UNIVERSIDADE FEDERAL DE SERGIPE

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PATRICIA: um sintetizador de canto em tempo real para o português brasileiro

Dissertação de Mestrado

Leonardo Araujo Zoehler Brum

São Cristóvão – Sergipe

2023

UNIVERSIDADE FEDERAL DE SERGIPE

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Leonardo Araujo Zoehler Brum

PATRICIA: um sintetizador de canto em tempo real para o português brasileiro

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Edward David Moreno Ordonez
Coorientador(a): Eduardo Aparecido Lopes Meneses

São Cristóvão – Sergipe

2023

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

B893p

Brum, Leonardo Araujo Zoehler
Patricia: um sintetizador de canto em tempo real para o português brasileiro / Leonardo Araujo Zoehler Brum ; orientador Edward David Moreno Ordonez. - São Cristóvão, 2023.
112 f.; il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2023.

1. Software de aplicação. 2. Música. 3. MIDI (Normas). I. Ordonez, Edward David Moreno orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: LEONARDO ARAUJO ZOEHLER BRUM

Em 17 dias do mês de novembro do ano de dois mil e vinte três, com início às 14h00min, realizou-se na Sala de Seminários do PROCC da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Marcus Vinicius Santana Poletti**, que desenvolveu o trabalho intitulado: **"*PATRICIA: um sintetizador de canto em tempo real para o português brasileiro*",** sob a orientação do Prof. Dr. **Edward David Moreno Ordonez e coorientação do prof. Dr. Eduardo Aparecido Lopes Meneses (SAT - Societé des arts technologiques)**. A Sessão foi presidida pelo Prof. Dr. **Edward David Moreno Ordonez** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **Carlos Humberto Llanos Quintero (UnB)** , posteriormente o Prof. Dr. **Gilton José Ferreira da Silva** (Procc/UFS) e, em seguida, Dr. **Eduardo Aparecido Lopes Meneses** (McGill) . Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) Aprovado "*(aprovado/reprovado)*". Atendidas as exigências da Instrução Normativa 05/2019/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 04/2021/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos",* 17 de novembro de 2023.

gov.br Documento assinado digitalmente
EDWARD DAVID MORENO ORDONEZ
Data: 17/11/2023 18:15:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Edward David Moreno Ordonez
(PROCC/UFS)
Presidente

gov.br Documento assinado digitalmente
GILTON JOSE FERREIRA DA SILVA
Data: 18/11/2023 01:01:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dr.Gilton José Ferreira da Silva
(PROCC/UFS)
Examinador Interno

Prof. Dr. Eduardo Aparecido Lopes Meneses
(SAT - Societé des arts technologiques)
Examinador Externo

Prof. Dr. Carlos Humberto Llanos Quintero
(UnB)
Examinador Externo

Leonardo Araujo Zoehler Brum
Candidato

A Patricia, minha amada esposa e musa inspiradora deste projeto.

"Todos os afetos do nosso espírito, cada um segundo a sua diversidade, têm na voz e no canto as suas próprias melodias, não sabendo eu qual é a oculta afinidade com essas melodias que os desperta."(Santo Agostinho, Confissões X, 33)

Resumo

A síntese de canto consiste na geração artificial de uma canção, dadas a sua letra e notas musicais, por meio de métodos computacionais. O presente trabalho descreve o desenvolvimento de PATRICIA — acrônimo para Programa que Articula em Tempo Real o Idioma Cantado Inscrito em Arquivo — um sistema que realiza a síntese de canto em tempo real para o português brasileiro. Um mapeamento sistemático foi conduzido a fim de prover diretrizes para o projeto e implementação de PATRICIA. Além disso, experimentos foram conduzidos no intuito de demonstrar as capacidades musicais do sistema e perfazer uma análise comparativa de seu desempenho em um computador pessoal e num dispositivo Raspberry Pi. Como resultados, o mapeamento incidou a escolha da abordagem de síntese concatenativa baseada em *samples* e arquivos de texto provendo a letra da canção antecipadamente. Uma versão inicial do sintetizador foi implementada na linguagem SuperCollider, a título de prova de conceito. Os experimentos conduzidos validaram tal prova de conceito, demonstrando a viabilidade da ideia, sobretudo em termos de desempenho, que mostrou indicadores excelentes em ambas as plataformas escolhidas. Algumas melhorias são apontadas no intuito de que o sintetizador supere suas atuais limitações em versões futuras.

Palavras-chave: Síntese de voz cantada, TTS, MIDI, Computação Musical

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 of PATRICIA, a system that performs real-time singing synthesis for Brazilian Portuguese. A systematic mapping was conducted to provide guidelines for the design and implementation of PATRICIA. Additionally, experiments were carried out to demonstrate the musical capabilities of the system and perform a comparative analysis of its performance on a personal computer and a Raspberry Pi device. The mapping resulted in the choice of a concatenative synthesis approach based on samples and text files providing the song lyrics in advance. An initial version of the synthesizer was implemented in the SuperCollider language as a proof of concept. The conducted experiments validated this proof of concept, demonstrating the feasibility of the idea, particularly in terms of performance, which showed excellent indicators on both chosen platforms. Some improvements are pointed out with the aim of making the synthesizer overcome its current limitations in future versions.

Keywords: Singing synthesis, TTS, Computer music, MIDI.

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de algoritmos

Sumário

1 Introdução

A finalidade da síntese de voz cantada (*Singing Voice Synthesis*, SVS) é a geração de canto por meio de métodos computacionais. Alguns autores consideram que SVS seja um ramo da tecnologia *Text-to-Speech* (TTS) (ALIVIZATOU-BARAKOU et al., 2017; KHAN; LEE, 2015), uma vez que diversos sintetizadores de canto lidam com uma entrada de dados textual, correspondente à letra da canção a ser sintetizada. Entretanto, seria mais preciso classificar tal entrada de dados como *fonética* em lugar de *textual* pois, como pode ser visto ao longo do presente trabalho, este tipo de dado pode ser fornecido por outros meios, como um sinal de áudio de voz, por exemplo. O outro tipo de entrada de dados é o *musical*, que provê as qualidades do som, como altura e duração, à voz sintetizada.

