Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 27, Num. 04 (2020) 118-126

RESEARCH ARTICLE

Challenges and Perspectives on Real-time Singing Voice Synthesis

Síntese de voz cantada em tempo real: desafios e perspectivas

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

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

Resumo: Este trabalho trata do estado da arte do campo da síntese de voz cantada em tempo real, apresentando seu conceito, aplicações e aspectos técnicos. Realiza um mapeamento tecnológico uma revisão sistemática de literatura no intuito de apresentar os últimos desenvolvimentos na área, perfazendo uma breve análise comparativa entre os trabalhos selecionados. Por fim, discutem-se os desafios e futuros problemas de pesquisa.
Palavras-Chave: Síntese de Voz Cantada em Tempo Real — Síntese Sonora — TTS — MIDI — Computação Musical

[1] *Universidade Federal de Sergipe, São Cristóvão, Sergipe, Brasil*
*__Corresponding author__: leonardo.brum@dcomp.ufs.br
DOI: http://dx.doi.org/10.22456/2175-2745.107292 • **Received:** 30/08/2020 • **Accepted:** 29/11/2020

1. Introduction

From a computational vision, the aim of singing voice synthesis is to 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 applications on singing voice synthesis technology have been increased by the development of real-time synthesizers, like Vocaloid Keyboard [4], whose virtual singer was 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 that serve as base for singing voice synthesis in general; Section 3 discusses about singing synthesis techniques; Section 4 presents a technological mapping of the patents registered for this area; in Section 5 the systematic review of literature is shown. Section 6 contains a comparative analysis among the selected works; Section 7 discusses the challenges and future tendencies for this field; finally, Section 8 presents a brief conclusion.

2. Theoretical framework

The problem domain on singing voice synthesis is multidisciplinary: beyond computer science, it depends on concepts from acoustics, phonetics, music theory and signal processing. The following subsections present some concepts from each mentioned knowledge area.

2.1 Elements of Acoustics

Sound is a physical phenomenon produced by a variation of air pressure levels over time, coming from a vibratory source, for example, and a guitar string. Due to its undulatory

nature, sound is measured by physical quantities like period, frequency, and amplitude. Period consists in the duration of a complete wave cycle; frequency is the inverse of period, and indicates the quantity of cycles per second that the sound wave presents; amplitude is the maximum value of the pressure variation in relation to the equilibrium point of the oscillation [5]. The period and amplitude of a simple soundwave can be viewed in Figure 1.

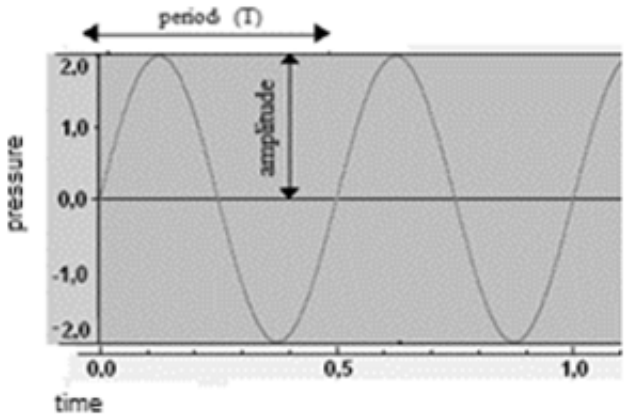

Figure 1. A simple soundwave.

The simplest soundwaves are the so-called sinusoids, which consist in a single frequency, but this kind of sound is produced neither by nature, nor by conventional musical instruments. Such sources generate complex sounds, composed by several frequencies. The lowest of them is called fundamental frequency. When the values of the other frequencies are multiples of the fundamental one, the sound is said to be periodic. When the opposite is true, the sound is called aperiodic. The superposition of frequencies results in a waveshape, which is typical for each sound source. It describes a curve called envelope, obtained from the maximum values of the waveshape oscillation.

The envelope shape is commonly decomposed into four phases, denominated by the following abbreviation ADSR. ADSR means:

(i) attack, which corresponds to the time between the beginning of the execution of the sound and its maximum amplitude;

(ii) decay, the necessary time from the maximum amplitude to a constant one;

(iii) sustain, the interval of time where the amplitude keeps such constant state; and

(iv) release, from the end of the constant state to the return to silence [5].

