Música Hodie | ISSN: 2317-6776 DOI: 10.5216/mh.v26.85249

Development and evaluation of PATRICIA: a real-time singing synthesizer for the Brazilian Portuguese language

PATRICIA: implementação e validação de prova de conceito para um sintetizador de canto em tempo real

Leonardo Araujo Zoehler Brum
LaSiD (Universidade Federal da Bahia, Salvador, Bahia, Brazil)
leonardobrum@ufba.br

Eduardo A. L. Meneses
METALAB (Société des Arts Technologiques, Montreal, Quebec, Canada)
emeneses@sat.qc.ca

Edward David Moreno Ordonez
PROCC (Universidade Federal de Sergipe, São Cristóvão, Sergipe, Brazil)
edward@dcomp.ufs.br

Abstract: Singing voice synthesis involves the artificial generation of a song given its lyrics and musical notes through computational methods. This paper describes the development and evaluation of PATRICIA, a proof-of-concept prototype of a system that performs real-time singing voice synthesis (SVS) for Brazilian Portuguese. It is the first singing synthesizer specifically designed for this idiom. A literature review was conducted to study the latest developments in real-time singing synthesis and to give directions for PATRICIA's design and implementation. The implemented architecture is described, and, for the validation process, a series of video recordings of the system working was made as a musical demonstration. This demonstration was evaluated by musical educators. A performance analysis with two CPU usage indicators in two different devices was also conducted. The validation results

are discussed, and future enhancements are outlined to overcome the system's current limitations.

Keywords: computer music. singing synthesis. real-time systems. Brazilian Portuguese.

Resumo: A síntese de canto consiste na geração artificial de uma canção, dadas a sua letra e notas musicais, por meio de métodos computacionais. Este artigo descreve o processo de desenvolvimento e avaliação de PATRICIA, um protótipo em fase de prova de conceito de um sistema que realiza a síntese de canto em tempo real em português brasileiro. Trata-se do primeiro sintetizador de canto especificamente projetado para esse idioma. Uma revisão de literatura foi conduzida a fim de estudar os mais recentes desenvolvimentos na área de síntese de canto em tempo real e fornecer diretrizes para o projeto e implementação de PATRICIA. A arquitetura implementada é descrita e, para o processo de validação, foi realizada uma série de gravações de vídeo do sistema em funcionamento, a título de demonstração musical. Esta demonstração foi avaliada por educadores musicais. Uma análise de desempenho com dois indicadores de uso de CPU em dois dispositivos distintos também foi conduzida. Os resultados da validação são discutidos e apontadas futuras melhorias, a fim de que o sistema supere suas atuais limitações.

Palavras-chave: computação musical. síntese de voz cantada. Sistemas de tempo real. português brasileiro

Submitted on: February 23, 2026

Accepted on: March 17, 2026

Published on: May 2026

1. Introduction

The goal of singing voice synthesis (SVS) is to generate singing by means of computational methods. Regarding the input, SVS systems deal with two types of data: phonetic data, corresponding to the lyrics of the song to be synthesized, and musical data, which provides sound qualities, e.g., pitch and duration, to the synthesized voice.

Singing synthesizers can be applied, for example, in the educational field (Geogarki, 2004, p. 5; Li, 2022, p. 1). Digital files containing singing can be easily created, shared and modified for learning purposes, dispensing with the human presence of a singer as a reference, or even recordings. Another application of this type of synthesis is in the artistic field itself. Investments in virtual singers, such as Hatsune Miku, in Japan, have been made, creating a new cultural phenomenon (Kärki, 2021; Hui, 2022). In addition, singing voice synthesizers can assist the process of musical composition, acting as simulators (Kenmochi, 2012, p. 5385). The game and entertainment industries have also demanded innovations in the field of SVS (Hwang; Lee; Lee, 2025, p. 1).

The possibilities of use of SVS have been extended by the development of real-time synthesizers. This kind of SVS allows the use of the singing synthesizer as a musical instrument for live artistic performances. However, only a small number of real-time synthesizers have been developed in the last years compared to the universe of SVS, in general (Brum; Moreno, 2019).

Furthermore, it was observed both in industry and academia the absence of tools that provide SVS for the Brazilian Portuguese language. This paper describes the implementation and validation processes of a proof-of-concept prototype of a real-time singing synthesizer called PATRICIA, which was designed for the idiom.

This work is organized as follows: Section 2 provides a background for SVS, discussing its technical approaches and evaluation strategies. Section 3 presents some real-time SVS systems as related work. Section 4 describes PATRICIA's

implementation. Section 5 presents the experiments conducted to validate the proof-of-concept prototype. Section 6 discusses the results of the validation process. Finally, Section 7 presents a brief conclusion and points out future work.

2. Singing synthesis background

SVS has had a series of paradigm changes over the years. The several techniques developed to perform this task can be classified according to three paradigms or approaches (Kim, 2008, p. 370-371): rules-based approach, concatenation synthesis, and data-driven approach, in chronological order. These approaches and the evaluation strategies for SVS systems are discussed in this section.

2.1 Rules-based approach

The rules-based approach emerges in the early era of SVS. In this approach, the singing voice is generated in a completely artificial manner. It considers the way sound is produced, generating voice by means of physical models, or describes perceptual aspects of the human voice through signal models.

The first case is on the basis of the articulatory synthesis technique, which consists of controlling a set of parameters related to the production of voice by the organs of the vocal apparatus. Therefore, an articulatory synthesizer produces voice by changing the physical mechanisms modelled on the system, such as position of the tongue and lips, opening of the oral cavity, and throat wall radiation, among others. An implementation of this technique can be found in systems such as SPASM (Cook, 1993), which has an interface with a cross-sectional representation of a human head.

On the other hand, formant synthesis is an example of a vocal synthesis technique by rules based on signal models. Formants are the result of the resonances produced by the vocal tract. They correspond to the value of the central frequency of each energy peak verified when an audio signal is analysed from its spectrum, in

frequency domain. From a perceptual perspective, formants allow the listener to distinguish the several vowels of a given language from each other.

CHANT (Rodet; Potard; Barriere, 1984) is a pioneering system among formant-singing synthesizers. It uses a model called FOF (*Formes d'Ondes Formantiques*, French for "formant wave functions" in a loose translation) (Rodet, 1980), which consists in short sinusoids whose frequency is the formant of a given vowel. FOFs are generated sequentially in time domain at a rate which corresponds to the fundamental frequency of a given musical note. The result is a singing voice with the phonetic and musical characteristics of the frequencies involved in the process.

2.2 Concatenation synthesis

In concatenation synthesis, a set of pre-recorded sample voices is stored in a database (Brum, 2012, p. 3). These samples are commonly recorded by a professional singer. The voice database consists of units which can be modelled to contain one or more phonemes. The units are successively selected and concatenated according to the phonetic and musical inputs of the system.

An example of a system that performs concatenative SVS is Vocaloid (Kenmochi; Ohshita, 2007), developed by Yamaha, which has achieved great commercial success. It has a piano-roll interface, consisting of a virtual keyboard attached to a table whose filling corresponds to the chosen musical notes. From this musical editor, the lyrics of the song and the notes associated with them are transmitted to the synthesis engine, where they will serve as criteria for unit selection and concatenation from the system's voice database, denominated as the singing library.

2.3 Data-driven approach

In a data-driven approach, machine learning models are trained to apply the behaviors of a target audio signal to the synthesized voice. The target signal is commonly a real singing performance.

One of the first tools widely used to perform this task was Hidden Markov Models, or HMMs (Oura, 2015), which are also successfully used in speech synthesis systems (Zen; Tokuda; Kitamura, 2004). A remarkable SVS system which uses HMMs is SinSy (Oura, *et al.*, 2010), developed by the Nagoya Institute of Technology. The system is available on a website where it is possible to upload a MusicXML file. It returns the synthesized song in WAV format. The languages supported by SinSy are English and Japanese.

More recently, other machine learning models, such as neural networks, have been used to create data-driven SVS systems (Cho, *et al.*, 2021), including a new version of SinSy (Hono, *et al.*, 2018). Other examples are NNSVS (Yamamoto; Yoneyama; Toda, 2023) and DiffSinger (Liu, *et al.*, 2022).

2.4 Evaluation strategies

Evaluation strategies for SVS systems can be made from an objective or subjective perspective (Umbert, *et al.*, 2015, p. 69). In objective evaluation, a comparison between the synthesized voice and a real singing performance taken as reference is made by algorithms which analyse both audio signals and compute an error.

Subjective evaluation consists of a series of listening tests in which the participants, according to their perception, should evaluate some psychoacoustic characteristics of the synthesized singing. In this process, standards such as Mean Opinion Score (Streijl; Winkler; Hands, 2016) and Likert scale (Joshi, *et al.*, 2015) are used to measure the quality assessment.