Os sintetizadores de voz cantada podem ser aplicados, por exemplo, na área educacional (PABON et al., 2019). Arquivos digitais contendo o canto podem ser facilmente criados, compartilhados e modificados para fins de aprendizagem, dispensando-se a presença humana de um cantor como referência, ou mesmo gravações.

Outra aplicação desse tipo de síntese é no ramo propriamente artístico (KENMOCHI, 2012). Investimentos na "carreira" de cantores virtuais, como Hatsune Miku (KENMOCHI, 2010), no Japão, têm sido feitos, inclusive com apresentações ao vivo, sendo a voz cantada gerada pelo sintetizador Vocaloid, da Yamaha, e a imagem por hologramas. Além disso, os sintetizadores de voz cantada podem auxiliar o processo de composição musical, na qualidade de simuladores. A indústria áudio visual e de jogos também tem demandado inovações no campo da síntese de canto (KHAN; LEE, 2015).

1.1 Motivação

Em softwares como o Vocaloid (KENMOCHI; OHSHITA, 2007), o usuário descreve as entradas (letra e notas musicais) para que o sistema num momento posterior gere o canto, de

modo análogo ao que acontece geralmente nos ambientes de programação, onde o tempo de projeto e o tempo de execução são distintos.

Essa limitação tem sido superada pela produção de sintetizadores de canto em tempo real, sistemas embarcados que geram o canto artificialmente no mesmo instante em que as entradas são indicadas pelo usuário, o que permite a ele o uso do sintetizador de canto como se fosse um instrumento musical (CHAN et al., 2016), ampliando as aplicações dessa tecnologia.

1.2 Justificativa

Os mais recentes desenvolvimentos na área de síntese de canto (BRUM; MORENO, 2019) em tempo real têm empregado diversos tipos de dispositivos, técnicas de síntese e métodos de entrada. Um dos principais desafios é o de conciliar a entrada simultânea de dados fonéticos e musicais num cenário de performance musical. Tal conciliação será tão mais complexa quanto maior forem as combinações fonéticas possíveis dentro de uma sílaba do idioma para o qual o sintetizador é projetado. Até o momento atual, a literatura não apresenta o desenvolvimento de sintetizadores de canto projetados para o português brasileiro e esta é a principal contribuição para o campo da computação musical que o presente trabalho pretende trazer.

1.3 Objetivo Geral

Desenvolver um sistema de síntese de voz cantada em tempo real no idioma português brasileiro a partir de um modelo de prototipação evolucionária, de modo que o protótipo incial sirva como núcleo minimamente funcional do sistema por sobre o qual outros requisitos poderão ser futuramente implementados, fazendo com que o sintetizador evolua.

1.4 Objetivos Específicos

O alcance do objetivo geral descrito depende do cumprimento das metas apresentadas a seguir:

- Desenvolver um protótipo minimamente funcional, como prova de conceito. Espera-se que o sinterizador seja capaz de articular corretamente os fonemas e controlar as qualidades fundamentais do som, quais sejam altura, duração, intensidade e timbre.
- Conciliar as entradas fonética e musical do sintetizador num contexto de performance em tempo real.
- Embarcar o sistema num computador de placa única Raspberry Pi como plataforma alvo, no intuito de possibilitar a criação de um protótipo de instrumento musical.
- Proteger a propriedade Intelectual do sistema.

1.5 Metodologia

A presente dissertação de mestrado refere-se a uma pesquisa aplicada com caráter bibliográfico e experimental, consistindo em diversas partes, descritas a partir de agora.

A primeira parte da dissertação consta de uma revisão bibliográfica que permitirá fornecer a fundamentação teórica necessária ao desenvolvimento do trabalho.

Na segunda parte, são apresentados um mapeamento tecnológico e uma revisão sistemática, de acordo com os métodos propostos por (PETERSEN et al., 2008) e (KITCHENHAM, 2004), respectivamente, que se propõem a analisar as técnicas e métodos de entrada utilizados nos mais recentes sintetizadores de canto em tempo real desenvolvidos, a fim de subsidiar os rumos da implementação do protótipo proposto.

A terceira parte consiste no desenvolvimento do protótipo de sintetizador, envolvendo a implementação do software no ambiente SuperCollider versão 3.13.0, a definição do inventário de canto e a configuração do hardware necessários para alcançar os objetivos da pesquisa. Na fase de validação e experimentação do trabalho, são avaliadas as capacidades musicais do sistema e seu desempenho em dois dispositivos diferentes: um computador pessoal com sistema operacional Windows 11 Pro e um computaor de placa única Raspberry Pi Model B Rev 1.5.

1.6 Organização do documento

A organização da presente proposta está em conformidade com a descrição a seguir:

- Capítulo 2 – Fundamentação teórica: aborda os requisitos teóricos que fundamentam a síntese de voz cantada em geral;
- Capítulo 3 – Trabalhos relacionados: apresenta um mapeamento tecnológico de patentes relacionadas à área e uma revisão sistemática de literatura, além de uma análise comparativa entre os trabalhos selecionados;
- Capítulo 4 – O sintetizador PATRICIA: descreve o projeto e implementação do objeto de estudo da pesquisa;
- Capítulo 5 – Considerações finais e trabalhos futuros: contém as conclusões que foram tiradas até o momento atual da pesquisa, além de apresentar as possibilidades e desafios para futuras pesquisas

2

Fundamentação Teórica

O domínio do problema da síntese de voz cantada é multidisciplinar: para além da própria computação, tem como requisitos conceitos provindos da acústica, fonética, teoria musical e processamento de sinais. Cada um desses aspectos é tratado nas seções a seguir.

