symmetry

MDPI

Article

Metric of Highlighting the Synchronicity of Time Series and Its Application in Analyzing the Fundamental Frequencies of the Speaker's Speech Signal

Elena Kataeva, Alexey Yakimuk, Anton Konev and Alexander Shelupanov *

Department of Security, Tomsk State University of Control Systems and Radioelectronics, 40 Lenina Prospect, 634050 Tomsk, Russia; kes@keva.tusur.ru (E.K.); yay@keva.tusur.ru (A.Y.); kaa1@keva.tusur.ru (A.K.)
* Correspondence: saa@tusur.ru; Tel.: +7-(3822)-70-15-29

Received: 13 October 2020; Accepted: 22 November 2020; Published: 25 November 2020

Abstract: As part of the study, the problem of developing and applying a metric for assessing the degree of similarity of time series is considered, which makes it possible to consider the assumptions about the model of a series when comparing, as well as to compare the values of the corresponding characteristics of the series. Characteristics can be values that describe the structure of a series, or directly the values of the approximating function, which can be obtained using nonparametric statistics methods. One of the directions in which this approach can be applied to assessing the similarity of time series is the study of vocal performances. In practice, the degree of similarity in the performance of melodies by several speakers was analyzed. It was determined that, using the synchronicity metric, it is possible to implement an exercise in which students need to repeat the melody after the teacher. In addition, this approach was applied in the segment identification module with an abrupt change in the sounding of the fundamental frequency. This work is devoted to the modification of the program complex for vocal recognition in order to identify notes with a sharp change in the fundamental frequency. The complex is aimed at carrying out additional independent training in teaching vocals. The use of the software package will allow, in real time, providing feedback to the user with an assessment of the quality of their singing. This should allow students to study not only under the supervision of a teacher, but also independently in the early stages of learning. The basic algorithm of the program recognizes notes without sharp changes in frequencies with high accuracy, which is confirmed by experiments. In order to recognize by the algorithms of the program notes sung vibrato and glissando in singing, a new analysis method based on the metric of time series synchronicity is proposed.

Keywords: synchronicity; time series; vocal performance; note recognition; speech technologies; fundamental frequency; learning analytics

1. Introduction

Often when solving problems, it becomes necessary to assess the similarity of objects. In some cases, this problem is solved by correlation analysis methods. Moreover, methods of cluster analysis have become widespread, dividing a set of objects according to the values of their attributes into groups. The most famous and simple methods of cluster analysis use the following approach: objects are combined into clusters based on the values of the distances between them, and the merging process is regulated by the selected method, i.e., when solving the problem of cluster analysis, it is necessary to choose a metric and a partitioning method. Different partitioning methods and metrics are better suited for different types of objects and different clustering goals. In particular, for clustering time

series, separate types of metrics are required that consider the peculiarities of such data: the presence of a dependence on time, which is expressed in the dynamics of the series, and the complexity of the structure of the series, which is measured by various indices and coefficients. In recent decades, for particular problems, a large number of metrics have been proposed, which either consider a priori assumptions about the form of a model that well describe the dynamics of the series, or measure the distance between objects based on various characteristics of the series. Thus, there is no generally accepted metric for clustering time series; for each task, it is necessary to select an individual approach.

One of the directions in which the analysis of the similarity of time series can be applied is the sphere of speech technologies. A lot of research in this area is aimed at analyzing the parameters of the speech signal. One of these parameters is the fundamental frequency (F_0). Knowing the F_0 of a signal at a particular point in time is important in many areas of speech technology.

Determining the features in the fundamental frequency of a speech signal is an important task in the field of studying the features of the language. Some languages and accents are characterized by the appearance of an ascending or descending tone, which in its characteristics may be similar to vocal performance. The works of [1,2] consider the possibility of using the detection of an ascending–descending tone in speech as a marker for deciding whether a speaker has a Welsh accent. The authors evaluated intonation based on whether the voice was raised or lowered. The resulting F_0 spread was obtained in the range from 169 to 358 Hz. The study of [3] raised the issue of determining the characteristics in the speech of young people. Based on the assessment of the F_0 of the recordings of speech signals, it was determined that intentional stretching of vowels is characteristic of young people. The authors of works [4–6] conducted a study of the pitch-melodic parameter of speech, namely the influence of age-related changes on the rhythmic characteristics of the language. It was determined that with an increase in the speaker's age, a decrease in the maximum and minimum values of the F_0 is manifested. One of the features of [7] is the determination of the emotional state of the speaker based on the identification of the characteristics of the melody form in speech. The author drew attention to such characteristics as direction, character and range of tone movement. In this study, the values used in the study of musical melodies were used to measure the F_0. This made it possible to visually represent the features of changes in the speaker's speech when emotions or intentions are displayed. A similar task was posed in [8], where, based on the change in the difference in the acoustic values of the average and maximum frequency of a phrase, the evaluativeness in the speech of reporters was determined in order to convey the atmosphere of what was happening.

In addition, in the direction related to the identification of speakers, the F_0 of the speech signal also plays a role. In addition to formants reflecting the individual characteristics of a person, speaker identification systems also analyze the fundamental frequency of the signal under study. This is confirmed by the fact that the determination of the basic frequency of the speech signal, considering the peculiarities of speech formation and speech perception associated with human anatomy and physiology, is extremely important not only in studies on user identification [9–11], but also in the field of rehabilitation for cancer patients after resection of the larynx [12,13]. The further choice of material for the study of the metric is due to the fact that it is easier to determine the reference frequencies when analyzing vocal performance than when studying speech.

Singing can be thought of as a special form of speech that is created in the same way, but there is additional control to create the musical aspect. The natural melody of speech (prosody) differs in different languages and determines the pitch contour, volume variation, rhythm and tempo of emotional expression. In singing, the pitch, volume and timbre are primarily determined by the composition due to the fact that to match the duration of the note, the vowels sound longer than usual. Speech vowels are characterized by special formant positions. In singing, the position of the formant can be radically altered by changing the length and shape of the vocal tract and the position of the articulators. A review of studies on music processing has shown that from the point of view of signal processing, a melody can be represented by sequences of F_0 determined at the moments of sounding,

that is, in areas where the instrument that creates the melody is active. This allows us to consider the study of the dynamics of the parameters of a speech signal as a time series.

The review of articles in the field of application of feedback from software for teaching singing showed that this topic is relevant. These studies provide different approaches to providing feedback to a singer based on the melody performed by them. Some of the more popular approaches include spectrogram displays, piano keyboards and frequency graphs [14–17]. Among other possibilities, the study of [16] provides an additional opportunity provided by teaching with a software tool. The program can accumulate the results of the student, which can later be viewed and used to evaluate the changes in development that have occurred. This can have a positive effect on student motivation. Since the provided result is an assessment of physiological parameters (fundamental frequency), in further work, we will call this biofeedback [18]. However, most of the works concentrate on the teaching method without giving detailed information about the quality of vocal performance processing. The algorithms used in the programs do not allow calculating the value of the fundamental frequency in a vocal performance with high accuracy due to the presence of a high percentage of gross errors in them or are limited by a narrow spectrum of covered frequencies.

The purpose of this research is to test the applicability of the metric of highlighting the synchronicity of time series for assessing the type of vocal performance with a sharp change in the fundamental frequencies. Previous research has concluded that the program complex for vocal recognition is able to identify notes with smooth frequency changes. A detailed description of the applied algorithm for identifying notes in vocal performance and the obtained test results will be published as part of a further study. As a result, vibrato and glissando singing styles were perceived as noise. Thus, the task of identifying these segments appeared.

In this issue, the value of the fundamental frequency is used as estimated information at the recognition of music sung by a speaker. This information is sufficient for identifying notes at each estimated moment in accordance with music theory. Further stages of note recognition include segmentation, taking into account the minimum sound duration and filtering out noise. Speaker identification is not a purpose of this study. Moreover, one of the tasks of the system development is maximizing the objectivity of the evaluation. The recorded note is determined by the fundamental frequency. Accordingly, the individual characteristics of the singer determined by their articulators do not matter for note identification. With further research of adjacent directions associated with more stringent requirements for the set of parameters, it is possible to expand the processed dataset.

Accordingly, the data analysis method should enable determining the type of singing based on the values of the fundamental frequencies in the investigated period of time. The developed software package is aimed at helping the user in individual singing lessons. Biofeedback provided by the program should tell the speaker if they sang the note incorrectly. In this matter, the promptness of the assessment is important. Accordingly, the processing of singing should be done in real time or for a comparable duration. Thus, the less information needs to be processed to get an estimate of the sung melody, the better. This method should not require large numbers of computations and has to be easy to implement.

Otherwise, real-time signal processing would be unattainable. In addition, the system should not depend on the speaker. The use of neural networks will require training the system on examples of singing with a change in the values of the frequencies of the main tone when playing a single note. Due to the fact that the system is being developed to teach speakers how to sing correctly, it is incorrect to set up the system based on the speaker's initial exercises. There is a possibility that the system will evaluate the user non-objectively, since conclusions about the correct performance will be based on the speaker's unformed ear for music.

2. Overview of Similarity Analysis Methods

Throughout history, humans have sought and accumulated knowledge about the world around them. Based on the knowledge gained, people ordered objects in accordance with their similarity. Often, such a grouping was essential for survival, for example, the division of plants into edible and inedible, animals based on danger and so on. This process of grouping objects by a person on some grounds is called classification.

A preliminary task for grouping is to determine the degree of similarity of objects by a particular attribute (or a set of attributes). In particular, the study of the degree of similarity of large amounts of data is a popular problem at the intersection of mathematics and climatology. The presence of such a similarity was considered, in particular, in [19]. The author described the presence of similarities in the dynamics of the magma column shift in Iceland and the Hawaiian Islands and investigated the influence of these shifts on other factors: temperature, continental climate, amount of atmospheric dust, etc. The so far popular task of studying the influence of solar activity on temperature and pressure was considered in the study of [20]. A large amount of values was investigated when determining the similarity of the behavior of atmospheric pressure at different stations in [21].

In modern geophysics and climatology, determining the presence of similarity between different data is an important problem. In the books [22,23], many problems are considered, one way or another related to the problem of similarity. For example, in [24–26], the determination of similarity is necessary to construct a function of the dependence of the maximum density of the annual ring on the total ozone content (TOC) in the atmosphere and for further reconstruction of unknown TOC values from this function. This task of the so-called bioindication is very important, since the series of ring densities are very long, while the TOC values have been measured only since 1927, and it is still impossible to carry out a correct analysis for them.

With the development of science, classification has also developed as one of its fundamental elements. However, until the last century, this process was based, first of all, on the natural capabilities of a person to recognize patterns and group objects. The 20th century is characterized by a sharp increase in scientific knowledge. Effective analysis of information becomes almost impossible for a person due to its volume and complexity and requires new approaches. Under these conditions, the automation of various areas of human activity through the introduction of computer technology has also affected the classification process. This is based on the idea of using mathematical methods to group objects in the real world. It is necessary to conduct research on methods for determining the similarity of objects.

Some of the simplest and most well-known methods for assessing the similarity of time series are correlation analysis methods. For the first time, the term correlation was introduced into scientific circulation by the French paleontologist Georges Cuvier in the 18th century. He deduced the principle of correlation of parts and organs of living beings, with the help of which it is possible to restore the appearance of a fossil animal, having only a part of its remains at disposal. He introduced the principle of correlation between parts and organs of animals. The principle of correlation helps to restore the appearance of an organism from the skull, bones, etc., found in excavations. The appearance of the entire animal and its place in the system can be assumed as follows: if the skull is with horns, then it was a herbivore, and its limbs had hooves; if the paw is with claws, then it is a predatory animal without horns, but with large fangs.

