State of art of real-time singing voice synthesis

Leonardo Araujo Zoehler Brum[1], **Edward David Moreno**[1]

[1]Programa de Pós-graduação em Ciência da Computação – Universidade Federal de Sergipe
Av. Marechal Rondon, s/n, Jardim Rosa Elze – 49100-000 São Cristóvão, SE

leonardo.brum@dcomp.ufs.br, edwdavid@gmail.com

Abstract

This paper describes the state of art of real-time singing voice synthesis and presents its concept, applications and technical aspects. A technological mapping and a literature review are made in order to indicate the latest developments in this area. We made a brief comparative analysis among the selected works. Finally, we have discussed challenges and future research problems.

Keywords: Real-time singing voice synthesis, Sound Synthesis, TTS, MIDI, Computer Music.

1. Introduction

The aim of singing voice synthesis is to computationally generate a song, given its musical notes and lyrics [1]. Hence, it is a branch of text-to-speech (TTS) technology [2] with the application of some techniques of musical sound synthesis.

An example of application of singing voice synthesizers is in the educational area. Digital files containing the singing voice can be easily created, shared and modified in order to facilitate the learning process, which dispenses the human presence of a singer as reference, or even recordings.

This kind of synthesis can also be used for artistic purposes [3]. Investments on the "career" of virtual singers, like Hatsue Miku, in Japan, have been made, which includes live shows where the singing voice is generated by Vocaloid synthesizer, from Yamaha, and the singer image is projected by holograms.

The singing voice synthesis technology applications have been increased by the development of real-time synthesizers, like Vocaloid Keyboard [4], whose virtual singer is implemented by an embedded system into a keytar, allowing its user the execution of an instrumental performance.

The present article presents a review about real-time singing voice synthesis embedded systems, through the description of its concept, theoretical premises, main used techniques, latest developments and challenges for future research.

This work is organized as follows: Section 2 describes the theoretical requisites which serve as base for singing voice synthesis in general; Section 3 presents a technological mapping of the patents registered for this area; in Section 4 the systematic review of literature is shown; Section 5 contains a comparative analysis among the selected works; Section 6 discuss the challenges and future tendencies for this field; finally, Section 7 presents a brief conclusion.

2. Theoretical framework

Singing voice synthesis has two elements as input data: the lyrics of the song which will be synthesized and musical parameters that indicate sound qualities. The lyrics can be inserted according to the orthography of the respective idiom or through some phonetical notation, like SAMPA, while the musical parameters can be given by MIDI messages or other file formats, such as MusicXML. The expected output is a digital audio file which contains the specified chant.

In respect of data processing, the main singing voice synthesis techniques consist in a combination of text-to-speech (TTS) and musical synthesis. In the early years of this area, two approaches were developed: rule-based synthesis, which computationally generates sound according to its physical characteristics and sample-based syntheses, that uses pre-recorded audio. The data-driven approach has been developed recently. It uses statistical models.

Singing voice synthesis problem domain is, therefore, multidisciplinary: beyond computer science, it depends on concepts from acoustics, phonetics, musical theory and signal processing. The following subsections present how each mentioned knowledge area interact in the main singing synthesis technical approaches, with a brief description of real-time synthesis at the end of this article.

2.1. Rule-based approaches

Rule-based singing synthesis considers the way sound is produced, by the analysis of its physical characteristics, which are applied in the artificially generated signal.

Sound is a physical phenom generated by the variation, through time, of atmospheric pressure levels provided by a vibratory source. Given its wavy nature, sound has physical quantities such as period, frequency and amplitude. **Period** consist in the duration of a complete wave cycle. **Frequency** is the inverse of period and indicates how many cycles per second has the sound wave. Finally, **amplitude** is the maximum value of pressure variation in relation to the equilibrium point of the oscillation [8]. The period and amplitude of a simple sound wave can be seen in the Figure1.

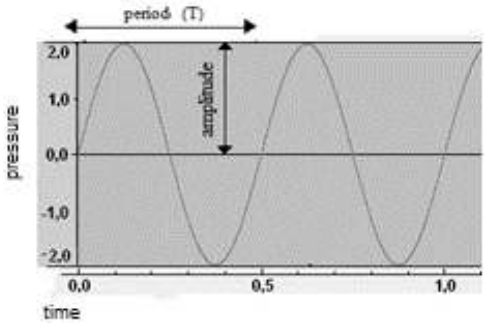


Figure 1. Simple sound wave.