3. Related work

All the approaches described in Section 2 are used in the development of performative, real-time singing synthesizers. In this type of SVS system, the voice is produced at the very moment the input data — phonetic, musical, or both — is provided by the

user. The present work performed a systematic literature mapping (SLM) that updates the results of a previous one about real-time SVS systems (Brum; Moreno, 2019). This method consists of a systematic protocol for searching and selecting relevant studies with the goal of extracting data and mapping results to a specific research problem (Petersen, *et al.*, 2008). The SLM process was conducted with the aid of two other complementary methods: a technological mapping (Orjuela, *et al.*, 2019) and a systematic literature review (Kitchenham, 2004).

In the technological mapping, we searched for patents related to real-time singing synthesis. The systematic literature review performs the search in scientific databases. These two methods provide, respectively, a better comprehension of the state of the technique and the state-of-the-art of the real-time branch of singing voice synthesis. Both processes are described in the following two subsections.

3.1 Technological mapping

The search for patents related to real-time singing voice synthesizers was performed in the World Intellectual Property Organization database, called Patentscope. The following search string was used:

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

This search provided eight patent deposits as a result. All of them were the property of Yamaha Corporation, from Japan, and involves an apparatus, method, and storage medium for real-time singing synthesis. However, most of the patents were registered outside Japan, probably to warrant international legal protection.

This product is denominated Vocaloid Keyboard and consists of a musical keyboard with an embedded singing synthesizer, allowing its user to make an instrumental performance. A prototype of the instrument, designed to synthesize singing in the Japanese language, was presented in a scientific paper (Kagami, *et al.*, 2012).

It has alphabetic buttons on the left, organized as follows: two horizontal rows with diacritical signs and consonants, and bellow them, five buttons for the vowels, organized in a cross shape. With left hand, the user could press these buttons to generate syllables; meanwhile, the musical keyboard could be played by the right hand to provide the musical notes. A display with *katakana* Japanese characters shows the generated syllables. Figure 1 presents the prototype described.

Figure 1 – Vocaloid Keyboard Prototype.

Source: Kagami ***et al.*** *(2012).*

Nevertheless, the commercial version of Vocaloid Keyboard, the VKB-100 keytar, abandons its prototypical phonetic buttons in favor of giving the song lyrics in advance instead of providing them in real-time (Kashiwase, 2017). The phonetic buttons were, thus, replaced by others that add expression to the song, controlled by the user's left hand.

3.2 Systematic literature review

To describe how the search and selection process for the primary studies was performed, it is necessary to define the research questions, the search and selection strategy, and the criteria for selection as follows.

3.2.1 Research questions

The proposed systematic mapping focuses on identifying and analyzing existing tools for performing real-time singing synthesis.

For that, two questions were elaborated to be answered through the analysis of the selected works:

1. What is the SVS approach employed by most of the real-time systems?
2. How do such systems retrieve the phonetic input data?

3.2.2 Search and selection strategy

The following databases of Computer Science were selected for the execution of the search: Scopus, IEEE Xplore, and Web of Science. The CAFE service, available in the CAPES (Coordination for the Improvement of Higher Level of Education Personnel) journals portal, was used to access all the works present in the three databases.

A string was developed according to the terms and keywords of the research objective. This standardizes the search in the mentioned databases. The string as follows:

```
((“singing”) AND (“voice synthesis” OR “vocal synthesis” OR “singing synthesis”) AND (“real time” OR “real-time” OR “performative”))
```

These searches were carried out in November 2022 and returned a total of 56 works found, following the distribution: Scopus (30), Web of Science (19) and IEEE Xplore (7). In these works, the selection phase was initiated. The paper that presents Vocaloid Keyboard (Kagami, *et al.*, 2012), found in the technological mapping described in the previous subsection, was manually added to the SLM, totaling 57 works. To filter the articles relevant to this systematic mapping, inclusion and exclusion criteria were defined. These are presented in Frame 1.

Frame 1 – Inclusion and exclusion criteria for studies selection

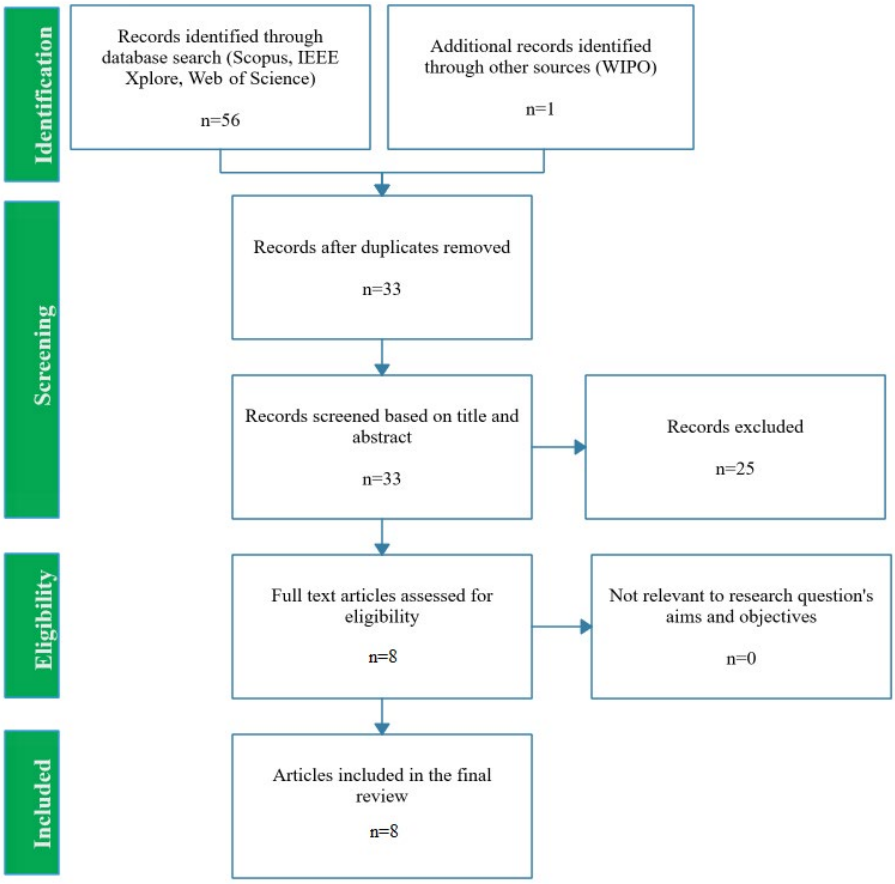
Inclusion Criteria (IC)	Exclusion Criteria (EC)
IC1- Works developed in the last 10 years (2012-2022).	EC1- Duplicated works.

IC2- Works that describe a new singing synthesizer.	EC2- Duplicated subject (same synthesizer described).
	EC3 - Works which are not fully available.

Source: Production by the author.

To illustrate the application of the criteria shown Frame 1 in this study, Figure 2 presents the selection process of the papers. After the search was performed in the scientific databases, the title and abstracts of the papers were read, and the inclusion and exclusion criteria were applied. A PRISMA — *Preferred reporting items for systematic reviews and meta-analyses* flow diagram (Moher, *et al.*, 2009) is used to present each stage of this process, the number of eligible, excluded studies, and the respective justifications.

Figure 2 - PRISMA diagram describing the article selection process.



Source: Production by the author.

3.2.3 Selected works

The SLM process resulted in eight works included in the final review. The real-time singing synthesizers described by such works are presented below, ordered by the most recent ones.

The MLP Singer (Tae; Kim; Lee, 2021) uses a multi-layer perceptron (MLP) to provide vocal synthesis in the Korean language. The perceptron network was trained with 45 songs performed by a professional singer. Each song was sung twice in different keys, accompanied by MIDI data and textual annotations, which are, respectively, the musical and phonetic input formats of the system. The syllables are divided into three parts: onset, nucleus and coda, with the nucleus occupied by the vowel, receiving the largest share of the time. The inputs pass through the layers of the neural network, which apply various transformations to them, resulting in an audio output projected in a spectrogram.

Full-Band LPCNet (Matsubsara, *et al.*, 2021) is a speech and singing synthesizer that works like a vocoder, that is, it modulates the voice signal according to another signal that provides the frequency, called a carrier wave. It uses a neural network trained with 48 songs performed acapella by a Japanese singer.

Cantor Digitalis (Feugère, *et al.*, 2017) is the only rule-based real-time singing synthesizers found by the review. It uses formant synthesis to generate vowels from a gestural input method called chironomy (D'Alessandro; Rilliard; Le Beux, 2011), described as follows by the authors: with one hand, the user activates a tablet with a stylus (Wang, *et al.*, 2013), in order to inform the desired melodic contour; at the same time, with the fingers of the other hand, he gesturally indicates on the tablet the vowel to be synthesized. The integration of several instances of Cantor Digitalis with the aim of forming a virtual choir was named Chorus Digitalis (Le Beux; Feugère; D'Alessandro, 2011).

VOKinesiS (Delalez; D'Alessandro, 2017) is a synthesizer which integrated the input interface of Cantor Digitalis with pedals. In this

system, pre-recorded voice samples are transformed using the pitch control provided by Cantor Digitalis, while the pedals provide temporal parameters that alter the rhythm of the original samples.

