criteria. These include: calculation of the proximity of two-dimensional probability density function curves for the frequency of the main tone and the location in the spectrum of three frequency ranges that are extracted from the speech recorded in the speech signal; calculation of the proximity of the probability density function curves for each of these features separately; calculation of the degree of closeness of the absolute maxima of the formant spectra extracted from the speech recorded in the speech signal.

Keywords: speaker identification software system, wavelet, diagrams, UML, speech recognition.

ЯНА БЄЛОЗЬОРОВА, КАТЕРИНА ЯЦКО
Національний авіаційний університет

ОСОБЛИВОСТІ РЕАЛІЗАЦІЇ ПРОГРАМНОЇ СИСТЕМИ ІДЕНТИФІКАЦІЇ МОВЦЯ

Обробка мовного сигналу з метою ідентифікації мовця є найбільш актуальною і популярною в задачах, пов'язаних з мовною обробкою. Постійний і високий попит на програмні реалізації систем ідентифікації дикторів існує в різних сферах: від контролю доступу користувачів до голосових послуг виявлення злочинців. Проте, враховуючи відсутність чіткого наукового обґрунтування алгоритмів ідентифікації, значну складність їх реалізації, а також точність ідентифікації особистості, можна відзначити, що ці завдання в цілому ще далекі від свого остаточного вирішення.

Запропонована архітектура програмної системи ідентифікації у вигляді діаграм класів і послідовностей. Досліджено основні критерії оцінки точності ідентифікації мовця та виявлено можливі джерела втрати точності ідентифікації мовця, які можуть бути використані при побудові системи ідентифікації. Створена програмна система на основі запропонованої архітектури та раніше розроблених алгоритмів і методів ідентифікації.

На основі проведених досліджень можна зробити наступні висновки: розглянуто підходи до побудови існуючих систем ідентифікації диктора; досліджено основні критерії оцінки точності ідентифікації диктора та визначено основні джерела втрати точності при ідентифікації диктора; розглянуто структурну побудову системи ідентифікації диктора з урахуванням виявлених джерел втрати точності при ідентифікації диктора; запропонована архітектура системи ідентифікації мовця мовою UML у вигляді діаграм класів і послідовностей; побудовано програмний комплекс, що реалізує функції ідентифікації мовного сигналу за методами та алгоритмом, запропонованими в попередніх роботах.

Програмна система використовує метод ранжування на основі трьох різних критеріїв. До них відносяться: розрахунок близькості двовимірних кривих функції щільності ймовірності для частоти основного тону і розташування в спектрі трьох частотних діапазонів, які виділяються з мови, записаної в мовний сигнал; розрахунок близькості кривих функції щільності ймовірності для кожної з цих ознак окремо; розрахунок ступеня близькості абсолютних максимумів формантних спектрів, виділених з мови, записаної в мовному сигналі.

Ключові слова: програмна система ідентифікації мовця, вейвлет, діаграми, UML, розпізнавання мови.

Introduction

Processing of the speech signal for the purpose of speaker identification is the most relevant and popular in tasks related to language processing. Constant and high demand for software implementations of announcer identification systems exists in various areas from user access control to criminal detection services by voice. However, taking into account the lack of a clear scientific basis for identification algorithms, the significant complexity of their implementation, as well as the accuracy of personal identification, it can be noted that these tasks in general are still far from their final solution.

Related works

The task of automatic speaker verification is considered to be the creation of a mathematical model, a set of algorithms and, as a result of their application, a software or software-hardware implementation that would allow identification of a person with the same accuracy and reliability as is available to a person.

Research efforts in the field of speech technology have led to the appearance of a large number of commercial speech recognition systems. Such companies as Nuance, IBM, ScanSoft offer a large set of software solutions for both server and desktop applications.