Figure 2 shows a schematic chart with an envelope shape and its four phases indicated by its initial letters.

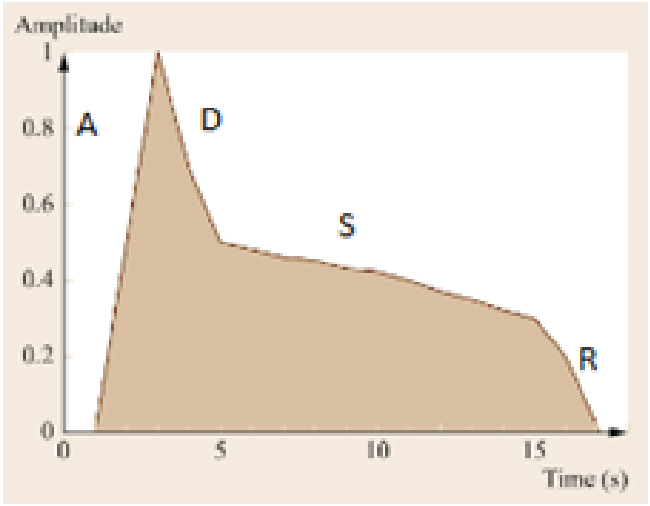

Figure 2. Envelope shape and its four phases.[5]

2.2 Sound qualities and music theory

In regard of the perception of the sound phenomenon by the human sense of audition, the physical components previously described have a relation with the so-called sound qualities:

pitch, which permits us to distinguish between bass and treble sounds, is directly proportional to the fundamental frequency;

intensity, which depends on amplitude and allows us to stablish the difference between strong sounds and weak ones;

The timbre, related to the waveshape and its envelope, is the quality that makes possible the perception of each sound source as they have their own “voice”.

The sounds of a piano and a guitar, for example, are distinguishable by their timbre, even if they have the same pitch and intensity. The fourth quality of sound is its duration over time.

It is important to remark that complex periodic sounds are usually perceived as they have a definite pitch, which corresponds to the fundamental frequency, for these sounds are called musical sounds. On the other hand, aperiodic sounds do not have a distinguishable pitch and they are denominated as noise, although they are also employed in music, especially in percussion instruments.

Since the development of the music theory in the Western World, some ways of selection and organization of musical sounds for artistic purposes are established, as well graphical representations of them, according to their qualities. The sensation of similitude between a certain sound and another one with twice the value of the fundamental frequency of the first, allowed the division of the range of audible frequency into musical scales, composed by individual sounds called by the music theory as notes. The succession of sounds with different pitches or, in other words, the succession of different musical notes is the part of music denominated melody.

The rhythm is another part of music and it can be defined as the movement of sounds regulated by their duration. The duration of notes is measured by and arbitrary unit called

tempo, based on the proportion that the notes keep in relation to each other's durations. The tempos are grouped in equal portions into structures called bars.

In modern musical notation, melodies are written in two dimensions: symbols with different shapes, called musical figures, indicate the duration of sounds. Their succession is made horizontally in a staff, which is a set of five lines and four spaces that indicate, according to the position that the musical figures occupy in it vertically, the different pitches. Figure 3 shows a melody written in a staff.

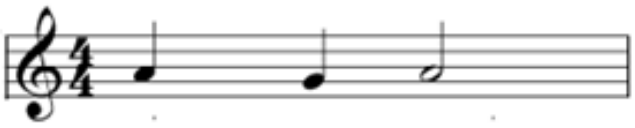

Figure 3. Melody written in a staff.

The quality of intensity is treated by a part of music theory called dynamics. It is indicated in musical notation by symbols located below the staff. A more complete exposition about music theory can be found in [6].

2.3 The human voice: notions of phonetics

Human voice is produced by the phonatory apparatus, which is composed by respiratory system, vocal tract, and vocal cords. The air coming from lungs serve as energy source, meanwhile the vocal cords, which consist in muscles, membranes, and mucosa, plays the role of a vibratory element in order to produce sound. The space between the vocal cords, called glottis, adjusts the frequency and amplitude of such sound, which permits us to sing and speak with intonation. By its turn, vocal tract, which is a set of cavities (nasal cavity, mouth, pharynx, and larynx), serve as resonance structure for voice. A scheme of the vocal tract is shown in Figure 4.