SERAPHIM (Chan, *et al.*, 2016) aims to overcome certain limitations of Cantor Digitalis — which only synthesizes vowels — by providing a gestural input that enables real-time phoneme-by-phoneme synthesis of both vowels and consonants. The technique used is concatenative synthesis, based on samples, which form a singing database in indexed wavetables. The system provides speech and singing synthesis for Mandarin and only singing for Japanese.

I^2R Speech2Singing (Dong, *et al.*, 2014) instantly converts a voice input into a song, applying the characteristics of the voices of professional singers present in its database to the user's voice.

MAGE/pHTS (Veaux, *et al.*, 2013) is a distributed system that uses trained HMM models to perform real-time sung voice synthesis through gesture control using the Kinect device. It uses the PureData (Puckette, 1997) real-time programming environment and allows multiple users to interactively and simultaneously control system inputs.

The Vocaloid Keyboard (Kagami, *et al.*, 2012) is a prototype implemented in the form of a musical keyboard for which the Vocaloid concatenative synthesizer works as an embedded system that generates singing in Japanese. With the left hand, the user can press alphabetic buttons to generate syllables; meanwhile, the musical keyboard can be played with the right hand to provide the musical notes. A display with Japanese katakana characters shows the generated syllables.

3.2.4 Mapping results

The first research question established in the SLM process was: "What is the approach employed by most real-time singing synthesis systems?"

In order to answer this, Frame 2 presents the technical approaches used by the singing synthesizers described for each of the selected papers. The works appear in the same chronological order as used in the previous subsection.

Frame 2 – Technical approaches used by the synthesizers studied.

Synthesizer	Rules-based	Concatenative	Data-driven
MLP Singer (Tae; Kim; Lee, 2021)			✓
Full-Band LPCNet (Matsubara, *et al.*, 2021)			✓
Cantor Digitalis (Feugère, *et al.*, 2017)	✓		
VOKinesiS (Delalez; D'Alessandro, 2017)		✓	
SERAPHIM (Chan, *et al.*, 2016)		✓	
I²R Speech2Singing (Dong, *et al.*, 2014)			✓
MAGE/pHTS (Veaux, *et al.*, 2013)			✓
Vocaloid Keyboard (Kagami, *et al.*, 2012)		✓	

Source: Production by the author.

It was verified that most of the studied singing synthesizers use the sample-based the data-driven approach (Tae; Kim; Lee, 2021; Matsubara, *et al.*, 2021; Dong, *et al.*, 2014; Veaux, *et al.*, 2013), totalling four systems, followed by the concatenative synthesizers, which are three (Delalez; D'Alessandro, 2017; Chan, *et al.*, 2016; Kagami, *et al.*, 2012). Only one synthesizer uses the rules-based approach (Feugère, *et al.*, 2017), which is the oldest of them. This fact could explain why it is less prevalent among the most recent works.

The other proposed research question was: "How do such systems retrieve the phonetic input data?". This is a critical element in systems that provide real-time performance. The studied works present three input methods: text files, audio signals, and manual controls. Frame 3 shows a comparison of this aspect in the selected papers. The order of the works remains the same.

Frame 3 – Input method for phonetic parameters.

Synthesizer	Text files	Audio signals	Manual controls
MLP Singer (Tae; Kim; Lee, 2021)	✓		
Full-Band LPCNet (Matsubara, *et al.*, 2021)		✓	
Cantor Digitalis (Feugère, *et al.*, 2017)			✓
VOKinesiS (Delalez; D'Alessandro, 2017)		✓	
SERAPHIM (Chan, *et al.*, 2016)			✓
I^2R Speech2Singing (Dong, *et al.*, 2014)		✓	
MAGE/pHTS (Veaux, *et al.*, 2013)		✓	
Vocaloid Keyboard (Kagami, *et al.*, 2012)			✓

Source: Production by the author.

Most of the synthesizers, i.e. four of them, use an audio signal to provide the phonetic input (Matsubara, *et al.*, 2021; Delalez; D'Alessandro, 2017; Dong, *et al.*, 2014; Veaux, *et al.*, 2013). Three rely on some manual control (Feugère, *et al.*, 2017; Chan, *et al.*, 2016; Kagami, *et al.*, 2012), and only one synthesizer retrieves phonetic data from a text file (Tae; Kim; Lee, 2021). A hypothesis which could explain such a result is that text files need to be previously prepared, which means that the phonetic input cannot be given in real-time. At the same time, the provided controls are very limited to represent all the possible phonemes of the several existent languages. On the other hand, an audio signal from a human voice could give the phonetic input on the fly.

4. Development of PATRICIA

To the best of our knowledge, the system presented in this paper is the first singing synthesizer specifically designed for singing in the Brazilian Portuguese language. An outline of the system's architecture was proposed in a previous work (Brum; Moreno, 2019), and a proof-of-concept prototype was presented (Brum; Meneses, Moreno, 2023). The synthesizer is called PATRICIA, an acronym for *Programa que Articula em Tempo Real o Idioma Cantado Inscrito em Arquivo* (Portuguese for Program that Articulates in Real-Time the Singing Idiom Written in a File) and was

implemented in SuperCollider (McCartney, 1996), an environment and programming language for real-time audio synthesis. GitHub was used for version control.

The development of PATRICIA sought to follow the evolutionary software prototyping model, in which the best-understood functional requirements are implemented, as opposed to rapid or throwaway prototyping (Sherrell, 2013). Hence, it is expected that the synthesizer, from a phonetic point of view, correctly selects the syllables that make up the lyrics of the song. Musically, what is desired is that the system adequately controls the qualities of the sound such as pitch and duration.

PATRICIA was initially conceived as an embedded system that performs vocal synthesis on a MIDI keyboard to provide real-time performance. This initial conception of PATRICIA found a greater correspondence with Vocaloid Keyboard among the systems selected by the scientific mapping presented in Section 3. Therefore, a concatenative synthesis approach was chosen for the implementation of PATRICIA.

However, for the proof-of-concept prototype, the complexity of the concatenative synthesis technique and the construction of the voice database were transferred to an existing system, the MBROLA speech synthesizer (Dutoit, *et al.*, 1996), which performs concatenative synthesis. MBROLA was chosen because, in addition to being free software with a wide range of voice databases available, it works through command lines, responding satisfactorily to real-time calls made by the SuperCollider audio server.

PATRICIA's architecture counts on three modules: musical processor, phonetic processor, and audio processor, which are described in detail in this section.

4.1 Musical processor

The *Note On* MIDI message has two data bytes: one of them indicates the MIDI note number, whereas the other provides the velocity, i.e., how hard the key was pressed in the musical

instrument. The musical processor module of PATRICIA deals with these two data bytes as input to generate a fundamental frequency and a loudness value on its output.

Equation 1 shows the formula used to calculate the fundamental frequency f_0, in Hertz (Hz), for each *Note On* MIDI message, given its corresponding MIDI note number *n*:

$$f_0 = \sqrt[12]{2^{n-81}} \cdot 440 \quad (1)$$

In *Note On* messages, velocity varies between 1 and 127. Since the loudness of the audio depends on the force used to play the instrument, the velocity parameter can be mapped into an intensity measured in decibels (dB), a logarithmic scale unit. The MIDI protocol specifications establish the formula shown in Equation 2 to map a velocity *V*, which varies between 1 and 127, into a loudness *L*, measured in decibels:

$$L = 40 \cdot log(V/127) \quad (2)$$

The formulas presented by Equations 1 and 2 were implemented in the musical processor module, and their results were attributed as inputs to PATRICIA's audio processor.

4.2 Phonetic processor

In the Vocaloid Keyboard prototype, both phonetic and musical inputs are given in real time by the user. This kind of performance was possible because this product was designed for the Japanese language, whose syllables structures are simpler than those of western languages. Most of Japanese syllables consists of a consonant and a vowel, in that order. Thus, when the user presses, for example, buttons "T" and "A" simultaneously in the Vocaloid Keyboard panel, the synthesis engine necessarily interprets this as a "TA" since an eventual "AT" syllable does not exist in Japanese (Labrune, 2012, p. 144). In contrast, Brazilian Portuguese has

syllables whose structure increases the challenges for using manual controllers to map a real-time phonetic input.

However, even the commercial version of Vocaloid Keyboard, the VKB-100 keytar, renounces this phonetic input method for giving the song lyrics in advance (Kashiwase, 2017). Thus, giving a previously prepared phonetic input in a text file appeared to be the most suitable method for the implementation of a Brazilian Portuguese SVS system, although it is the least used by the most recent real-time synthesizers according to the results of the scientific mapping presented in Frame 3.

