# Technical aspects of concatenation-based singing voice synthesis

Aspectos técnicos da síntese de voz cantada baseada em concatenação

L. A. Z. Brum

*Núcleo de Música, Universidade Federal de Sergipe, 49100-000, São Cristóvão-Se, Brasil*

*lazb18@yahoo.com.br*

O presente artigo discorre sobre alguns aspectos técnicos do método de síntese digital de voz cantada baseada em concatenação, tomando como referência para estudo de caso o canto na língua portuguesa falada no Brasil. Com este intuito, serão abordados certos aspectos da fonética do português do Brasil e suas relações com a acústica do canto, para que se possa então descrever o método propriamente dito, que consiste basicamente em concatenar fonemas pré-gravados em formato digital, aplicando-se *loops* nas vogais de sustentação de acordo com duas entradas de dados principais: mensagens do protocolo MIDI e a letra da canção desejada em notação fonética adequada.

Palavras-chave : síntese de canto; MIDI; TTS

This article discusses some technical aspects of the concatenation-based method for digital singing voice synthesis, taking as a reference for case study singing in Brazilian Portuguese language. In order to do it, certain aspects of Portuguese phonetics and its relations with singing acoustics will be presented, so a description of the method itself will be possible. Such method basically consists of a concatenation of digitally prerecorded phonemes with the application of loops on the sustained vowels according to two main data inputs: MIDI protocol messages and the lyrics of the chosen song in a suitable phonetic notation.

Keywords: singing synthesis; MIDI; TTS

## 1. INTRODUCTION

In a previous work, the importance of the MIDI protocol has been emphasized. Among several applications of this technological standard, sequencer systems, "able to handle human voice samples and even to articulate chant with the aid of the protocol[1]"[1] have been mentioned. The goal of the present work is to detail how such systems work through the presentation of one of the techniques used by them. This technique is the *concatenation-based synthesis*.

However, so as to make the problem domain more comprehensible, it is convenient to establish a comparison between the sonorous manifestation of speech and the acoustical characteristics of musical sounds. Thus, the two next sessions will determinate the premises that will guide the presentation of the technique which is the subject of this article.

## 2. BASIC NOTIONS OF PHONETICS

*Phoneme* is the name given to the smallest distinguishable sonorous unit in human speech. Phonemes can be classified into two great groups: *consonants*, which are aperiodic vibrations (noises) caused by a partial or total obstruction of the airstream during speech due to the action of the so-called *articulators*, such as lips and tongue; and *vowels*, which "are opposed to

1 "Existem seqüenciadores capazes de manipular amostras de voz humana e, inclusive, articular canto com o auxílio do protocolo."

consonants because 1) they are, acoustically, complex periodic sounds; 2) they form the syllable nucleus and a tone or intensity accent can occur over them[2]"[2].

A *syllable* is a unit which is difficult to conceptualize, but its idea is widely used to establish the distinction between the two great groups of phonemes. Vowels occur as syllable center, whereas consonants occur as syllable margins. Thus, the syllable in its acoustical manifestation can be described as follows:

> From the point of view of perception, the sonorous chain is considered to be composed by acclivities, peaks and declivities of sonority, so each syllable has a peak, which is its nucleus or center, occupied by sounds with high sonority, for example, vowels. The acclivities and declivities are "slopes" of sonority that determinate the *syllabic boundaries* or margins, the preferential place for consonants[3][2].

However, certain vocalic sounds can also occur at the margin of a syllable, such as happens in diphthongs and triphthongs. Such sounds are denominated *non-syllabic vowels* or *semivowels*.

Once these basic notions of phonetics were given, it is convenient to present an important concept of musical acoustics, the *envelope curve*, and then relate the later to the former in order to infer certain aspects of singing acoustics that will be the premises for the description of the proposed voice synthesis technique.

## 3. THE ENVELOPE CURVE AND ITS PHASES

The wave motion caused by the sound of a musical instrument or by human voice performing a note varies in amplitude along time. The curve described by this variation is called *envelope* and "can be decomposed into four phases: attack, decay, sustain and release[4]"[3].

The *attack* phase corresponds to the time between the start of the execution of a note and its maximum volume. *Decay* is the time needed for the note to run down from the maximum level to a constant volume. While this constant volume is kept, the curve will be at the *sustain* phase, whereas the *release* phase is the time between the end of the sustain phase and the return to silence. The terminology in Portuguese language to denominate such phases can vary according to the author, but the English terms are already well established in literature. Figure 1 presents an envelope curve and its four phases are indicated by the respective initial letters of each one.

