Applying the principle of distribution in the program complex for vocal recognition

IOP Conf. Series: Materials Science and Engineering **597** (2019) 012072 doi:10.1088/1757-899X/597/1/012072

Applying the principle of distribution in the program complex for vocal recognition

A Yu Yakimuk, A A Konev, Yu V Andreeva and M M Nemirovich-Danchenko
Faculty of Security, Tomsk State University of Control Systems and Radioelectronics, Tomsk, Lenin str., 40, 634050, Russia

E-mail: yay@keva.tusur.ru

Abstract. This article is devoted to studying the possibility of improving the quality of the identification of notes in vocal performance. In the present work, the authors described the transition of the program complex to a distributed functioning model. At this stage, this allowed organizing the storage of results for students and the introduction of an assessment mode for teachers. A study was conducted to improve the quality of note recognition through the use of other algorithms. During the tests, it was determined that the presence of vibrato in singing makes the greatest contribution to the number of unidentified notes. In view of the findings scheduled implementation in a program complex the paalgorithm which is responsible for determining the quality of vibrato in singing. The purpose of this modification is to get rid of novice singers from problems in singing, such as tremor.

1. Introduction

Nowadays there are many applications designed to record various musical performances, but only a few of them provide the ability to identify notes of musical performances, in particular, vocal. Services that have the ability to identify notes involve the use of specific algorithms for determining notes without the possibility of replacing them with others algorithms. These services are used for various purposes. From getting information about audio recordings and organizing music information retrieval systems [1] to vocal training programs. The main task for all these systems is to obtain information from audio recordings and visualize [2] it for comfortable use by users.

In the developed algorithm, as well as in [3, 4], the assumption is taken that the speech signal is described by a harmonic model. The idea of an accurate estimate of the fundamental frequency, developed in [5], is based on the decomposition of a signal into narrowband components and using their instantaneous frequencies as input data. This approach has been used in the analysis of speech and singing voice [6], and in creating an error-resistant algorithm for determining fundamental frequency [7]. The idea of the algorithm for determining the fundamental frequency of the speech signal uses the effect of simultaneous masking, which consists in the following. Each point along the main membrane of the inner ear, which converts mechanical vibrations into nerve impulses, is assigned a sound frequency that causes the maximum response at a given point. The greater the distance from this point, the lower the amplitude of the response. And if the amplitude of the response to a component with a natural frequency is lower than to others, then this component will not be perceived by the auditory system. The main membrane is considered here as a set of frequency resonant filters.

The purpose of this work is to improve the quality of recognition notes sung by the singer. To achieve this, the study used a modification of the recognition score to give the ability to remotely perform studies

vocal performances and apply various algorithms for recognizing notes. The modification will allow changes in the operation of the algorithms, without affecting the performance of the complex.

2. Problem statement

In a study [8], a program complex for recognizing notes was tested, which showed that in the range of 70-400 Hz, the algorithm used can correctly identify 85% of the notes sung by the singer. The used algorithm [9] for measuring the fundamental frequency, the results of which are used in determining notes, has an error in calculating fundamental frequency less than 1%, and the reliability of determining voiced sections during speech processing is 89-93%. The predicted accuracy of the note recognition algorithm with the above characteristics can exceed 95%. The resulting difference in the number of errors became the basis for additional research on finding the sources of errors in the work of the developed algorithm used in the program complex, which determines the notes in vocal performance.

It was decided to conduct a separate testing of the definition of fundamental frequency, segmentation and recognition of notes. In this way, it will be possible to understand at which of the problems arise that lead to 15% of errors in the recognition of notes.

The current limit of the upper threshold for determining the fundamental frequency of 400 Hz was used to study the signal parameters in continuous speech and in cancer diseases [10]. The increase in the upper threshold for determining the fundamental frequency is associated with the use of the considered approach for recognizing a melody sung, where the fundamental frequency can reach values of 1,400 Hz [8]. In order to increase the upper limit of determining the fundamental frequency, the algorithm was improved, but the study of the correctness of the work after the improvement was not carried out.