The simplest sound waves are called sinusoids and have a single frequency. However, this kind of sound is not produced neither by nature, nor by the conventional musical instruments. Such sources generate **complex sounds**, composed by several frequencies. The lowest frequency in a complex sound is called **fundamental frequency**. In case of the other frequency values of the sound be multiples of the fundamental frequency, the sound is said to be **periodic**. If not, the sound is called **aperiodic**. The superposition of frequencies results in the **waveform** of each sound source and describes a shape called **envelope**, obtained from the maximum oscillation values of the waveform.

The envelope shape is commonly decomposed into four stages, indicated by the ADSR acronym: attack, which corresponds to the period between the beginning of the sound execution and its maximum amplitude; decay, the time needed for the sound depart form is maximum amplitude towards a constant one; sustain, the interval in which such constant state persists; release, whose duration is between the constant state and the return to silence.

In respect of speech, specifically, its smallest distinguishable unit is called **phoneme**. Phonemes are classified into two main groups: **consonants** and **vowels**. From an acoustic perspective, both groups consist in a set of complex sounds: consonants are aperiodic vibrations that result from an obstruction of the air flow by body parts such as lips or the tongue. On the other hand, vowels have periodic nature.

Another factor of differentiation between vowels and consonants is the role that each type of phoneme performs in the syllable. According to the frame/content theory [5], speech is organized in syllabic frames which consist in cycles of opening and closure of mouth. In each frame there is a segmental content, the phonemes. This content has three structures: attack, nucleus and coda. Consonantal phonemes are located in attack or coda, while a vowel form the syllable nucleus [6]. Beyond consonants, there are phonemes called semivowels, which form diphthongs and can also appear in attack or coda. Acoustically speaking, syllable is a waveform in which consonants and semivowels are found in attack and release (coda) stages, while vowels are in sustain (nucleus) stage of the envelope shape. Figure 2 schematically presents the relation between the structure of the Brazilian Portuguese syllable “pai” and the envelope shape stages.

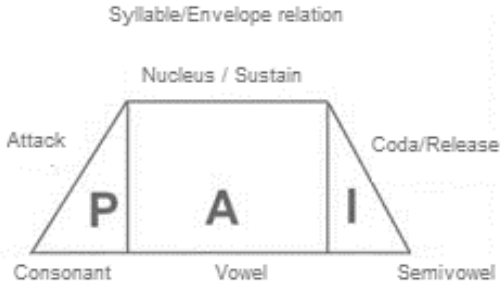


Figure 2. Relation between a syllable structure and the envelope shape stages.

Furthermore, vowels have another important feature, caused by resonances produced by the human vocal apparatus. These elements are called **formants** and appear as energy peaks verified when an acoustic signal is analyzed from its spectrum, in frequency domain. The spectral representation put in evidence the most relevant frequencies of a complex sound in relation to amplitude. Formants have the value of the central frequency in each energy peak that appears in the shape called spectral envelope. They are commonly enumerated from the lowest frequencies, as F1, F2, F3 and so on [7]. Figure 3 presents the spectrum of a voice signal where four formants are indicated.

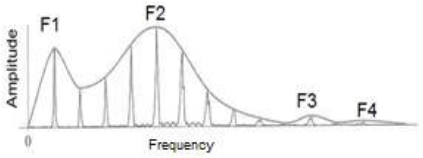


Figure 3. Spectral envelope of a voice signal with the indication of four formants.[7]

The presence of formants allows one to perceive the difference between vowels, since the sound qualities, such as pitch and loudness, of "a" and "o", for example, can be identical.

Formant synthesis is an example of rule-based singing voice synthesis approach, which consists in the generation of units called *Forme d'Onde Formantique* (FOF, French for formant waveform). FOFs are sinusoids with a very short duration whose frequencies are equal to the value of the formants of the phoneme to be synthesized. Each FOF is then repeated according to the periodicity of the fundamental frequency of the musical note that is intended to synthesize. This process produces a series of sequences which are summed in order to generate the synthesized singing voice [7].