To analyze the work of software systems for speaker identification, it is necessary to consider the main approaches to performing evaluations of the work of such systems. The US National Institute of Standards and Technology (NIST) coordinates evaluations of various speech signal analysis systems: automatic speech recognition systems, key word extraction from speech, and speaker recognition. A description of some annual system evaluations can be found in [1]. The Institute develops research methodologies for comparing different systems, which include a clear statement of the task, the definition of evaluation metrics, carefully selected and uniform sets of training and test data for all participants, clear requirements for conducting and providing test results.

For the problems of speech signal identification, there is always a separate decision-making issue [1]. This question should establish the degree of relationship between the declared model and the characteristics of the speech signal being tested. Identification system based on the provided speech signal with language parameters Z person A, must accept one of the following variants of accepted hypotheses:

$H0$: Z defined as A (is taken as a null hypothesis)

$H1$: Z not defined as A

The conclusion regarding the choice of a certain hypothesis is based on the criterion of plausibility based on the assessment of the probability of obtaining differences between the samples:

$$\Delta(Z) = \frac{p(Z|H0)}{p(Z|H1)} \tag{1}$$

where $p(Z|H0)$ and $p(Z|H1)$ - probability density functions (also called likelihood) associated with person A ("own") and "not person" A - («alient»).

In the study of linguistic information identification systems of the well-known US Institute of Standards (NIST), instead of criterion (1), the criterion is recently used [6].

$$\Delta(Z) = \log\left(\frac{p(Z|H0)}{p(Z|H1)}\right) \tag{2}$$

There are many methods for describing a "non-person" model, but we will be most interested in the following two methods. The first method will be based on the selection for each person A of certain standard of templates $\overline{A_1}, \ldots, \overline{A_N}$ [5]. Due to the fact that these templates will be created for each person, it can be concluded that the templates refer to the "non-person" model. The second method is based on the selection as a benchmark of persons falling out of the general distribution, who will correspond to the "not a person" model. This method requires long-term training on test data, but due to adjustment to a large set of input data, it is considered the most effective and is used more often than other methods when building language information identification systems [6].

When developing practical systems for the identification of language information, a certain threshold value θ s most often determined to make a choice about accepting or rejecting a person. The correct choice of θ is always a difficult task that requires a number of experimental studies.

Before using speech signal identification systems, it is necessary to evaluate the accuracy of identification. For such systems, there are three main accuracy factors:

1. crossover error rate (CER);
2. false reject rate (FRR);
3. false acceptance rate (FAR).

A false rejection occurs when the identification system rejects the "correct" option. A false deviation is usually called Type I errors. Such false deviations are quite important in real identification because they lead to a loss of trust in the systems and to the need for additional time for authorization. False acceptance occurs when the "wrong" option is accepted as the "correct" one. False acceptance is also very important, because with this option, an unauthorized person will gain access to resources. It is customary to call a false acceptance Type II errors. Type II errors are thought to be more important than Type I errors because it is better to not admit the real person than to admit the wrong person when making decisions.

The crossover error rate (Fig. 1) is the point of intersection of the error curves of Type I error (FRR) and Type II error (FAR) [7].

The false acceptance rate is also called equal error rate (EER). Considering the fact that this coefficient accumulates the values of false rejection and false acceptance coefficients, it is most often found in the description of speech signal identification systems [8].