3. The advantages of a distributed model of the program complex

One of the changes in the program complex [8] was the transition to the use of the client-server model. This modification allowed to facilitate the introduction of changes in the applied algorithms. From the user using the program, you will not need to install new versions. As the user records the audio file in the client part of the application, the audio file conversion process is presented, as shown in Figure 1.

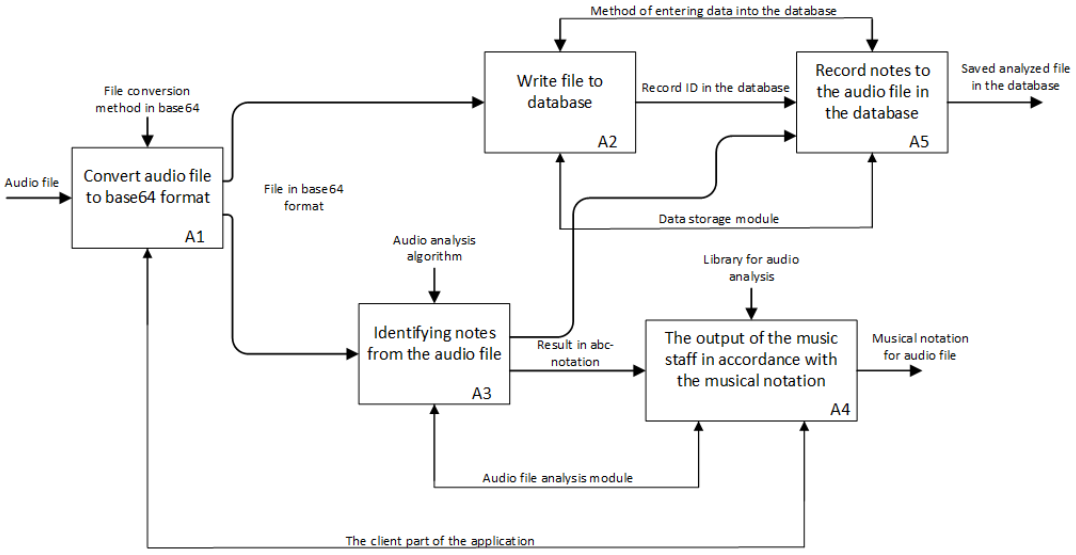

Figure 1. The process of converting the audio file for inclusion in the database

Next, the recorded file is sent to the server and processed there. After that, a message is sent to the client machine in abc notation with the results. The audio file is converted to base64 format for subsequent storage in the database using the built-in JavaScript method btoa (). In order to decode a base64 string on a server, you must likewise use the standard atob () method for the required string.

To determine the notes in the audio file, a library integrated into the project is used, which implements filtering, identifies voiced and unvoiced sections, determines the fundamental frequency and compares the received frequencies to the notes, and then outputs a string in the Abc format if the notes were found. After the user has recorded the audio, the file is converted to base64 and sent to the server. The mentor will be able to see this user in the list of his "students", and when disclosing the form, he will be able to download an audio file, rate it and leave a review. In turn, the user, after receiving feedback from the mentor, will be able to see him.

4. Conducting an experiment

The first task was to test the hypothesis of the presence of problems in the identification of notes. In the current version, the basis is the application of the Fourier transform. The hypothesis was that the transition to a wavelet transform or a hybrid algorithm can improve note recognition quality. The behavior of the analysis on the studied program complex using a hybrid algorithm showed recognition accuracy in the range of 65-70%. It was experimentally determined for test recordings that the algorithm worked better for some audio recordings. However, in most cases, the quality of identification has decreased.

The material used for experiment was audio recordings with a sequence of notes sung by the singer in the range covered by the program complex. The requirements for testing files were as follows: a duration of no more than 5 seconds, the *.wav format.

The crucial point in determining the quality of the identification of notes was the discovery in the vocal of the speakers of vibrato, perceived by experts as a correctly played note, and the program as noise. The specificity of one of the stages of the note identification algorithm, which is responsible for determining whether a voiced section belongs to a note, is a focus on pure performance.