Systems based on this kind of approach, such as CHANT, developed by *Institut de Recherche et Coordination Acoustique/Musique* (IRCAM) at the early 1980's, are among the first ones in respect to the use of synthesized voices for artistic purposes. They are capable of synthesize realistic vowels, but it costs a big studio effort to analyze and adjust parameters [2].

2.2. Sample-based approaches

In regard of perception, by human audition, of the sound phenom, the physical quantities previously mentioned are related to the so-called **qualities of sound**: **pitch**, that allows humans to distinguish between high and low sounds and is proportional to fundamental frequency; **loudness**, which depends on amplitude and indicate the difference between loud and soft sounds; timbre, related to the waveform, and for consequence, to the envelope shape, is the quality perceived as the proper "voice" of each sound source, which permits to distinguish, for example, the sound of piano from the guitar one, even if both have the same pitch and loudness. A fourth quality which can be cited is the **duration** of sounds [8].

Periodic complex sounds are usually perceived as having a defined pitch, which corresponds to the fundamental frequency. These sounds are called **musical sounds**. On the other hand, aperiodic sounds, which do not have a clearly distinguishable pitch, are denominated as **noises**, although they also are employed in music, specially through percussion instruments.

Since the development of musical theory in Western World, certain ways to select and organize sounds for artistic purposes and to graphically represent them according to its qualities, were created. The sensation of likeness when two sounds are heard and one of them has two times the fundamental frequency of the other, allowed the division of the audible frequencies range into musical scales, which consist in a set of individual sounds called **musical notes** [8]. The succession of sounds with different pitches or, in other words, the succession of different musical notes is the part of music denominated **melody**.

Musical sounds can be synthesized by means of sample-based approach, where recordings form real musical instruments are stored and handled according to the needs of the musician. From the recorded samples, other musical notes whose pitch is near are generated, while timing is treated as follows: if the note duration is less than the sample duration, the execution is interrupted; if not, there are two possibilities that depend on the instrument which is intended to be synthesized. For some instruments, like piano, the recorded sample will be executed till its end, returning to silence; for other instruments, like organ or flute, it is wanted for its execution to be prolonged as the note stays activated, either by a keyboard or some software. This indefinite prolongation is a result of the application of the looping technique, where a specific part of the sustain stage of the waveform is continuously repeated. The end of the note activation makes execution gets out of the loop towards the release stage [9].

One of the technologies widely employed to perform sample-based synthesis is MIDI (acronym for Music Interface Digital Instrument) protocol, developed by Yamaha at the 1980's. This protocol provides communications between electronic musical instruments and computers. Its messages and file format do not contain audio signals, but only musical parameters that correspond to the sound qualities: musical note/pitch, dynamics/loudness, musical instrument/timbre, duration, among others. Such parameters serve as base for an on-demand handling of the sound samples, which can be stored in a computer or even in a musical instrument.

Sample-based synthesis can be also applied in order to perform the conversion of text into speech (TTS, text-to-speech). The greatest challenge of this approach is that how much larger is the size of the samples, more natural the result will sound, but the range of possible expressions will be smaller. Thus, for example, in case of developing a speech

synthesizer for Brazilian Portuguese language taking word by word as samples, it will demand hundreds of thousands of recordings. On the other hand, if samples were at phoneme level, they would be less than a hundred and could generate any word, but the result of their concatenation will sound invariably "robotic" [2].

The challenges of sample-based TTS technique are naturally transposed for singing synthesis field, in case of performing it through the same approach. The song lyrics associated to a melody serves as input data, where each syllable corresponds to a single musical note. The phonetic samples are then being concatenates as the input is read by the system.

The looping technique, previously described, is applied on vowels, because they are periodic, musical sounds which correspond to the musical notes and to the sustain stage of each syllable. This process prolongs the syllable duration according to the musical parameters of the input. Consonants and semivowels are concatenated at the vowel's margins [9].

Pre-recorded samples are commonly stored in a singing library which consists in units that can be modeled to contain one or more phonemes. In singing voice, the pitch variation among vowels is much less than in speech, because the first one is driven by the musical notes, but this fact does not exclude the difficulties to obtain a "realistic" result from samples in singing synthesis [10].