2.1 Elementos de acústica

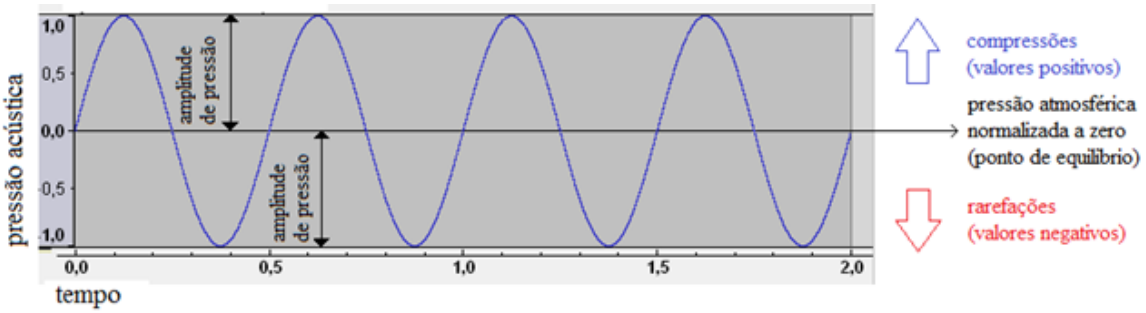
O som é um fenômeno físico que pode ser descrito como uma onda mecânica, ou seja, "uma perturbação que se desloca através de um material chamado de **meio**, no qual [...] se propaga" (YOUNG; FREEDMAN, 2008, p. 103). O movimento associado à perturbação provocada por uma onda qualquer é chamado movimento ondulatório que, por sua vez, enquadra-se na categoria mais geral dos movimentos oscilatórios, caracterizados por serem movimentos "de vaivém em torno de uma posição de equilíbrio" (HENRIQUE, 2002, p. 45) Em tais movimentos, denomina-se **ciclo** uma oscilação completa, ou seja, "o percurso efetuado, ao fim do qual o movimento repete as mesmas características" (HENRIQUE, 2002, p. 45). O ciclo é um número abstrato, uma grandeza adimensional, isto é, não possui unidade de medida. À duração de um ciclo dá-se o nome de **período**, que costumeiramente é representado pela letra T e normalmente medido em segundos. Seu inverso, a frequência, que se pode representar por f, corresponde ao número de ciclos por segundos do movimento, tendo por unidade de medida o Hertz (Hz). Outra grandeza associada aos movimentos oscilatórios é a **amplitude**, valor máximo de deslocamento em relação à posição de equilíbrio, sempre positivo. As unidades de medida da amplitude podem ser diversas a depender da natureza da onda, sendo que o caso específico das ondas sonoras é tratado mais adiante.

Embora a onda sonora possa se propagar em meios líquidos e sólidos, interessa ao presente trabalho estudar sua propagação num meio gasoso, a saber, o ar. Tomando-se o ar como meio de referência para o estudo das ondas sonoras, tem-se que a perturbação provocada por elas

corresponde a flutuações de pressão, ou seja, compressões e rarefações em relação à pressão atmosférica do local. É nesse sentido que se pode afirmar que o som "surge e é transmitido como minúsculas variações de pressão no ar" (TAYLOR; CAMPBELL, 2001, p. 760)[1] . É certo que as referidas flutuações provocadas pelas ondas sonoras causam alterações não somente na pressão das moléculas de ar, mas também em sua posição — o que se traduz em deslocamento — e em sua velocidade. Porém, como é a pressão a grandeza que pode ser mais facilmente medida de maneira direta (HENRIQUE, 2002, p. 242), é ela que deve ser tomada como referência para o estudo do movimento ondulatório do som.

Em concordância com o que já foi estabelecido até aqui, a onda sonora, ao propagar-se no ar, é classificada como uma onda de pressão ou onda de compressão/rarefação, sendo a pressão atmosférica a posição de equilíbrio do movimento ondulatório a ela associado. Por simplificação, pode-se normalizar a pressão atmosférica a zero, de modo que as compressões assumam valores positivos, enquanto as rarefações são sempre negativas (HENRIQUE, 2002, p. 203). Desta maneira, podemos definir a amplitude de pressão de uma onda sonora como sendo "a maior variação de pressão acima ou abaixo da pressão atmosférica" (HENRIQUE, 2002, p. 242). A grandeza que mede o aumento de pressão em relação à pressão atmosférica chama-se pressão acústica, sendo sua unidade de medida, bem como da amplitude de pressão, o Newton por metro quadrado (N/m^2) ou Pascal (Pa). A Figura 1, a seguir, mostra a representação gráfica de uma onda sonora simples conforme a descrição dada até aqui, ou seja, estando os valores de pressão acústica representados no eixo vertical, a pressão atmosférica corresponde ao valor numérico zero, sendo este o ponto de equilíbrio em torno do qual a onda oscila, tendo uma amplitude de pressão positiva quando das compressões e negativa, quando das rarefações ocorridas ao longo do tempo, representado no eixo horizontal.

Figura 1 – Representação gráfica de uma onda sonora simples.



Fonte: elaboração própria.

A Subseção 2.1.1, a seguir, abordará em maiores detalhes esse modo de representar o fenômeno sonoro, bem como as grandezas físicas nele envolvidas, por meio da análise de uma

[1] Tradução nossa do original, em inglês: "[Sound, then,] arises and is transmitted as tiny pressure changes in the air.

onda sonora similar à apresentada.

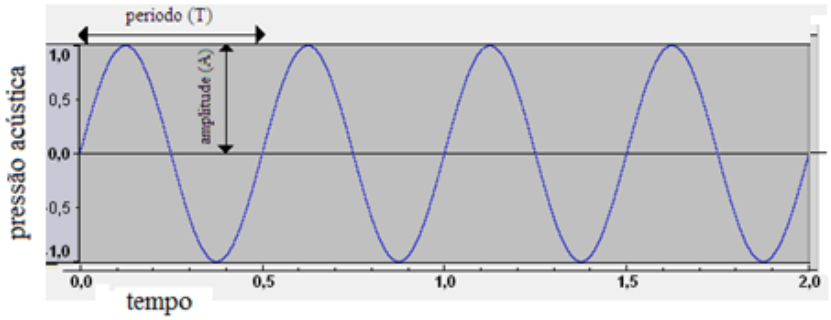
2.1.1 Ondas sonoras senoidais

As ondas sonoras mais simples são as chamadas senoidais (YOUNG; FREEDMAN, 2008, p. 289). Esta designação provém do fato de que a forma dessas ondas pode ser matematicamente representada por uma função seno, como pode ser visto mais adiante. O estudo das ondas senoidais tem grande importância científica para a acústica e para o processamento de sinais, sobretudo porque " qualquer onda, não importa o quão complexa, pode ser representada pela adição dos efeitos de um grande número de ondas senoidais"(TAYLOR; CAMPBELL, 2001, p. 760)[2].