Once the phonetic input method has been determined, it is necessary to discuss the format of such input. Syllables could be written according to the official Portuguese orthography, but a phonetic notation was preferred to give the user two types of control over the voice synthesis:

- The lyrics of the song are not a text in prose, but in verse, so certain phenomena that occur in versification such as elision, where two syllables merge into one, must be considered. The syllables of the verses do not necessarily coincide with the grammatical ones (Bandeira, 1960, p. 3241) and are better controlled at the phonetic level.
- Brazilian Portuguese has several regional variants. For example, the first syllable of the word "tia" (Portuguese for "aunt") can be pronounced as [ti] or [tʃi][1], depending on the region of the Brazilian speaker (Silva; Leite, 2015, p. 19), so that an orthographic syllable "ti" in the system input would be transformed by the synthesis mechanism into a specific phonetic variant chosen arbitrarily, excluding other possibilities.

In PATRICIA, each line of the lyrics text file must have a syllable written in a phonetic notation with its phonemes separated by hyphens. The phonetic notation used is similar to SAMPA (Wells, 1997), which uses only ASCII characters. It is possible to alternate

1 The syllables into brackets are represented according to the International Phonetic Alphabet notation.

among distinct text files to synthesize several songs. The selection of phonetic files is made using the Program Change MIDI message, allowing the system to handle up to 128 different song lyrics. Once a lyrics text file is selected, for each time a MIDI Note On message is received, the phonetic processor reads a line from this file and retrieves a syllable. The correspondence between a syllable and a musical note is the basis of singing synthesis.

The phonetic processor also indicates which MBROLA voice database must be used. The selection of the voice databases is made by the MIDI channel number, which can take on values between 1 and 16. MIDI channel changes can be made in real time, which makes it possible to use up to three voice timbres while playing a single song. The feasibility of this change depends on how MIDI controls are available on each device or musical instrument. Although this means that PATRICIA supports up to sixteen different MBROLA databases, only four of them are effectively used by the system. They are called *br1*, *br2*, *br3*, and *br4* and consists of voice samples in Brazilian Portuguese controlled by two different sets of phonetic symbols. The set used by *br1*, *br2*, and *br3* databases are different from the *br4* set.

The syllable retrieved from the lyrics file and the voice database selected according to the MIDI channel are also sent as input data to the audio processor, described in the next subsection.

4.3 Audio processor

As mentioned, PATRICIA uses the synthesis engine of an external program, a speech synthesis system called MBROLA. This synthesizer performs a concatenative synthesis by diphones, that is, a conjunction of two phonemes. A set of diphones of a given language is recorded and stored in a file that serves as a voice database for the system. For its turn, the speech to be synthesized can be stored in a text file with the *.PHO* extension. Each line of this file contains a phoneme written according to the set of phonetic symbols used, a duration in milliseconds, and a series of pitch

targets composed of two numbers: the position of the target within the phoneme, which is a percentage of its total duration, and the frequency value in Hertz at this position.

For each MIDI note and its corresponding syllable, it is generated a *.PHO* file. Thus, each line of this file contains these four parameters:

- A phoneme from the syllable retrieved by the phonetic processor (the *.PHO* file will have as many lines as phonemes in this syllable).
- A phoneme duration, arbitrarily set as 50 ms for semi-vowels and consonants and 200 ms for vowels.
- A pitch target position that is always 100%, which means that the pitch of the phonemes will be the same during all their duration.
- A fundamental frequency to provide the pitch, calculated by the musical processor according to Equation 1.

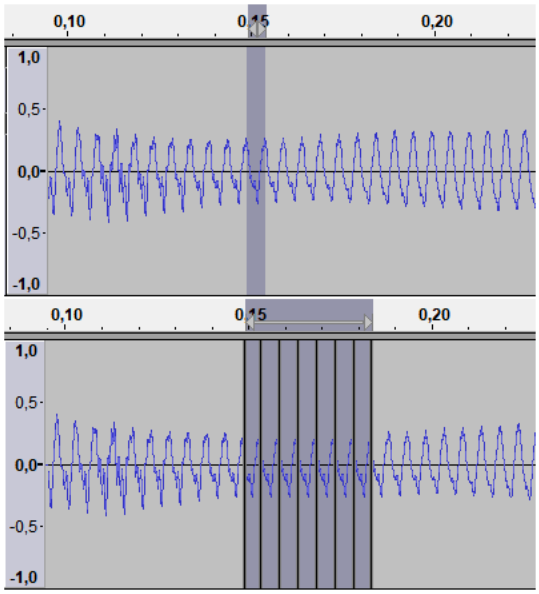
Once the *.PHO* file was created, PATRICIA uses the SuperCollider engine to send an MBROLA command line to the operating system of the host computer. This instruction references both the *.PHO* file and the MBROLA Brazilian Portuguese voice database selected by the phonetic processor. This creates the corresponding audio file in *AU* format.

Since each audio file generated by MBROLA has a constant duration, PATRICIA needs to extend it as long as the respective key is pressed in the MIDI instrument. To achieve this result, for each active MIDI note, the corresponding *.AU* file is played into a loop. The start position of the loop is in the middle of the file in the time domain, where the vowel is. The ending position is calculated by adding the period in seconds, i.e., the inverse of the fundamental frequency, to the starting position. The interval between the start and ending positions corresponds to one cycle or pitch period of the waveform. This cycle is repeated while the corresponding MIDI note is kept active, extending the duration of the syllable. When a

MIDI Note off message is received, the loop that corresponds to this note is released, and the rest of the *.AU* file is played.

Figure 3 shows a partial view of a waveform generated by MBROLA for the first syllable of the word "quando" (Portuguese for "when") singing a G_4 note. The region selected in (a) corresponds to a period of the waveform calculated by the audio processor, starting from the center of the sample. The repetition of such a segment, which appears seven times in (b), extends the duration of the audio.

Figure 3 - (a) Loop segment calculated by PATRICIA; (b) Repetition of the segment, extending the duration of the sample.

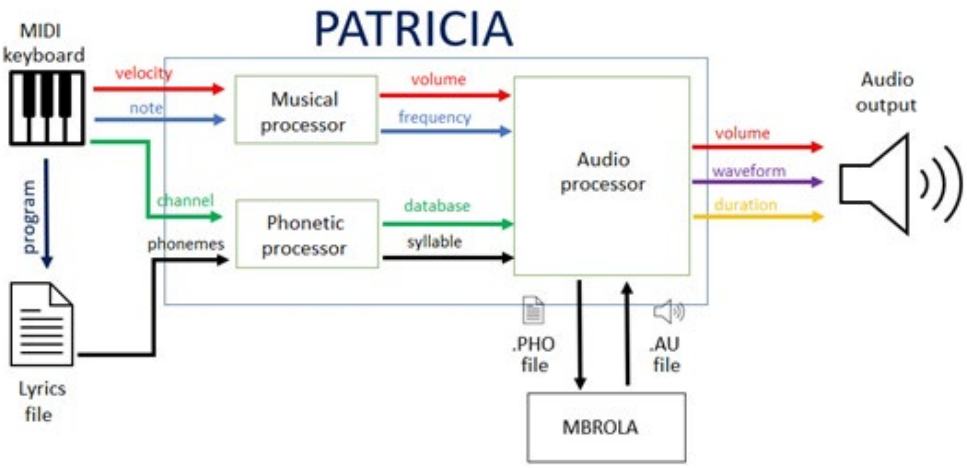


Source: Production by the author.

The successive repetition of the process described above results in generating a synthesized singing voice in the audio output of the host computer. An amplification is applied to this output according to the loudness value calculated by the musical processor as shown in Equation 2.

A system's architecture diagram of PATRICIA is shown in Figure 4, highlighting the interaction between the internal modules of PATRICIA and its external components, i.e., the input sources, the MBROLA synthesizer, and the audio output.

Figure 4 - PATRICIA system architecture diagram.



Source: Production by the author.

5. System evaluation

As a proof of concept, the singing synthesizer described in this work has only a minimum set of functional requirements implemented. These requirements are validated by the experiments described below, which sought to evaluate the musical capabilities of the system and to make a comparative analysis between its performance on a personal computer and on a single-board computer.

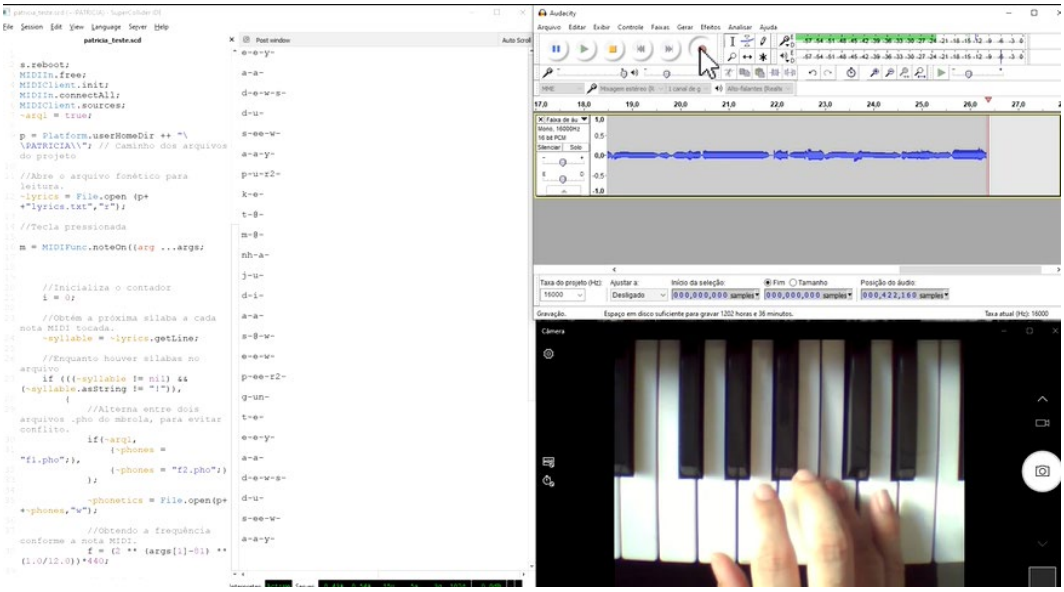
5.1 Musical demonstration

A brief demonstration of PATRICIA's musical capabilities was recorded and the resulting videos are available on the system's channel on YouTube[2]. The videos show PATRICIA running in the SuperCollider IDE. At the same time, a picture-in-picture image shows the MIDI keyboard being played at a mirrored angle, while the Audacity audio editor records the output signal. A different voice database was used in each of the videos. Figure 5 presents a frame of one of these videos.