An example of system that performs concatenative singing voice synthesis is Vocaloid [3], which has achieved great commercial success. Vocaloid has a piano roll-type interface, composed by a virtual keyboard associated to a table whose filling is correspondent to the chosen musical notes. Input can be made by means of conventional peripherals, such a mouse, or through electronic musical instruments that support MIDI protocol. The song lyrics is associated to musical notes as it is typed into the piano roll. Input data is sent to the synthesis engine, serving as reference to the selection of samples stored in the singing library. A system diagram of Vocaloid is shown by Figure 4.

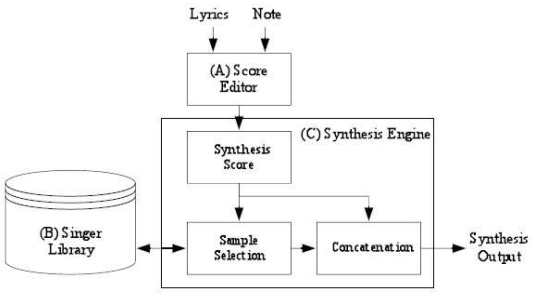


Figure 4. System diagram of Vocaloid [3].

2.3. Data-driven approaches

In the last years, some singing synthesizers have been developed based on probabilistic models, which differ from the deterministic nature of the rule-based approach. Tools like Hidden Markov Model (HMM) [1], successfully employed in TTS systems, are useful, for example, to apply in samples that contains a single phoneme the behavior provided by a statistical analysis of the voice of a singer. This decreases the size of the singing library and minimizes the lack of "naturalness" of the synthesizes voice in a more efficient way than the concatenative approach. The adjust of parameters performed by this kind of model is commonly called **training**, while the signal from which the parameters are extracted is denominated **target**.

The first HMM-based singing synthesizer was SinSy [11], developed by Nagoya Institute of Technology. This system is available on a website, where the upload of a MusicXML file, a format generated by most of the music score editors, can be made as input. SinSy provides as output a WAV file that contains the synthesized singing voice. The idioms supported by SinSy are English and Japanese.

2.4. Real-time singing voice synthesis

Users of synthesizers like Vocaloid define input data (song lyrics and musical notes) for the system in order to generate the singing voice later, in such a way analog to an IDE engine, where design time and run time are distinct.

This limitation has been overcome by the development of real-time singing voice synthesizers. They are embedded systems that artificially produce chant at the very moment the input data is provided by the users, which allows to use the synthesizer as a musical instrument [12].

In order to achieve a better comprehension of this new branch of singing synthesis, the present work performed a scientific mapping, according to the methodology proposed in [13]. The research questions that must be answered are the following ones: (i) What are the singing synthesis techniques employed by most of the real-time systems? and (ii) What are the input methods that such systems use in order to provide the phonetic and musical parameters?

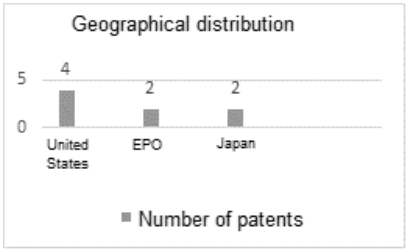
The scientific mapping consists in a technological mapping, where the patents related to real-time singing synthesis are searched, and a literature review. Both parts of the scientific mapping are described by the next two sections.

3. Technological mapping

The search for patents related to real-time singing voice systems was performed in two different databases: WIPO (World Intellectual Property Organization) and INPI (Instituto Nacional da Propriedade Industrial, from Brazil). The INPI database returned no results, even for more generic search keys in English and Portuguese, like "síntese de voz cantada" or "singing synthesis". The WIPO database, for its turn, provided some patent deposits from the following search string:

FP:(FP:((" SINGING SYNTHESIS " OR "SINGING VOICE SYNTHESIS" OR "SINGING SYNTHESIZING") AND ("REAL TIME" OR "REAL-TIME")))

The research presented eight records as result. All of them were property of Yamaha Corporation, from Japan, and their author was Hiraku Kayama, except by one, whose author was Hiroshi Kayama. However, most of the patents were registered outside Japan, probably in order to warrant international legal protection. Graphic 1 presents the geographical distribution of the patents, where EPO is the European Patent Office.



Graphic 1. Patents geographical distribution.