German psychiatrist G.T. Fechner (1801–1887) proposed a measure of the tightness of communication in the form of the ratio of the difference in the number of pairs of coinciding and non-coinciding pairs of signs to the sum of these numbers:

$$K_{Fechner} = \frac{C - H}{C + H} \tag{1}$$

where C is the number of pairs in which the signs of deviations of values from their means coincide, and H is the number of pairs in which the signs of deviations of values from their means do not coincide.

The Fechner coefficient is a rather rough indicator of the tightness of communication, which does not consider the magnitude of deviations of features from the mean values, but it can serve as a certain guideline in assessing the intensity of communication.

One of the most popular correlation indicators is Spearman's coefficient, calculated by the formula

$$r_{ij} = \frac{\sum_{k=1}^{n}\left(X_k{}^{(i)} - \overline{X}^{(i)}\right)\cdot\left(X_k{}^{(j)} - \overline{X}^{(j)}\right)}{\sum_{k=1}^{n}\left(X_k{}^{(i)} - \overline{X}^{(i)}\right)\cdot\sum_{k=1}^{n}\left(X_k{}^{(j)} - \overline{X}^{(j)}\right)} \quad (2)$$

where $X^{(i)} = \begin{pmatrix} X_1^{(i)} \\ \dots \\ X_n^{(i)} \end{pmatrix}$ and $X^{(j)} = \begin{pmatrix} X_1^{(j)} \\ \dots \\ X_n^{(j)} \end{pmatrix}$ are values of two time series of the same length.

When assessing the degree of similarity by calculating the correlation coefficients, a serious inaccuracy is possible, since the correlation coefficient shows the presence of a linear relationship between the data, while the relationship may not be linear or there will be partial similarity of objects.

In econometric problems, such kind of similarity of time series as cointegration is considered. Cointegration [27] refers to a cause-and-effect relationship in the levels of two (or more) time series, which is expressed in the coincidence or opposite direction of their tendencies and random fluctuations. According to this theory, cointegration exists between two time series if the linear combination of the time series is a stationary time series (i.e., a series containing only a random component and having constant variance over a long period of time).

Additionally, such an approach to assessing the similarity of data as cluster analysis has received great development. The beginning of the development of cluster analysis dates back to the first half of the 20th century. A detailed description of cluster analysis is given in the book [28]. Cluster analysis has several advantages over other methods of data classification. First of all, this is due to the fact that it allows you to split objects not one by one, but by a whole set of features. Moreover, the influence of each of the parameters can be rather simply strengthened or weakened by introducing the corresponding coefficients into the mathematical formulas. In addition, cluster analysis does not impose restrictions on the type of objects to be grouped and allows considering a variety of initial data of almost arbitrary nature. Another feature of clustering is that many algorithms are able to independently determine the number of clusters into which the data should be split, as well as highlighting the characteristics of these clusters without human intervention, only using the algorithm involved.

The great advantage of cluster analysis is that it allows you to split objects not by one parameter, but by a number of features. In addition, cluster analysis does not impose any restrictions on the type of objects under consideration, and allows considering a variety of initial data of almost arbitrary nature.

On the other hand, one of the key problems of cluster analysis is determining the number of clusters. Some methods require a priori determination of clusters, while in others, their number is determined in the process of agglomeration or division of a set of existing objects. One of the most frequently used methods for determining the number of clusters is the application of the "scree test". The process of grouping objects in hierarchical cluster analysis corresponds to a gradual decrease in the average distance between clusters with an increase in the number of clusters. At the stage where this measure of distance stops decreasing abruptly, the process of partitioning into clusters must be stopped, since otherwise the partitioning of clusters is unnecessary. The optimal number is considered to be the number of clusters equal to the difference between the number of observations and the number of steps, after which the coefficient decreases smoothly.

Time series clustering methods are divided into two groups:

(1) Methods based on estimating the complexity of a time series;
(2) Methods based on model assumptions.

The methods of the first group are considered in the works of D. Sankoff and J. Kruskal [29], D. J. Berndt and J. Clifford [30], T. Oates, L. Firoiu and P.R. Cohen, E.A. Maharaj, M. Corduas and D. Piccolo, Chouakria A. Douzal, and PN Nagabhushan [31]—these works propose an algorithm for dynamic-scale transformation, including using the theory of hidden Markov processes, using autocorrelation coefficients, spectral characteristics, wavelet coefficients and the adjustable difference index. In the studies of K.Y. Staroverov and V.M. Bure [32,33], they solve the narrower problem of clustering short time series. It is also possible to single out the method of identifying structures described in the works of S.G. Kataev [34], which is applicable in various problems, including clustering of time series. Among the methods of the second group, the transition to autoregression coefficients is mainly considered as to the features—in particular, in [35] for classification, in addition to autoregressive coefficients, the characteristics of a series are used—mean value, variance, etc. In [36], the clustering after transformation and smoothing (CATS) approach is presented, in which it is proposed to construct approximating functions for the compared time series (transformation) before clustering, where the resulting curves are additionally smoothed (smoothing) and then the set of coefficients obtained are used as clustering objects. It is stated further that any clustering method can be applied; the k-means method is used in the work itself.

At the Institute for Monitoring Climatic and Ecological Systems of the Siberian Branch of the Russian Academy of Sciences in the laboratory of bioinformation technologies under the leadership of V.A. Tartakovsky, an approach to assessing the similarity of time series was developed considering the hypothesis of synchrony factors. The results are presented in the works of [37,38]. It is believed that the external forcing effect synchronizes natural and climatic processes and manifests itself in the similarity of their essential features. The essence of the approach is that orthogonal components of processes are introduced, which differ in the coincidence and non-coincidence of essential features, orthogonal expansion of the series under study is performed, similar to the principle of trigonometric Fourier transform, and then the correlation coefficient is calculated for the obtained components. If the correlation coefficient is high, the presence of synchronicity in the dynamics of the series is recognized. This method allows one to examine series of different lengths for synchronous behavior.

The concept of synchronicity of time series is also used in this study. We will call time series synchronous if the dynamics of their development are similar. Then, we can consider the problem of assessing the similarity of time series as an assessment of the degree of synchronicity of the series, that is, the synchronicity of the dynamics of time series is a special case of their similarity.

The review of articles has shown that standard classification methods are not suitable for solving this problem. To apply clustering methods, there should be several speakers, while in our task, the singing of one performer is studied, represented by one time series. For classification, it is necessary to preset the characteristics of the vibrato or glissando class, which at this stage of the study is not possible due to the lack of necessary information about the general laws of singing with such effects for any performer. Using Gaussian mixture model (GMM) also involves handling multiple speakers. This parametric probability density function is characterized by a feature data vector with the characteristics of the weights of these features. Accordingly, this approach is not suitable for this study. This is due to the fact that the work of the software package is not aimed at identifying the speaker from several options, but only the value of the fundamental frequency is used in the assessment.

3. Synchronicity Extraction Method

The novelty of the proposed method lies in the possibility of adjusting the parameters when assessing the similarity. Earlier, the software package implemented the ability to independently regulate the system's requirements for the purity of the note performance. By default, each note is assigned a sound band with boundaries, determined from data from music theory. In this case, the lower boundary of each note starts from the last value that was not included in the previous note's band. If desired, on the part of the speaker, it is possible to increase the system's requirements for the quality of performance. This will narrow the boundaries for each note. As a result, the boundary

frequencies between the notes will be perceived by the system as noise. This idea can be applied to assess the similarity of multiple time series. In contrast to traditional metrics, the proposed method presents the degree of similarity of time series as a percentage, which is convenient for a qualitative interpretation of the degree of closeness. The requirement for the rigor of the system in assessing similarity can be adjusted taking into account E, which allows increasing (or decreasing, if necessary) the complexity in the learning process.

The synchronicity of the dynamics of time series is a rather subjective concept, which follows from the ambiguity of solutions to the problem of assessing the similarity of time series—there is no single indicator of either the degree of similarity of the series or the degree of their synchronicity. The concept of synchronicity can be best illustrated with graphs. Let there be given three time series X, Y and Z. Series X and Y have synchronous dynamics, and the behavior of series Y and Z is different. Figure 1 shows their comparative graphs with synchronous and non-synchronous dynamics.

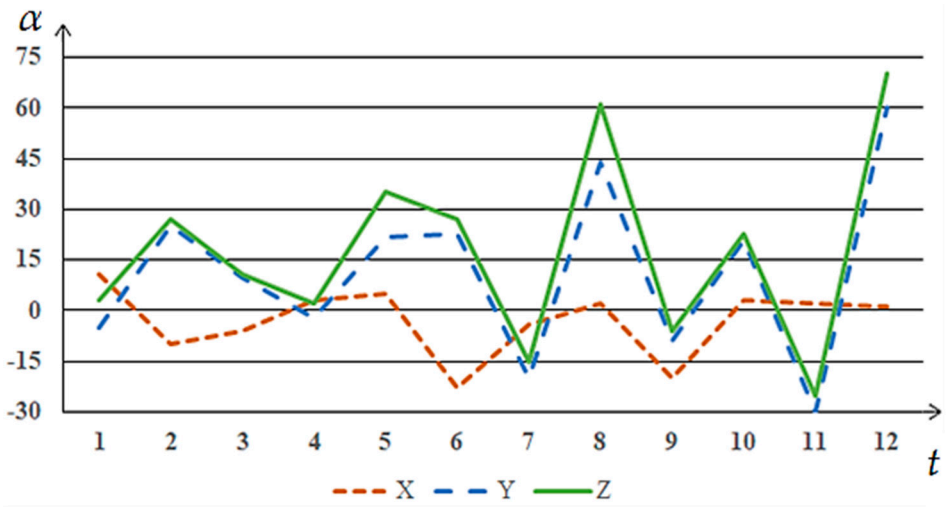

Figure 1. Time series plots with synchronous and asynchronous dynamics.

To simplify the understanding of the concept of synchronicity, we will assume that this graph shows the dependence of a certain value of alpha on time. The timeline in this example is split into 12 units representing months. The ordinate is the alpha value that changes over time. This value may have positive or negative values. The behavior of a quantity is determined by the research subject. In the presented study, the ordinate will be the fundamental frequency values, measured in Hz, and the abscissa will be the time scale in seconds. Each second of the processed audio recording contains 44,000 values. This number of values per second is determined by the sampling rate of the audio recordings made. The choice of the fundamental frequencies as the object of research is determined by the results of the previous research stages.

One of the most famous models of the time series Y_t assumes its representation as the sum of two quantities:

$$Y_t = f(x) + e_t \tag{3}$$

where $f(x)$ is a deterministic component, which includes the entire pattern of the formation of the dynamics of a series under the influence of time, and e_t is a random component that describes all deviations caused by the influence of other factors.

Since in defining synchronicity the emphasis is placed on the dynamics of the development of the series, we will investigate only the deterministic component. Further, since the deterministic component can be described by some selected approximation function, we will consider the transition from assessing the similarity of the values of the series themselves to comparing the coefficients or values of the approximating function.

Consider the formal statement of the problem of assessing the degree of synchronicity. Let there be given two time series $X_t,\ Y_t,\ t = \overline{1,n}$. Suppose that the dynamics of series X_t and Y_t are described, respectively, by the approximating functions $f_1(t)$ and $f_2(t)$ of the form

$$\begin{array}{l} f_1(t) = \sum\limits_{i=1}^{m} a_i \cdot \varphi_i(t) \\ f_2(t) = \sum\limits_{i=1}^{m} b_i \cdot \varphi_i(t) \end{array} \tag{4}$$

where $\varphi_i(t)$ are some elementary functions, $i = \overline{1,m}$, and a_i, b_i are coefficients, $i = \overline{1,m}$.

