The program complex for vocal recognition

Anton Konev, Evgeny Kostyuchenko, Alexey Yakimuk

Tomsk State University of Control Systems and Radioelectronics, 40 Lenina Ave., Tomsk, 634012 Russia

E-mail: kaa1@keva.tusur.ru, key@keva.tusur.ru, yay@keva.tusur.ru

Abstract. This article discusses the possibility of applying the algorithm of determining the pitch frequency for the note recognition problems. Preliminary study of programs-analogues were carried out for programs with function "recognition of the music". The software package based on the algorithm for pitch frequency calculation was implemented and tested. It was shown that the algorithm allows recognizing the notes in the vocal performance of the user. A single musical instrument, a set of musical instruments, and a human voice humming a tune can be the sound source. The input file is initially presented in the .wav format or is recorded in this format from a microphone. Processing is performed by sequentially determining the pitch frequency and conversion of its values to the note. According to test results, modification of algorithms used in the complex was planned.

1. Introduction

The problem of the of pitch frequency determined for the speech signal is one of the main objectives of the study of voice characteristics. However, the focus in this regard is given to a range of frequencies corresponding to the speech signal. During singing, the human voice occupies a much wider range than in it does in a conversation. With the transition from speech signal processing to the evaluation of human's singing, it becomes possible to analyze the material with brand new parameters. With the expansion of the definition of the boundaries of the frequency, it will be possible to use the sound information processing algorithms for a range of new applications. For example, the non-professional musicians and beginner singers often have a problem with the necessity to match their vocal and musical creativity with the score (notes), relevant to the rules of musical notation. Automatic recognition of sounding music with the help of the special software would speed up and improve the convenience of recording the scores. In the software market, it is possible to find a variety of solutions to recieve the notes from tunes. However, most of them are designed to handle musical instruments with clear sound. As to the voice, it is dificult, especially for a non-professional singer, to achieve a long-term stability of the sound in one note. As a result, high-quality processing of the singing with the help of such software is not expected. On the other hand, the task of systems, operating with a voice, is to identify a tune (not a single note).

2. Theoretical aspects of the study

Key information for the task of note identification is the frequency at which it is performed. So the idea for the algorithm of note recognition was put forward. Its use determines the pitch frequency for note classification. Basing on the knowledge of the pitch frequency that was performed at each time point it will be possible to understand which of the music sound was closest at that time. Frequency

sound values of each note are known and constant. Accordingly, the problems with the determination of correlation of sound frequencies and the proposed note do not arise. It should be noted that for a person, as opposed to a musical instrument, it is dificult to tune out the sound on a particular frequency for a long period of time. Therefore, based on the frequencies corresponding to the notes, the list of the confidence intervals for each note was formed. It was decided to identify the vocal site as some note if the result sound frequency belonged to the confidence interval of alleged notes. Another meaningful factor in the problem of identification of the notes is the duration of the sound. Choosing a value that defines the minimum duration of the note, we can weed out the noise and get a more accurate value.

3. Planning of experiments and analysis of analogs

During evaluation of the quality of programs working with vocal music recognition, the function was performed in the ”black box” format. The program input data were notes, reproduced by a speaker. Some strategies to programs test were selected:

- testing of short sounds (staccato);
- sound testing with a minimal interval;
- testing of sounds that are in the middle interval;
- testing of sounds with pronunciation of words;
- testing of tunes with pronunciation of words.

According to the defined strategy, a test plan was drawn up. It consists of 9 entries with different variations of the play:

- location: small octave. Range: C(Do) - B(Si). 12 notes. Interval: semitone. Duration: 2 notes per second;
- location: The first octave. Range: C(Do) - B(Si). 12 notes. Interval: semitone. Duration: 1 note per second;
- location: small octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Duration: 2 notes per second;
- location: The first octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Duration: 1 note per second;
- location: small octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Staccato;
- location: The first octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Staccato;
- location: small octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Duration: 2 notes per second. With pronunciation of the names of the notes;
- location: The first octave. Notes: C(Do), E(Mi), G(Sol), C(Do), G(Sol), E(Mi), C(Do). Duration: 2 notes per second. With the pronunciation of the names of the notes;
- melody with the words.

Thus, 9 recordings, which contain 82 notes, were made on the basis of the stated strategies. Each audio was recorded by a female voice and has the following parameters:

- extension: wav;
- sampling frequency: 12 kHz;
- checksum: 16 bits;
- channel: mono.