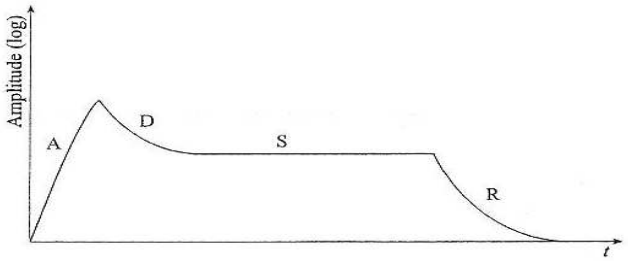


*Figure 1. Idealized envelope curve and its four phases[4].*

2 "[As vogais] se opõem as consoantes por 1) serem acusticamente sons periódicos complexos; 2) constituírem núcleo de sílaba e sobre elas poder incidir acento de tom e/ou intensidade"

3 "Do ponto de vista da percepção, considera-se a cadeia sonora como composta de aclives, ápices e declives de sonoridade, cada sílaba sendo constituída de um ápice, que é seu núcleo ou centro ocupado por sons de alta sonoridade, como, por exemplo, as vogais. Os aclives e declives constituem "vales" de sonoridade que determinam as *fronteiras silábicas*, suas margens, lugar preferencial das consoantes"

4 "(...) é a curva de envoltória, que pode ser decomposta em quatro fases: ataque, decaimento, sustentação e relaxamento"

However, there are authors who tend to considerate the decay phase as the less important for the analysis of the envelope curve [4,5], placing greater emphasis on the three other phases. Luís L. Henrique, for example, says that there are three periods to be considered during a sound: two transient periods, attack and release, and one stationary period, which is called by him "stability period" and corresponds to the sustain phase. It is important to remark that the transient periods are formed by a noise, whereas "certain characteristics of sound, such as pitch and intensity, are fixed during the stability period[5]"[4]or sustain phase.

*The New Grove dictionary of music and musicians* discusses the importance of the envelope curve to human speech in its entry about sound. Such assertions are equally important for singing:

> Envelope shapes play an essential part in human speech. The consonants are usually fairly drastic changes in envelope shape. A plosive, like 'p', makes a fairly rapid initiation of random noise (air escaping when the lips are opened) leading on to a vowel, a steady note. If the noise is allowed to rise in amplitude more slowly, the result is an 'f' [6].

From here, it is possible to establish relations between the three main phases of the envelope curve and the acoustical manifestation of the syllable in order to deduce some principles of the concatenation-based singing voice synthesis technique. The sonorous acclivities and declivities which occur as margins in a syllable correspond to attack and release phases of the envelope curve, respectively. Therefore, consonants and semivowels appear in these phases when a syllable is sung. Meanwhile, the vowel that forms the syllable nucleus occurs at the sustain phase. Accordingly, the noisy nature of consonants, proper to the transient periods at attack and release phases, opposite to the periodic nature of vowels ("steady notes"), characteristic of the stationary period, also coincide. Thus, it is possible to come to a very important conclusion: the pitch of a sung sound is determined during the occurrence of the vowel; it is on the vowel that the musical note concentrates, even when there are melismas, that is, pitch variations during the prolongation of a same vowel.

Given these premises, it is possible to describe the concatenation-based singing synthesis techniques itself. It will be made in the next session.

## 4. CONCATENATION-BASED SINGING VOICE SYNTHESIS TECHNIQUE

Singing acoustics was presented till here as a combination of the characteristics of musical acoustics and speech. In an analogous way, the concatenation-based singing synthesis method joins musical and speech synthesis techniques. Below, there will be a description of how such techniques are joined with the same goal through the presentation of each one and an implementation proposal of a singing synthesizer.

### 4.1. A COMBINATION OF TWO TECHNIQUES

Concatenation-based singing voice synthesis is a type of sample-based musical synthesis method. Although such technique is not, in the strict sense, a synthesis, since it does not generate sound signals from artificial reconstructions of its elements, but handles prerecorded audio samples, this is the habitual designation of the process, so the present article has kept it.

Generally, the sample-based musical synthesis technique consists of recording some samples of the sound which is to be handled, generating then the corresponding sounds of the notes that are closer to these samples: "A **C** is recorded, for example, and from this **C** some contiguous

5 "É durante o período de estabilidade que se fixam certas características do som tais como a altura e a intensidade."

notes are generated: **C♯**, **D**...[6]"[4]. Another general characteristic of sample-based synthesis is the application of the looping technique on the stationary part of the sound, that is, a sound cell of the sustain phase is continuously played in order to prolong the sound as much as is desired. The starting and end points of the loop on the stationary part of the waveform must be well defined, so there will not be a sensation of discontinuity when a prolonged execution occurs. The attack and release phase also must be generated from prerecorded samples. MIDI-based sequencers use this technique.