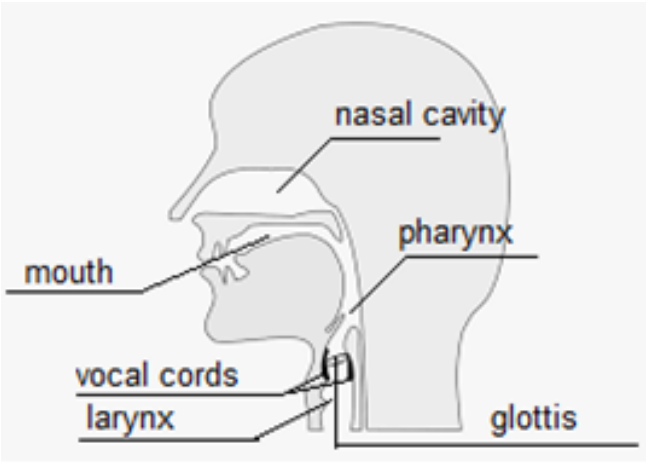

Figure 4. Basic scheme of the vocal tract.

In regard of speech, its minimum distinguishable unit is called phoneme. Phonemes are classified into two great groups: consonants and vowels. Acoustically, both groups are composed by complex sounds, but consonants are aperiodic vibrations that result from the obstruction of the airflow by body parts such as lips or tongue. On the other hand, vowels have a periodic nature.

Another difference between vowels and consonants is related to the role that each one plays in the syllable, which is a unit whose definition is difficult to be given, but it can be described as a sonorous chain composed by acclivities, peaks and declivities of sonority. According to the frame/content theory [7], 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 [8]. Furthermore, there are phonemes called semivowels, which are present in diphthongs and can be found in the syllabic boundaries. In an acoustic perspective, the syllable is a waveshape where consonants and semivowels occupy the phases of attack (acclivity) and release (declivity), while vowels are in sustain phase (peak or nucleus). In the singing voice, musical notes are present in vowels, for their periodic nature. Figure 5 shows the relation between the Brazilian Portuguese syllable "pai" and the phases of the envelope shape.

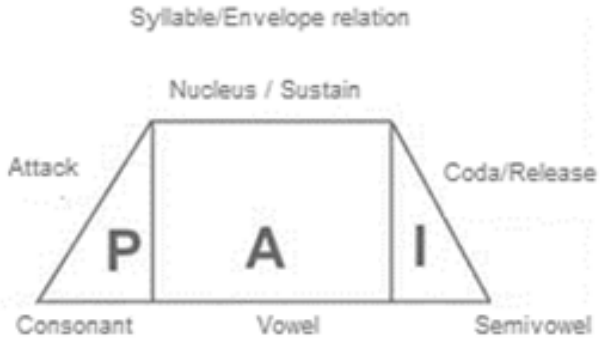

Figure 5. Relation between a syllable structure and the envelope shape stages.

Vowels also have another important acoustic feature, which is a result of the resonances produced by the vocal tract. It is called formant and appears as energy peaks when the acoustical signal is analyzed in frequency domain, related to its spectrum. The spectral representation puts in evidence the most relevant frequencies in a complex sound in terms of amplitude level. Formants correspond to the value of the central frequency of each energy peak that appears in a curve called spectral envelope. It is common to designate formants as F1, F2, F3 and so on, from the lower central frequencies to the higher ones. [9] Figure 6 shows the spectrum of a voice signal where four formants are indicated.