Then, we can go from studying the similarity of time series X_t and Y_t to studying the similarity of functions $f_1(t)$ and $f_2(t)$ and give a definition of the similarity metric of the time series. According to this, we call the synchronicity extraction metric the quantity.

$$\Delta_{XY}^{an} = \sum_{i=1}^{m} I(|a_i - b_i| > E) \tag{5}$$

where $I(q)$ is the indicator function defined as follows: $I(q) = \begin{cases} 1, q - true; \\ 0, q - false. \end{cases}$, $E > 0$,where q is some condition that may or may not be hold, and then the indicator function argument is set to true or false.

Thus, to assess the degree of synchronicity of the dynamics of time series, the number of coefficients that differ by an amount less than a given critical value is calculated—in the framework of this study, we will call such coefficients "equal". Since the distance is calculated using coefficients, it is necessary to consider the difference in the order of values and the contribution to the total value of the function of the corresponding components. If considering the contribution of the components requires additional information and features of calculating the metric, then the differences in values can be reduced by converting to a single scale using normalization by the formula

$$\widetilde{a}_{ij} = \frac{a_{ij} - c_j}{S_j}, \tag{6}$$

where $\widetilde{a}_{ij}$ is the normalized value of the j-th coefficient in the approximating function for the i-th series, $j = \overline{1,m}$; m is the number of coefficients; a_{ij} is the original j-th coefficient in the function for the i-th series; c_j is the sample mean value of the j-th coefficient for the functions of all considered series; and S_j is the sample standard deviation of the j-th coefficient for the functions of all considered series.

Then, the value E can be regarded as the proportion of the variances of the coefficients of the compared series or the variance of the values of the function. This value is selected from subjective considerations during the study of the compared time series.

For clarity, the value of the metric can be converted into a percentage by dividing its value by the total number of coefficients in the expansion—as a result, the percentage of "equal" coefficients in the total number of coefficients will be obtained. Let us consider the proposed estimate using an example. Again, let there be given two time series $X_t,\ Y_t,\ t = \overline{1,n}$. For the series, approximating functions were obtained, consisting of 14 terms, which were then normalized by Formula (6). Further, to assess the degree of synchronicity of their dynamics, the value of the metric $\Delta_{XY} = 8$ at $E = 0.3$ was calculated; divide 8 by 14 and get that in percentage terms, then the number of "equal" coefficients is 57 percent of the total number of coefficients. Then, we can describe the result of calculating the metric for series X and Y as follows: "the series X and Y are 57% synchronous with a difference threshold equal to 30% of the variance of the coefficients."

If the compared series cannot be qualitatively approximated by functions of form (4), but it is possible to obtain an approximation of the series by functions $f_1(t)$ and $f_2(t)$ in tabular form, that is, to each given moment of time t corresponds the value of the functions $f_1(t)$ and $f_2(t)$, $t = \overline{1,n}$, then in

calculating the metric, (5) can be used as values of a_j and b_j values of approximating functions at given points in time. Then, we get the form of the metric

$$\Delta_{XY}^{tab} = \sum_{t=1}^{n} I\Big(\big|f_1(t) - f_2(t)\big| > E\Big) \tag{7}$$

Further, using the metric, it is possible to simultaneously compare both the expansion coefficients and indicators characterizing the dynamics of the series—for example, the autocorrelation coefficients. Then, to calculate the metric, it is necessary for each series to calculate the values of the compared quantities, normalize them, collect them in one array and carry out the above calculations. Thus, the metric of synchronicity can be considered as a distance, which can consider the coefficients of the time series model, its characteristics or both at the same time.

4. Determining the Similarity of Vocal Performance in Multiple Users

When assessing the degree of similarity of vocal performances of the same melody by different singers, it is not always possible to rely on ear, so the issue arises of a programmatic solution to this problem. To solve it, it is convenient to investigate a time series composed of a sequence of fundamental frequencies (F_0) corresponding to the sung notes. The authors analyzed the modern scientific literature related to the assessment of the similarity of vocal performances [39–44]. Publications related to the assessment of signals from birds and other animals were carried out in order to study the approaches used in them. These signals have a similar structure in comparison to vocal performance, which allows them to be considered as a special form of singing. As a result, it was revealed that correlation analysis is mainly used for this purpose.

The authors of [39] consider the analysis of the similarity of vocal performance by the same adult bird (finch). The finches sang a song composed of repeated syllables that form motifs that form a strumming. After calculating the spectral similarity, hierarchical clustering was applied, and a dendrogram was generated, which shows the similarity between the syllables. To check the similarity, after clustering the syllables, the Pearson correlation coefficient was calculated for each cluster. Their colleagues in the work of [40] studied the effect of noise on bird singing by assessing the similarity of sparrow singing outside the city and in the urban structure. For the purity of the experiment, the records were mixed. To test the hypothesis on the effect of noise on birds singing, the researchers conducted cluster analysis using the k-means method to group sparrows according to their average song characteristics, and as a result, two groups were identified. After that, the residual plots were studied, a correlation test was performed for diagrams with a normal probability plot of residuals and a test for the constancy of error variance was conducted. The similarity of the designs was compared in terms of maximum frequencies and bandwidths.

The study described in article [41] uses methods to quantify spatio-temporal synchronization and causality between populations of insect pests. To determine the relationship between the time series, statistical methods were used: cross-correlation, partial cross-correlation and Granger causality indices. Synchronicity was assessed using the correlation coefficient.

The study described in [42] investigated the similarity of vocalization in domestic mice. For this, the Pearson correlation coefficient was calculated, and the slopes of the regression line were also investigated.

The work of [43] evaluates synchrony and asynchrony based on a smoothing method, a moving average (over a given number of points), where the desired value is obtained by averaging several values directly adjacent to the central value of the current group. Synchronicity is determined visually according to the graph. To determine the degree of synchrony (asynchrony), a study of the parameters was carried out. The purpose of the study was to assess the relationship between the parameters using the "delta" calculation, as a result of which it can be concluded that the parameters change according to a certain criterion. How the delta was calculated, the authors do not indicate.

In addition to research aimed at determining patterns in the sounds that representatives of the fauna emit, there are several works that study human singing. For example, in the work of [44], the assessment of choral singing was carried out using the autocorrelation function and the calculation of non-stationary "correlation portraits" of the sound of the choir and their visual comparison.

The analysis showed that the existing approaches do not allow for a numerical assessment of the degree of similarity of several series. Assessment by the synchronicity metric allows one to set, among other things, a threshold value for comparing series. The synchronicity detection metric allows one to consider the dynamics of the series. In order to test the applicability of the metric to the analysis of vocal performances, it was decided to conduct an experiment comparing the vocal performances of several speakers.

To compare the similarity of vocal performances, a standard was chosen—the performance of the melody by a person with musical education and seven versions of performances by other people who listened to the audio recording of the standard and tried to reproduce it. All performers were women aged 22 to 28. In total, eight locations were recorded—four types of melodies were performed with smooth and abrupt sounds.

To determine the F_0 of vocal performances, the program "Amadeus" [45] was chosen, which determines the frequencies of the main tone of vocal performances and on the basis of which software for singing training will be developed. This program uses a fundamental frequency identification algorithm based on a model of the human auditory system [46].

Since that, in the process of applying the method of extracting synchronicity, it was not the F_0 values themselves that were compared, but rather the coefficients or the values of the functions approximating them, it was necessary to choose an approximation method and make an estimate. For greater versatility, a group of nonparametric statistics methods was chosen, since the true form of the approximating function is unknown, and it cannot be parameterized due to strong non-periodic fluctuations in values. As a result, due to its simplicity and positive results, one of the most famous nonparametric estimates, the Nadaraya–Watson regression, was chosen.

The problem of approximating the time series is posed as follows. A number of F_0 values Y_t, where $t = \overline{1, n}$, are times. The true form of the dependence of the F_0 values on time is unknown. It is required to construct a function $f(t)$ that approximates the unknown dependence.

For the calculation, it was decided to use the Nadaraya–Watson estimator [47]. This estimator is the most famous among the nonparametric estimates and is carried out according to the formula

$$f(x, h_n) = \frac{\sum_{i=1}^{n} Y_i \cdot K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)}, \tag{8}$$

where K is a non-increasing, smooth, bounded function called the kernel; and h_n is the blur parameter.

Thus, it was decided to use the most known nuclear functions to solve the problem. The work uses six nuclear functions:

1. Triangular:

$$\begin{cases} 1 - |u|, & |u| \le 1 \\ 0, & |u| > 1 \end{cases} \tag{9}$$

2. Optimal Epanechnikov:

$$\begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{3}\right), & |u| \le 0.5 \\ 0, & |u| > \sqrt{5} \end{cases} \tag{10}$$

3. Fisher:

$$\frac{1}{2\pi} \frac{\sin \frac{u}{2}}{\frac{u}{2}} \tag{11}$$

4. Vallée-Poussin:

$$\frac{1}{2\pi}\left(\frac{\sin\frac{u}{2}}{\frac{u}{2}}\right)^2 \tag{12}$$

5. Trisquare:

$$\frac{35}{36}\left(1-|u|^2\right)^3,\ |u|\le 1 \tag{13}$$

6. Tricubic:

$$\frac{70}{81}\left(1-|u|^3\right)^3,\ |u|\le 1 \tag{14}$$

The estimation of the accuracy of the obtained values will be carried out by calculating the absolute and relative average errors of the approximation.

For the mean absolute approximation error, the following formula is used [48]:

$$E_{abs} = \frac{\sum_{i=1}^{n}\left|Y_t - f(t)\right|}{n} \tag{15}$$

where Y_t is the value of the original time series at time t; $f(t)$ is the time series value obtained after approximation; and n is the length of the time series.

To calculate the average relative error of approximation, the following formula is used:

$$E_{abs} = \frac{\sum_{i=1}^{n}\left|Y_t - f(t)\right|}{n\cdot\overline{Y}} \tag{16}$$

where $\overline{Y}$ is the mean of the original time series.

As a result, the software implementation of the algorithm for assessing the similarity of vocal performances was performed (Figure 2).

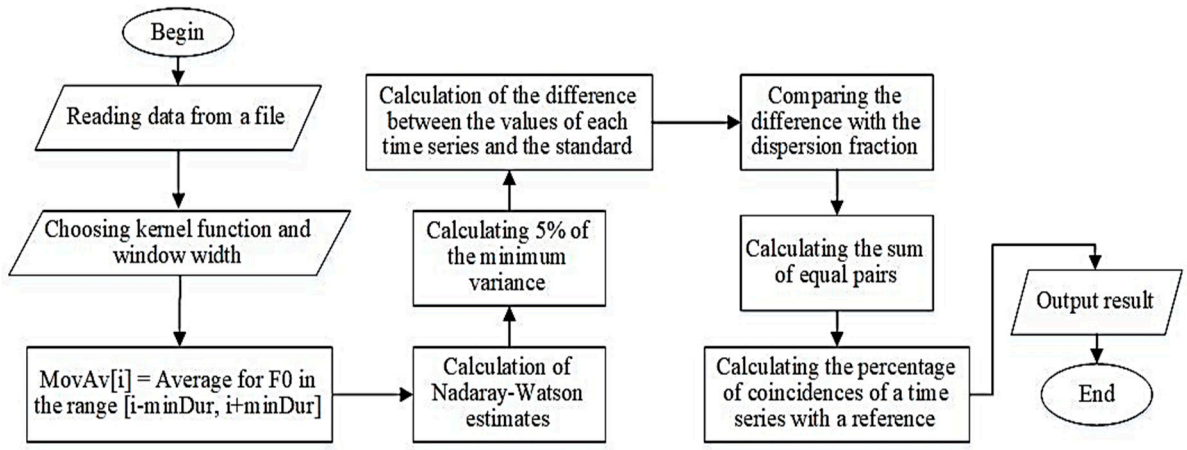

Figure 2. Algorithm for assessing the similarity of vocal performances.

