ALGORITHMS AND COMPUTER SYSTEM OF SINGING SYNTHESIS FOR RUSSIAN LANGUAGE

ALIAKSANDR LOMAU
Submited by Liliya I. Tsirulnik

This paper gives an outline of basic principles and methods for functioning of system which is meant for synthesis of a singing voice. The aspects of the system usage are shown. The given study detects features of natural professional singing voice.

Introduction. Nowadays the usage of a synthesized sounding of musical instruments is very widespread in different areas. The development of this field of study resulted in the creation of natural sounding digital systems. So far the singing voice, which is the leading instrument of musical expression in lots of musical genres, cannot be convincingly simulated through synthesizing techniques. This is mainly attributed to the complex nature of the singing voice production mechanism. Even small changes of any of its components can seriously affect acoustic properties of the resulting waveform as well as the listener's perception. As yet there are software products that are able to solve this problem, but they cannot work with the Russian language. Thus, for example, the software suite Vocaloid [1] developed by Yamaha company works with the Japanese and the English languages and it is widely used by sound-recording companies for producing various sound effects.

This research touches upon the issues of development and usage of singing voice synthesis. Also it gives information about the features of natural professional singing, describes algorithms and methods which are used for creation of a singing synthesizer.

Essential information about singing voice features. Among all the variety of sound characteristics which are applied to convey the melody of a musical composition there exist three basic characteristics: voice pitch frequency (F_0), sound vibration amplitude (A) and the duration of the sound (T). In speech these characteristics are expressed by intonation or prosody of the speech. The vibration of the vocal folds which takes place during the talking produces a periodic signal. This signal changes in the vocal track under the influence of the articulators (which work as resonators of speech wave) and results in the creation of a voice waveform. The interval of a sound wave which corresponds to one period of vocal folds' vibration is called "the pitch" or "the period of the fundamental tone". Sounds which have an explicit periodical interval are referred to as voiced sounds. Sounds, which don't have computable fundamental frequency, are called unvoiced sounds.

Singing voices are often grouped by gender of the singer or fundamental frequency range. As it is shown in table 1, fundamental frequency of classical singing voices can be varied in range 75-1100 Hz [2]. Every voice group covers from two to two and a half octave. In contrast to this, a regular speech diapason usually covers about one octave.

Table 1

Classification of singing voices according to frequency range.

Voice group	Frequency range, Hz	Voice group	Frequency range, Hz
Bass	75-330	Contralto	150-710
Baritone	110-450	Mezzo-soprano	170-700
Tenor	120-500	Soprano	230-1100

To reach natural sounding of a synthesized vocal we have to take into account different additional effects of a real professional vocal. One of the widespread and frequently used singing effects is vibrato. Vibrato can be described as a sinusoidal modulation of the fundamental frequency during some fragments of singing. The rate of vibrato is typically 5-8 Hz and the modulation depth varies between 50 and 150 cents[1]. The greater the depth and the regularity of

[1] Cent (in music) is a logarithmic unit of measure which is used for measurement musical intervals, 1200 cents are equal to one octave.

vibrato the more professional skills it demands from a singer. It is well known that baritones with the most aesthetically pleasing voices maintained vibrato during more than 80% of their singing [3].

Professional singers' formant structure also differs from the structure of ordinary singing. In professional voices three first formants are more clearly expressed. The fourth and the fifth formants are also often well defined on spectrograms of professional singers' voices. [3]

Algorithms of synthesis of singing voice. The singing voice has been shown to be different from normal speech, and thus speech synthesis techniques that were designed for regular speech are not always suitable for the singing voice. However, many of these techniques can provide a basis from which algorithms specific to the synthesis of the singing voice can be derived.

It should be noticed that in singing prosodic characteristic are rigidly bounded to the melody and are described by musical notation of a song. In case of text to speech synthesis prosodic information is not a priori defined. To obtain this information number of algorithms is used. [4]

The general structure for singing synthesis system is illustrated on fig. 1. Musical notation of a song is used as input data for the system. This notation consists of lyrics divided into syllables, where every syllable is associated with one or more notes. At the first stage, lyrics and note representation of a song are singled out from the musical notation. Lyrics divided into syllables come to the input of phonemic translator. Note representation is passed through the converter of musical notation.