2 **Available at** https://www.youtube.com/@ProjetoPATRICIA .

In another recording, it was possible to demonstrate potential pedagogical use of the PATRICIA system. An arrangement for a three-voice choir of the first two verses of the song “Asa Branca”, by Luiz Gonzaga and Humberto Teixeira, was made. PATRICIA was used to generate the audio for each of the voices, using the *br1* database for the tenor, who performs the main melody, *br2* for the contralto voice and *br3* for the bass. The audios, in the Audacity editor, are played separately and then together in another video on the project’s channel. When teaching choral singing, it is common for the teacher or conductor to present the melodic lines of each voice to the choir members using a musical instrument, such as a piano or keyboard. By using the PATRICIA synthesizer, it will be possible to use a musical instrument to generate not only the melodic line, but also the audio of the voice itself to be sung. Three of the songs chosen for the video recordings were taken from Brazilian children’s folklore. A fourth popular song, “Asa Branca”, by Luiz Gonzaga and Humberto Teixeira, is part of Brazilian northeastern culture. A summary of the recordings is presented in Frame 4.

Figure 5 – Frame from one of the musical demonstration videos of the PATRICIA system.



Source: Production by the author.

Frame 4 – Video recordings for demonstration

Video number	Song title	MBROLA database	Duration
1	O cravo brigou com a rosa	*br1*	56″
2	Atirei o pau no gato	*br2*	35″
3	Terezinha de Jesus	*br3*	55″
4	Asa branca	*br4*	1'35″
5	Asa branca (three-voice arrangement)	*br1, br2, and br3*	1'18″

Source: Production by the author.

5.2 Evaluation by music educators

The five musical demonstrations described in the previous subsection were made available for a subjective evaluation by five music educators who work in the State of Sergipe, Brazil.

The evaluation was carried out using a form containing twenty questions, four for each demonstration, which the educator had to watch before responding. For the demonstrations corresponding to videos 1 to 4, according to Frame 1, the four questions were as follows:

1. Do you consider that the sung syllables are understandable?
2. Do you consider that the generated song is natural?
3. Do you consider that the musical pitches generated correspond to the musician's performance?
4. Do you consider that the duration of the generated notes corresponds to the musician's performance?

These four questions aim to evaluate, respectively, the following psychoacoustic characteristics of the generated song: intelligibility (Benoît; Grice; Hazan, 1996), i.e., not only the correct selection of syllables from a technical point of view, but their effective understanding; naturalness (Morton, 1990), which would be the assessment of how close the synthesized voice would be

to a natural human voice; pitch and duration control according to the keyboard performance watched on the videos. The possible answers to the questions, corresponding to a Likert scale with four options, giving up a neutral position, are shown in Frame 5.

Frame 5 – Likert scale score for each answer.

Answer	Score
I totally agree.	4
I partially agree.	3
I partially disagree.	2
I totally disagree.	1

Source: Production by the author.

Regarding the demonstration of video 5, the last two questions were modified. The penultimate of them asked "Do you consider that the singing generated corresponds to the arrangement shown in the score above?", considering the score used for the three-voice arrangement. Finally, the last question aimed to evaluate the pedagogical usefulness of the synthesizer, by posing the following question to the educators: "Do you consider that the PATRICIA synthesizer would be a useful tool for teaching choral singing?".

The average of the scores given by each of the five users in each question corresponds to a Mean Opinion Score (MOS) for each requirement. Tables 1 to 5, below, present the results obtained for each of the demonstrations. In them, educators are identified as E1... E5 and the questions as Q1...Q4.

Table 1 – Evaluation of video 1 demonstration

	E1	E2	E3	E4	E5	MOS
Q1(intelligibility)	2	4	3	3	3	3.0
Q2(naturalness)	2	3	2	1	1	1.8
Q3(pitch control)	3	4	4	4	3	3.6
Q4(duration control)	3	4	4	3	3	3.4

Source: Production by the author.

Table 2 – Evaluation of video 2 demonstration

	E1	E2	E3	E4	E5	MOS
Q1(intelligibility)	3	4	3	3	3	3.2
Q2(naturalness)	2	3	3	2	1	2.2
Q3(pitch control)	3	4	4	4	4	3.8
Q4(duration control)	3	4	4	4	4	3.8

Source: Production by the author.

Table 3 – Evaluation of video 3 demonstration

	E1	E2	E3	E4	E5	MOS
Q1(intelligibility)	2	3	2	3	1	2.2
Q2(naturalness)	2	3	2	1	1	1.8
Q3(pitch control)	2	4	4	4	4	3.6
Q4(duration control)	3	4	4	4	4	3.8

Source: Production by the author.

Table 4 – Evaluation of video 4 demonstration

	E1	E2	E3	E4	E5	MOS
Q1(intelligibility)	3	4	3	3	3	3.2
Q2(naturalness)	3	3	2	1	1	2.0
Q3(pitch control)	3	4	4	4	4	3.8
Q4(duration control)	2	4	4	3	3	3.2

Source: Production by the author.

Table 5 – Evaluation of video 5 demonstration

	E1	E2	E3	E4	E5	MOS
Q1(intelligibility)	3	3	3	3	3	3.0
Q2(naturalness)	3	3	3	2	1	2.4
Q3(conformity to the arrangement)	3	4	4	4	4	3.8
Q4(pedagogical usefulness)	4	4	4	3	2	3.4

Source: Production by the author.

5.3 Performance analysis

The experiments that aimed to analyse the system performance were conducted on two devices: a personal computer and a Raspberry Pi single-board computer. Frame 6 presents, in

greater detail, the specifications of these computers. The musical keyboard used in these experiments was the M-AUDIO Keystation Mini32 MK3 MIDI controller.

Frame 6 – Specifications of the computers used in experiments.

Computer	Operating system	CPU	RAM size
Positivo Master N8140 Blackstone Laptop	Windows 11 Pro v.22H2 64 bits	Intel® Core™ 5-7300U CPU @ 2.60GHz	8 GB
Raspberry Pi 4 Model B Rev 1.5	Raspberry Pi OS Debian 11 (bullseye) 64 bits	Broadcom BCM2711 SoC @ 1.8GHz	8 GB

Source: Production by the author.

The proposed comparative analysis is based on the measurement, on each device, of the following parameters related to the SuperCollider audio server, the *scsynth* process:

a. Average CPU Usage: represents the average CPU load over time while the audio server is running. The average CPU usage by the audio server is calculated over a given time interval. This is a useful metric for evaluating the server's overall performance.

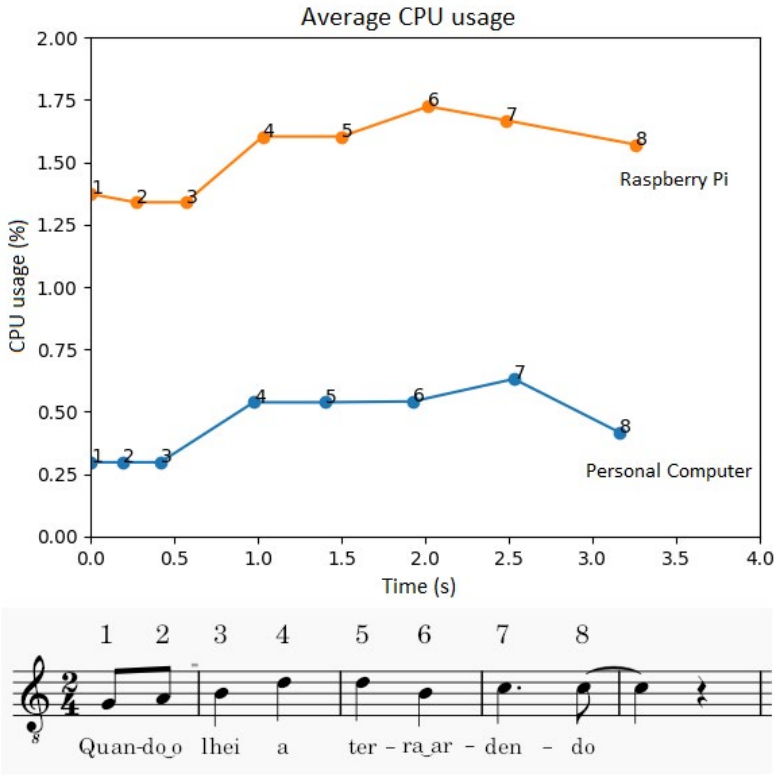
b. CPU Usage Peaks: indicates the highest CPU usage peak while the audio server is running, recording the maximum value reached by the CPU load while performing audio processing. This measurement can be used to identify moments when the CPU load reaches critical levels or when peaks in processing demand occur.