The developed algorithm for assessing the similarity of vocal performances, based on the use of the synchronicity extraction metric, contains two stages. At the first stage, the reference and estimated frequency series are approximated; the second stage consists of comparing each value of the approximated estimated series with the corresponding value of the approximated reference series, determining the number of equal pairs and calculating the proportion of coincidences of the time series with the reference. Previously, this algorithm was used to analyze meteorological data [49].

When investigating the error values choosing different kernels, the following recommendations were made: for triangular, trisquare and tricubic kernels, the window width should be 2. For the rest of the kernels, the window width is 1. The minimum errors were obtained with the Vallée-Poussin kernel and were less than 5%.

Table 1 shows the results of assessing the similarity of vocal performances of each location by each performer. As can be seen from the results obtained, locations 1, 2, 7 and 8 are most accurately executed—the degree of similarity is over 90%, and the results are almost the same for all types of kernels. The differences in the choice of kernels are especially noticeable for evaluating the performance of locations 4 and 6—the use of the Vallée-Poussin and Fisher kernels noticeably increases the degree of similarity for all performers. The best core results for each location sung by the performer are highlighted in bold text.

Table 1. Results of assessments of the similarity of vocal performances.

Location	Kernel	Perf.1	Perf.2	Perf.3	Perf.4	Perf.5	Perf.6	Perf.7
1	Triangular	99.87%	100%	99.74%	99.74%	100%	100%	99.61%
	Epanechnikov	99.61%	100%	99.08%	98.56%	100%	100%	99.61%
	Fisher	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Vallée-Poussin	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Trisquare	99.87%	100%	99.87%	99.87%	100%	100%	99.74%
	Tricubic	99.87%	100%	99.61%	99.61%	100%	100%	99.61%
2	Triangular	99.17%	96.99%	96.83%	96.83%	97.33%	97.997%	98.16%
	Epanechnikov	96.66%	93.49%	90.82%	92.32%	93.82%	95.16%	95.16%
	Fisher	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Vallée-Poussin	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Trisquare	99.33%	97.99%	97.33%	98.49%	98.498%	98.83%	98.83%
	Tricubic	98.33%	95.66%	95.993%	96.49%	95.83%	97.66%	97.496%
3	Triangular	76.53%	82.05%	70.81%	73.57%	74.36%	81.26%	79.68%
	Epanechnikov	71.2%	78.69%	66.27%	70.02%	70.81%	80.47%	76.33%
	Fisher	91.32%	93.49%	88.17%	92.5%	91.32%	95.86%	91.72%
	Vallée-Poussin	**98.42%**	**98.03%**	**95.46%**	**97.24%**	**96.06%**	**98.62%**	**97.44%**
	Trisquare	77.32%	82.84%	71.99%	75.15%	75.54%	82.25%	79.49%
	Tricubic	77.32%	82.84%	71.99%	75.15%	75.54%	82.25%	79.49%
4	Triangular	44.11%	45.02%	35.95%	24.47%	45.32%	51.06%	32.93%
	Epanechnikov	39.58%	36.25%	32.93%	17.52%	38.67%	47.73%	23.57%
	Fisher	73.41%	72.81%	71.9%	68.88%	74.02%	80.36%	74.62%
	Vallée-Poussin	**83.99%**	**79.15%**	**83.69%**	**76.13%**	**80.97%**	**88.52%**	**83.99%**
	Trisquare	45.92%	47.13%	38.97%	26.59%	48.04%	54.38%	34.74%
	Tricubic	46.53%	41.69%	35.05%	23.26%	45.02%	51.66%	29.3%
5	Triangular	79.35%	79.92%	81.64%	73.42%	87.19%	85.85%	77.25%
	Epanechnikov	72.66%	69.79%	77.63%	65.01%	86.81%	82.6%	69.22%
	Fisher	98.47%	97.32%	97.13%	97.51%	99.24%	99.43%	98.85%
	Vallée-Poussin	**99.23%**	**100%**	**98.47%**	**99.81%**	**100%**	**100%**	**100%**
	Trisquare	80.31%	81.45%	82.22%	77.44%	88.34%	85.85%	80.11%
	Tricubic	78.39%	77.63%	80.49%	72.66%	86.62%	85.09%	76.48%
6	Triangular	29.91%	21.99%	29.03%	13.49%	30.49%	32.55%	20.82%
	Epanechnikov	25.22%	18.48%	22.29%	9.09%	28.45%	29.33%	18.18%
	Fisher	62.76%	60.41%	63.34%	43.99%	73.31%	68.33%	63.05%
	Vallée-Poussin	**77.71%**	**77.13%**	**73.31%**	**68.04%**	**80.35%**	**80.35%**	**78.59%**
	Trisquare	31.38%	23.46%	28.45%	13.19%	31.67%	35.48%	21.7%
	Tricubic	29.32%	20.23%	28.15%	12.02%	32.55%	34.89%	20.23%

Table 1. *Cont.*

Location	Kernel	Perf.1	Perf.2	Perf.3	Perf.4	Perf.5	Perf.6	Perf.7
7	Triangular	99.89%	**100%**	**100%**	99.56%	99.56%	99.56%	99.67%
	Epanechnikov	99.89%	**100%**	**100%**	99.67%	99.56%	99.56%	**100%**
	Fisher	100%	**100%**	**100%**	**99.89%**	**100%**	**99.89%**	**100%**
	Vallée-Poussin	100%	**100%**	**100%**	**99.89%**	**100%**	**99.89%**	**100%**
	Trisquare	100%	**100%**	**100%**	99.67%	99.56%	99.56%	99.67%
	Tricubic	99.89%	**100%**	**100%**	99.45%	99.56%	99.56%	99.67%
8	Triangular	99.87%	99.6%	99.87%	**100%**	99.87%	99.87%	99.87%
	Epanechnikov	**100%**	99.47%	**100%**	**100%**	**100%**	**100%**	99.6%
	Fisher	99.87%	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Vallée-Poussin	99.87%	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	Trisquare	99.87%	99.74%	**100%**	**100%**	99.87%	**100%**	99.87%
	Tricubic	**100%**	99.6%	99.87%	**100%**	99.74%	99.74%	99.87%

When listening to the performances, the fourth and sixth locations are indeed performed worse by almost all singers, but their degree of similarity is quite high—more than 50%. Considering that the accuracy of the assessment with the Vallée-Poussin kernel is higher than that of the others, the results obtained show that this method of assessing the degree of similarity using Vallée-Poussin or Fischer kernels can be further developed and used to determine the similarity of vocal performances.

One of the options for using this method of assessing similarity can be to check the quality of the task of performing a sequence of notes when teaching vocals. As indicated in [50], hitting a note depends on the pitch of the voice, which is the sum of the physiological characteristics of a person [51]. It is because of these features that a person may not fall into a certain octave of a note. In the initial stages of learning to sing, this should not be taken as a mistake, since the note is sung correctly. Therefore, at the initial levels of training, it is necessary to assess the dynamics of F_0.

5. Vibrato and Glissando Identification in Vocal Performance

5.1. Approach to Vocal Performance Analysis

The tests of the algorithm showed that when processing audio recordings, including various approaches to playing notes in the range from 70 to 800 Hz, the software complex allows you to recognize more than 95% of the notes sung by the speaker. It was determined that when singing arpeggio, crescendo and decrescendo, there is no effect on the quality of the algorithms in the software. The lack of influence on the quality of note identification lies in the fact that the algorithm considers only the frequency of the fundamental tone in the analysis and the specifics of the operation of one of the stages of the note identification algorithm, which is responsible for determining the belonging of a voiced section to a note, lies in the focus on pure performance.

It is quite difficult to play a note so that all frequencies at the moment of singing are within the band allocated for it. For this reason, adjacent notes above and below the played one are considered. In Figures 3–5, the frequency spectrum is divided into sections corresponding to the boundaries of the notes. Each area is highlighted with a horizontal bar in the background of the frequency graph. The *X*-axis is time, and the *Y*-axis is the fundamental frequencies.

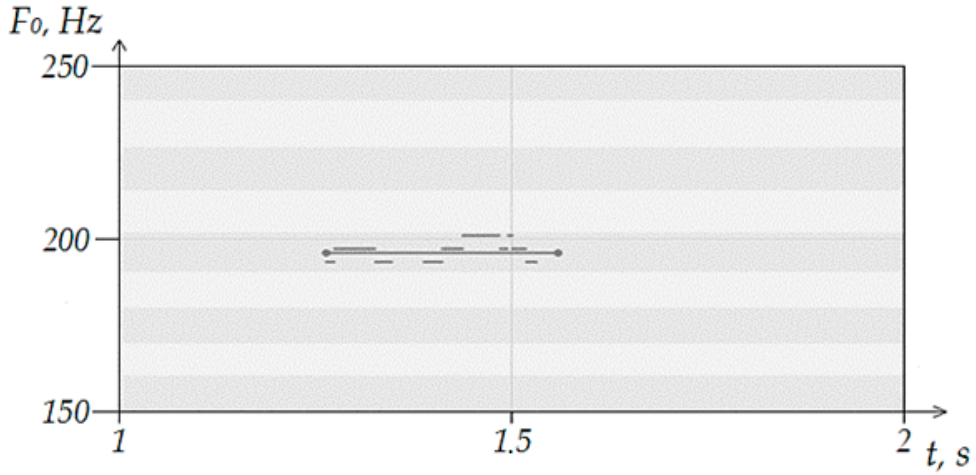

Figure 3. Segment with singing within one note.

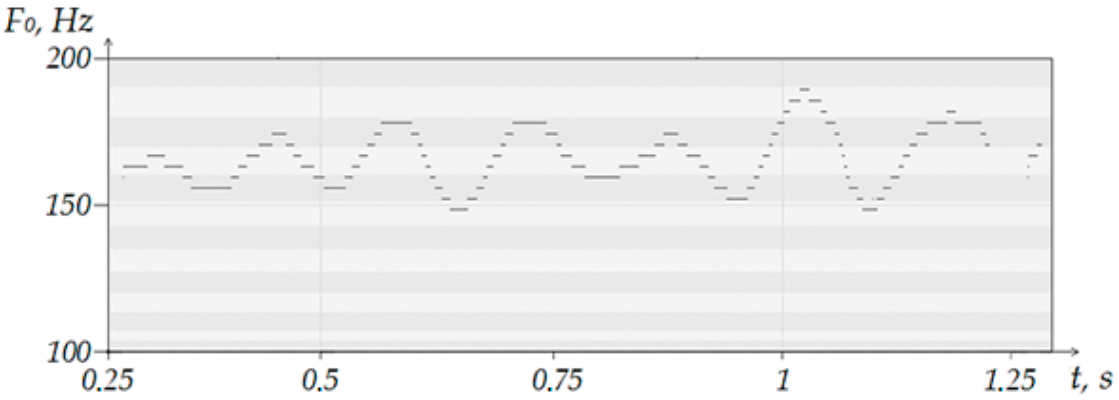

Figure 4. Segment with a note sung with vibrato in the voice.