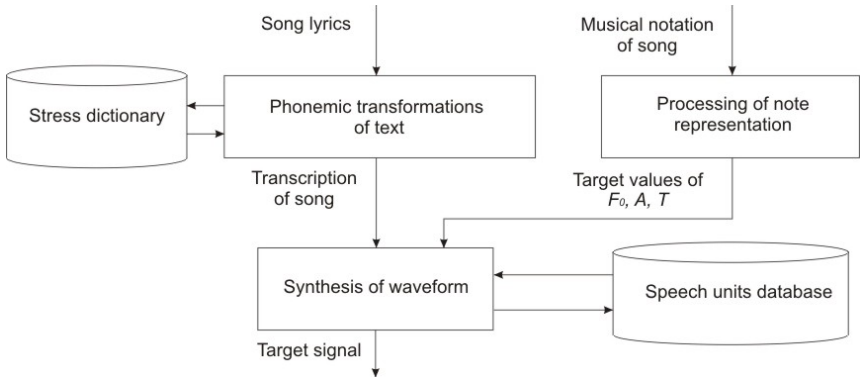

Fig. 1. General scheme of computer system of synthesis.

At the first step of phonemic translation the algorithm sets words stresses. It helps to carry out the following letter-to-phoneme[2] and phoneme-to-allophone[3] transformations. The result of text processing is the creation of an allophones sequence divided into syllables [5].

Note processor's gets sequence of notes as an input. To calculate target prosodic values special functions are used. Target fundamental frequency is calculated using formula $F_0 = f_0 \cdot 2^{n/12}$, where f_0 is a frequency of basic degree and n is a distance in degrees from current note to the basic degree [6].

To get source voice signal a concatenative method is used. It is based on the compiling of elementary units of natural speech. The set of elementary units of natural speech, which are necessary and sufficient for singing synthesis, compose speech units' database. To improve naturalness of transition between speech units the base also includes records of pairs of units used together. Searching for proper pairs is performed by the next strategy: if a syllable includes several allophone pairs, which are present in base, the preference is given to a consonant-vowel pair, then to vowel-consonant pairs and only then to consonant-consonant pairs. The obtained set of speech units' records are bounded together in source signal, which is processed then with the help of algorithms in order to reach target values of frequency F_0, amplitude A and duration T.

[2] Phoneme is a smallest unit of speech distinguishing one word (or word element) from another.
[3] Allophone is a conditioned realization of a phoneme.

The order of algorithms is important because during the changing of fundamental frequency the duration of pitch and whole allophone is changed. That is why the frequency changing algorithm takes place at the very beginning.

Changing of fundamental frequency is implemented only for vowels and sonorant consonants /m/, /m'/, /n/, /n'/, /r/, /r'/, /l/ and /l'/. Sonorant sounds affect the perceptible fundamental frequency of the syllable [4].

Duration changing of pitches is accomplished by «soft lacing» algorithm [7].One of the main advantage of this algorithm is that it doesn't change the period of a signal which corresponds to the closing of vocal folds, thus preserving the individual characteristics of voice. Another noticeable advantage is that the algorithm has a linear computational complexity as opposed to other algorithms of such kind like TD-PSOLA [7] and frequency domain algorithms. The algorithm assumes the existence of pitch marking, where the start of every pitch corresponds to the closing of vocal folds.

When the pitch is enlarged the modification of the speech wave is held according to formula (1).

$$\tilde{s}(n) = s(n)L_1(n) + s(n+T-T_0)L_2(n)\ , \qquad (1)$$

where $s(n)$ is the source signal taken from allophone base, T_0 - fundamental period of source wave, T - target meaning of fundamental period. $L_1(n)$ and $L_2(n)$ are piecewise linear functions that can be determined by the following formulas:

$$L_1(n) = \begin{cases} 1, n \le T - T_0 \\ 1 - \dfrac{n-(T-T_0)}{2T_0 - T}, n > T - T_0, \end{cases} \quad L_2(n) = 1 - L_1(n) \qquad (2)$$

When the duration of the pitch is reduced, the final piece of it, which doesn't exceed the needed length, is deleted. The removed piece is overlaid onto the foregoing section in the same way as in the case of duration enlarging.

This algorithm imposes constraints on frequency range. It changes the pitch with quite a good result in the interval 50%-150% of source pitch duration. It conforms to fundamental frequency changing in range from 70% to 200% of source frequency.