2.4 Audio signals processing

Sound, as a variation of a physical quantity, can be treated as a signal and be handled, stored, and transmitted by means

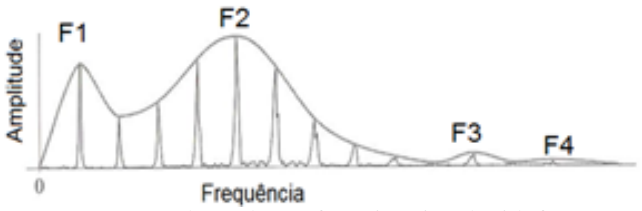

Figure 6. Spectral envelope of a voice signal with four formants indicated. [9]

of physical processes. One way to do so is perform the conversion of the acoustic signal into an analog electrical one, which is called audio signal, through a device called transducer, for example, a microphone. The air pressure variations are converted by the transducer to voltage level variations, resulting in a signal that have the same duration, waveshape and frequency composition. Such electrical signal can by transformed again into sound by the excitation of a speaker [10].

The electrical signal of sound has its own envelope, where rise time, stationary regime, and fall time correspond, respectively to attack, sustain and release phases of the envelope of the acoustic signal, so the timbre of the sound can be electrically handled. Pitch is given by the fundamental frequency of the electrical signal, while intensity is proportional to the output power of the speaker.

In 1964, Robert Moog develops an analog synthesizer composed by a set of modules, each one of them with a specific function: VCO (Voltage Controlled Oscillator), which were activated by a musical keyboard. It generates a wave with a certain fundamental frequency, according to the musical note that was played, giving pitch to the sound; VCA (Voltage Controlled Amplifier), that were connected to the VCO output and amplified its signal, controlling intensity; EG (Envelope generator), whose function was modify the signal coming from VCA, according to ADSR parameters controlled by a panel, performing an amplitude modulation in order to stablish the sound timbre. Figure 7 shows a diagram of modules of the Moog's analog synthesizer.

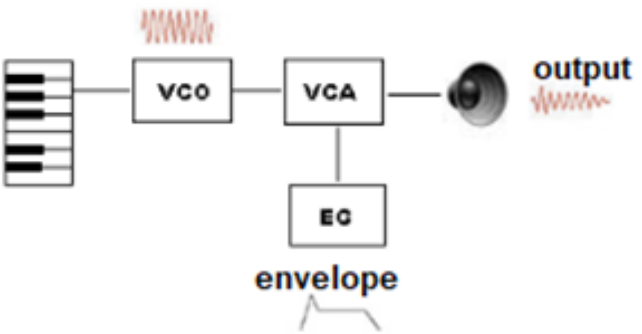

Figure 7. Diagram of modules of Moog's analog synthesizer. [9]

The analog audio signal is continuous and requires a discretization in order to be computationally treated. A device denominated analog-to-digital converter, which is present, for example, in the sound card of personal computers, makes such processing. This device makes periodic measures of the variations of the continuous electrical signal and generate an approximated discrete signal according to the samples collected. To avoid significative loses in discretization Analog-to-digital converters perform their measurements based on the sampling theorem, developed by Shannon and Nyquist. According to this theorem, the sample rate must have at least twice the value of the highest frequency present in the original signal. Since the maximum limit of human audition is about 20 kHz, a sample rate about 40000 samples per second or 40 kHz is sufficient for a digital file to faithfully represent any audible sound. The standard sample rate used by audio CDs, for example us 44100 Hz [10].

The digitalization of sounds increases the possibilities of manipulation of this kind of signal, including the development of new synthesis techniques. Even the musical synthesizer keyboards begin to embed digital circuits in opposition to the old analog modules. However, digital sound synthesis is not performed directly by hardware devices. Several kinds of software have been developed over the last decades with this end.

Speech synthesis or text-to-speech (TTS) is a field of digital sound synthesis with a great variety of applications, from translation software to accessibility interfaces. In TTS, a textual input is converted into a digital audio file that simulates human voice. One way to perform this kind of synthesis is to work with real samples of pre-recorded voice. The greatest challenge of this approach is that bigger either the size of the samples, more natural the result will sound, but the range of possible expressions will be smaller, or the number of recordings must grow considerably up. Therefore, for example, a speech synthesizer for Brazilian Portuguese that has samples word by word will demand hundreds of thousands of recordings. On the other hand, if the samples are recorded phoneme by phoneme, the recordings will be less than one hundred and could represent any word, but the result of their concatenation will sound invariably "robotic". [2]

A similar technique is used for musical purposes. It is known as sample-based synthesis. While the techniques previously presented consist in an imitation, applying in a generated signal the features of the original timbre, sample-based synthesis handles real musical instruments recordings that are stored and executed according to the necessities of the musician.