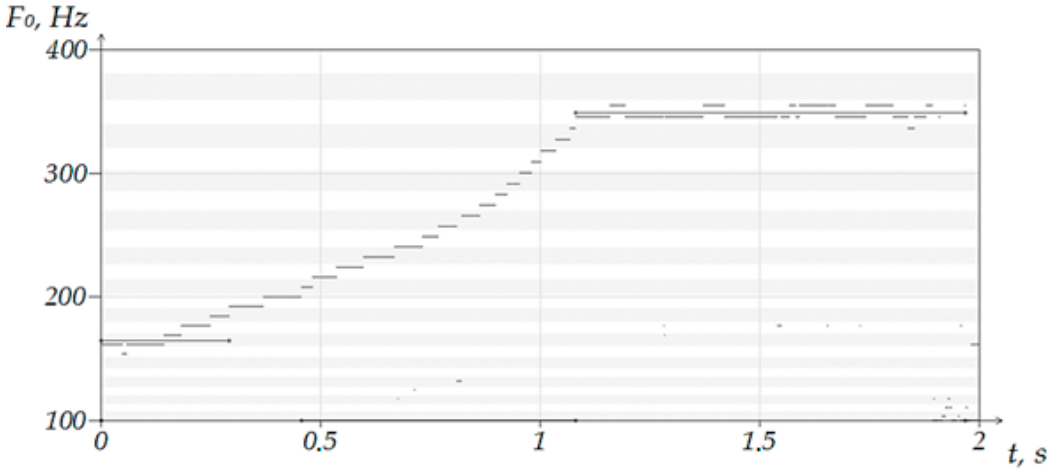

Figure 5. Segment sung with an ascending glissando.

As you can see in Figure 3, a segment sung within one note was recognized by the program. The number of moments found in adjacent notes is insignificant and did not affect the identification result. The overwhelming majority of recognized fundamental frequencies were in the range of the sung note. Considering the requirements for the accuracy of execution prescribed in the algorithm, the program was able to draw a conclusion about the sung note. If the number of fragments in each of the three sections turned out to be approximately the same, the algorithm perceives the entire voiced segment as noise.

When singing vibrato, fluctuations occur in the frequency of the sound relative to a certain main note within a semitone. It should be noted that there are varieties of vibrato-like vibrations. This includes tremolation and voice swing. Due to the fact that the algorithm is set to detect clean notes, such segments are perceived as noise. As can be seen, the fluctuations during the performance of the main note (in this case, the note "minor octave E") cover four notes at once. Considering the minimum duration of the sounding of a note leads to the fact that for each of the four notes, the singing within its boundaries took an insufficient amount of time. In this case, the return to this note occurs after the transition to other notes, which reduces the proportion of frequencies related to the main note.

Another example of adding a coloristic effect to singing is such a technique as glissando. As is known from music theory, when singing a glissando, there is a smooth glide from one note to another. With this singing, the transition occurs too quickly in order to be able to identify each individual note at the moment of sliding, since there is less than 0.1 s for each segment of the covered notes. Glissando can be either ascending (as shown in Figure 5) or descending. As can be seen in the figure, the algorithm is able to determine the start and end notes between which it is sliding. However, the coverage of 12 notes in between is perceived as noise.

Thus, it becomes necessary to identify areas perceived by the program as noise. We will assume that a directed transition from one note to another is a glissando, and oscillations relative to one note within a single segment are tremolation. Determining the type of tremolation will not be considered at this stage.

Since the array of found fundamental frequencies, within which it was not possible to identify a pure note, can be perceived as a time series, it was decided to conduct a preliminary analysis of the data. Among the preliminary analysis procedures, there are anomaly search, time series smoothing, trend checking and calculation of process dynamics indicators.

From the point of view of the considered subject area, anomalous observations can be perceived as bursts of frequencies determined outside the sound of the main melody. This includes any noise that the algorithms filter out during the note segmentation stage. Accordingly, this procedure can be considered completed.

As part of smoothing the time series, the true levels of the series are computed by the calculated values that have smoother dynamics than the initial data. It was decided to focus on the methods of mechanical smoothing, since in the problem under study it is necessary to evaluate each individual leveling of the series considering the actual values of the levels adjacent to it. The weighted moving average method is inapplicable for the problem under study, since it cannot contain a quadratic or cubic trend. Further, there is no need to use the exponential smoothing method, since it is used in the problems of predicting the development of the process after the study area. In this regard, the simple moving average method will be used according to the formula

$$Y_t = \frac{1}{m} \sum_{i=t-p}^{t+p} Y_t, \tag{17}$$

where m is the number of observations included in the smoothing interval; and p is the number of observations on opposite sides from the smoothed.

As the number of observations p included in the smoothing interval, it was decided to use the value of the minimum duration of the sounding of a note, used in the algorithm for the segmentation and identification of notes. The choice is due to the fact that when singing vibrato and glissando, the duration of singing within such sections is significantly higher than this parameter, but the identification of individual clean notes was not carried out by the basic algorithm. The disadvantage of the method is that the first and last p observations will remain unsmoothed, which can be neglected due to the total duration of the estimated area.

Testing for the presence of a trend is essentially a test of the hypothesis that the mean value of the time series remains unchanged. For testing, a series can be divided into n parts, each of which

is considered as a separate set. The resulting averages for each population will be compared. If the average values increase or decrease, we will assume that the segment under study may contain a glissando (ascending or descending, respectively).

All preliminary calculations were carried out in Microsoft Excel. The built-in functionality contains the "Data Analysis" function, in which a moving average can be plotted for the selected series.

Figure 6 shows the result of applying this method. In the built-in method, only the previous values are analyzed, without considering the following after the estimated one, which makes it possible to obtain an envelope curve that is uninformative within the framework of the study. In order to solve this problem, the classical formula of sliding smoothing was used, considering the measurements before and after the smoothing one, with 0.5 * p values preceding the smoothing one, and the same number of subsequent values was included in the smoothing interval (Figure 6). As a result, at the interval from 0 to 0.5 * p, the values of the series will coincide, but starting from the next value, we will be able to see the averaging. The solid line indicates the values of the analyzed fundamental frequencies, and the dashed oscillating line indicates the smoothed values.

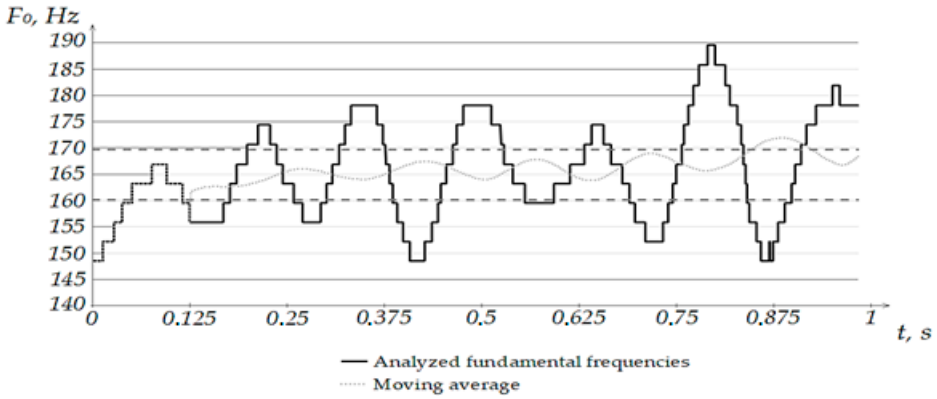

Figure 6. Moving average for a vibrato-like area.

In the figure presented, for the note "minor octave E", parallel dashed lines indicate the lower (160.121 Hz) and upper (169.643 Hz) sounding limits, determined using the logarithmic average. Note that in the studied segment, there are seven oscillations in the frequency of the fundamental tone, judging by the graph. On further investigation of the type of vibration in singing, this aspect can be used as a criterion for distinguishing vibrato from tremolation.

As mentioned earlier, the presence of a trend is determined by comparing the obtained mean values for each of the sets of the studied series. Further, by means of Excel, on the graph of the fundamental frequencies for the series, one can set the construction of a linear trend (Figure 7). At the stage of analyzing the applicability of the method to the problem under study, we will use the built-in function in Excel, and at the stage of implementation in the software package, we will compare the average values for segments of the studied range.

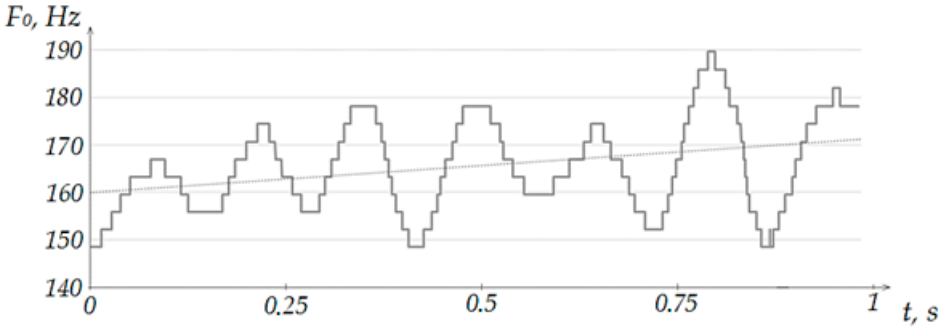

Figure 7. Linear trend for a vibrato-like area.

As can be seen from the figure above, the linear trend allows one to determine that in most of the surveyed area, the average frequency values refer to one note. However, it is impossible to assess the presence of fluctuations in the fundamental frequencies from a linear trend. In this regard, it is not possible to use only a linear trend for assessing areas with tremolation. The trend allows you to estimate the number of notes covered in the study area, but not the type of performance. In turn, the moving average simultaneously interprets both the presence of vibrations and the covered notes, which makes this estimation method more preferable for vibrato-like areas.

To determine the glissando in the studied area, one must complete another task. The basic algorithm for the segmentation and identification of notes determines two notes in such situations: the one with which the slip began, and the one at which the slip ended. Accordingly, the first task to be solved to determine the presence of glissando is to compare notes at the boundaries of the area under study. If the notes turn out to be identical, then the area between them cannot correspond to the glissando and should be considered for attribution to vibrato. In other cases, we will observe a trend line passing through several notes. The greater the difference between the notes and the faster the transition between them, the greater the angle of the resulting trend line relative to the time axis.

As for the tremolation section, consider the moving average for the range under study. As can be seen in Figure 8, there are no obvious fluctuations in the frequency of the main tone in the transition from the note "minor octave G" to the note "1st octave E". The solid line on the graph indicates the fundamental frequencies, and the dashed line indicates the moving average for the range under study. Parallel dotted lines correspond to note boundaries. The resulting moving average line goes through nine notes. Thus, the data obtained indicate an ascending glissando. Below (Figure 9) is a graph for a linear trend obtained on the studied range. In fact, the linear trend for the range under study is close to the result obtained when plotting the moving average.

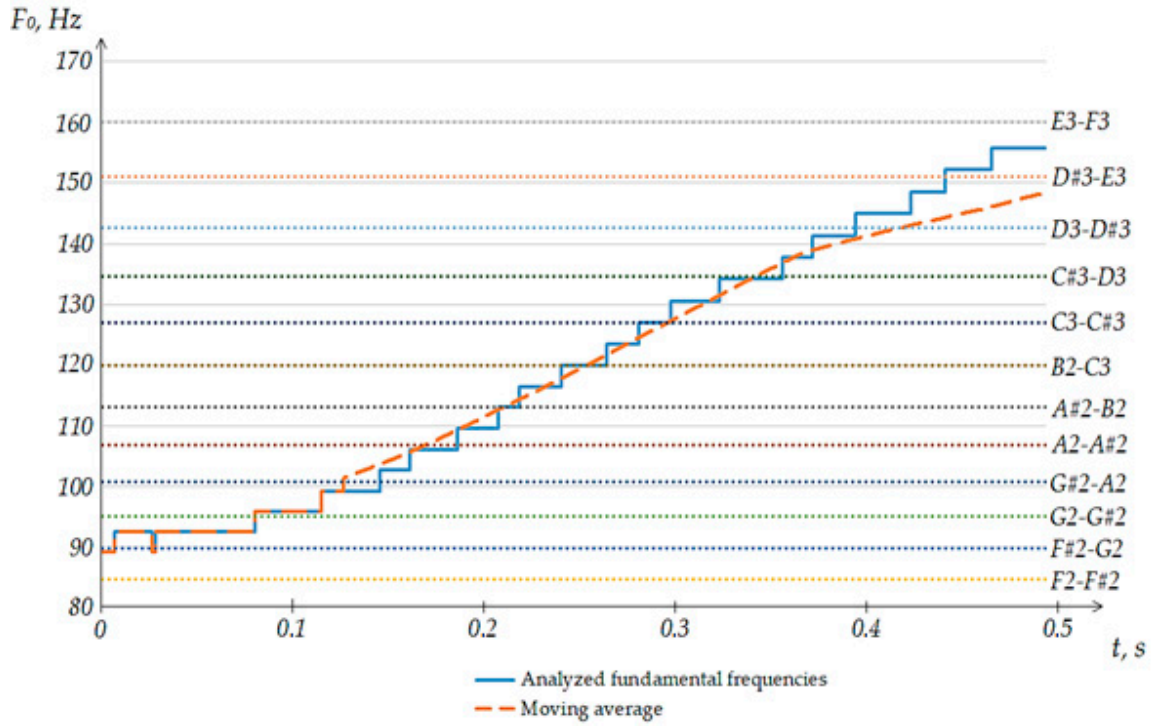

Figure 8. Moving average for a section with an ascending glissando.