Another technique related to concatenation-based singing synthesis is *text-to-speech* (TTS), whose goal is "To create, from a text message typed on a computer, the corresponding reading voice, allowing its transmission[7][4]". A brief description of such technique is given by O'Sullivan and Igoe, who also point some of its difficulties:

> To synthesize speech, you have to break language down into prerecorded parts. If you break it down into larger parts, like words and sentences, it will sound better, but the range of possible utterances will be smaller. If you break speech all the way down to the phonemes of a language, you can theoretically synthesize any text, but the result always sounds artificial.[7]

Such artificiality referred by the authors occurs especially due to the *prosody* of the language, that determines mainly the frequency, duration and intensity variations of the pronounced sounds. In the case of speech, controlling such parameters is a complex task if a synthesis with maximum naturally is to be obtained. However, such difficulties are widely reduced when singing voice is handled, because the melodic line, rhythm — associated to tempo — and metrical accent of the song will previously determinate the values of such three variables, respectively, in a more precise way.

### 4.2. IMPLEMENTATION PROPOSAL

According to the explanation above, to implement a system which generates singing voice using the concatenation-based synthesis technique it is necessary to create a database with phoneme samples recorded in digital format and establish ways of representing both musical parameters and phonemes itself in order to give the input data to the algorithm that will handle the samples. The most widely used standard in the world to represent the musical sounds characteristics in order to control prerecorded samples is, doubtless, the MIDI protocol, which allows integration with other sequencers and musical instruments. The phonemes can be represented in text format in SAMPA (*Speech Assessment Methods Phonetic Alphabet*) notation, which is a convenient standard because it uses the 7-bit ASCII character set, available in any common computer keyboard, whereas the International Phonetic Alphabet uses special characters which are not always at the hand of users. SAMPA notation was originally developed at the end of the 1980's by European Economic Community and it is used by professional singing synthesis systems such as *Virtual Singer*, by Myriad. In Appendix I there is a table with SAMPA symbols for the phonemes pronounced in Brazilian Portuguese.

The mentioned database could contain consonants and semivowels recorded once, whereas vowels would be recorded by singing several different notes, extracting from them the sound cell to be used on the loop and generating, for each note, the neighbor notes *ad hoc* at run time.

A singing synthesis system must also provide a data structure and an interface that allow the internal and external (to users) association between each note and certain syllable. The data structure must be compatible to the MIDI messages, so the structure and the messages are mutually convertible. This kind of conversion was already described in a previous work[8], and

---

6 "Grava-se por exemplo um **dó** e com esse dó geram-se algumas notas contíguas: o **dó♯**, o **ré**..."

7 "[O objetivo da técnica TTS é] Criar a partir de uma mensagem de texto gravada em computador a correspondente voz na leitura dessa informação permitindo a sua transmissão."

a very interesting alternative is presented by Paul Hudak using the *Haskore* library, developed in *Haskell* functional programming language[9]. Thus, the messages from an imported MIDI file or from a musical instrument could be interpreted and reflected into the data structure and on the system interface or users could employ the interface itself in order to define the desired song, while the system fills the data structure. Furthermore, the system could save a file with the ".kar" (*MIDI karaoke*) extension, which already contains the association between the song and its lyrics.

Syllables must be typed by users in SAMPA notation on the interface, and then associated to the musical notes as Figure 2 shows. This figure presents as an example the interface of a program called *Harmony Assistant*, whose singing synthesis module is the *Virtual Singer*, already mentioned. Other singing synthesizers, as *Vocaloid*, by Yamaha, have a piano roll-based interface[8] instead of using the traditional musical notation.