Musical notes are generated from other ones by manipulation of pitch. Duration is treated in the following way: if the execution of the note is smaller than the duration of the sample, the execution is stopped; in the opposite case, there are two possibilities, depending on the instrument whose sound is to be reproduced. For some instruments, like piano, the recording will be played until its end, with the extinction of the sound; for other instruments, like organ or flute, it is de-

sirable that their execution could be indefinitely extended, as long as a musical keyboard or some software activates the musical note. This indefinite extension is a result of the looping technique, where a certain part of the sustain phase of the waveshape is continuously played. The end of the activation of the note makes the reproduction finish the loop towards the release phase. [10]

One of the most widely used technologies to perform sample-based synthesis is MIDI (acronym for Music Interface Digital Instrument), developed by Yamaha in the decade of 1980. MIDI is a protocol for communication between electronic musical instruments and computers. Its messages and its file format do not carry any audio signal, only musical parameters corresponding to sound qualities: musical note/pitch; dynamics/intensity, instrument/timbre, duration, and so on. Such parameters serve to do an ad hoc manipulation of the sound samples, which can be stored in a computer or even in the musical instrument.

3. Singing synthesis techniques

Singing voice synthesis has two elements as input data: (i) the lyrics of the song that will be synthesized and (ii) 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 that contains the specified chant.

In respect of data processing, the main techniques on singing voice synthesis 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, which uses pre-recorded audio. The data-driven approach has been developed recently. It uses statistical models.

3.1 Rule-based approaches

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

Formant synthesis is an example of rule-based singing voice synthesis approach. It consists in the generation of units called Forme d'Onde Formantique (FOF, French for formant waveform). FOFs are sinusoid 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 that are summed in order to generate the synthesized singing voice [9].

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].

3.2 Concatenative synthesis

Concatenative singing synthesis is a kind of sample-based synthesis. Its input data are the lyrics of the song associated to a melody, where each syllable correspond to a single musical note, so the phonetic samples are successively concatenated 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 [11].

Pre-recorded samples are commonly stored in a singing library that 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 musical notes, but this fact drive the first one does not exclude the difficulties to obtain a "realistic" result from samples in singing synthesis [12].

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. Figure 8 shows a system diagram of Vocaloid.

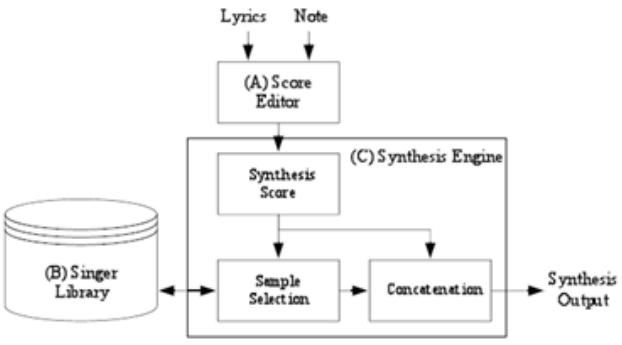

Figure 8. System diagram of Vocaloid [3]

3.3 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 [13], 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.

3.4 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 [14].

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 [15]. 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.

4. 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. Figure 9 presents the geographical distribution of the patents, where EPO is the European Patent Office.

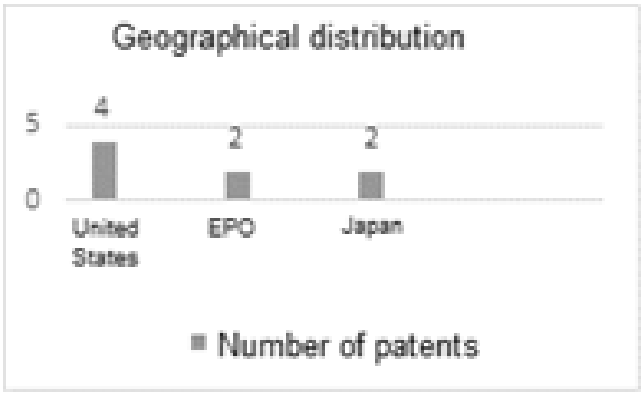

Figure 9. Patents geographical distribution.