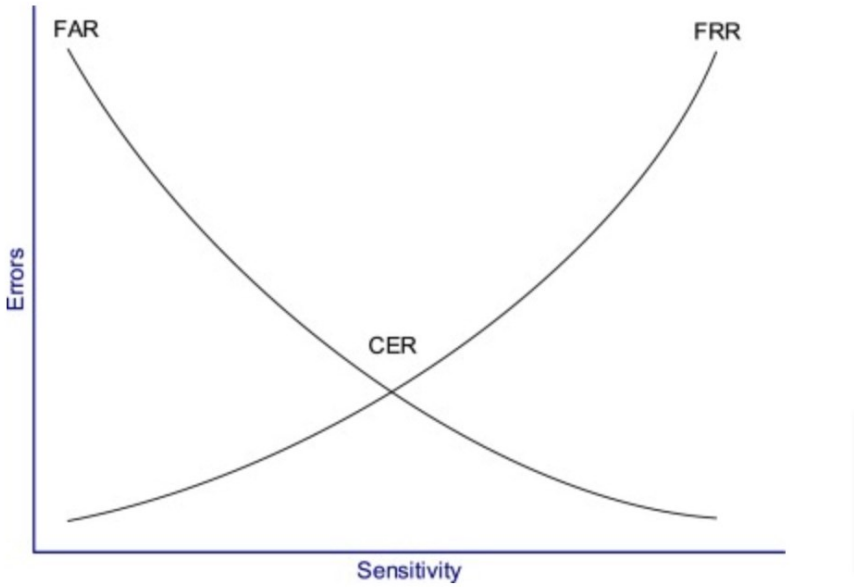

Fig.1. Graph of dependencies of errors of the I and II Types

In the competition among systems of identification of speech signal systems, the learning cost function is also adopted, which determines the weighted sum of the probabilities of false acceptance and false rejection:

$$C_{Det}(\theta) = C_{Miss} \times P_{Target} \times P_{Miss}(\theta) + C_{FalseAlarm} \times \\ \times \left(1 - P_{Target}\right) \times P_{FalseAlarm}(\theta) \quad (3)$$

where the parameters of the cost function are C_{Miss} (missed detection cost factor=1) i $C_{FalseAlarm}$ (false detection cost factor=1), and P_{Target} (a priori probability of the specified target person=0,05), P_{Miss} (probability of missed detection).

Making a decision on the identification of a speech signal is the main indicator and result of research when comparing the speech signals of individuals. Therefore, special attention should be paid to the correct selection of the parameters of the identification system in order to ensure the necessary values of errors of the I and II Types. Most biometric systems have a flexible threshold that controls the balance between these two types of errors. In each program, the optimal threshold is found empirically.

NIST evaluations of various speaker identification systems showed [5]:

– comparison of announcers' voices on the basis of a limited set of data - The point of equality of errors of the I and II types lies within 5-10%. The degree of confidence of the classifier in the obtained result is approximately 95%;

– verification of the announcer based on an extended data set - the point of equality of errors of the I and II types is much lower, in the region of 1.3 - 2%, which roughly corresponds to a relative decrease in the number of errors by 74-80%;

– comparison of announcers' voices based on an extended data set - point of equal probability of errors - 12-15%.

Experiments

In view of the presented identification errors, it can be concluded that the existing announcer identification systems cause fair complaints from users related to the objectivity of examination results. The conducted studies [2] showed that the expressed doubts are fully justified. This conclusion is due mainly to the fact that in most modern means of conducting identification studies of voice signals, the Fourier transformation is used, which is an artificial mathematical method of decomposing a complex signal into periodic components. But the mechanism of perception and transformation of sound vibrations by the human hearing apparatus is arranged differently and such artificial transformations cannot exist in it. It was also established that the main processes of information transmission to the brain, contained in sound signals, are of an impulse nature, and the duration of these impulses ranges from tens to hundreds of milliseconds [4], in connection with the above, it was concluded that the need the use of a multifractal approach to build a speaker identification system [5]. The study of signals in the time domain is necessary because all phonemes have a well-defined fractal character that is preserved and is individual for each phoneme, that is, the form of the phoneme signal in the time domain is the same in all languages and approximately the same when it is pronounced by any individual. It is this uniformity that allows us to recognize the language of any person. The main

difference, which determines the individuality of the speaker, is the individuality of the frequency composition of the signals that make up this sound when it is pronounced by a specific person. This individuality, in our opinion, is determined by the frequency of the main tone and is modulated by the parameters of this frequency. Both the frequency of the main tone and these parameters are determined by the individuality of the components of the vocal tract of any person [2].

During the implementation of the software, two interrelated tasks arose - automatic segmentation of the phonogram and selection, calculation and determination of the degree of proximity of fractal formations contained in the investigated signals of the controversial and exemplary phonograms. Both of these tasks are solved in [5, 6].