One of the complications in the realization of the system with such nature is automatic detection of the duration of allophones in syllables. In current realization of the system only the length of vowels is changes. The duration of consonant allophones are left unchanged. It allows calculating the duration of a vowel allophone using the following formula:

$$T_{tv} = T_{ts} - \sum_{i=1}^{N_c} T_{sc_j}\ ,$$

where T_{tv} - target length of vowel allophone, T_{ts} - target length of syllable, T_{sc_j} - source duration of consonant allophone with sequence number j in syllable, N_c number of consonants in the syllable [which is equal to the number of all allophones without one].

Time scaling of allophones in both algorithms is realized by adding or removing pitches. It is necessary to add, that the changing of an allophone starts from the central part of the wave to save transition points between allophones [7]. If a large increase of duration is done the adding of noise and frequency modulation in a monotonous piece makes sense.

The perception of sound volume depends on signal amplitude *A*. The changing of signal amplitude can be made by simple multiplication of signal by a constant. It should be noted that, subjective dependence of perception of loudness on amplitude is not linear. To accomplish relative changing of loudness decibel measurement is often used [6].

Realization of the system for singing synthesis. The described above system was realized using C++ language for working on Windows family operation systems. The program uses DirectSound as an audio library. The system is realized as stand-alone and as client-server application. Client-server architecture (fig. 2) gives the user an opportunity to use the synthesizer

without any installations and always get the latest updates. Front-end is implemented with flex technology; it allows the user to work with the synthesizer using a standard web browser.

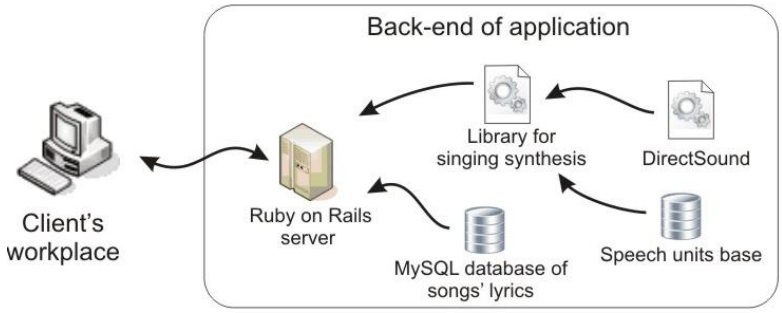

Fig. 2. Client-server architecture of singing synthesizer.

MusicXML is used as the inner format of data storing and transferring. Being an XML extension, MusicXML is an open format and is supported by various note musical editors [8]. Songs' lyrics and users' information is stored on the server using MySQL database. User can create, change and save songs on server; he also can save song in MusicXML format on his computer. To perform singing synthesis the server uses C++ library. Further signal is sent to the client, where it can be played, changed and saved.

Conclusion. The developed algorithms and program realization can find application in singing learning, in mass media, in advertisement and entertainment business.

The further work will be directed to increasing of frequency range and speech naturalness. There is several ways to do this. There are enhancement of algorithms and expansion of the speech base among them.

The algorithms described can be applied for singing synthesis not only for the Russian language, but also for another language. To realize that, speech base and phonemic translator for corresponded language are needed [7].

BIBLIOGRAPHY

1. Vocaloid official web site [Electronic resource] / Yamaha Corporation. Mode of access: http://www.vocaloid.com/en/index.html - Date of access: 10.03.2010
2. A.P. Ivanov. «The Art of Singing» [in Russian]. / A.P. Ivanov. Golos-press, 2006. — 436 p.
3. Acoustic Models for the Analysis and Synthesis of the Singing Voice. / Matthew E. Lee; Georgia Institute of Technology – 2005 – 127 p.
4. Boris M. Lobanov "Text-To-Speech Synthesis" [in Russian], Proc. IV International School-Seminar of Artificial Intelligence, Minsk, BSU, 2000, pp. 57-76.
5. L. Tsirulnik "Algorithm of Generation of Phonemic Sequence by Orthographic Text in TTS-Synthesis System" [in Russian], Informatics, 2006, N°4, pp. 61–70.
6. Steven, W.S. «The Scientist and Engineer's Guide to Digital Signal Processing» / Steven W. Smith, Ph.D. – Science Book, 1998. – 842 p.
7. Lobanov B.M., Tsirulnik L.I. "TTS-synthesis and Voice Cloning" [in Russian], Minsk, Belorusskaya Nauka, 2008, 342 p.
8. MusicXML 2.0 Tutorial // Recordare LLC [Electronic resource]. – 2008 – Mode of access: http://www.recordare.com/xml/tutorial.html. – Date of access: 09.03.2010.