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 10 shows the described prototype.

Figure 10. 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 [16].

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 hard-

ware, an Arduino board was one of the used technologies, at least in the prototyping phase [4].

5. Systematic review

The systematic review of literature, made in 2019, 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) [17] 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) [18] 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) [8] 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) [14] 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 synthesizing 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) [19], 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) [20] 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) [21] 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) [22] 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 that 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.

6. Comparative analysis

The research questions of this work were presented in Section 3. 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 table 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.

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 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 [18] and [17] were produced in European institutes that, under the influence of IRCAM, developed Cantor Digitalis synthesizer using the formant synthesis technique. The paper [8] also employed features of Cantor Digitalis, but in order to overcome the limitation of providing only vowels, it needed to use samples. Therefore, the Cantor Digitalis interfaces only served to control certain parameters.

In Asia, data-driven approach is prevalent, as works [20], [21], [19] and [14] indicate. For its turn, sample-based ap-

Article	Rule-based approaches	Sample-based approaches	Data-driven approaches
MORISE *et al.* (2009) [20]			✓
GU; LIAO (2011) [21]			✓
LE BEUX *et al.* (2011) [18]	✓		
KAGAMI (2012) [4]		✓	
DONG *et al.* (2014) [19]			✓
CHAN *et al.* (2016) [14]		✓	
DELALEZ (2017) [8]		✓	
FEUGÈRE (2017) [17]	✓		
YU (2017) [22]			✓

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

Article	Static files	Musical instruments	Electronic devices	Voice signal
MORISE *et al.* (2009) [20]				✓
GU; LIAO (2011) [21]	✓			
LE BEUX *et al.* (2011) [18]			✓	
KAGAMI (2012) [4]		✓		
DONG *et al.* (2014) [19]			✓	✓
CHAN *et al.* (2016) [14]			✓	
DELALEZ (2017) [8]			✓	✓
FEUGÈRE (2017) [17]			✓	
YU (2017) [22]	✓			

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

proach continues to be promoted by Yamaha, with the development of Vocaloid Keyboard [4]. The SERPHIM synthesizer [14] 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. Table 2 presents a comparison among the articles in relation to this aspect.

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 [21] and a 3D facial animation in [22].

In works [20], [19] e [8], 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 [18], [19], [14], [8], [17], other hardware devices were employed as input method.

7. 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.

The development of a real-time singing synthesizer prototype for Brazilian Portuguese language is the aim of a work in progress at Federal University of Sergipe (UFS). The input data for the synthesizer is provided by a static file with phonetic data written in SAMPA, a notation that is used in commercial singig synthesizers, like Virtual Singer [11], while a MIDI keyboard provides the musical parameters during a performance.

Both inputs, musical and phonetic, will be read by a computational device that executes the synthesizer, called PATRICIA (a Portuguese acronym for "program that articulates in real-time the singing idiom written in a file"). For each musical note activated by the MIDI keyboard, a syllable is read from the text file. The system performs a concatenative synthesis with such parameters by the retrieving of sound samples stored in a singing library. When these samples are played, the singing voiced is generated. The implementation of the system will be made in a environment for real-time audio synthesis and the device intended for execution is a Raspberry Pi computer. The system diagram of PATRICIA is shown in Figure 11.

8. 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. The real-time synthesizers in general also employ all the main approaches used by

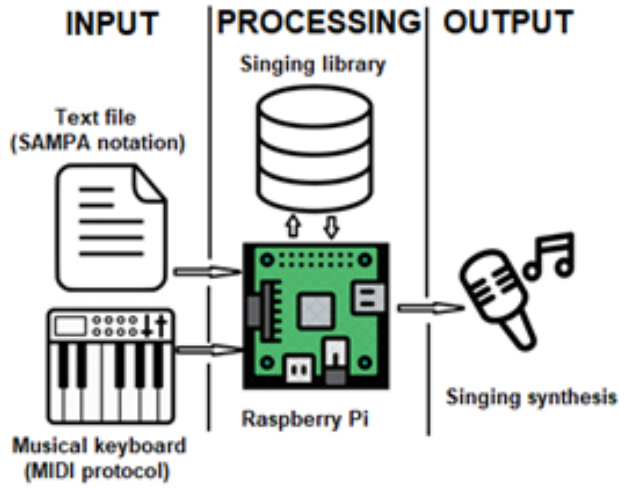

Figure 11. System diagram of PATRICIA.

singing synthesis and several solutions are adopted in order to overcome the challenges that are inherent to the input methods.