Figure 6 shows the average CPU usage measurements by the audio server on each of the devices, indicated in the graph by their respective operating systems. The performance of the first verse of the song "Asa Branca", in 120 bpm, was chosen to complete the experiment. The various points represent the moments in which each note was played and are numbered from 1 to 8 so that they correspond to the score placed at the bottom of the graph. A larger number of points could be generated to cover a longer duration of the musical performance, but the note-to-note correspondence

between the two performances would be less clear, and the score would tend to either become illegible or take up too much space.

It was observed that, although the average CPU usage was higher on the Raspberry Pi than on the personal computer, on both devices the variation of the measurement over time behaved similarly. On the Raspberry Pi, the average CPU usage of the SuperCollider audio server varied within the range of 1.25 to 1.75%, while the variation on the personal computer occurred within the range of 0.25 to 0.75%.

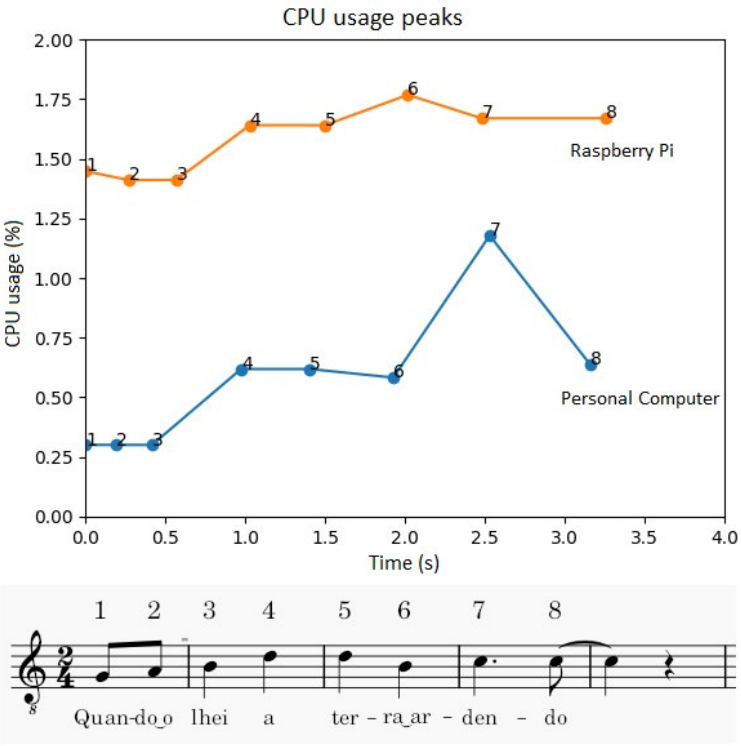
Figure 6 - Average CPU usage on each device.



Source: Production by the author.

In turn, Figure 7 presents measurements of CPU usage peaks by the *scsynth* process on both computers, relating them to the musical notes present in the score displayed at the bottom, similar to the pattern adopted in Figure 6.

Figure 7 - CPU usage peaks on each device.



Source: Production by the author.

Here, the graph's behavior was similar to that of the previous measurement. The CPU usage peaks were more pronounced on the Raspberry Pi, but the variation in measurements within each level was similar across devices, except for a higher CPU usage value observed when the seventh note was activated on the personal computer. CPU usage peaks varied between 1.25% and 1.75% on the Raspberry Pi and between 0.25% and 1.25% on the personal computer. It is believed that exploring SuperCollider's parallelism features can improve the synthesizer's performance on the Raspberry Pi.

6. Discussion of the results

The experiments conducted validated the current version of the synthesizer as a proof of concept. Real-time control of the basic sound qualities — pitch, intensity, duration, and timbre — as

well as the correct selection of diphones, proved to be feasible, as they could be verified both aurally and visually, by analyzing the behavior of the audio signal generated in the Audacity editor as the songs were played. It is possible to verify that even the imprecisions in the performance were reflected in the audio output, which is a desired characteristic when dealing with a real-time synthesizer.

Regarding the subjective evaluation by music educators, the requirement that received the least favorable evaluation was naturalness. The main cause of this is probably PATRICIA's dependence on MBROLA, which was not designed to synthesize the singing voice. Other causes are the absence of voice expression control and a duration extension algorithm that generates a very short loop segment, covering only one period of the synthesized voice waveform. In addition, "click" noises were noticeable in some syllabic transitions.

Regarding the intelligibility of the generated syllables, most evaluations were positive, but this may be because the music educators likely already knew the lyrics of the evaluated songs. Furthermore, the expected control at the phonetic level of the synthesized singing was not at all possible, given the configurations of the voice databases used, since, at least in the cases of br1 and br3, some diphones, such as [ti] and [di], had already been programmed to correspond to a specific regional variant of the Brazilian Portuguese language, resulting in the palatalized syllables [t i] and [d i], respectively.

The control of pitch and duration was also well evaluated by the educators, including in relation to the demonstration of three-voice singing, where these parameters were evaluated together under the heading "conformity to the arrangement". The pedagogical usefulness, in turn, divided opinions, although three of the five educators stated that they fully agreed that the PATRICIA synthesizer would be a useful tool for teaching choral singing.

Finally, both measurements of system performance described in Section 5 yielded very low CPU usage and, consequently, quite

satisfactory performance. This result indicates that the choices, in terms of software, were correct: SuperCollider proved to be quite efficient in real-time audio processing and MBROLA responded well to calls made by the server. As for the hardware, however, the experiments lead us to believe that the synthesizer could be implemented in a device with much lower computing power, making the solution cheaper, especially when developing its final version as an embedded system.

7. Conclusion and future work

The system proposed here, PATRICIA, is a real-time vocal synthesizer, and is the first to be specifically designed to generate vocals in Brazilian Portuguese. This is a work in progress, and the synthesizer described is a proof of concept with a limited set of functional requirements implemented in the SuperCollider language and environment.

Concatenation synthesis and lyrics provided in advance in a text file proved to be, respectively, the most suitable technical approach and phonetic input method for the system, as discussed. Musical input is given in real time via MIDI, and synthesis is performed by MBROLA, a speech synthesizer.

For future versions of the system, the implementation of a series of modifications is planned:

- The creation of PATRICIA's own voice database with better audio quality. The samples will be recorded by a professional female singer rather than using speech samples. This will improve the naturalness and intelligibility of the synthesized singing voice, reducing reliance on MBROLA's databases.
- The implementation of an SVS algorithm into PATRICIA's code is another factor that contributes to this system's independence from MBROLA.
- More comprehensive treatment of MIDI messages in order to provide the system with artistic expression controls.

- Optimization of the synthesizer performance by using the supernova process instead of *scsynth*, which will allow to exploit the parallelism capabilities of the device's multicore processor.
- The integration between PATRICIA and a MIDI keyboard to make the synthesizer work as an embedded system for the instrument. In the current version, the keyboard acts as an external input peripheral to the system.
- Transformation of PATRICIA into an augmented instrument, that is, one equipped with controls that go beyond the conventional technique by which it would be played. In such a context, gestural controls that allow real-time phonetic input could be implemented in PATRICIA, which would be beneficial for its use in more artistic environments.

Among the challenges to be faced with future implementations, the following can be mentioned:

- The discussion regarding the patents of the Vocaloid Keyboard, which is a similar product and property of Yamaha Corporation.
- Constructing a voice database from recordings of a singer raises ethical questions about research involving human subjects.
- The systematization of tests that allow an evaluation of the product by artists who, when using it, will validate the functional requirements considered to be more subjective.
- The establishment of a hardware platform that is more suitable, in terms of cost/benefit, for the implementation of the synthesizer.

Finally, it can be asserted that PATRICIA is a system classified in a research area with few developments — real-time SVS — and boasts a pioneering feature: it is designed for Brazilian Portuguese. Consequently, its limitations do not diminish its significance as a valuable contribution to Computer Music. Moreover, it opens a series of possibilities and challenges for future research.

References

BANDEIRA, Manuel. A versificação em língua portuguesa. In: _____ **Enciclopédia Delta Larousse**. [S.l.]: [s.n.], v. VI, 1960. p. 3239-3249.