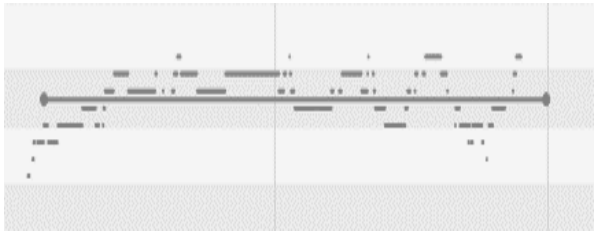

Figure 2. Note sung without vibrato in voice

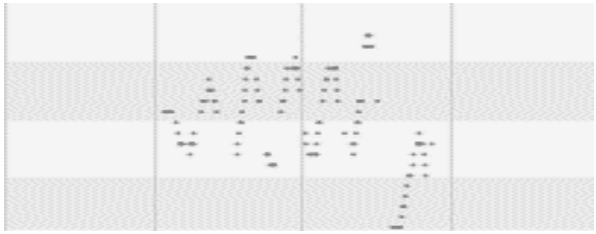

Figure 3. Note sung with vibrato in voice

To execute a note so that all frequencies at the moment of singing are within the band reserved for it is quite difficult. For this reason, the adjacent notes above and below the executable are taken into account. In Figures 2-3, the frequency spectrum is divided into sections corresponding to the boundaries of the notes. Each section is highlighted by a horizontal bar on the background of the frequency graph. On the X axis - the time, on the Y axis - the frequency.

As can be seen in Figure 2, a segment sung within a single note was recognized by the program. The presence of moments caught in the neighbouring notes is insignificantly large and did not affect the result of identification. If the fragments in each of the 3 sections turned out to be approximately equal, the algorithm perceived the entire segment as noise. This situation can be seen in Figure 3. The note sung by the speaker was played with a vibrato in the voice. By definition, vibrations within a semitone (neighbouring notes) with an oscillation frequency of 5 to 7 Hz are a sign of presence in vibrato singing [11]. As you can see on the plot, sung with the use of vibrato, there are vibrations in 4 adjacent notes. In addition, the assessment of the contribution to each of the notes in the sung segment is less than the required amount.

It should be noted that the presence of fluctuations in the voice when singing may indicate not only the professionalism of the singer. If fluctuations occur more often (from 8 to 9 Hz) - this effect is considered tremor. Also, the disadvantage is that the frequency variation (3-4 Hz) in singing is too slow,

perceived as “sound swinging”. One of the objectives of improving the quality of singing can be the refinement of the algorithm in order to determine the frequency of changes in the singing of a note.

To assess the accuracy of the algorithm, sinusoidal signals were generated with a known frequency in the range from 70 to 700 Hz. Graphs for signals with frequencies from 70 Hz up to 600 Hz inclusive looks like a straight line at generated frequency level.

The results of the algorithm for sinusoidal signals with different frequencies were analyzed for errors. Table 1 presents the values of the relative errors in determining the fundamental frequency. It is seen that the relative error increases with increasing frequency of the signal.

Table 1. The relative errors in determining the fundamental frequency

Fundamental frequency of a sinusoidal signal, Hz	Relative error, %
70	0
90	0
120	0
150	0
200	0
250	0.4
300	0.33
350	0.29
400	0.5
450	0.44
500	0.6
550	0.73
600	0.83
650	50
700	48.14

As for signals with frequencies of 650 and 700 Hz, the errors for them are calculated on the basis of a specific channel number, since the frequency at each time reference is defined as different (Fig. 4). The channel number for 650 Hz approximately corresponds to the channel number for a signal of 325 Hz, and the channel number for 700 Hz approximately corresponds to the channel number for a signal of 363 Hz. The abscissa is the time in milliseconds, and the ordinate is the fundamental frequency in Hz.