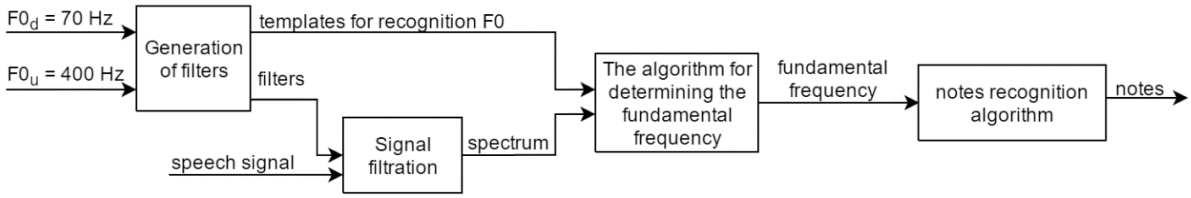

Figure 1. Stages of the note definition.

The sequence of notes obtained from each of the program, were compared to the reference sequences for each of the records. As a result, we have obtained a conclusion that the program does not always provides the user with the music recognition function and satisfies the declared quality or copes with the tasks. Most often, before recognition it is necessary to carry out the setting, indicating what sequence of notes will be taken right now by the system. Therefore, the system often perceive notes differently than the notes performed in reality. Some programs can not process the voice signal in real time, which is an extremely important aspect in teaching of vocal prowess. There are several options in description of the software that can be understood as a function of the ”recognition of the music”. Generalized conclusions about these options are presented in Table 1.

Table 1. Options of the “music recognition” function.

Input data	Output data
Files with images of notes	The sequence of notes in the program window
Musical format midi files	The sequence of notes, that is encoded in the program file explicitly
Clicking on the button with the answer choice	The verdict on the correctness of the answer, that was selected by the user
Vocal performance	The sequence of notes, that was sung by the user

Most of the programs with the music recognition feature works with peak values of frequency. For this reason, the correct recognition of the note is not always possible.

4. The software package

Since the pitch frequency at the time of its execution is the key information for the identification of the notes - the idea of using algorithms for determining the pitch frequency has been put forward. The task was to study the possibilities of accurate recognition of music that was sung by the speaker in these conditions. To solve it, it was required to write a program, whose algorithm of pitch frequency measurement was presented in detail in [1]. The input data of used module are only an audio signal in the .wav format, or converted to this format in real time during the sound recording. The parameters were selected at the early stages of research related to the evaluation of the accuracy of determining the pitch frequency [2] and provide results with an error of no more than 0.6. This greatly exceeds the accuracy for methods based on peaks, which use fast Fourier transform [3], chalk-cepstral coefficients [4] and other classical methods [5]. Schematically, the music recognition process using the program is shown in Figure 1.

Specifications applicable to the algorithm for determining the pith frequency in the range of 70 Hz to 400 Hz:

- determination of the pitch frequency in real-time
- the error of the pitch frequency calculation is less than 1

- the reliability of the voiced portions that were determined during the speech signal processing is 89 - 93

The developed music recognition software is based on two basic stages. At the first stage, we take into account the result of the algorithm of pitch frequency measuring (Figure 2), when voiced areas are defined (Figure 3), based on the use of the algorithm for determining the voiced and unvoiced regions [2].

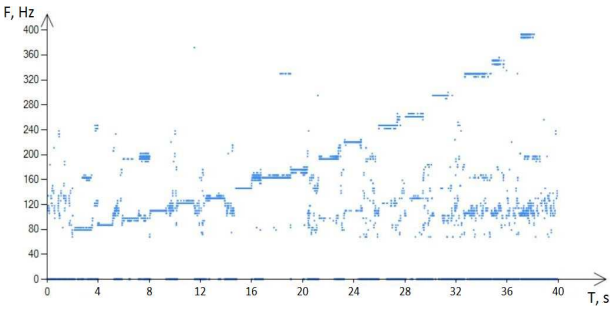

Figure 2. the notes recognition phase - determining the pitch frequency.

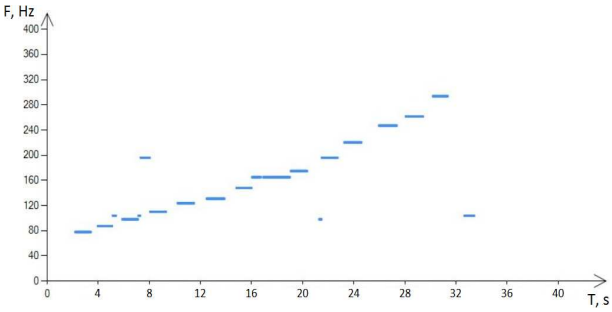

Figure 3. The notes recognition phase - determination of voiced portions.