All patents had as object a method, apparatus and storage medium for real-time singing voice synthesis. It is a product, developed by Yamaha, which consists in a musical keyboard with an embedded singing synthesizer, allowing the user to make an instrumental performance with its virtual singer. The product was denominated Vocaloid Keyboard [4].

A prototype of the instrument, presented in 2012, had alphabetic buttons at left, organized as follows: two horizontal rows with consonants and diacritical signs and, bellow them, five buttons with vowels, organized in a cross shape. With left hand, the user could activate these buttons to generate syllables, meanwhile the musical keyboard could be played by the right hand to indicate the musical notes. The generated syllables were shown in a display with *katakana* Japanese characters. Figure 5 shows the described prototype.

Figure 5. Vocaloid Keyboard prototype [4].

This device was designed to synthesize singing in Japanese and the phonetic limitations of such idiom favored this kind of interface. The prevalent structure of Japanese syllables is consonant-vowel, which means that, for example, when "S" and "A" buttons are simultaneously activated, the systems generates the "SA" syllable, since a syllabic structure like "AS" does not exist in Japanese [14].

The singing synthesis technique employed was the concatenative one, the same of the Vocaloid software, and the instrument is already in commercialization. In respect to hardware, an Arduino board was one of the used technologies, at least in the prototyping phase [4].

4. Systematic review

The systematic review of literature consisted, in first place, in a search performed on the Scopus scientific database, with the following search string:

TITLE-ABS-KEY (("singing synthesis " OR "singing voice synthesis") AND ("REAL TIME" OR "REAL-TIME"))

This search returned nineteen records, and the works were selected according to the following criteria: (i) The work must have been published in the last ten years; (ii) The work must describe a new product.

Six works were excluded by the chronologic criterion; two of them did not describe a new product, but evaluation methods; finally, other three records were excluded because they were repeated. Searches were made in other scientific databases, like IEEE Xplore and ACM, but they did not return different results. Hence, eight articles were selected and evaluated. A brief description of each one of them follows.

FEUGÈRE *et al.* (2017) [15] present a system called Cantor Digitalis, whose input method in denominated chironomy and consists in an analogy between hand movements and the phonetic and musical parameters required by singing synthesis. The system performs formant synthesis, which

produces only vowels. With one of the hands, the user touches a tablet with a stylus, in order to indicate the wanted melodic line; simultaneously, the vowel to be synthesized is indicated by gestures made by the fingers of the other hand on the tablet.

LE BEUX *et al*. (2011) [16] proposed the integration of several instances of Cantor Digitalis by means of an environment called Méta-Mallete, which allows to execute simultaneously several computational musical instruments on the same computer through USB interfaces. Since several singing synthesizers could be used at the same time, it was possible to present a virtual choir, which was denominated Chorus Digitalis.

DELALEZ and D'ALESSANDRO (2017) [6] used the interface of Cantor Digitalis and connected it to pedals in order to build another system, called VOKinesiS, which transforms pre-recorded voice samples by means of a pitch control provided by Cantor Digitalis, while the pedals indicate timing parameters that change the rhythm of the original samples.

The work of CHAN et al. (2016) [12] describes the development of a real-time synthesizer called SERAPHIM that intends to overcome certain limitations of Cantor Digiatlis — which produces only vowels — and of Vocaloid Keyboard, whose real-time synthesis capabilities are at frame (syllable) level, but not at content (phoneme) level. SERAPHIM system provides a gestural input that allows to synthesize phoneme by phoneme, either vowels or consonants, in real time. The technique employed is sample-based concatenative synthesis, with a singing library stored in indexed structures denominated wavetables.

The I²R Speech2Singing system, developed by DONG et al. (2014) [17], instantly converts a voice input into singing, through the application of characteristics of the voices of professional singers, stored in its database, over the user voice. Hence, this system employs a data-driven approach, where the parameters are extracted from a signal and applied into another one.

MORISE *et al.* (2009) [18] developed an interface denominated v.morish'09, which also provides the transformation of a voice signal that serves as input according to characteristics extracted from a professional singer's voice.

The synthesizer of GU and LIAO (2011) [19] is a system embedded in a robot designed to present singing abilities. The system uses the harmonic plus noise model (HNM) in order to adjust parameters. The pre-recorded 408 syllables of Mandarin Chinese language serve as target signals.