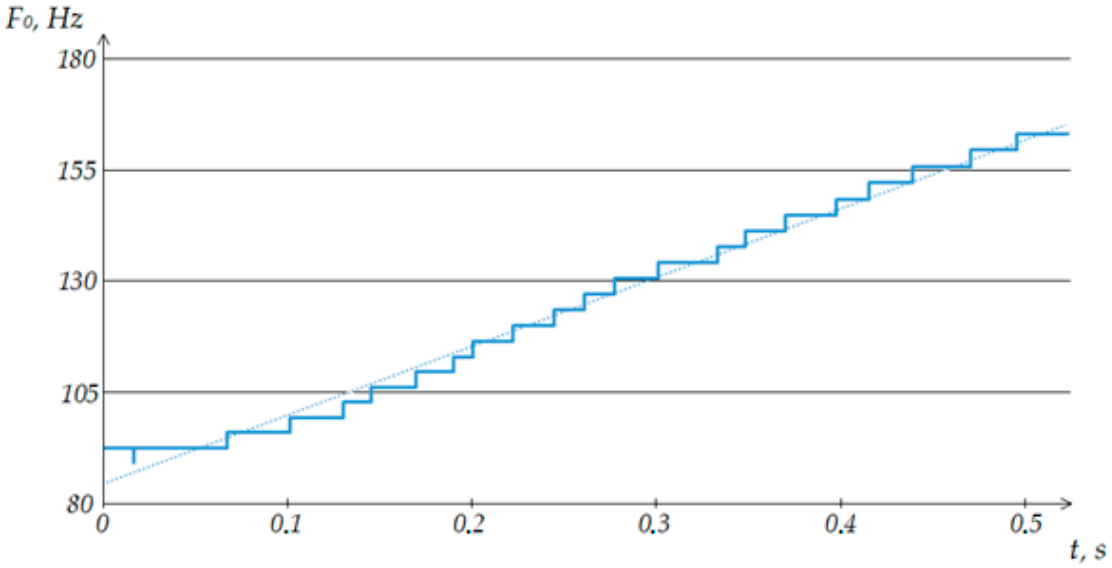

Figure 9. Linear trend for a segment with an ascending glissando.

As can be seen from the results obtained, when using a moving average for areas with tremolation, an oscillation effect is observed that is absent in the linear trend for this area. On the other hand, for the glissando, the moving average and the linear trend behave in a similar way. This allows us to apply both methods in an integrated manner to determine the current situation.

The proposed idea is to obtain both estimates for the segment under study. The resulting datasets will be compared for the degree of similarity between their estimates for a certain parameter N. The size of the value of this parameter requires a separate experiment. Within the framework of this experiment, mixed situations should be considered, in which the time series behave ambiguously, in order to reduce the number of errors of the first and second kinds when determining the type of the studied fundamental frequency ranges.

If there was a glissando in the singing, the results of the moving average and linear trend for the given time series will be close. For this situation, it will be necessary to check for the number of notes through which the transition was made, and the direction of the transition (ascending or descending glissando). As a result, it will be possible to filter out areas with noise between adjacent segments corresponding to one note and areas with a burst of noise outside the range between notes.

For situations in which the difference in the estimates of the time series will exceed the value of the criterion, we will assume that there was singing with vibrato in the voice. As noted earlier, the moving average has a wave-like structure in the case of fluctuations in the studied area. Counting the number of oscillations per unit of time can be used in the classification problem of the tremolation type. This parameter will be useful when teaching vocal performance in case of too frequent or rare fluctuations in the performance of a note. In addition, in the case of detecting vibrato in singing, it will also be necessary to control the number of notes through which the trend passed. As can be seen in Figure 10, the oscillations for the sung note slide downward. In fact, in this example, glissando and vibrato are mixed. However, this section cannot be attributed to any of the types for two reasons: the presence of vibrations in the voice is not typical for glissando, and for vibrato, there should be no slips to another note.

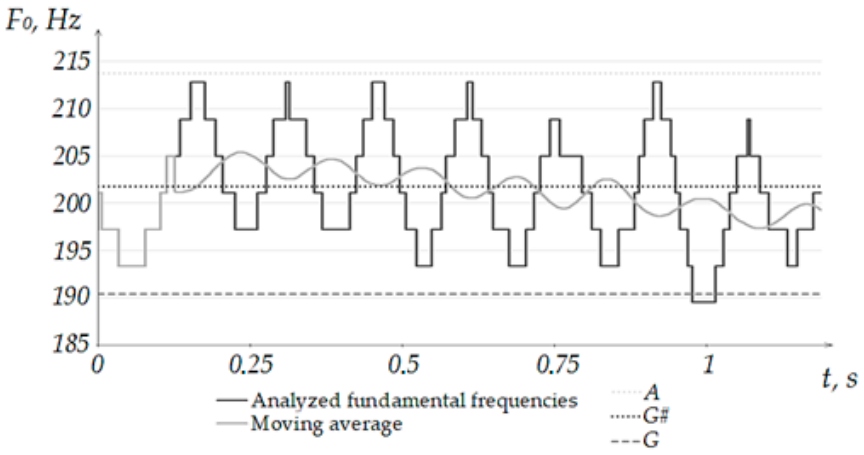

Figure 10. Moving average for the combined area.

5.2. Description of the Collected Audio Database

As a material for calculating the values of the moving average and linear trend, audio recordings with vocal performances by students of a music studio were considered. Each student was given the task to listen to the reference recording and make five recordings with the singing of the assigned task. In total, 17 master records and 740 student records were collected, containing a total of 13,078 sung notes. The database contains recordings with different ranges of notes, sung by both male and female voices of different ages. The duration of the audio recordings varies from 2 to 20 s, depending on the exercise performed by the speakers.

Each of the collected audio recordings has the following parameters:

- Extension: wav;
- Sampling frequency: 44 kHz;
- Checksum: 16 bits;
- Channel: stereo.

Each of the audio recordings was segmented into separate notes and sections, within which the notes were not identified. Sections with vibrato-like or glissando singing were separately identified. Each recording was rated by vocal teachers from the music school. The consistency in expert estimates for the audio recordings was checked using the Kendall coefficient. Further evaluation was carried out in accordance with the exhibited expert assessments.

Figure 11 shows an example with four highlighted areas corresponding to singing with vibrato in the voice. As can be seen, in Section 1, fluctuations occur within one note (Figure 12), which made it possible to identify the note. A similar situation is observed in the fourth section.

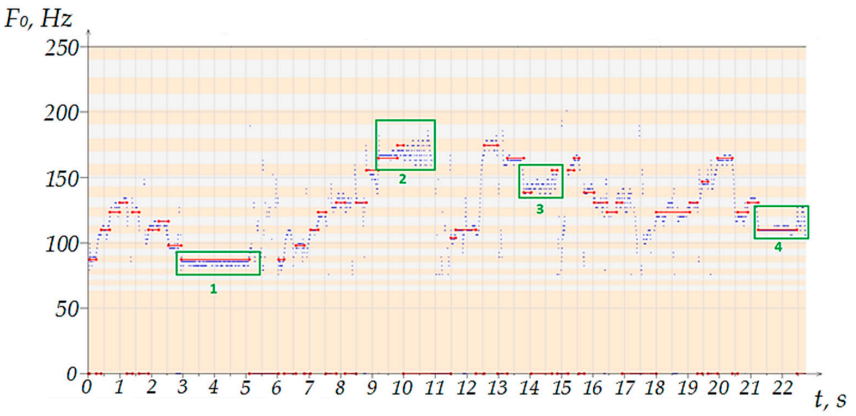

Figure 11. Frequency domain with recognized notes for the 4th test recording sung by speaker №9.

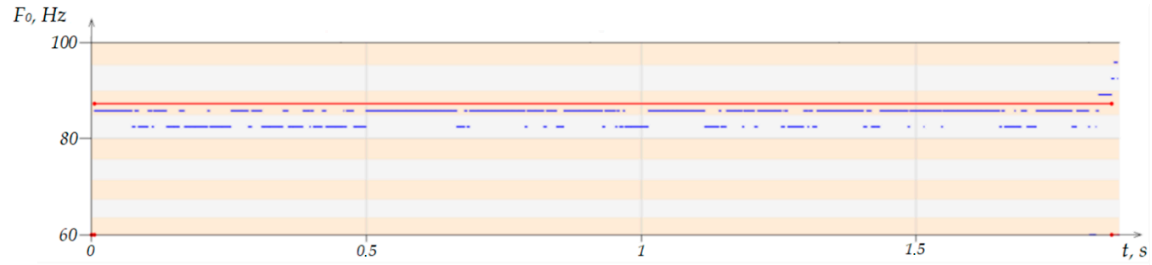

Figure 12. Segment №1 of the 4th investigated audio recording sung by speaker №9.

Despite the fact that the presence of vibrations in the voice did not affect the accuracy of recognizing notes in the situations considered, it was decided to re-examine such areas in order to find possible distinguishing features that would allow us to classify similar segments. This will help to not only more accurately recognize notes with fluctuations, but also consider the intentional transitions to adjacent notes when assessing the quality of singing. In addition, similar segments will not be overlooked in the analysis phase of the vibrato-like performance type.

Figure 13 shows an example with segmented sections containing an upward glissando. As can be seen for segments 1 and 3, the note recognition algorithm was able to identify the initial and final values of the notes between which the slip occurs, and the frequencies between them are perceived as noise. For segment №2, the starting moment of the ascending glissando is also the final one for the descending one from the previous note.

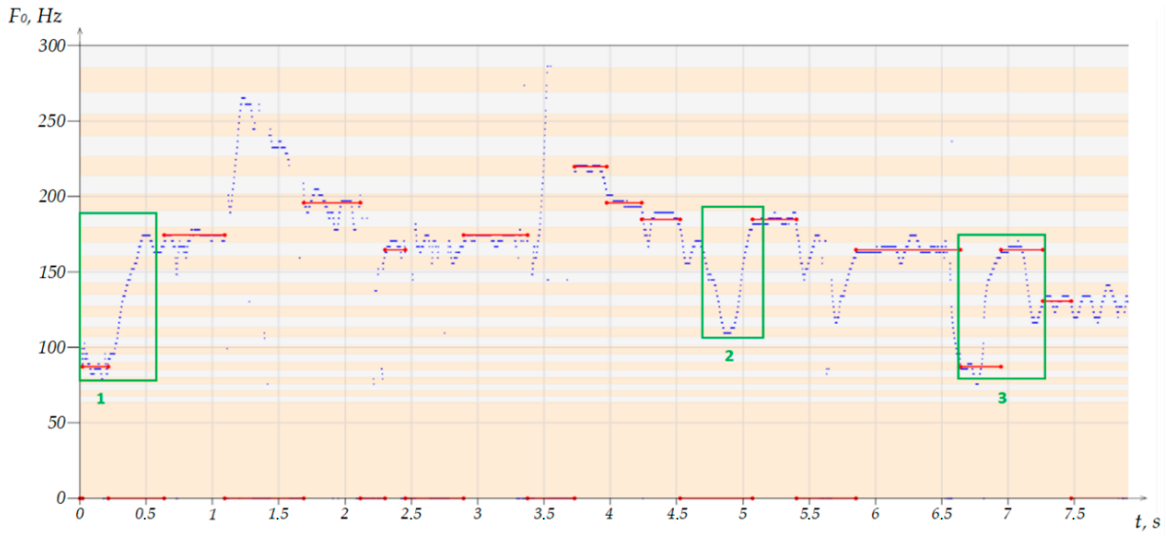

Figure 13. Frequency domain with recognized notes for the 8th test recording sung by speaker №5.