BENOÎT, Christian; GRICE, Martine; HAZAN, Valerie. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. **Speech communication**, 18, n. 4, 1996. 381-392. DOI: https://doi.org/10.1016/0167-6393(96)00026-X. Available at: https://www.sciencedirect.com/science/article/abs/pii/016763939600026X?via%3Dihub. Accessed on: Feb. 19, 2026.

BRUM, Leonardo A. Z. Technical aspects of concatenation-based singing voice synthesis. **Scientia Plena**, 8, n. 3(a), March 2012. 039904. Available at: https://scientiaplena.org.br/sp/article/view/919.

BRUM, Leonardo A. Z.; MENESES, Eduardo A. L.; MORENO, Edward D. **Patricia:** a real-time singing synthesizer prototype for the Brazilian Portuguese language. Proceedings of the International Computer Music Conference 2023. [S.l.]: [s.n.]. 2023. p. 176 – 180. Available at: https://www.fulcrum.org/epubs/0c483n30f?locale=en#page=208.

BRUM, Leonardo A. Z.; MORENO, Edward. **State of art of real-time singing voice synthesis**. Anais do XVII Simpósio Brasileiro de Computação. Porto Alegre: SBC. 2019. p. 50–57. DOI: https://doi.org/10.5753/sbcm.2019.10422. Available at: https://sol.sbc.org.br/index.php/sbcm/article/view/10422.

BRUM, Leonardo A. Z.; MORENO, Edward D. Challenges and perspectives on real-time singing voice synthesis. **Revista de Informática Teórica e Aplicada**, 27, n. 4, 2020. 118-126. DOI: https://doi.org/10.22456/2175-2745.107292. Available at: https://seer.ufrgs.br/index.php/rita/article/view/RITA_VOL27_NR4_118.

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

isca-archive.org/interspeech_2016/chan16_interspeech.html. Accessed on: Feb. 19, 2026.

CHO, Yin-Ping *et al.* **A Survey on Recent Deep Learning-driven Singing Voice Synthesis Systems**. 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). [S.l.]: [s.n.]. 2021. p. 319-323. DOI: https://doi.org/10.1109/AIVR52153.2021.00067. Available at: https://ieeexplore.ieee.org/document/9644369/. Accessed on: Feb. 19, 2026.

COOK, Perry R. SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. **Computer Music Journal**, 17, n. 1, 1993. 30-44. DOI: https://doi.org/10.2307/3680568. Available at: https://www.jstor.org/stable/3680568?origin=crossref. Accessed on: Feb. 19, 2026.

D'ALESSANDRO, Christophe; RILLIARD, Albert; LE BEUX, Sylvain. Chironomic stylization of intonation. **The Journal of the Acoustical Society of America**, 129, n. 3, 2011. 1594-1604. DOI: https://doi.org/10.1121/1.3531802. Available at: https://pubs.aip.org/asa/jasa/article-abstract/129/3/1594/978416/Chironomic-stylization-of-intonation?redirectedFrom=fulltext. Accessed on: Feb. 19, 2026.

DELALEZ, Samuel; D'ALESSANDRO, Christophe. **Adjusting the Frame:** Biphasic Performative Control of Speech Rhythm. Proceedings of Interspeech 2017. [S.l.]: [s.n.]. 2017. p. 864-868. DOI: http://doi.org/10.21437/Interspeech.2017-396. Available at: http://doi.org/10.21437/Interspeech.2017-396.

DONG, Minghui *et al.* **I²r speech2singing perfects everyone's singing**. Proc. Interspeech 2014. [S.l.]: [s.n.]. 2014. p. 2148–2149. Available at: https://www.isca-archive.org/interspeech_2014/dong14_interspeech.html. Accessed on: Feb. 19 2026.

DUTOIT, Thierry *et al.* **The MBROLA project:** Towards a set of high quality speech synthesizers free of use for non commercial purposes. Proceeding of Fourth International Conference on Spoken

Language Processing. [S.l.]: [s.n.]. 1996. p. 1393-1396. DOI: https://doi.org/10.1109/ICSLP.1996.607874. Available at: https://ieeexplore.ieee.org/document/607874. Accessed on: Feb. 19, 2026.

FEUGÈRE, Lionel *et al.* Cantor digitalis: chironomic parametric synthesis of singing. **EURASIP Journal on Audio, Speech, and Music Processing**, 2017(1), 2017. 1–19. DOI: https://doi.org/10.1186/s13636-016-0098-5. Available at: https://link.springer.com/article/10.1186/s13636-016-0098-5. Accessed on: Feb. 19, 2026.

GEORGAKI, Anastasia. **Virtual voices on hands:** Prominent applications on the synthesis and control of the singing voice. Journées d'informatique musicale. [S.l.]: [s.n.]. 2004. Available at: https://hal.science/hal-03354297v1. Accessed on: Feb. 19, 2026.

HONO, Yukiya *et al.* **Recent Development of the DNN-based Singing Voice Synthesis System — Sinsy**. Proceedings of APSIPA Annual Summit and Conference 2018. [S.l.]: [s.n.]. 2018. p. 1003-1009. DOI: https://doi.org/10.23919/APSIPA.2018.8659797. Available at: https://doi.org/10.23919/APSIPA.2018.8659797.

HUI, Jessica T. L. Reconfiguring Voice in The End: Virtuosity, Technological Affordance and the Reversibility of Hatsune Miku in the Intermundane. **Cambridge Opera Journal**, 34, n. 3, 2022. 364-379. DOI: https://doi.org/10.1017/S0954586722000301. Available at: https://www.cambridge.org/core/journals/cambridge-opera-journal/article/reconfiguring-voice-in-the-end-virtuosity-technological-affordance-and-the-reversibility-of-hatsune-miku-in-the-intermundane/4B1FF899767FB46A7F010040049D4765. Accessed on: Feb. 19, 2026.

HWANG, Ji-Sang; LEE, Sang-Hoon; LEE, Seong-Whan. Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models. **Neural Networks**, 181, 2025. 106762. DOI: https://doi.org/10.1016/j.neunet.2024.106762. Available at: https://www.sciencedirect.com/science/article/pii/S0893608024006865?via%3Dihub. Accessed on: Feb. 19, 2026.

JOSHI, Ankur *et al.* Likert scale: Explored and explained. **British journal of applied science & technology**, 7, n. 4, 20 February 2015. 396-403. DOI: https://doi.org/10.9734/BJAST/2015/14975. Available at: https://journalcjast.com/index.php/CJAST/article/view/381. Accessed on: Feb. 19, 2029.

KAGAMI, Shota *et al.* **Development of Realtime Japanese Vocal Keyboard [**リアルタイム日本語歌唱鍵盤楽器**" VOCALOID** キーボード **"** の開発**]**. Information Processing Society of Japan INTERACTION (2012). [S.l.]: [s.n.]. 2012. p. 837-842. Available at: https://www.interaction-ipsj.org/archives/paper2012/data/Interaction2012/interactive/. Accessed on: Feb. 19, 2026.

KÄRKI, Kimi. Vocaloid liveness? Hatsune Miku and the live production of Japanese virtual idol concerts. In: ANDERTON, Crhis; PISFIL, Sergio **Researching Live Music:** Gigs, Tours, Concerts and Festivals. [S.l.]: Focal Press, 2021. p. 127-140. DOI: https://doi.org/10.4324/9780367405038. Available at: https://www.taylorfrancis.com/books/edit/10.4324/9780367405038/researching-live-music-chris-anderton-sergio-pisfil. Accessed on: Feb. 19, 2026.

KASHIWASE, Kazuki. An over-the-shoulder keyboard that extends the potential for VOCALOID performance, 2017. Available at: https://www.yamaha.com/en/tech-design/design/synapses/id_104. Accessed on: Feb. 19, 2026.

KENMOCHI, Hideki. **Singing synthesis as a new musical instrument**. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE. 2012. p. 5385-5388. DOI: https://doi.org/10.1109/ICASSP.2012.6289138. Available at: https://ieeexplore.ieee.org/document/6289138. Accessed on: Feb. 19, 2026.

KENMOCHI, Hideki; OHSHITA, Hayato. **VOCALOID - commercial singing synthesizer based on sample concatenation**. Proc. Interspeech 2007. [S.l.]: [s.n.]. 2007. p. 4009–4010. DOI: http://doi.org/10.21437/Interspeech.2007. Available at: https://www.isca-archive.org/interspeech_2007/. Accessed on: Feb. 19, 2026.

KIM, Youngmoo. Singing voice analysis, synthesis, and modeling. In: HAVELOCK, David; KUWANO, Sonoko ; VORLÄNDER, Michael **Handbook of Signal Processing in Acoustics**. Nova Iorque: Springer New York, 2008. p. 359–374. DOI: https://doi.org/10.1007/978-0-387-30441-0_23. Available at: https://link.springer.com/chapter/10.1007/978-0-387-30441-0_23. Accessed on: Feb. 19, 2026.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele University**, 33(2004), 2004. 1–26.

LABRUNE, Laurence. **The phonology of Japanese**. Nova Iorque: Oxford University Press, 2012.