Within the framework of the speaker identification task, two interrelated tasks of speaker identification and verification can be distinguished [6]. In the first task, the goal is to identify the audio component as pronounced by one of the announcers from the considered set, in the second - to establish the belonging of the audio component to a specific reference announcer.

Based on these tasks, systems are divided into three parts:

1. determination of individual features of the speech signal;
2. representation of the characteristic standard of the announcer;
3. making a decision about the announcer's personality.

On the basis of the above, it is possible to distinguish the following main stages of the implementation of the announcer recognition system:

Measurement of the fractal dimension of signal components. A stage that is simple to implement, but quite effective in the set of all discriminability measures. Its implementation is possible both with a permanent window and with an adaptive type of window.

Definition of phrase boundaries. To solve this problem, it is most rational to use language segmentation algorithms based on the multifractal approach. Based on this approach, in those elements of the signal, where the change in fractal dimension exceeds some set threshold, it is assumed that a phrase begins.

Selection of the main tone. To solve the task of selecting the main tone, there is a need to develop an interference-resistant method of selecting the main tone for each period. An algorithm based on the use of Morle wavelet approximation of the signal with subsequent statistical analysis of the distribution of wavelet maxima, which is physically explained by the presence of self-similar structures characteristic of signals associated with resonators, can be taken as a basic algorithm for the selection of the main tone.

At the stage of measuring the main tone on the signal sections, it makes sense to compare not the absolute values, but the normalized values - this makes it possible to more accurately distinguish announcers by intonation color.

Selection of characteristic parameters of the main tone. To solve this problem, you can use the finding of only some of the considered parameters during the analysis for each fragment: the average frequency and dispersion of the main tone; distribution of periods of the main tone; amplitude modulation of the main tone; frequency modulation of periods of the main tone.

Comparison of signal parameters with reference parameters. After carrying out the process of comparing speech parameters with the reference ones, you need to select the most "close" speaker from the database. To do this, it is necessary to compare the selected parameters of the main tone from the database based on a probabilistic approach.

In the process of conducting research, the following method of examination was found [4]:

1. To conduct speaker identification research, two or more phonograms are provided for speaker matching, as a rule, in a set of phonograms, at least one phonogram clearly belongs to the voice of a particular speaker.
2. Each phonogram is segmented into fragments based on the fractal dimension [4].
3. For each fragment of each phonogram, the frequency distribution of the main tone over the entire length of the phonogram is calculated based on the frequency distributions of the main tone obtained for the fragments.
4. Phonogram data accurately identifying a person (the given owner of the voice) is stored in the identification database.
5. For each of the phonograms, for which speaker identification must be performed, membership in one distribution for the frequencies of the main tone of each fragment selected from the phonogram is checked, with similar distributions stored in the identification database.
6. Based on the assessment of the degree of closeness between the frequency distribution of the fundamental tone, the announcer is set based on the degree of closeness to the considered distribution.

Let's consider the architecture of the implemented speaker identification system in the UML language in the form of class and sequence diagrams. The class diagram reflects the static structure of the system. It consists of a description of classes and relationships between them. A sequence diagram displays the dynamic relationships in the system, for example, the sequence of calls.

Figure 2 shows a diagram of calls during preliminary preparation for highlighting the characteristics of the announcer's speech. In pre-language recognition mode, the system boots with a prepared configuration file and an input signal. Recognition will be done through the configuration manager.