Uma maneira de viabilizar o estudo científico das ondas sonoras é tratá-las como um sinal, que pode ser definido como "a variação de uma grandeza física, que pode ser armazenada, manipulada e transmitida por processos físicos" (HENRIQUE, 2002, p. 268). O sinal correspondente a uma onda sonora será designado, a partir de agora, **sinal acústico**. Por agora, interessa analisar o sinal acústico das ondas senoidais, que se enquadra na categoria dos sinais harmônicos, isto é, aqueles matematicamente representados por senos e cossenos. Este é o ponto de partida para o estudo das grandezas físicas envolvidas no fenômeno sonoro no presente trabalho.

Entre as diversas formas de representação dos sinais, a representação temporal é a mais básica e "consiste no traçado da variação de uma determinada grandeza em função do tempo" (HENRIQUE, 2002, p. 257). Assim, uma possível representação temporal do sinal acústico de uma onda sonora senoidal apresentaria as flutuações de pressão no ar, ou seja, a oscilação dos valores da pressão acústica ao longo do tempo, conforme foi apresentado pela Figura 1. Por sua vez, a Figura 2 exibe esse mesmo tipo de representação, dando destaque à amplitude e ao período da onda senoidal tomada como exemplo.

Figura 2 – Representação temporal do sinal acústico de uma onda senoidal.



Fonte: elaboração própria.

[2] Tradução nossa do original, em inglês:"any wave, no matter how complicated, can be represented by adding up the effects of a large number of sine waves".

A onda apresentada na Figura 2 tem uma amplitude $A = 1{,}0$ e um período $T = 0{,}5$ s. Logo, sua frequência f é igual a $\frac{1}{T} = \frac{1}{0{,}5} = 2$ Hz. Ondas sonoras senoidais possuem uma frequência única. À razão entre as frequências de dois sons dá-se o nome de **intervalo acústico** (HENRIQUE, 2002, p. 926).

A partir do que foi exposto até agora acerca da natureza das ondas sonoras mais simples, as senoidais, passa a ser possível passar a estudar os chamados sons complexos, resultantes da combinação de duas ou mais ondas senoidais.

2.1.2 Sons complexos

De acordo com o que foi anteriormente estabelecido, as ondas sonoras senoidais, ou sons puros, possuem uma frequência única, porém "em rigor, um som puro não existe"(HENRIQUE, 2002, p. 177). Isso se deve ao fato de que "cada som concreto corresponde na realidade não a uma onda pura, mas a um feixe de ondas, uma superposição intrincada de frequências de comprimento desigual" (WISNIK, 2001, p. 23). Tais sons, constituídos por mais que uma frequência, são chamados **sons complexos**, sendo o modo pelo qual se compõem estudado a partir de agora.

Cada uma das frequências constituintes de um som complexo designa-se componente ou parcial, sendo a mais baixa delas designada **frequência fundamental**. Por sua vez, os parciais são classificados em harmônicos ou não-harmônicos, conforme suas frequências sejam ou não múltiplos inteiros da frequência do som fundamental, respectivamente. Quando todos os parciais de um som complexo são harmônicos, tal som denomina-se **periódico** e, se ao menos um dos parciais for não-harmônico, o som complexo é dito **aperiódico**.

O estudo da composição dos sons complexos por parciais puros é de grande importância para o presente trabalho à medida que "qualquer som é constituído por sons sinusoidais" (HENRIQUE, 2002, p. 177), o que servirá de base para os métodos de síntese de sons estudados a partir da Seção 2.4.

2.1.3 A curva envoltória e suas fases

Os sons complexos, além de contarem com mais de uma frequência em sua composição, variam também em amplitude ao longo do tempo, como é possível perceber observando-se representação gráfica dos sinais acústicos dos exemplos dados na Subseção anterior. Tal variação dá a cada onda sonora uma forma característica. Chama-se envoltória [3] a curva "obtida pelos valores máximos da forma de onda" (WEBER, 2003, p. 205). Trata-se de um traçado imaginário do contorno dado pelos diversos picos de amplitude de uma determinada onda sonora.

A curva envoltória é convencionalmente dividida em quatro segmentos (LOY; CHOWNING, 2011a, p. 35), também chamados **fases** (WEBER, 2003, p. 205), correspondentes a quatro períodos diferentes de tempo:

[3] Também chamada "envolvente", ou por sua designação em inglês, *envelope*.

a) ataque (*attack*): tempo transcorrido entre o início da propagação da onda sonora e o alcance de sua amplitude máxima;

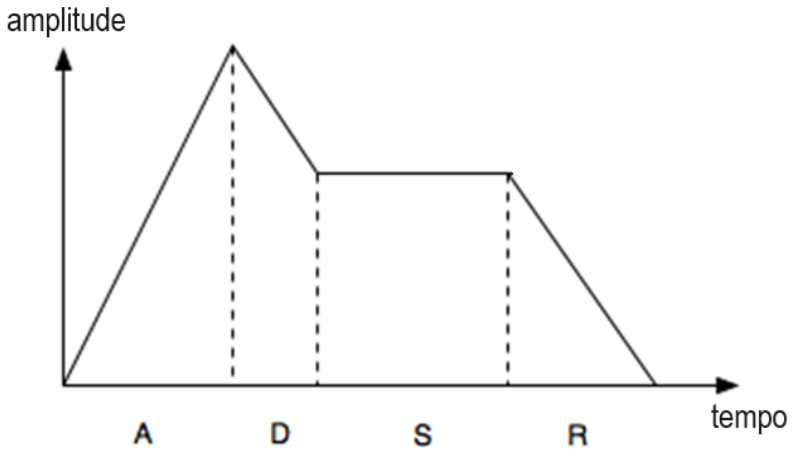
b) decaimento (*decay*): intervalo entre o fim do período de ataque, quando a amplitude chega ao máximo e depois decai para assumir um valor constante;

c) sustentação (*sustain*): período de tempo no qual os valores de amplitude se mantém constantes, também chamado de regime estacionário da onda;

d) relaxamento (*release*): fase que dura desde o fim do período estacionário até a extinção da onda.