LE BEUX, Sylvain; FEUGÈRE, Lionel; D'ALESSANDRO, Christophe. **Chorus Digitalis:** experiment in chironomic choir singing. Annual Conference of the International Speech Communication Association (INTERSPEECH 2011). [S.l.]: [s.n.]. 2011. p. 2005-2008. DOI: https://dx.doi.org/10.21437/Interspeech.2011-527. Available at: https://www.isca-archive.org/interspeech_2011/beux11_interspeech.html. Accessed on: Feb. 19, 2026.

LI, Rong. Intelligent analysis of music education singing skills based on music waveform feature extraction. **Mobile Information Systems**, 2022. n/a - n/a. DOI: https://doi.org/10.1155/2022/9747342. Available at: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/9747342. Accessed on: Feb. 19, 2026.

LIU, Jinglin *et al.* **Diffsinger:** Singing voice synthesis via shallow diffusion mechanism. Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22). [S.l.]: [s.n.]. 2022. p. 11020-11028. DOI: https://doi.org/10.1609/aaai.v36i10.21350. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/21350. Accessed on: Feb. 19, 2026.

MATSUBARA, Keisuke *et al.* Full-Band LPCNet: A Real-Time Neural Vocoder for 48 kHz Audio With a CPU. **IEEE Access**, 9, 2021. 94923-94933. DOI: https://doi.org/10.1109/ACCESS.2021.3089565. Available at: https://ieeexplore.ieee.org/document/9455356. Accessed on: Feb. 19, 2026.

MCCARTNEY, James. **SuperCollider, a New Real Time Synthesis Language**. Proceedings of the International Computer Music Conference (ICMC) 1996. [S.l.]: [s.n.]. 1996. p. 257-258. Available at: http://hdl.handle.net/2027/spo.bbp2372.1996.078. Accessed on: Feb. 19, 2026.

MOHER, David *et al.* Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. **PLOS Medicine**, 6(7), Julho 2009. 1-6. DOI: https://doi.org/10.1016/j.ijsu.2010.02.007. Available at: https://www.sciencedirect.com/science/article/pii/S1743919110000403. Accessed on: Feb. 19, 2026.

MORTON, Katherine. Naturalness in synthetic speech. **Proceedings of the Institute of Acoustics**, 12, n. 10, 1990. 125-132.

ORJUELA, Alejandro *et al.* Methodological proposal for the identification of incremental innovation in SMES. **European Research Studies Journal**, XXII(4), 2019. 199-214. DOI: doi.org/10.35808/ersj/1506. Available at: https://ersj.eu/journal/1506. Accessed on: Feb. 19, 2026.

OURA, Keiichiro *et al.* **Recent development of the HMM-based singing voice synthesis system — Sinsy**. Proceedings of the 7th ISCA Workshop on Speech Synthesis (SSW). [S.l.]: [s.n.]. 2010. p. 211-216. Available at: https://www.isca-archive.org/ssw_2010/oura10_ssw.html. Accessed on: Feb. 19, 2026.

OURA, Keiichiro. The State-of-the-art of Singing Voice Synthesis Based on Statistical Model. **The Journal of the Institute of Electronics, Information and Communication Engineers**, 98, n. 6, 01 jun. 2015. 460-466. Available at: https://www.journal.ieice.org/summary.php?id=k98_6_460&year=2015&lang=en. Accessed on: Feb. 19, 2026.

PETERSEN, Kai *et al.* **Systematic mapping studies in software engineering**. Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, EASE'08. Swindon: BCS Learning Development Ltd. 2008. p. 68–77. Available at: https://dl.acm.org/doi/10.5555/2227115.2227123. Accessed on: Feb. 19, 2026.

PUCKETTE, Miller S. **Pure Data**. Proceedings of the International Computer Music Conference 1997. [S.l.]: [s.n.]. 1997. Available at: http://hdl.handle.net/2027/spo.bbp2372.1997.060. Accessed on: Feb. 19, 2026.

RODET, Xavier ; POTARD, Yves; BARRIERE, Jean-Baptiste. The CHANT project: From the synthesis of the singing voice to synthesis in general. **Computer Music Journal**, 8(3), 1984. 15–31.

RODET, Xavier. **Time-domain formant-wave-function synthesis**. Spoken Language Generation and Understanding: Proceedings of the NATO Advanced Study Institute. [S.l.]: [s.n.]. 1980. p. 429-441. DOI: https://doi.org/10.1007/978-94-009-9091-3_21. Available at: https://link.springer.com/chapter/10.1007/978-94-009-9091-3_21. Accessed on: Feb. 19, 2026.

SHERRELL, Linda. Evolutionary Prototyping. In: RUNEHOV, Anne L. C.; OVIEDO, Lluis **Encyclopedia of Sciences and Religions**. Dordrecht: Springer, 2013. p. 803. DOI: https://doi.org/10.1007/978-1-4020-8265-8_201039. Available at: https://link.springer.com/rwe/10.1007/978-1-4020-8265-8_201039. Accessed on: Feb. 19, 2026.

SILVA, Thaïs C.; LEITE, Camila. Padrões sonoros emergentes:(oclusiva alveolar+ sibilante) no português brasileiro. **Caderno de Letras**, 24, 2015. 15-36. DOI: https://doi.org/10.15210/cdl.v0i24.7270. Available at: https://periodicos.ufpel.edu.br/index.php/cadernodeletras/article/view/7270. Accessed on: Feb. 19, 2026.

STREIJL, Robert C.; WINKLER, Stefan; HANDS, David S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. **Multimedia Systems**, 22, n. 2, 2016. 213-227. DOI: https://doi.org/10.1007/s00530-014-0446-1. Available at: https://link.springer.com/article/10.1007/s00530-014-0446-1. Accessed on: Feb. 19, 2026.

TAE, Jaesung; KIM, Hyeongju; LEE, Younggun. **MLP singer:** Towards rapid parallel Korean singing voice synthesis. 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing

(MLSP). [S.l.]: [s.n.]. 2021. p. 1-6. DOI: https://doi.org/10.1109/MLSP52302.2021.9596184. Available at: https://ieeexplore.ieee.org/document/9596184. Accessed on: Feb. 19, 2026.

TESTE2. **teste2**. [S.l.]: [s.n.]. 2026.

UMBERT, Marti *et al.* Expression Control in Singing Voice Synthesis: Features, approaches, evaluation, and challenges. **IEEE Signal Processing Magazine**, 32, n. 6, 2015. 55-73. DOI: https://doi.org/10.1109/MSP.2015.2424572. Available at: https://ieeexplore.ieee.org/document/7298564/. Accessed on: Feb. 19, 2026.

VEAUX, Christophe *et al.* **Gesture Control of HMM-Based Singing Voice Synthesis**. Proceedings of 8th ISCA Speech Synthesis Workshop. [S.l.]: [s.n.]. 2013. p. 247-248. Available at: https://www.isca-archive.org/ssw_2013/veaux13_ssw.pdf. Accessed on: Feb. 19, 226.

WANG, Johnty *et al.* **PENny:** An Extremely Low-Cost Pressure-Sensitive Stylus for Existing Capacitive Touchscreens. Proceedings of the International Conference on New Interfaces for Musical Expression. [S.l.]: [s.n.]. 2013. p. 466-468. DOI: https://doi.org/10.5281/zenodo.1178686. Available at: https://zenodo.org/records/1178686. Accessed on: Feb. 19, 2026.

WELLS, John C. SAMPA computer readable phonetic alphabet. In: GIBBON, Dafydd; MOORE, Roger; WINSKI, Richard **Handbook of Standards and Resources for Spoken Language Systems**. [S.l.]: [s.n.], v. 4, 1997. p. 684–732.

YAMAMOTO, Ryuichi; YONEYAMA, Reo; TODA, Tomoki. **NNSVS:** A Neural Network-Based Singing Voice Synthesis Toolkit. Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: [s.n.]. 2023. p. 1-5. DOI: https://doi.org/10.1109/ICASSP49357.2023.10096239. Available at: https://ieeexplore.ieee.org/document/10096239/. Accessed on: Feb. 19, 2026.

ZEN, Heiga; TOKUDA, Keiichi ; KITAMURA, Tadashi. **An introduction of trajectory model into HMM-based speech synthesis**. Proceedings of the 5th ISCA Workshop on Speech Synthesis (SSW). [S.l.]: [s.n.]. 2004. p. 191-196. Available at: https://www.isca-archive.org/ssw_2004/zen04_ssw.html. Accessed on: Feb. 19, 2026.

Authorship contribution

Leonardo Araujo Zoehler Brum: Conceptualization; Software; Data collection and analysis; Writing – original draft. Eduardo A. L. Meneses: Conceptualization; Validation; Writing – review. Edward David Moreno Ordonez: Methodology; Supervision; Validation; Writing – review.

Financing

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Publisher

Federal University of Goiás. School of Music and Performing Arts. Graduate Program in Music. Publication in the Portal of Periodicals UFG.

The ideas expressed in this article are the responsibility of their authors and do not necessarily represent the opinion of the editors or the university.