A total of 740 audio recordings were segmented into 149 vibrato-like singing sections and 68 glissando sections, of which 54 were ascending and 14 were descending. The duration of the audio recordings varies from 0.5 to 3 s. For each audio recording, the values of the fundamental frequencies were obtained, which were preliminary estimated manually in MS Excel using the functions described above.

5.3. Analysis of Audio with Vibrato and Glissando Singing

Within the framework of the software package [52], a module was developed that is responsible for the analysis of the selected fundamental frequency range. The algorithm of the module is shown in Figure 14. The module receives an array with the fundamental frequencies of a segment that is not recognized by the note recognition algorithm and is considered for the presence of tremolation

or glissando transition between notes. In addition, it is necessary to inform the module about the minimum duration of the sounding of notes, corresponding to the considered audio recording.

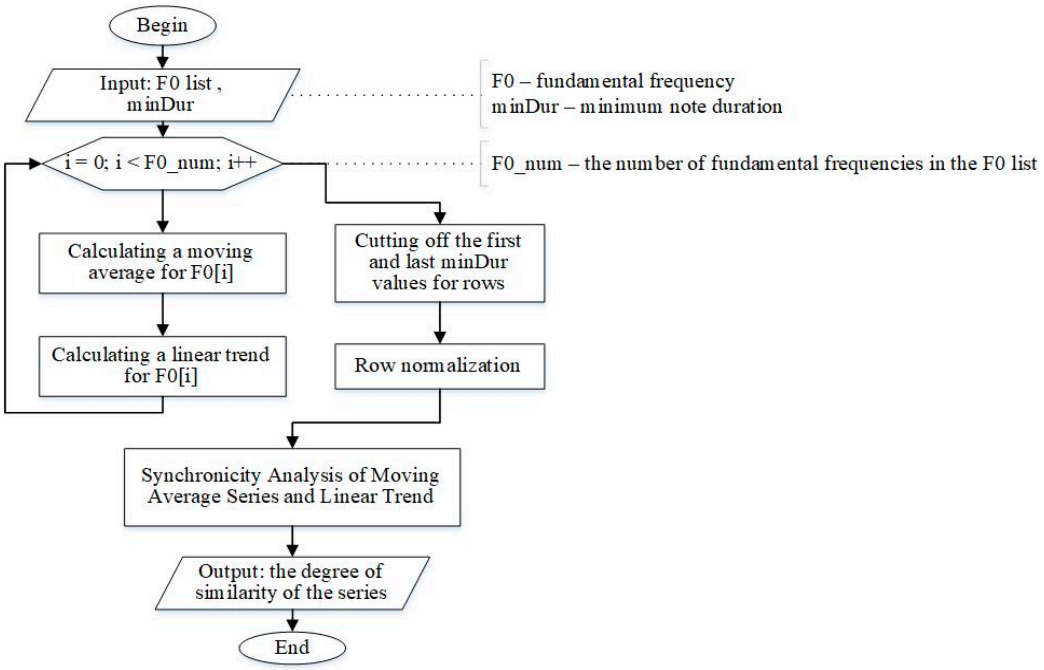

Figure 14. Algorithm of the module for analyzing segments with a sharp change in the fundamental frequencies.

For the investigated array of frequencies, using Formula (17), an array is formed with the values of the moving average at each moment of time in the studied range. When calculating the values of the linear trend at each time point, the least squares method was used.

Since the initial and final p observations of the moving average remained equal to the initial values of the fundamental frequencies, the series of the moving average and linear trend were cut off at the first and last p values. The remaining values were normalized and processed using the synchronicity extraction metric. The evaluation of the similarity of time series by the metric is launched by calling the subroutine. For each selected audio recording, the fundamental frequencies were calculated and analyzed using the developed module.

5.4. Audio Processing Results

A fragment of the experimental results is presented in Table 2.

Table 2. Fragment of the results of the synchronicity detection experiment.

Record Number	Segmented from	Singing Type	FF	Degree of Difference
001	Performer1_Song1.wav	vibrato	31.985	67.78%
002	Performer1_Song1.wav	vibrato	41.473	73.2%
003	Performer1_Song1.wav	vibrato	47.265	68.36%
004	Performer1_Song1.wav	mixed	36.071	48.7%
005	Performer2_Song1.wav	mixed	48.217	49.9%
006	Performer2_Song1.wav	vibrato	32.969	80.4%
007	Performer2_Song1.wav	vibrato	32.170	68.16%
008	Performer1_Song2.wav	vibrato	47.424	81.44%
009	Performer2_Song2.wav	glissando	59.169	15.04%

Analysis of 217 audio recordings showed that glissando has a 5 to 15% synchronicity rate between a linear trend and a moving average, while vibrato singing has a range of 65 to 85%. At the same time, the cleaner the vibrato singing, the higher the percentage of discrepancy between the series. For glissando, on the contrary, the lower the percentage, the less side frequencies there were when switching to singing. In addition, 67 recordings with a combination of vibrato and glissando in singing were processed. The mixed types of singing described before Figure 10 are perceived by the method as 45–50% different.

Figure 15 shows an example in which a mixture of vibrato in singing with a transition to a lower note leads to the impossibility of identifying separately each of the applied singing methods. The presence of a trend towards a decrease or increase in the overall sound of a segment in the frequency domain leads to a smoother change in the moving average for the area under study.

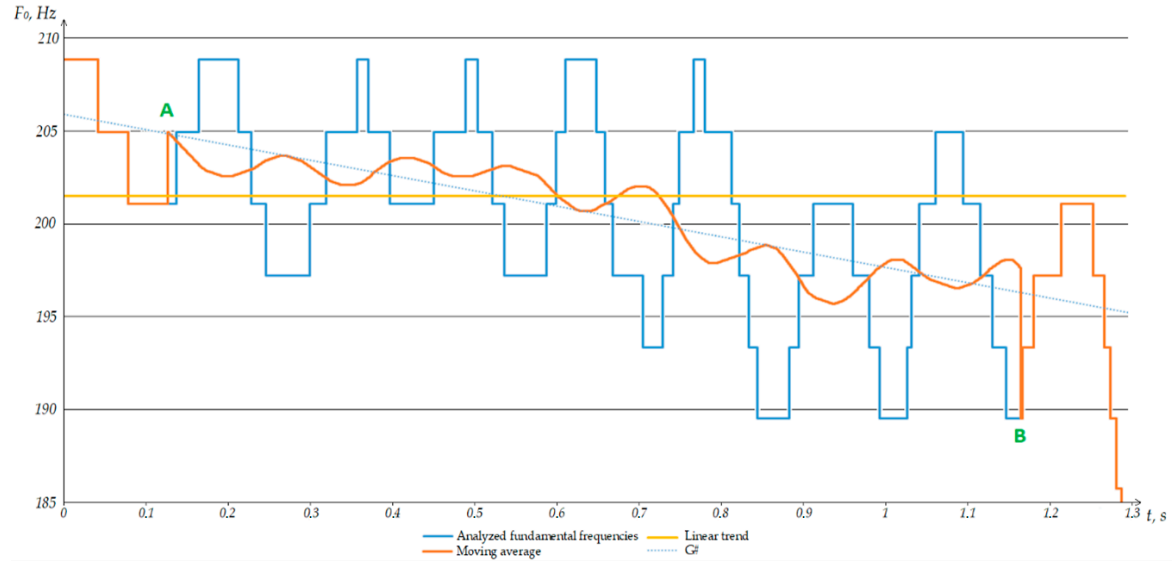

Figure 15. Graph with an unidentifiable segment.

The results obtained make it possible to automate the assessment of areas that were not recognized by the basic algorithm for recognizing notes in vocal performance. By dividing the range of accepted results from the method of separating the synchronicity of time series into sections, it is possible to unambiguously classify areas with a sharp change in the fundamental frequency. If the difference is no more than 15%, one can assume that there is a glissando in the singing of the treated area. With estimates in the range from 65% to 85%, we will assume that the studied segment was sung with vibrato-like vibrations. Intermediate values (from 16 to 64%) will be perceived as noise.

6. Conclusions

In the course of this study, a metric for assessing the degree of similarity of time series was developed and applied. This metric makes it possible to consider the assumptions about the model of the series and to compare the values of the corresponding characteristics. The sphere of speech technologies was chosen as an area of application of this approach for assessing the similarity of time series. Tests were carried out with the values of the fundamental frequencies of vocal performances in several directions.

One of the directions for assessing the similarity of vocal performances was the comparison of the functions of the dependence of fundamental frequencies on time for several speakers. As a test, one speaker was given the task to sing a sequence of notes, and the rest of the speakers were given the task to reproduce the melody they heard. It was determined that the use of the synchronicity allocation metric allows for an exercise in which students need to repeat the melody after the teacher.

This teaching model is the closest to that implemented in the framework of teaching singing in music schools.

Another application of the synchronicity extraction metric is the implementation of an algorithm for identifying segments with a sharp change in fundamental frequencies. The tests of the software complex showed that the note recognition algorithm is able to identify up to 95% of the notes sung by the speaker. The exceptions are notes played with glissando and vibrato. The basic algorithm of the program considered only clean notes, which did not allow identifying vocal performances with a sharp change in frequencies. For the analysis of unrecognized areas, the synchronicity detection metric was applied to the time series estimates. It was determined that glissando is characterized by a high degree of similarity for the linear trend and the moving average for the study area, while the opposite is true for vibrato. For vocal performances with tremolation, differences from 65 to 85% are characteristic.

The results obtained show that the use of the synchronicity extraction metric in the analysis of speech signals allows one to determine the similarity of both several speech signals and estimates of the behavior of the fundamental frequency for one recording. This will allow the described metric to be considered in other studies. One of the applications can be the use of the metric in the task of identifying a speaker. In this direction, it is possible to evaluate the similarities of the reference recording with the speaker's voice to the recordings of a legal user and intruders, as well as the signal parameters determined by the articulators of different speakers.

This work is devoted to the study of the possibility of identifying singing with a sharp change in fundamental frequencies. Within the framework of the described study, the features of the software package are redundant. A detailed description of the applied algorithm for identifying notes in vocal performance and the obtained test results will be published as part of a further study.

Author Contributions: Funding acquisition, A.K. and A.S.; project administration, A.K. and A.S.; software, E.K. and A.Y.; writing—original draft, E.K. and A.Y.; writing—review & editing, E.K. and A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of Russia, Government Order for 2020–2022, project no. FEWM-2020-0037 (TUSUR).

Conflicts of Interest: The authors declare that they have no competing interest. The sponsors had no role in the design, execution, interpretation or writing of the study.