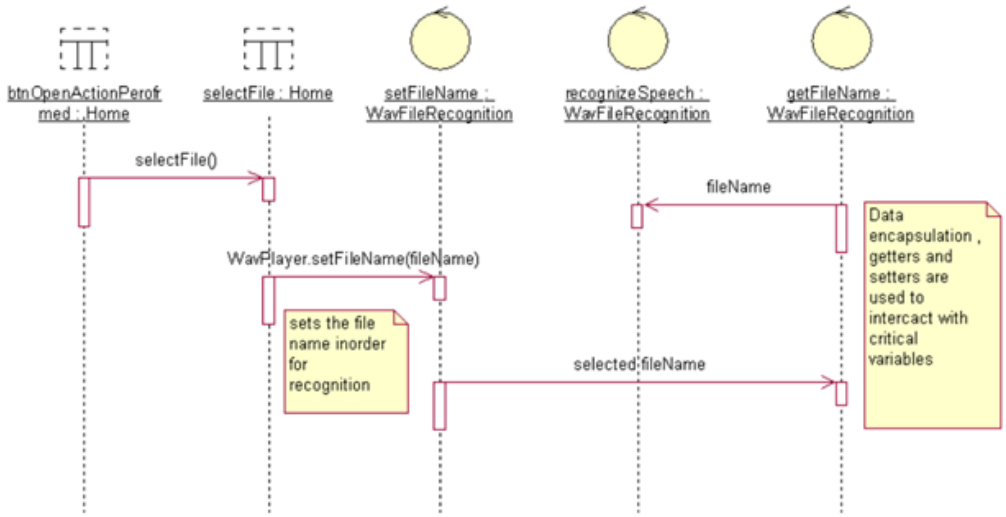

Fig.2. Sequence of preprocessing calls

Figure 3 shows the call diagram during post-processing during speaker recognition. At this stage, the input digital signal will go through the process of dividing it into vocalized and non-vocalized parts, decomposition by Morley wavelet followed by statistical analysis of the distribution of wavelet maxima and determination of the frequency of the main tone for the segments. The AudioFileDataSource and Recognizer classes implement functions to perform these tasks. The result of the sequence of calls are the labels of the class of the announcer and the language to which the classifier classified the input speech signal.

In figure 4 presents a diagram of entity classes, which are object representations of data managed by the identification system.

Home acts as a graphical interface of the software system, which directly interacts with DBSpeaker and WavFileRecognizer. DBSpeaker performs the functions of presentation and description of saved recordings of announcers. WavFileRecognizer is designed to implement the process of reading a sound signal (from a stream or from a file) and identifying the speaker. AudioFileDataSource implements the function of reading the audio signal, and Recognizer implements the speaker identification.

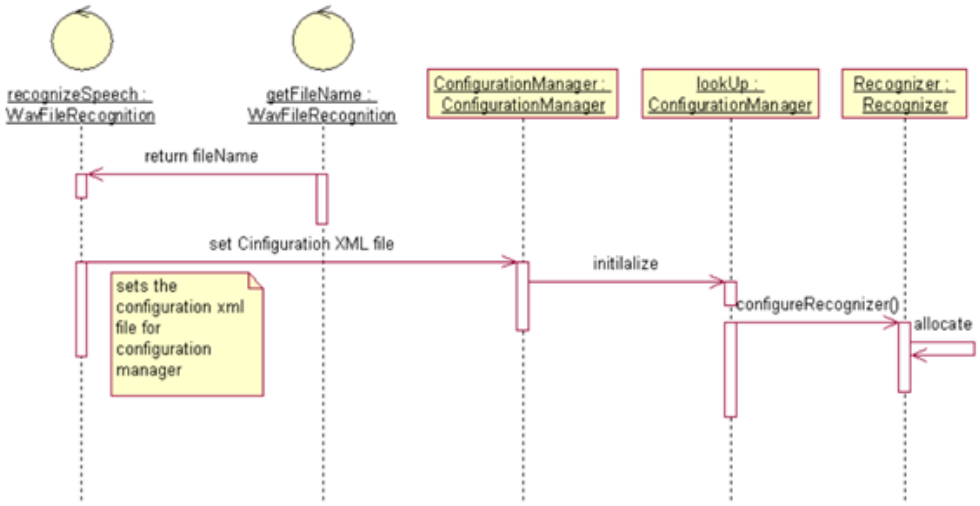

Fig.3. Sequence of calls of the speaker recognition process

The VoiceFeatures abstract class is designed to store and calculate the features of the input speech signal. The class consists of an array of VoiceFeatureValue objects and the ExtractFeatures obtaining method, which performs feature extraction from the input speech signal. The inheritors of the class are the classes performing fragment-by-fragment analysis: average frequency and dispersion of the main tone; distribution of periods of the main tone; amplitude modulation of the main tone; frequency modulation of periods of the main tone.