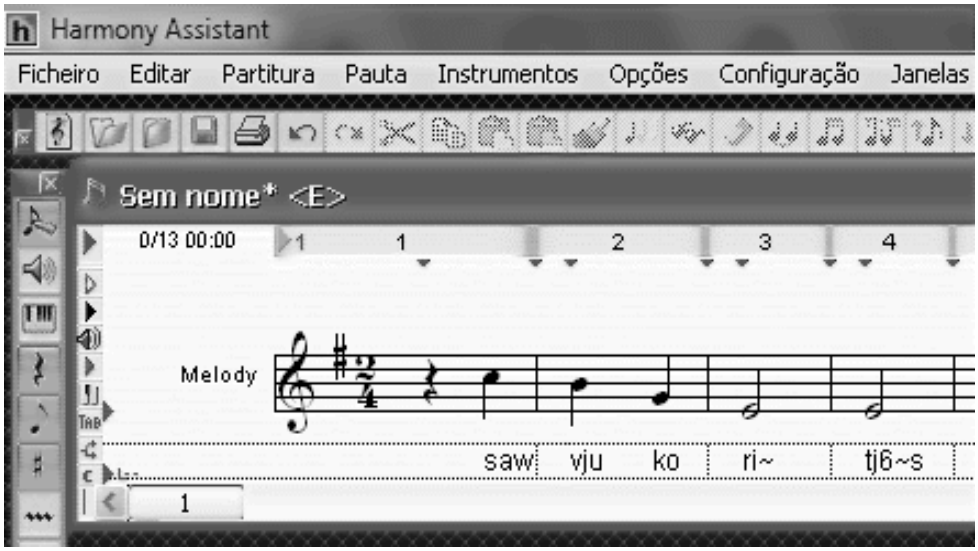


*Figure 2. Syllables typed in SAMPA notation and associated to musical notes on the interface of* Harmony Assistant.

It is important for the typed syllables to contain one, and only one, vowel, at least in the case of Brazilian Portuguese. Such vowel can be preceded and/or succeeded by semivowels or consonants. Thus, when the generation of chant occurs, the system will do the concatenation of the consonants and semivowels only according to the text typed in SAMPA notation, taking into account the order of appearance of the phonemes, whereas when a vowels occurs, the system will identify not only what was typed, but also the note associated to the syllable in order to select the most suitable audio sample according to its pitch. The system also must calculate how many loops will be needed to fit the indicated duration, so the same vowel sound cell will be repeatedly concatenated as many times as needed. At the end of the process, some digital audio, resulting from the concatenations will be obtained. Such audio will be the synthesized singing voice, which will have musical notes with transient periods of attack and release formed by the eventual consonants and semivowels and sustain phases generated by loops applied to vowels. The attack and release phases of each vowel can also be part of the database, being concatenated when there is no other phonemes preceding or succeeding the vowel in the syllable. The resulting audio can be saved as a file in formats like MP3.

An example of implementation similar to the above proposal was presented in 1997 by Michael W. Macon, among others. It is the LYRICOS system, which "employs a concatenation-

8 A piano roll, such as it was described in a former work, "consists of a virtual keyboard aggregated to a table, whose filling corresponds to the chosen musical notes"[8]. More details about this kind of interface can be found in the same work.

based text-to-speech method to synthesize arbitrary lyrics in a given language"[10], using a MIDI file generated by a commercial sequencer in order to provide the necessary musical parameters. The block diagram of LYRICOS is showed by Figure 3.

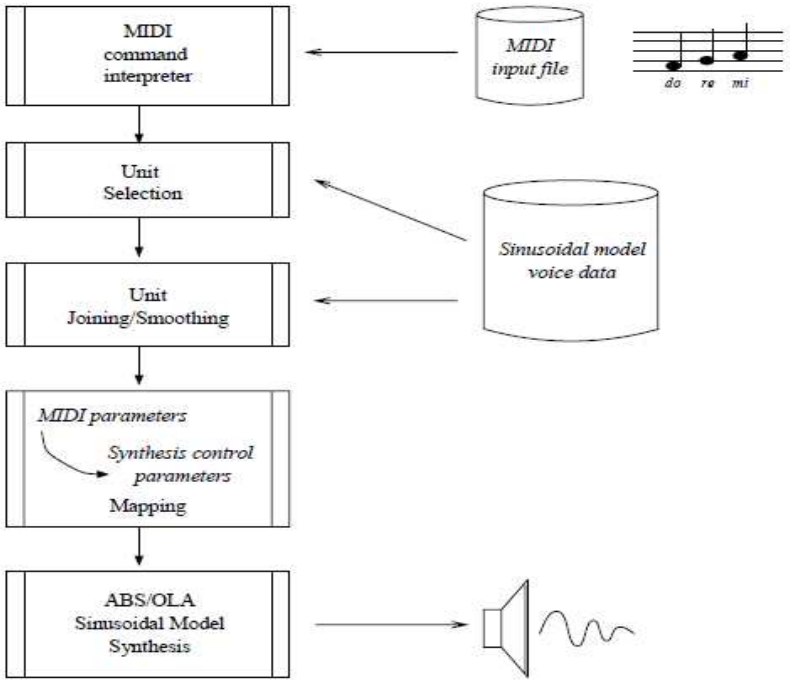


*Figure 3. LYRICOS system block diagram. [10]*

## 5. CONCLUSION

Concatenation-based singing voice synthesis, as explained above, combines sample-based musical synthesis and TTS to generate chant in digital format. The present article, besides providing a brief description of this method, has proposed, in general, the implementation of a singing synthesizer using MIDI and SAMPA standards for the musical synthesis and the TTS technique, respectively. Such proposal can be realized in a future work by developing a basic synthesis system which can be even a module or extension of MIDIBrum sequencer, presented in a previous work [8] . It is important to remark that the proposed implementation does not intend to generate an innovator product with great advantages over the existing singing synthesizers, but only to apply in a practical way the concepts that were presented with didactic purposes.