YU (2017) [20] uses data-driven approach with HMM in order to develop his synthesizer, with an additional feature: the system it is integrated to a 3D animation which articulates mouth movements.

In the last two works the musical parameters are provided by a static file which contains musical notes. The real-time nature of these systems is related to the operations of the robot and the 3D animation.

In next section, a brief comparative analysis among the selected works will be made in order to answer the proposed research questions.

5. Comparative analysis

The research questions of this work were presented in Section 2. The first of then was "What are the singing synthesis techniques employed by most of the real-time systems?".

To answer it, the following comparative chart present the technical approaches used by real-time singing voice synthesizers described by each one of the works selected in the systematic review, with the addition of Vocaloid Keyboard, which was the result of the technological mapping, having as reference the paper of KAGAMI *et al*. (2012) [4]. The articles appear in chronological order.

Article	Rule-based approaches	Sample-based approaches	Data-driven approaches
MORISE *et al.* (2009) [18]			✓
GU; LIAO (2011) [19]			✓
LE BEUX *et al*. (2011) [16]	✓		
KAGAMI *et al*. (2012) [4]		✓	
DONG *et al*. (2014) [17]			✓
CHAN *et al*. (2016) [12]		✓	
DELALEZ (2017) [6]		✓	
FEUGÈRE (2017) [15]	✓		
YU (2017) [20]			✓

Chart 1. Technical approaches for real-time singing synthesis discussed by the selected works.

Among the nine evaluated works, four employed a data-driven approach; three used a sample-based one; finally, two of them used a rule-based approach. In a such restricted universe, it is possible to assert that all the main approaches of singing synthesis in general are relevant for the specific branch of real-time singing synthesis.

A geographical hypothesis could explain such equilibrium: works [16] and [15] were produced in European institutes that, under the influence of IRCAM, developed Cantor Digitalis synthesizer using the formant synthesis technique. The paper [6]

also employed features of Cantor Digitalis, but in order to overcome the limitation of providing only vowels, it needed to use samples, so the Cantor Digitalis interfaces only served to control certain parameters.

In Asia, data-driven approach is prevalent, as works [18], [19], [17] and [12] indicate. For its turn, sample-based approach continues to be promoted by Yamaha, with the development of Vocaloid Keyboard [4]. The SERPHIM synthesizer [12] was developed using sample-based approach as it takes Vocaloid Keyboard as reference.

The other research question proposed by the present work was about the input methods employed by the synthesizers. It is a critical element in systems that provide a real-time performance. The selected works present four basic input types: static files, musical instruments, electronic devices (tablets, for example) and voice signals. Chart 2 present a comparison among the articles in relation to this aspect.

Article	Static files	Musical instruments	Electronic devices	Voice signal
MORISE *et al.* (2009) [18]				✓
GU; LIAO (2011) [19]	✓			
LE BEUX *et al*. (2011) [16]			✓	
KAGAMI *et al*. (2012) [4]		✓		
DONG *et al*. (2014) [17]			✓	✓
CHAN *et al*. (2016) [12]			✓	
DELALEZ (2017) [6]			✓	✓
FEUGÈRE (2017) [15]			✓	
YU (2017) [20]	✓			

Chart 2. Input method used by the singing synthesizers described in the selected works.

The option for static files was made by systems where the synthesized singing voice worked as a real-time controller of other elements: a robot in [19] and a 3D facial animation in [20].

In works [18], [17] e [6], a voice signal acts as input in order to provide simultaneously the phonetic and musical parameters required for singing synthesis. The systems presented by these works provide as output a synthesized voice that change or "correct" the musical imperfections of the input.

The only work whose interface consisted in a conventional musical instrument was [4], because of the nature of the proposed commercial product. It is important to remark that the combination between the musical keyboard and the textual buttons was possible because of the phonetic limitations of Japanese idiom, for which this synthesizer was designed.

In more than a half of the works [16], [17], [12], [6], [15], other hardware devices were employed as input method.

6. Challenges and future works

The main challenge of singing voice synthesis in general is to achieve naturalness to the generated chant, because, beyond any subjective aspect, the adjust of parameters that provide such characteristic requires a more complex processing than the simple extraction of data from the musical input.