The PersonClassifier abstract class is designed to implement the classification algorithm. The class consists of the Train and Classify methods, as well as the Parameters object, which contains all the parameters necessary for the work of the classifier. The Train method accepts as input a dictionary, in which the key is a class label, and the

value is an object of type Features, and returns a Parameters object. The PersonClassify method takes a VoiceFeatureValue object and returns the value of the decision feature, as well as the decision class label.

The SpeechUtils class contains helper methods needed for feature computation and classification, such as, for example, computation of vocalized/unvocalized segmentation and denoising.

Thus, a software architecture for speaker identification tasks using a multifractal approach in describing the structure of speech is proposed. The use of a similar architecture and the use of a multifractal approach will generally improve the accuracy of speaker identification.

Based on the proposed architecture, a software system was developed in the Python programming language using the SQLlite database.

The software system for identifying a person's speech signal is searchable, as it is the result of ranking according to the degree of proximity of individual parameters of the speech signal.

The software system for digital recording of informational messages automatically calculates the parameters of language characteristics and further ranks these characteristics in the database of individuals.

The software system uses a ranking method based on three different criteria. These include:

- calculation of the proximity of two-dimensional probability density function curves for the frequency of the main tone and the location in the spectrum of three frequency ranges that are extracted from the speech recorded in the speech signal;
- calculation of the proximity of the probability density function curves for each of these features separately;
- calculation of the degree of closeness of the absolute maxima of the formant spectra extracted from the speech recorded in the speech signal.