As fases da envoltória são comumente denotadas pela sigla ADSR (LOY; CHOWNING, 2011a, p. 36), oriunda das letras iniciais de cada uma delas. A Figura 3 apresenta uma curva envoltória idealizada com a posição de suas fases indicadas no gráfico. É comum dar menor importância à fase de decaimento e descrever a envoltória apenas em termos de ataque, sustentação e relaxamento, como fazem Camilo, Yabu-Uti e Yano (1986, p. 391) e Henrique (2002, p. 171)[4]. Cabe ressaltar também que, em contraposição à estabilidade do regime estacionário da onda na fase de sustentação, as demais fases da envoltória também são designadas como períodos transitórios (HENRIQUE, 2002, p. 171), ou seja, neles os valores de amplitude não se mantém constantes.

Figura 3 – Curva envoltória e suas quatro fases.



Fonte: elaboração própria.

[4] Camilo acaba por trocar os nomes das fases de decaimento e relaxamento. Ele denomina “decaimento” a última fase da curva envoltória e de “soltura” ou em inglês, release, a fase que se sucede ao ataque. Henrique, por sua vez, denomina a última fase da curva envoltória de “extinção” ou “decaimento final”. Em todo caso, eles destacam, na curva envoltória, a primeira, a terceira e a última das quatro fases, tendo a segunda um papel menos importante para esses autores.

2.2 Qualidades do som e teoria musical

Na Seção anterior, o som foi estudado em si mesmo, em caráter objetivo, sob o crivo do ramo da física denominado acústica. A presente seção, estando por sua vez no domínio da psicoacústica, visa estudar os efeitos do fenômeno sonoro no sujeito humano, desde sua percepção por meio do sentido da audição até os resultados psicológicos deste mesmo ato de ouvir, procurando-se estabelecer relações entre as grandezas físicas anteriormente estudadas e as sensações delas resultantes. Tais relações, uma vez estabelecidas, permitirão uma melhor compreensão dos fundamentos da teoria musical ocidental e dos elementos constituintes da música, tais como melodia, harmonia, ritmo, dinâmica e instrumentação.

Para os fins desta seção, pode-se partir da afirmação comum de que o som tem quatro características ou qualidades principais: altura, intensidade, duração e timbre (MED, 1996, p. 11). Entretanto, essa classificação tradicional não é cientificamente rigorosa, pois "mistura características psicológicas, como altura e o timbre, com físicas, como a intensidade" (HENRIQUE, 2002, p. 169). Seria necessário distinguir, então, a intensidade como característica física do som, da sensação de intensidade, característica psicológica. Quanto à duração, esta pode ser tanto física — no que seria auferida por aparelhos como os cronômetros — quanto psicológica, uma vez que, subjetivamente, "o mesmo intervalo de tempo pode passar num ápice ou parecer uma eternidade" (HENRIQUE, 2002, p. 170).

Essas quatro características do som são abordadas a seguir em maiores detalhes na qualidade de sensações, buscando-se a causa de seus efeitos psicológicos nas características físicas do som e explicando de que modo tais efeitos viabilizaram a sistematização da arte musical no Ocidente. A primeira das características a ser estudada é a da altura.

2.2.1 Altura

O ouvido humano possui grande sensibilidade, sendo capaz de captar as flutuações de pressão causadas pelas ondas sonoras numa faixa de frequências que varia entre 16 e 20000 Hz (HENRIQUE, 2002, p. 167). Ondas cuja frequência ultrapassa o limite superior apontado são designadas ultrassom, enquanto aquelas que ficam aquém dos 16 Hz denominam-se infrassom. Fazendo-se uma analogia com o sentido da visão, há no espectro eletromagnético uma faixa de frequências correspondentes à luz visível ao olho humano. Ao se decompor a luz branca, que contém todas as cores, nas sete colorações do arco-íris, tem-se a cor vermelha correspondendo à frequência mais baixa e a cor violeta à mais alta. Abaixo dessa faixa de frequências, tem-se a radiação infravermelha e, acima dela, a ultravioleta. Tais radiações são invisíveis, assim como o infrassom e o ultrassom são inaudíveis.

A sensação de altura é determinada sobretudo pela frequência da onda sonora ouvida, podendo-se defini-la como uma característica psicológica que "traduz a sensação auditiva que nos permite ordenar os sons do grave ao agudo" (HENRIQUE, 2002, p. 862). Quanto maior é

a frequência, mais alto, ou agudo, será o som percebido. Se o som em questão for senoidal, a sensação de altura será clara e derivará da única frequência constituinte dele. Porém, quando se trata sons complexos, é preciso levar em consideração sua periodicidade, uma vez que "Vibrações periódicas [...] são associadas à percepção de uma determinada nota musical, e por isso denominadas sons de altura definida"(FILHO, 2004, p. 22). Tais sons são designados **sons musicais** (MED, 1996, p. 11), sendo a frequência fundamental aquela que se relaciona à sensação de altura. Por sua vez, "Objetos que vibram de maneira não-periódica ou simplesmente aperiódica [...] geram sons indefinidos em altura [...]. Tais fenômenos sonoros são denominados pela acústica, genericamente, por ruídos" (FILHO, 2004, p. 24). Este último tipo de sons é igualmente utilizado em música, sobretudo por meio dos instrumentos de percussão (MED, 1996, p. 11).

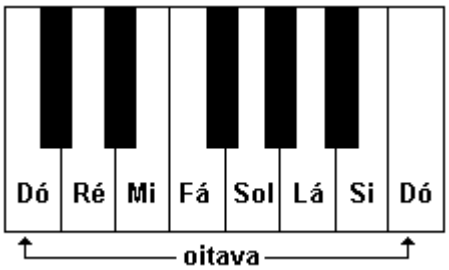
Em relação à arte musical observa-se que "as culturas precisam selecionar alguns sons entre outros" (WISNIK, 2001, p. 59). A altura tem um papel crucial em tal seleção, pois um ponto de partida bastante comum para o estabelecimento disso é a sensação de que dois sons, ao serem ouvidos simultaneamente, tendo um o dobro da frequência fundamental do outro "fundem-se de tal maneira que a maioria das pessoas sem aprendizagem musical pensa tratar-se de um único som" (HENRIQUE, 2002, p. 947) A partir daí, cumpre subdividir esse intervalo acústico de valor 2 (dois) em um determinado número de pontos, de modo a discretizar a faixa contínua de frequências entre os dois sons parecidos. Estabelecidos os pontos, à sequência de sons correspondentes às frequências selecionadas, dá-se o nome de **escala**.