References

1. Emelianova, N.A. Sociolinguistic conditions of functioning of the welsh variety of the English language (Wenglish). *Lang. Cult.* **2014**, *3*, 40–48.
2. Fedotova, M.V. Melodic structure of rising-descending tones as a marker of Welsh accent in English. *Bull. Mosc. State Linguist. Univ.* **2011**, *607*, 233–244.
3. Zharovskaya, E.V. Prosodic features of youth's speech. *Philol. Sci. Issues Theory Pract.* **2018**, 95–99. [CrossRef]
4. Reubold, U.; Harrington, J.; Kleber, F. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Commun.* **2010**, *52*, 638–651. [CrossRef]
5. Pozdeeva, D.; Shevchenko, T.; Abyzov, A. New Perspectives on Canadian English Digital Identity Based on Word Stress Patterns in Lexicon and Spoken Corpus. In Proceedings of the SPECOM 2019 Speech and Computer, Istanbul, Turkey, 20–25 August 2019; Volume 11658, pp. 401–413. [CrossRef]
6. Shevchenko, T.; Pozdeeva, D. Canadian English Word Stress: A Corpora-Based Study of National Identity in a Multilingual Community. In Proceedings of the SPECOM 2017 Speech and Computer, Hertfordshire, UK, 12–16 September 2017; Volume 10458, pp. 221–232. [CrossRef]
7. Scherer, K.R.; Sundberg, J.; Fantini, B.; Trznadel, S.; Eyben, F. The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *J. Acoust. Soc. Am.* **2017**, *142*, 1805–1815. [CrossRef]
8. Shuk, S. Acoustic cues of negative and positive reporter's attitude in British radio reports. *Vestn. Polotsk State Univ. Part A Humanit.* **2011**, *10*, 78–82.
9. Shelupanov, A.; Evsyutin, O.; Konev, A.; Kostyuchenko, E.; Kruchinin, D.; Nikiforov, D. Information Security Methods—Modern Research Directions. *Symmetry* **2019**, *11*, 150. [CrossRef]

10. Kostyuchenko, E.; Novokhrestova, D.; Tirskaya, M.; Shelupanov, A.; Nemirovich-Danchenko, M.; Choynzonov, E.; Balatskaya, L. The Evaluation Process Automation of Phrase and Word Intelligibility Using Speech Recognition Systems. In Proceedings of the SPECOM 2019 Speech and Computer, Istanbul, Turkey, 20–25 August 2019; Volume 11658, pp. 237–246. [CrossRef]
11. Rakhmanenko, I. Fusion of BiLSTM and GMM-UBM Systems for Audio Spoofing Detection. *Int. J. Adv. Trends Comput. Sci. Eng.* **2019**, *6*, 1741–1746. [CrossRef]
12. Kostuchenko, E.; Novokhrestova, D.; Pekarskikh, S.; Shelupanov, A.; Nemirovich-Danchenko, M.; Choynzonov, E.; Balatskaya, L. Assessment of Syllable Intelligibility Based on Convolutional Neural Networks for Speech Rehabilitation After Speech Organs Surgical Interventions. In Proceedings of the SPECOM 2019 Speech and Computer, Istanbul, Turkey, 20–25 August 2019; Volume 11658, pp. 359–369. [CrossRef]
13. Kharchenko, S.S.; Mescheryakov, R.V.; Volf, D.A.; Balatskaya, L.N.; Choinzonov, E.L. Fundamental frequency evaluation subsystem for natural speech rehabilitation software calculation module for cancer patients after larynx resection. *SIBIRCON* **2015**, 197–200. [CrossRef]
14. Hoppe, D.; Sadakata, M.; Desain, P. Development of real-time visual feedback assistance in singing training: A review. *J. Comput. Assist. Learn.* **2006**, *22*, 308–316. [CrossRef]
15. Tsai, W.H.; Lee, H.C. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE Trans. Audio Speech Lang. Proc.* **2011**, *20*, 1233–1243. [CrossRef]
16. Wilson, P.H.; Lee, K.; Callaghan, J.; Thorpe, C.W. Learning to sing in tune: Does real-time visual feedback help? *J. Interdiscip. Music Stud.* **2008**, *2*, 157–172.
17. Gaume, A.; Vialatte, A.; Mora-Sánchez, A.; Ramdani, C.; Vialatte, F.B. A psychoengineering paradigm for the neurocognitive mechanisms of biofeedback and neurofeedback. *Neurosci. Biobehav. Rev.* **2016**, *68*, 891–910. [CrossRef]
18. Barber, T.X. *Biofeedback & Self–Control: An Aldine Annual on the Regulation of Bodily Processing and Consciousness*; Aldine Publishing Company: Chicago, IL, USA, 1976; 581p.
19. Senese, V.P.; Venuti, P.; Giordano, F.; Napolitano, M.; Esposito, G.; Bornstein, M. Adults' Implicit Associations to Infant Positive and Negative Acoustic Cues: Moderation by Empathy and Gender. *Q. J. Exp. Psychol.* **2016**, *70*, 1–22. [CrossRef] [PubMed]
20. Vogt, P.R. Evidence for Global Synchronism in Mantle Plume Convection, and Possible Significance for Geology. *Nature* **1972**, *240*, 338–342. [CrossRef]
21. Landscheidt, T. *Cycles of Solar Flares and Weather. Climatic Changes on a Yearly to Millennial Basis*; Springer: Dordrecht, The Netherland, 1984; pp. 473–481.
22. Lockyer, W.J.S. The similarity of the short period barometric pressure variations over large areas. *Nature* **1903**, *67*, 224–226. [CrossRef]
23. Storch, H.; Zwiers, F.W. *Statistical Analysis in Climate Research*; Cambridge University Press: Cambridge, UK, 1999; 495p.
24. Ageev, B.G.; Nikiforova, O.Y.; Ponomarev, Y.U.N.; Sapozhnikova, V.A. Optoacoustic Gas-Analysis for Diagnostics of Biosystems. *J. Biomed. Photonics Eng.* **2019**, *5*, 10304. [CrossRef]
25. Ageev, B.G.; Sapozhnikova, V.A.; Gruzdev, A.N.; Golovatskaya, E.A.; Dukarev, E.A.; Savchuk, D.A. Comparison of Residual Gas Characteristics in Annual Rings of Scots Pine Trees. *Atmos. Ocean. Opt.* **2019**, *32*, 275–283. [CrossRef]
26. Babushkina, E.A.; Belokopytova, L.V.; Zhirnova, D.F.; Vaganov, E.A. Siberian spruce tree ring anatomy: Imprint of development processes and their high-temporal environmental regulation. *Dendrochronologia* **2019**, *53*, 114–124. [CrossRef]
27. Engle, R.; Granger, C. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica* **1987**, *55*, 251–276. [CrossRef]
28. Hennig, C.; Meila, M.; Murtagh, F.; Rocci, R. *Handbook of Cluster Analysis*; CRC Press: Boca Raton, FL, USA, 2015; 753p.
29. Sankoff, D.; Kruskal, J. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*; Addison Wesley Publishing Company: Toronto, ON, Canada, 1983; 382p.
30. Berndt, D.J.; Clifford, J. Using dynamic time warping to find patterns in time series. *KDD Workshop Knowl. Discov. Databases* **1994**, *10*, 359–370.
31. Douzal Chouarria, A.; Nagabhushan, P.N. Adaptive dissimilarity index for measuring time series proximity. *Adv. Data Anal. Classif.* **2007**, *1*, 1–43.

32. Bure, V.; Staroverova, K. Methods of Cluster Analysis for Detection of Homogeneous Groups of Healthcare Time Series. In Proceedings of the Constructive Nonsmooth Analysis and Related Topics (CNSA), Saint Petersburg, Russia, 22–27 May 2017; p. 7973944.
33. Staroverova, K.; Bure, V. Characteristics based dissimilarity measure for time series. *Appl. Math. Comput. Sci. Control Process.* **2017**, *13*, 51–58. [CrossRef]
34. Botygin, I.A.; Kataev, S.G.; Tartakovsky, V.; Sherstneva, A.I. Approach to clustering objects. *Bull. Tomsk Polytech. Univ. Geo Assets Eng.* **2015**, *326*, 78–85.
35. Kuznetsov, M.P.; Ivkin, N.P. Time series classification algorithm using combined feature description. *J. Mach. Learn. Data Anal.* **2014**, *1*, 1471–1483.
36. Serban, N.; Wasserman, L. CATS: Clustering After Transformation and Smoothing. *J. Am. Stat. Assoc.* **2005**, *100*, 990–999. [CrossRef]
37. Tartakovsky, V. Synchronicity as an essential property of solar–terrestrial relations: Latent components. *Nonlin. Proc. Geophys. Dis.* **2015**, *2*, 1275–1299. [CrossRef]
38. Tartakovsky, V.; Krutikov, V.; Volkov, Y.U.V.; Cheredko, N. Application of a principle of synchronicity to an analysis of climatic processes. *Earth Environ. Sci.* **2016**, *48*, 012002. [CrossRef]
39. Burkett, Z.D.; Day, N.F.; Peñagarikano, O.; Geschwind, D.H.; White, S.A. VoICE: A semi-automated pipeline for standardizing vocal analysis across models. *Sci. Rep.* **2015**, *5*, 10237. [CrossRef]
40. Job, J.R.; Kohler, S.L.; Gill, S.A. Song adjustments by an open habitat bird to anthropogenic noise, urban structure, and vegetation. *Behav. Ecol.* **2016**, *27*, 1734–1744. [CrossRef]
41. Damos, P. Using multivariate cross correlations, Granger causality and graphical models to quantify spatiotemporal synchronization and causality between pest populations. *BMC Ecol.* **2016**, *16*. [CrossRef] [PubMed]
42. Portfors, C.V.; Perkel, D.J. The role of ultrasonic vocalizations in mouse communication. *Curr. Opin. Neurobiol.* **2014**, *28*, 115–120. [CrossRef] [PubMed]
43. Ngenge, B.K. Parametric diagnostics based on synchronization and asynchrony of changing parameters of gas turbine engines of one aircraft. *Sci. Her. Mosc. State Tech. Univ.* **2014**, *208*, 48–52.
44. Bakaev, A.V. Correlation analysis of ensemble singing. *Eng. Bull. Don.* **2014**, *29*, 123–129.
45. Konev, A.; Kostyuchenko, E.; Yakimuk, A. The program complex for vocal recognition. *J. Phys. Conf. Ser.* **2017**, *803*, 012077. [CrossRef]
46. Konev, A.A.; Meshcheryakov, R.V.; Kostyuchenko, E.Y. Speech Signal Segmentation into Vocalized and Unvocalized Segments on the Basis of Simultaneous Masking. *Optoelectron. Instrum. Data Proc.* **2018**, *54*, 361–366. [CrossRef]
47. Bierens, H.J. The Nadaraya–Watson kernel regression function estimator. *Top. Adv. Econom.* **1994**, 212–247. [CrossRef]
48. Svetunkov, I.S. New coefficients of econometrics models quality estimation. *Appl. Econom.* **2011**, *24*, 85–99.
49. Kataeva, E.S.; Koshkin, G.M. The application of synchronism identification algorithm to meteorological time. *Russ. Phys. J.* **2015**, *56*, 229–231.
50. Kholopova, V.N. *Music as a kind of Art; Textbook*; Publishing Company «Planet of Music»: St. Petersburg, Russia, 2014; 320p.
51. Shvetsov, A.G. *Anatomy, Physiology, and Pathology of Hearing, Vision, and Speech*; Nelson Education: Toronto, ON, Canada, 2006; 68p.
52. Yakimuk, A.Y.; Konev, A.A.; Andreeva, Y.V.; Nemirovich-Danchenko, M.M. Applying the principle of distribution in the program complex for vocal recognition. *Mater. Sci. Eng.* **2019**, *597*, 012072. [CrossRef]