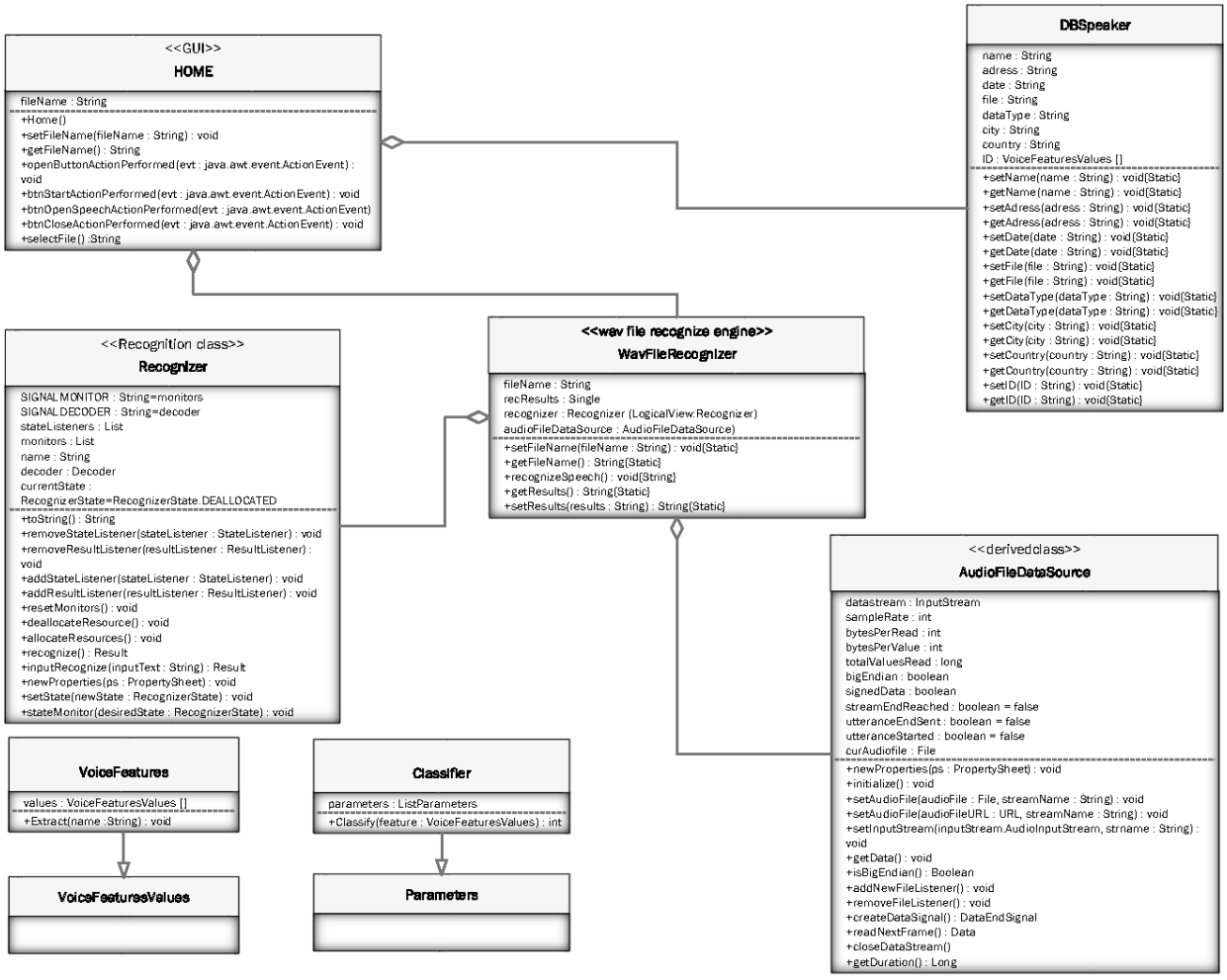

Fig. 4. Class-entity diagram

The result of making a decision of the developed program is a graphical representation of the proximity of the curves of the two-dimensional probability density functions for each of the signs (Fig. 5).

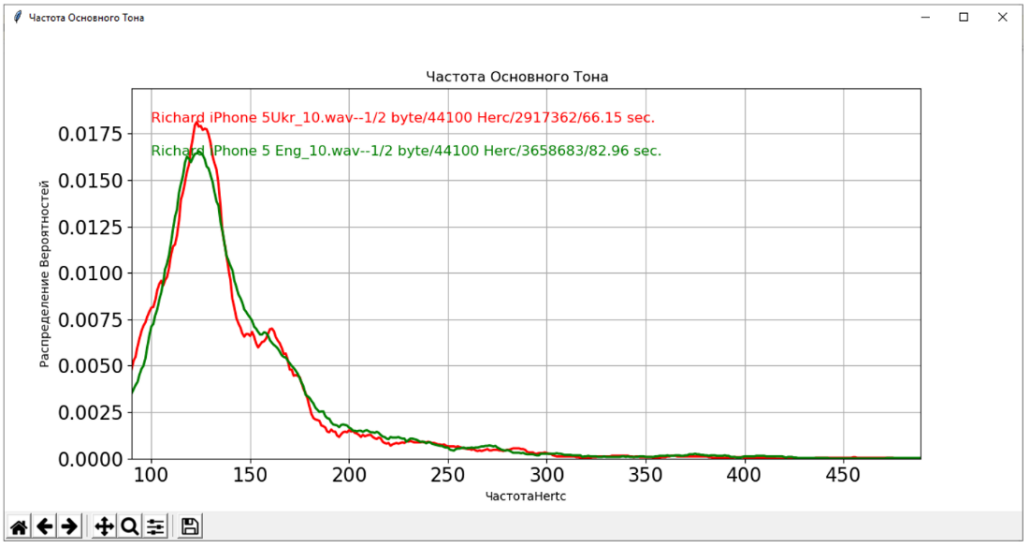

Fig. 5. The resulting graph of the comparison of speech signals of individuals by the frequency of the main tone