Para se ter uma compreensão básica da teoria musical ocidental, interessa que se estudem duas escalas. A primeira delas é a chamada **escala diatônica** ou **escala natural** e compõe-se de sete sons que são representados em música por elementos chamados **notas musicais**, cujos nomes são, em ordem crescente de altura, dó, ré, mi, fá, sol, lá e si. Em música, dá-se o nome de **melodia** à "alternância entre notas de alturas diferentes"(MED, 1996, p. 12).

Após a nota si, repetem-se os nomes, pois chega-se ao som cuja frequência fundamental é o dobro em relação à da primeira nota dó. Tal som, por soar de maneira similar, embora mais aguda, também é chamado "dó". Em seguida, tem-se novamente ré, mi, fá e assim por diante. O fato de que a segunda nota dó ocupa a oitava posição seguindo-se a sequência da escala diatônica fez com que o intervalo acústico de valor 2 (dois), entre duas notas de mesmo nome, fosse designado musicalmente como intervalo de **oitava**. Em instrumentos como o piano, os sons da escala diatônica são produzidos ao se tocarem as teclas brancas, como indica a Figura 4.

A escala diatônica, porém, contém intervalos acústicos desiguais entre algumas de suas notas. O intervalo acústico entre mi e fá, por exemplo, é cerca da metade do intervalo acústico entre sol e lá. Novas exigências composicionais postularam o surgimento de uma outra escala, na qual o intervalo acústico entre duas notas consecutivas fosse constante. Surgiu, assim a **escala cromática** ou **artificial** (MED, 1996, p. 87), que divide o intervalo de oitava em doze sons diferentes, entre os quais estão os sete sons da escala diatônica. Musicalmente, diz-se que o

Figura 4 – Notas da escala diatônica dispostas no teclado do piano.

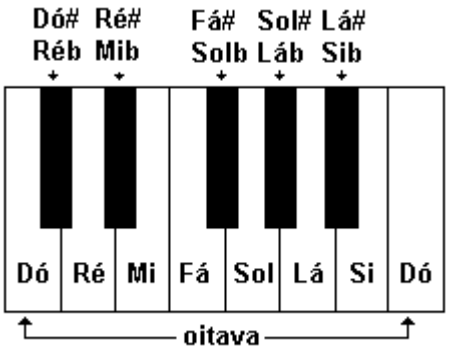


Fonte: elaboração própria.

intervalo entre dois sons consecutivos da escala cromática é de um **semitom**, sendo que dois semitons formam um **tom**. Os outros cinco sons da escala cromática são nomeados a partir das sete notas musicais acrescidas de sinais denominados alterações. O sinal ♯ (lê-se "sustenido") "eleva a altura da nota natural um semitom"(MED, 1996, p. 31). Já o sinal ♭ (lê-se “bemol”) "abaixa a nota natural um semitom"(MED, 1996, p. 32).

Porém, como o intervalo musical entre notas como dó e ré é de apenas um tom, tem-se que, na escala cromática, o som que fica um semitom acima de dó, designado como dó♯ (dó sustenido), e o som que fica um semitom abaixo de ré, denominado ré♭ (ré bemol), têm a mesma altura, ou, em outras palavras, são o mesmo som. Notas como essas, "de nome e grafia diferente, porém com o mesmo resultado auditivo"são chamadas **notas enarmônicas** (MED, 1996, p. 82). Esses sons com alterações são produzidos pelas teclas pretas de instrumentos como o piano. A Figura 5 exibe os doze sons da escala cromática dispostos num teclado de piano e designados por seus respectivos nomes, inclusive quando há notas enarmônicas. O sistema de afinação derivado da escala cromática, no qual há um intervalo acústico constante entre sons consecutivos é denominado **temperado**.

Figura 5 – Sons da escala cromática dispostos num teclado de piano.

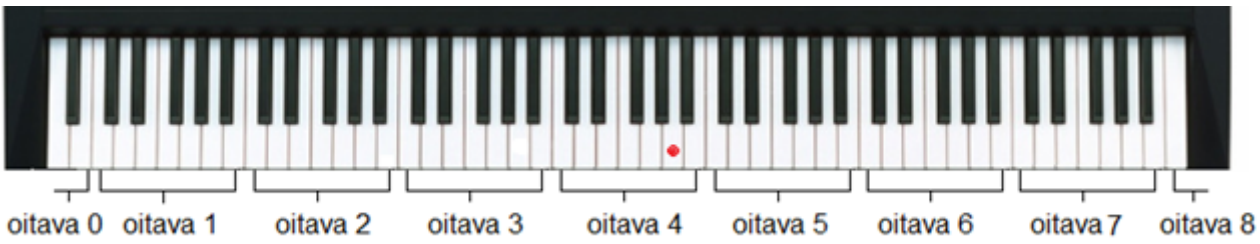


Fonte: elaboração própria.

Em relação à escala cromática, por extensão, o conjunto dos doze sons musicais abrangidos pelo intervalo de oitava é também chamado de "oitava". Assim, se há duas notas de mesmo nome e alturas desiguais, diz-se que elas se encontram em oitavas diferentes. Uma das formas

de designar as diversas oitavas é enumerá-las a partir de zero de acordo com a ordem que se encontram dispostas no teclado padrão de 88 teclas do piano (LOY; CHOWNING, 2011a, p. 40). Assim, a altura de um som musical é comumente designada por alguma nota acrescida de um número que indica a oitava em que o som se encontra. De acordo com esse padrão, o piano tem uma variedade de sons que vai do lá$_0$ ao dó$_8$, em ordem crescente de altura, conforme indica a Figura 6, onde a posição do som lá$_4$, a título de exemplo, é destacada por um ponto vermelho.

Figura 6 – Enumeração das oitavas nas 88 teclas do piano, destacando-se o som lá$_4$.