The second step is based on the note information (Figure 4). To this end, the program matched the frequency of the pitch, which recorded the voiced site with the most suitable frequency range.

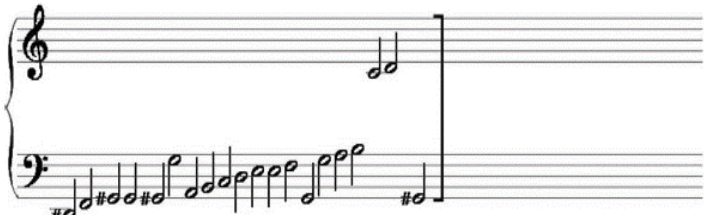

Figure 4. Recognized notes.

In addition to the written program, we also considered the Praat program, developed at the University of Amsterdam, that defines the pitch frequency recording sound [6]; the Melodyne program developed in Munich and is used for the recognition of music from an audio file and displaying the results in the frequency scale format [7]. Testing was carried out based on the same principles that were described in the preceding paragraph. For the new test developed for the program, 8 records of male and female voices were made. The variation in the performance of ”melody with the words” has not been considered in the test. In 16 tests, 114 notes were played, 58 of which were sung by a female voice, and 56 - by male. The best recognition results were obtained for records of sung notes in staccato [8]. In this case, the performance was discrete, and between sung notes, false notes did not appear. This situation is true for the developed program, as well as for the analog. On the other hand, analogs managed a little worse with the strategies executed by legato. For example, in the record based on the eighth strategies, notes E, D, F, E, G were executed as legato. Execution was carried out with the pronunciation of the names of the notes. Due to the nature of music performance, legato transition between notes that are not adjacent to each other can be erroneously interpreted for any note in-between. In the Praat program, we can see that the first fixed portion is in the range from 240 to 250 Hz that can be perceived as note B (Figure 5). After this burst at the beginning of the site, all other voiced portions fluctuate within the frequencies of notes, which were sung and determined correctly.

Record analysis using the Melodyne program revealed the existence of note B, which sounded for a short period of time at the beginning of the recording (Figure 6).

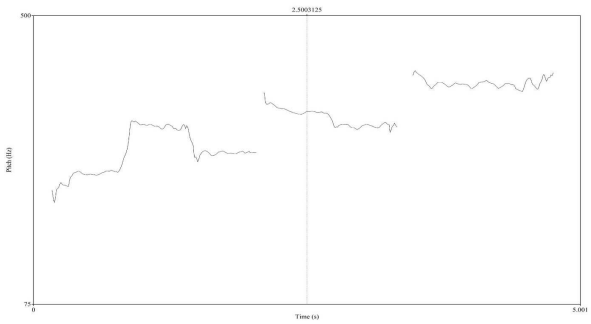

Figure 5. Review the 8-th strategy in Praat.

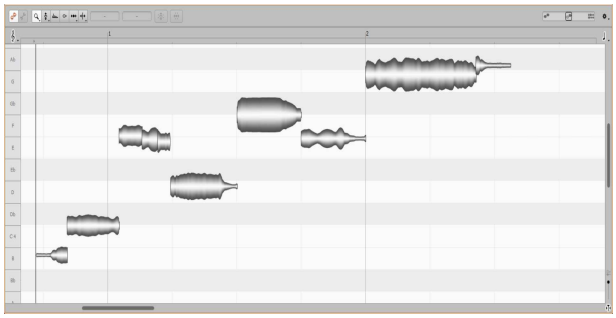

Figure 6. Review the 8-th strategy in Melodyne.

Using the criteria of the minimum voicing time allowed one to weed out the spike at the beginning of the record (Figures 7, 8, 9). As a result, the designed program was able to recognize all the notes sung by actors and weed out the false voiced area.

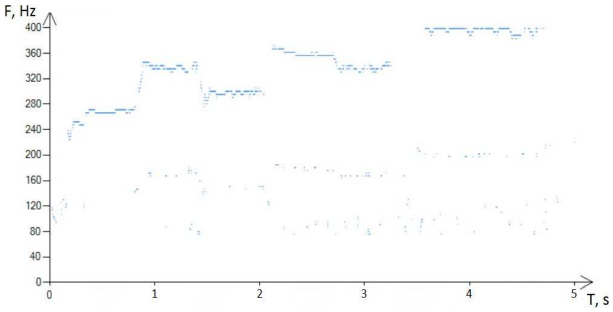

Figure 7. An array of pitch frequency for recording in the 8-th strategy.