Conclusions

The following conclusions can be drawn on the basis of the performed research:

1. approaches to the construction of existing announcer identification systems are considered;
2. the main criteria for assessing the accuracy of announcer identification were investigated and the main sources of loss of accuracy during announcer identification were identified;
3. the structural construction of the announcer identification system is considered, taking into account the identified sources of loss of accuracy during announcer identification;
4. the proposed architecture of the speaker identification system in the UML language in the form of class and sequence diagrams.
5. a software system was built that implements the functions of speech signal identification according to the methods and algorithm proposed in previous works [3-5, 10].

References

1. NIST Speaker Recognition Evaluation. URL:https://www.nist.gov/itl/iad/mig/speaker-recognition
2. Рибальський О.В., В.І. Соловйов, В.В. Журавель, Т.О. Татарнікова. Методика ідентифікаційних і діагностичних досліджень матеріалів та апаратури цифрового й аналогового звукозапису зі застосуванням програмного забезпечення «Фрактал» при проведенні експертиз матеріалів та засобів відео та звукозапису. Науково-методичний посібник : Київ: ДУІКТ, 2013. – 75 с.
3. Solovjov V.I., Byelozorova Ya.A. Multifractal approach in pattern recognition of an announcer's voice. // Teka. Commission of motorization and energetics in agriculture. – 204. – Vol. 15. – № 2. – P. 13-21.
4. Byelozorova Ya.A. The allocation of self-similar structures in voice signals for speaker identification tasks. // ScienceRise. – 2017. – № 5. – P.125-142
5. Бєлозьорова Я.А. Ідентифікація диктора на основі кратномасштабного аналізу// Інженерія програмного забезпечення. – 2017. – №1(29). – С. 15-24.
6. Greenberg S., Singer C., and Mason L. NIST 2020 CTS Speaker Recognition Challenge Evaluation Plan, NIST 2020 CTS Speaker Recognition Challenge, [online], https://www.nist.gov/itl/iad/mig/nist-2020-cts-speaker-recognition-challenge (Accessed June 25, 2021)
7. Hansen J.H.L. A Study on Universal Background Model Training in Speaker Verification / Hansen J.H.L., Hasan T. // Transactions on Audio, Speech, and Language Processing. vol. 19 (7). 2011. pp. 1890–1899.
8. False rejection. URL: https://www.sciencedirect.com/topics/computer-science/false-rejection
9. Татарникова Т.А. Рыбальский О.В., Соловьев В.И., Командина Т.В. Общие подходы к экспертизе оригинальности и подлинности материалов цифровой и аналоговой звукозаписи/ Науковий вісник НАВС. – К.2011, №4, с. 183-191
10. Zybin, S., Bielozorova, Y. Method of Extracting Formant Frequencies Based on a Vocal Signal. In: Hu, Z., Zhang, Q., Petoukhov, S., He, M. (eds) Advances in Artificial Systems for Logistics Engineering. ICAILE 2022. Lecture Notes on Data Engineering and Communications Technologies, vol 135. Springer, Cham. https://doi.org/10.1007/978-3-031-04809-8_40

Yana Bielozorova **Яна Бєлозьорова**	PhD, Associate Professor of Software Engineering Department, National Aviation University, Kyiv, Ukraine https://orcid.org/0000-0002-0688-3436 e-mail: bryukhanova.ya@gmail.com	кандидат технічних наук, доцент кафедри інженерії програмного забезпечення, Національний авіаційний університет, Київ, Україна
Kateryna Yatsko **Катерина Яцко**	Assoc. Prof. of Software Engineering Department, National Aviation University, Kyiv, Ukraine e-mail: kateryna.yatsko@npp.nau.edu.ua https://orcid.org/0000-0002-4389-6033	асистент кафедри інженерії програмного забезпечення, Національний авіаційний університет, Київ, Україна