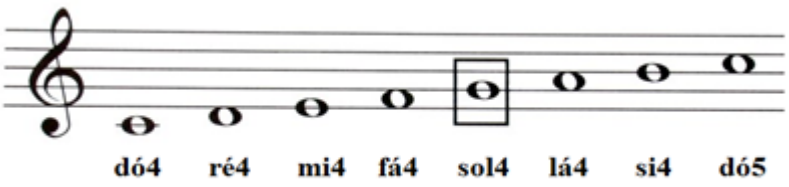


Fonte: elaboração própria.

A frequência fundamental do som lá$_4$ é de 440 Hz, o que serve como referência para a afinação de diversos instrumentos musicais (LOY; CHOWNING, 2011a, p. 14). A partir desse valor, é possível determinar a frequência fundamental de qualquer outro som musical do sistema de afinação temperado, conforme exposto posteriormente.

Na notação musical, a altura é indicada "pela posição da nota no pentagrama e pela clave" (MED, 1996, p. 12). O **pentagrama**, ou **pauta**, onde são assinaladas as notas é um conjunto de cinco linhas paralelas que formam entre si quatro espaços. Já a **clave**, é um sinal que se coloca no início da pauta para que se tenha uma referência das alturas representadas. Uma das claves mais utilizadas em música é chamada clave de sol, que, quando presente na pauta, indica que a nota assinalada na segunda linha — as linhas são contadas de baixo para cima — corresponde a sol$_4$. A Figura 7 apresenta as notas da escala diatônica entre as oitavas 4 e 5 dispostas numa pauta com clave de sol, com destaque para a nota sol$_4$, que nomeia a clave. Para uma explicação mais completa do funcionamento da notação musical em geral, recomendam-se as obras de Med (1996) ou Blatter (2016).

Figura 7 – Notas da escala diatônica dispostas em pauta com clave de sol.



Fonte: elaboração própria.

Uma vez apresentada a maneira com a qual a teoria musical ocidental lida com a percepção

das alturas e quais as relações entre as notas musicais e a grandeza acústica da frequência, pode-se passar a estudar uma outra característica sensorial do som, a duração.

2.2.2 Duração

A percepção da duração permite que se torne possível diferenciar os sons mais curtos dos mais longos. É possível dizer que a duração de um som pode ser encarada tanto pelo aspecto físico, no que pode ser medida em segundos por um cronômetro, quanto pelo aspecto psicológico, derivado das percepções sensíveis, sendo que a teoria musical elaborou um sistema de medida peculiar levando em conta esse último aspecto.

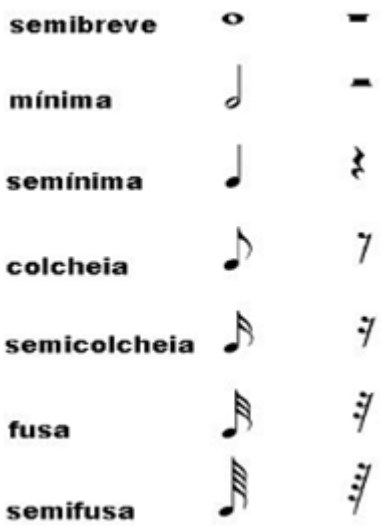
Musicalmente falando, dá-se o nome de **ritmo**, ao "movimento dos sons regulados por sua maior ou menor duração" (PRIOLLI, 1989, p. 6). Já o ramo teórico da música que trata do ritmo é chamado **métrica** (MED, 1996, p. 128). A duração de uma nota musical numa peça não precisa ter um valor matemático preciso; antes, ela é percebida por relações de proporcionalidade com as durações das demais notas. Desse modo, a duração já não precisa necessariamente ser medida em segundos, mas de acordo com uma unidade musical relativa denominada **tempo**.

Numa partitura, a duração das diversas notas — e também dos períodos de silêncio — é indicada por um conjunto de figuras musicais cujos valores guardam uma relação de proporcionalidade entre si. O valor em tempos de cada figura dependerá da fórmula de compasso, uma fração que geralmente figura no início da partitura. O numerador da fração indica a quantidade de tempos que os compassos terão, enquanto o denominador determinará qual das figuras musicais terá o valor equivalente a um tempo. As figuras musicais mais comumente utilizadas, são exibidas na Figura 8 em ordem descrescente de valor. Cada figura vale, em tempos, a metada daquela que a antecede e aparece acompanhada por seu nome à esquerda e, do lado direito, por uma figura de pausa de mesmo valor. As figuras de pausa são utilizadas para indicar a duração do silêncio na música.

Para que se obtivesse um valor exato, em segundos, para cada tempo musical, foi inventado um dispositivo chamado metrônomo que, ao acionar um pêndulo, determina o andamento ao marcar regularmente a duração dos tempos por meio de batidas (MED, 1996, p. 187). Assim, por exemplo, se o metrônomo é programado para oscilar de modo que se tenha 120 bpm — sigla para "batidas por minuto" — isso significa que cada tempo da música dura a 120ª parte de um minuto, ou seja 0,5 segundos. Atualmente, há dispositivos eletrônicos que fazem as vezes de metrônomo e mesmo alguns instrumentos musicais têm essa funcionalidade.

Porções iguais de tempos musicais são agrupadas em estruturas métricas designadas **compassos**. Os tipos de compasso mais comum são o binário, o ternário e o quaternário, cada um deles consistindo em dois, três e quatro tempos, respectivamente, valores estes que figuram no numerador da fórmula de compasso. Também é bastante comum que o denominador dessa fração tenha valor quatro, o que, por convenção, indica que a semínima é a figura musical equivalente a

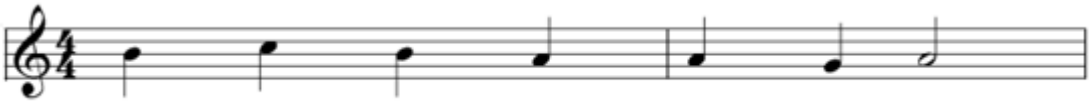
Figura 8 – Figuras musicais e suas figuras de pausa correspondentes.



Fonte: elaboração própria.

um tempo. Os compassos são separados entre si na pauta "por uma linha vertical, chamada barra de compasso ou travessão" (MED, 1996, p. 114). Uma pauta contendo uma melodia disposta em dois compassos do tipo 4/4 é mostrada, a título de exemplo, pela Figura 9.

Figura 9 – Melodia disposta numa pauta em dois compassos 4/4.



Fonte: elaboração própria.