Author contributions

Mr. Leonardo Araujo Zoehler Brum participated in all experiments, coordinated the data analysis and contributed to the writing of the manuscript. Dr. Edward David Moreno coordinated the experiments and contributed to the writing of the manuscript.

References

[1] KHAN, N. U.; LEE, J. C. HMM Based Duration Control for Singing TTS. In: *Advances in Computer Science and Ubiquitous Computing*. [S.l.]: Springer, 2015. p. 137–143.

[2] ALIVIZATOU-BARAKOU, M. et al. Intangible cultural heritage and new technologies: challenges and opportunities for cultural preservation and development. In: *Mixed reality and gamification for cultural heritage*. [S.l.]: Springer, 2017. p. 129–158.

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

[4] KAGAMI, S. et al. Development of realtime japanese vocal keyboard. *Information Processing Society of Japan INTERACTION*, p. 837–842, 2012.

[5] BADER, R. *Springer handbook of systematic musicology*. [S.l.]: Springer, 2018.

[6] BLATTER, A. *Revisiting Music Theory: Basic Principles*. [S.l.]: Taylor & Francis, 2016.

[7] MACNEILAGE, P. F. The frame/content theory of evolution of speech production. *Behavioral and brain sciences*, Cambridge University Press, v. 21, n. 4, p. 499–511, 1998.

[8] DELALEZ, S.; D'ALESSANDRO, C. Adjusting the frame: Biphasic performative control of speech rhythm. In: *Proceedings of Interspeech 2017*. [S.l.: s.n.], 2017. p. 864–868.

[9] LOY, G. *Musimathics: the mathematical foundations of music*. [S.l.]: MIT press, 2011. v. 2.

[10] RUSS, M. *Sound synthesis and sampling*. [S.l.]: Taylor & Francis, 2004.

[11] BRUM, L. A. Z. Technical aspects of concatenation-based singing voice synthesis. *Scientia Plena*, v. 8, n. 3 (a), 2012.

[12] HOWARD, D. Virtual choirs. In: *The Routledge Companion to Music, Technology, and Education*. [S.l.]: Routledge, 2017. p. 305–314.

[13] OURA, K. et al. Recent development of the hmm-based singing voice synthesis system—sinsy. In: *Seventh ISCA Workshop on Speech Synthesis*. [S.l.: s.n.], 2010.

[14] CHAN, P. Y. et al. SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms. In: *INTERSPEECH*. [S.l.: s.n.], 2016. p. 1225–1229.

[15] PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, Elsevier, v. 64, p. 1–18, 2015.

[16] KUBOZONO, H. *Handbook of Japanese phonetics and phonology*. [S.l.]: Walter de Gruyter GmbH & Co KG, 2015. v. 2.

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

[18] BEUX, S. L.; FEUGERE, L.; D'ALESSANDRO, C. Chorus digitalis: experiment in chironomic choir singing. In: . [S.l.: s.n.], 2011.

[19] DONG, M. et al. I2r speech2singing perfects everyone's singing. In: *Fifteenth Annual Conference of the International Speech Communication Association*. [S.l.: s.n.], 2014.

[20] MORISE, M. et al. v. morish'09: A morphing-based singing design interface for vocal melodies. In: SPRINGER. *International Conference on Entertainment Computing*. [S.l.], 2009. p. 185–190.

[21] GU, H.-Y.; LIAO, H.-L. Mandarin singing voice synthesis using an hnm based scheme. In: IEEE. *2008 Congress on Image and Signal Processing*. [S.l.], 2008. v. 5, p. 347–351.

[22] YU, J. A real-time 3d visual singing synthesis: From appearance to internal articulators. In: SPRINGER. *International Conference on Multimedia Modeling*. [S.l.], 2017. p. 53–64.