In the specific case of real-time singing synthesis, one of the most complex challenges is to provide an input method that conciliate phonetic and musical data simultaneously. The present work indicated that even a human voice signal has been used in order to perform this role. On the order hand, for specific idioms, like Japanese, a conventional musical interface was successfully adapted with buttons that provide phonetic parameters.

A future work, still inedited, would be the development of a real-time singing synthesizer for Brazilian Portuguese language. The input data could be provided by a static file with phonetic data, while a MIDI keyboard would be able to provide the musical parameters during a performance.

7. Conclusion

The field of real-time singing voice synthesis is still very restricted, with a small number of works developed in comparison to other areas where embedded systems are employed, such as IoT and neural networks. All the main approaches used by singing synthesis in general are also employed by the real-time synthesizers and several solutions are adopted in order to overcome the challenges that are inherent to the input methods.

References

[1] KHAN, Najeeb Ullah; LEE, Jung Chul. HMM Based Duration Control for Singing TTS. In: Advances in Computer Science and Ubiquitous Computing. Springer, Singapore, 2015. p. 137-143.

[2] ALIVIZATOU-BARAKOU, Marilena *et al.* Intangible cultural heritage and new technologies: challenges and opportunities for cultural preservation and development. In: Mixed Reality and Gamification for Cultural Heritage. Springer, Cham, 2017. p. 129-158.

[3] KENMOCHI, Hideki. Singing synthesis as a new musical instrument. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012. p. 5385-5388.

[4] KAGAMI, Shota *et al*. Development of Realtime Japanese Vocal Keyboard. In: Information Processing Society of Japan INTERACTION, pages 837-842, 2012.

[5] MACNEILAGE, Peter. The frame/content theory of evolution of speech production. Behavioral and brain sciences, v. 21, n. 4, p. 499-511, 1998.

[6] DELALEZ, Samuel; D'ALESSANDRO, Christophe. Adjusting the frame: Biphasic performative control of speech rhythm. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 864-868, 2017.

[7] LOY, Gareth. Musimathics: the Mathematical Fundamentals of Music., MIT Press, 2011.

[8] BADER, Rolf (Ed.). Springer handbook of systematic musicology. Springer, 2018.

[9] BRUM, Leonardo Araujo Zoehler. Technical aspects of concatenation-based singing voice synthesis. Scientia Plena, v. 8, n. 3 (a), 2012.

[10] HOWARD, David. Virtual Choirs. In: The Routledge Companion to Music, Technology, and Education. Routledge, 2017. p. 305-314.

[11] OURA, Keiichiro *et al*. Recent development of the HMM-based singing voice synthesis system—Sinsy. In: Seventh ISCA Workshop on Speech Synthesis. 2010.

[12] CHAN, Paul Yaozhu, *et al*. SERAPHIM: A wavetable synthesis system with 3D lip animation for real-time speech and singing applications on mobile platforms. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 1225-1229, 2016.

[13] PETERSEN, Kai *et al*. Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, v. 64, p. 1-18, 2015.

[14] KUBOZONO, Haruo (Ed.). Handbook of Japanese phonetics and phonology. Walter de Gruyter GmbH & Co KG, 2015.

[15] FEUGÈRE, Lionel *et al*. Cantor Digitalis: chironomic parametric synthesis of singing. EURASIP Journal on Audio, Speech, and Music Processing, v. 2017, n. 1, p. 2, 2017.

[16] LE BEUX, Sylvain *et al*. Chorus digitalis: Experiments in chironomic choir singing. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 2005-2008, 2011.

[17] DONG, Minghui *et al*. I²R speech2singing perfects everyone's singing. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH pages 2148-2149, 2014.

[18] MORISE, Masanori et al. v. morish'09: A morphing-based singing design interface for vocal melodies. In: International Conference on Entertainment Computing. Springer, Berlin, Heidelberg, 2009. p. 185-190.

[19] GU, Hung-Yan; LIAO, Huang-Liang. Mandarin singing voice synthesis using an HNM based scheme. In: 2008 Congress on Image and Signal Processing. IEEE, 2008. p. 347-351.

[20] YU, Jun. A Real-Time 3D Visual Singing Synthesis: From Appearance to Internal Articulators. In: International Conference on Multimedia Modeling. Springer, Cham, 2017. p. 53-64.