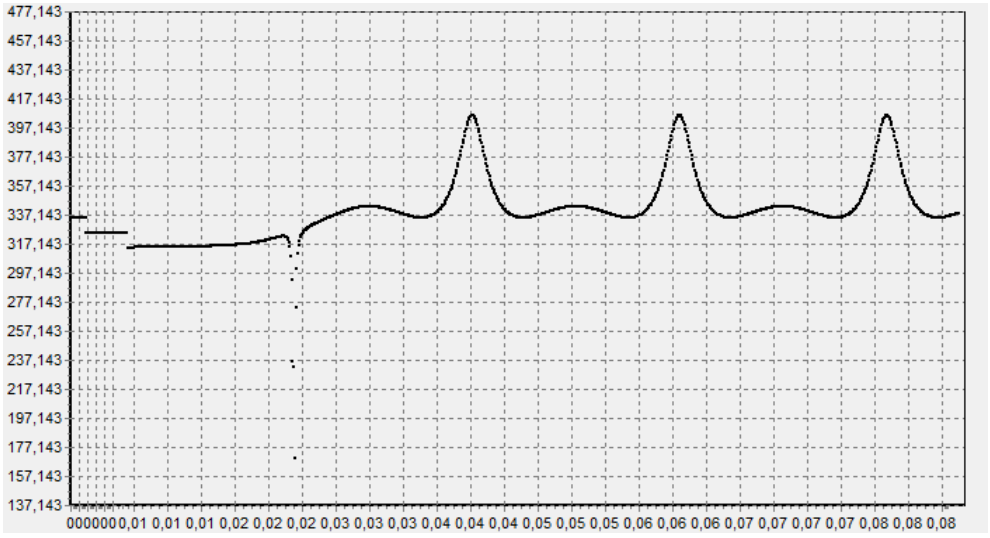

Figure 4. Graph of the fundamental frequency definition for a sinusoidal signal 650 Hz

It is worth adding that only voiced sections (having a periodic structure) are subject to analysis. This implies that the presence of a "dotted line" on the graphs is explained by the fact that the algorithm determines some parts of the signal as unvoiced.

5. Conclusion

The transition to a distributed model of the program complex made it possible to integrate new libraries that implement note recognition for the study of vocal performances without processing the source code of the program. Such an approach can become more effective in the vocal mastery model using the algorithms used in the work due to the possibility of entering new data without losing the ability to work with the previous ones. This is achieved by separating the modules responsible for data storage and audio file processing. This change will allow changing the algorithm for identifying notes without affecting the rest of the complex in order to determine the optimal parameters that will improve the quality of note recognition. One of the main directions of development of the note identification algorithm was the expansion of the existing range to 1,400 Hz. In the range of 70-400 Hz, the algorithm used can correctly identify 85% of the notes sung by the singer. Algorithms were tested at higher frequencies. As a result of the research, the algorithm showed incorrect operation for signals with frequencies from 650 Hz inclusive and requires adjustment. The magnitude of the errors tends to 50% for frequencies higher than 600 Hz, which is obviously unacceptable and requires further study.

Acknowledgments

This research was funded by the Ministry of Education and Science of Russia, Government Order no.2.8172.2017/8.9 (TUSUR).

References

[1] Dittmar C, Cano E, Abeßer J, Grollmisch S 2012 *Multimodal music processing* **3** 95–120
[2] McLeod P, Wyvill G 2003 *Computer Graphics International 2003*
[3] McAulay R, Quatieri T 1986 *IEEE Transactions on Acoustics, Speech and Signal Processing* **34(4)** 744–754
[4] Laroche J, Stylianou Y, Moulines E 1993 *IEEE International Conference on Acoustic, Speech, and Signal Processing* 550–553
[5] Abe T, Kobayashi T, Imai S 1995 *IEEE International Conference on Acoustic, Speech, and Signal Processing* 756–759
[6] Azarov E, Vashkevich M, Petrovsky A 2014 *IEEE International Conference on Acoustic, Speech, and Signal Processing* 7919–23
[7] Hotta K, Funaki K 2014 *Annual Summit and Conference AsiaPacific Signal and Information Processing Association* 1–4
[8] Konev A, Kostyuchenko E, Yakimuk A 2017 *Journal of Physics: Conference Series* **803(1)** 012077
[9] Bondarenko V, Konev A, Mescheriakov R 2007 *Proceedings of the XIIth International Conference "Speech and Computer" SPECOM'2007* 562-565
[10] Bondarenko V, Choinzonov E, Balatskaya L, Chizhevskaya S, Konev A, Meshcheryakov R 2007 *Biomedical Engineering* **41(4)** 154-156
[11] Leydon C, Bauer J, Larson C 2003 *Acoustical Society of America* **114(3)** 1575-1581