1. BRUM, L. A. Z. O Auxílio do protocolo MIDI na produção musical. In: SIMPÓSIO SERGIPANO DE PESQUISA E ENSINO EM MÚSICA, 1., 2009, São Cristóvão, Anais. São Cristóvão: Universidade Federal de Sergipe, 2009.
2. CALLOU, D.; LEITE, Y. Iniciação à fonética e à fonologia. 10. ed. Rio de Janeiro: Jorge Zahar, 2005.
3. WEBER, R. F. Arquitetura de computadores pessoais. 2. ed. Porto Alegre: Sagra Luzzato, 2003.
4. HENRIQUE, L. L. Acústica musical. Lisboa: Fundação Calouste Gulbenkian, 2002.
5. CAMILO, D.; YANO, Y.; YABU-UTI, J. B. Circuitos lógicos: teoria e laboratório: engenharia eletrônica. São Paulo: Livraria Ciência e Tecnologia, 1984.

6. TAYLOR, C.; CAMPBELL, M. Sound. In: SADIE, Stanley (org.). The New Grove Dictionary for Music and Musicians. 2. ed. v. 27. New York: Oxford University Press, 2001.
7. O'SULLIVAN, D.; IGOE, T. Physical computing: sensing and controlling the physical world with computers. Boston: Cengage Learning, 2004.
8. BRUM, L.A. Z. Sistema seqüenciador musical baseado no protocolo MIDI. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal de Sergipe, São Cristóvão, 2008.
9. HUDAK, P. The Haskell school of expression: learning functional programming through multimedia. New York: Cambridge University Press, 2007.
10. MACON et al. Concatenation-based MIDI-to-Singing Voice Synthesis. In: MEETING OF THE AUDIO ENGINEERING SOCIETY, 103., 1997, New York.

## APPENDIX I

**SAMPA notation symbols for Brazilian Portuguese phonemes[9]**

| Phoneme type | SAMPA symbol | Examples |
|---|---|---|
| Vowels | a | **Á**tomo, **a**rte . |
| | 6 | P**a**no, r**a**mo, l**a**nho. |
| | 6~ | **An**tes, **am**plo, maç**ã**. |
| | E | M**é**trica, p**e**ça. |
| | e | M**e**do, p**ê**ssego. |
| | e~ | S**em**pre, c**en**tro, tam**bém**. |
| | O | **Ó**tima, **o**va. |
| | o | R**o**lha, av**ô**. |
| | o~ | **Om**bro, **on**tem, c**ôm**puto, c**ôn**sul. |
| | i | **I**tem, silv**í**cola. |
| | i~ | S**im**ples, s**ím**bolo, t**in**ta, s**ín**crono. |
| | u | **U**va, **ú**tero. |
| | u~ | Alg**um**, pl**úm**beo, n**un**ca, ren**ún**cia. |
| Consonants | m | **M**arca. |
| | n | **N**ervo. |
| | J | Arra**nh**ado. |
| | b | **B**arco. |
| | p | **P**ato. |
| | d | **D**ata. |
| | t | **T**elha. |
| | g | **G**ato, **gu**erra. |
| | k | **C**arro, **q**uanto., **qu**eijo. |
| | v | **V**ento. |
| | f | **F**arelo. |
| | z | **Z**ero, ca**s**a, e**x**alar. |
| | s | **S**eta, **c**ebola, espe**ss**o, e**xc**esso, au**x**ílio, a**sc**eta. |
| | Z | **G**elo, **j**arro. |
| | S | **X**arope, **ch**uva. |
| | R | **R**ato, ca**rr**oça. |
| | r | Va**r**iação. |
| | L | Cava**lh**eiro |
| | l | **L**uz. |
| Semivowels | j | U**i**vo. |
| | w | A**u**tomático, móve**l**, p**ão**, freq**ü**ente |

9 Actually, there is not an official SAMPA table for Brazilian Portuguese, but only for European Portuguese. The above table is an adaptation for the phonemes pronounced in Brazil.