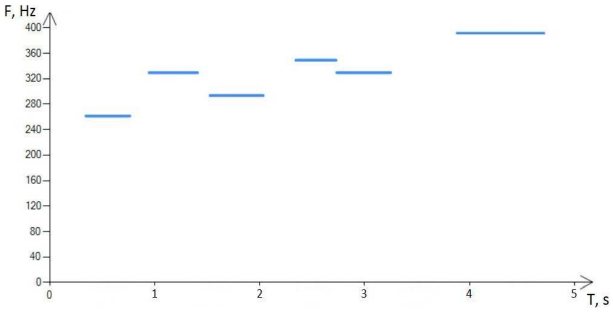

Figure 8. Voiced areas for recording in the 8-th strategy.

Thus, as a result of the study [9], test recordings were measured so that program Praat could correctly identify voiced areas for 70 notes.

Figure 9. The sequence of notes that was recognized in the record in the 8th strategy.

5. Conclusion

The work proposes an approach for determining the notes based on the algorithm for determining the pitch frequency. Testing has revealed high accuracy (more than 90%). The calculation module of the pitch frequency is a component of a complex software that was written to evaluate the voice information in the speech process. For this reason, the range of pitch frequencies was from 70 to 400 Hz. It is necessary to increase the range from the big octave to the third octave. This is because the average singers can have pitch frequency in the range of 65 to 1400 Hz. In this way, the program will cover the singing voices [10] from bass to soprano. The algorithm for determining the pitch frequency should be modified to increase the recognition frequency range. The first step of the modification is the generation of a new set of filters, through which it will be possible to obtain a full spectrogram for singing. After that new templates that are used to mask the spectrogram should be generated. As a result, the algorithm will allow receiving the pitch frequency for music located higher than note G of

the first octave. The next problem is increasing the speed of the music identification. It is necessary for training vocal skills in the real-time mode. Providing results in real time is very important for the tasks of learning vocal skills. Providing vocal training results to a user allows designing a complex software for teaching vocal skills by a biological feedback. In addition, the idea to modify the program for the client-server model of the application was formulated. Accordingly, there is a problem to separate the software into individual program modules. The client part of the application will be responsible for processing an audio signal and transmitting the calculated pitch frequency to the server component of the complex. This component will convert the pitch frequency sequence into the finished note frequency sequence. The server application will return this sequence to the client. As a result, the application on the client side will display the note representation of the music that was sung by the user. This option is considered as a possible version of the algorithm execution in mobile applications or websites.

6. Acknowledgments

This work was supported by the Ministry of Education and Science of the Russian Federation in the framework of the basic part of state task TUSUR, 2015-2016 (project number 3657).

References

[1] Bondarenko V, Konev A, Meshcheriakov R 2007 *Proceedings of the XIIth International Conference Speech and Computer (SPECOM-2007)* **2** 562-565
[2] Bondarenko V, Gitman A, Chabanec A, Ustyugova T 1989 *Automatic detection of auditory images (ADAI-15)* **247** 186-189
[3] Mitre A, Queiroz M, Faria R 2006 *Proceedings of the 4th AES Brazil Conference* **1** 113-118
[4] Qi Y, Hunt B 1993 *IEEE Transactions on Speech and Audio Proceeding* **2(1)** 250-255
[5] Gerhard D 2003 Pitch Extraction and Fundamental Frequency: History and Current Techniques. Technical Report Department of Computer Science University of Regina **23**
[6] Boersma P, Heuven V 2001 *Glot International* **9/10(5)** 341-347
[7] Lim K, Raphael C 2010 *Computer Music Journal* **34(3)** 45-55
[8] Gfeller K, Witt S, Adamek M 2002 *Journal of the American Academy of Audiology* **13(3)** 132-145
[9] Konev A, Onischenko A, Kostyuchenko E, Yakimuk A 2015 *Science Bulletin of the Novosibirsk State Technical University - Proceedings of Novosibirsk State Technical University* **60(3)** 32-47
[10] Yushenko N, Denisova E 2015 *Culture, art, education in the information space of the third millennium: Problems and prospects. Collection of scientific works* **1** 267-271