Perceba-se que no segundo compasso há duas semínimas e uma mínima. O valor quatro no denominador da fórmula de compasso significa que cada semínima equivale a um tempo, o que quer dizer que as duas semínimas que figuram no início do segundo compasso equivalem a dois tempos. Como a mínima tem o valor do dobro da semínima, num compasso 4/4, ela equivalerá a dois tempos. Assim, temos quatro tempos, respeitando-se o indicado pelo numerador da fórmula de compasso.

Cabe ressaltar que o agrupamento de tempos dentro de um compasso não se dá de maneira arbitrária, mas antes guarda relação com a variação periódica de intensidade dos sons de uma obra musical conforme o ritmo, uma vez que o primeiro tempo de um compasso possui intensidade maior que os demais. Tal variação é chamada **acento métrico**, e é abordada em maiores detalhes na seção a seguir, que trata da sensação de intensidade.

2.2.3 Intensidade

Enquanto a sensação de altura permite que se ordene os sons do grave ao agudo, permitindo o surgimento das escalas musicais e de sistemas de afinação como o temperado, é por meio da sensação de intensidade que é possível diferenciar os sons mais fortes dos mais fracos. Aqui pode haver certa confusão, pois é comum, em linguagem coloquial, que se designe por "alto" um som na realidade percebido como sendo forte.

Foi também estabelecido que a sensação de altura depende fundamentalmente da frequência da onda sonora ouvida. A sensação de intensidade, por sua vez, está relacionada com a grandeza física da amplitude, de modo que quanto maior for esta grandeza, mais forte tenderá a ser a sensação de intensidade (MED, 1996; HENRIQUE, 2002, p. 92). Em termos práticos, pode-se dizer que a sensação de intensidade "é determinada pela força ou pelo volume do agente que as produz" (MED, 1996, p. 12). Deste modo, o som produzido por um piano, por exemplo, será tão mais intenso quanto mais fortemente forem tocadas suas teclas.

Em música, dá-se o nome de **dinâmica** ao resultado da "alternância entre notas de intensidades diferentes" (MED, 1996, p. 12). Há nas peças musicais uma dinâmica dita natural, resultante do próprio desenvolvimento do discurso musical. Tal dinâmica é fruto das variações do acento métrico, não sendo indicada na partitura, pois decorre da própria estrutura rítmica apontada pelos compassos. Já a dinâmica chamada artificial, é resultado do chamado **acento agógico**, que faz as notas variarem de intensidade conforme a vontade do compositor como meio de expressão (MED, 1996, p. 213). A dinâmica artificial é representada na partitura por meio de indicações sob a pauta, na forma de abreviaturas de nomes italianos, tais como "***p***", representando o termo *piano* ("fraco"), ou "***f***", para o nome *forte*.

Abordadas até aqui as sensações de altura, duração e intensidade, a próxima seção tratará da última sensação sonora a ser estudada pelo presente trabalho, a de timbre.

2.2.4 Timbre

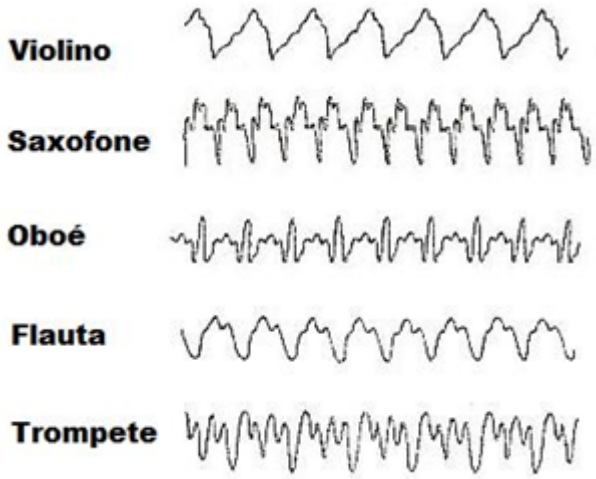
O timbre pode ser definido como "uma característica subjectiva [sic] do som que nos permite diferenciar sons de altura e intensidade iguais" (HENRIQUE, 2002, p. 871), de modo a ser possível distinguir "uma voz ou instrumento do outro" (MED, 1996, p. 95). Daí a alternância e combinação de diferentes timbres numa peça musical ser chamada **instrumentação** (MED, 1996, p. 12). No caso do canto, cada pessoa possui um timbre pelo qual se pode identificar sua própria voz.

Os instrumentos musicais, em geral, produzem sons complexos, sendo que o timbre de cada um deles deriva "da intensidade dos sons harmônicos que acompanham os sons principais" (MED, 1996, p. 12). Ora, já foi estabelecido que a intensidade está diretamente relacionada à amplitude e que a variação dessa grandeza ao longo do tempo, nos sons complexos, descreve a chamada curva de envoltória que, por sua vez, pode ser decomposta em quatro fases. Desta forma,

Henrique (2002, p. 171) afirma que "o que se passa durante os períodos transitórios é de máxima importância para o reconhecimento do timbre do som". Por períodos transitórios entendem-se aquelas fases da curva envoltória nas quais os valores de amplitude não se mantêm constantes, a saber, ataque, decaimento e relaxamento, conforme explicado na Subseção2.1.3. No que diz respeito ao timbre, Henrique (2002) destacará o papel da fase de ataque para sua identificação.

O resultado é que instrumentos musicais diferentes produzirão formas de ondas desiguais (CAMILO; YABU-UTI; YANO, 1986, p. 390) e são essas diversas formas físicas, causadas pelo modo com que cada instrumento produz o som, que causam a sensação de que cada um dos instrumentos tem a sua própria "voz", ainda que eventualmente soem com mesma altura, intensidade ou mesmo duração. A Figura 10 exibe distintas formas de onda que podem ser produzidas por alguns instrumentos musicais. Entretanto, é fato que certos instrumentos têm a capacidade de gerar mais de um timbre.

Figura 10 – Formas de onda geradas por diferentes instrumentos musicais



Fonte: adaptado de Weber (2003, p. 200).

Tendo sido apresentadas as características acústicas e sensoriais do fenômeno sonoro em geral, o próximo capítulo tratará de abordar essas mesmas características em relação ao caso específico da voz humana.

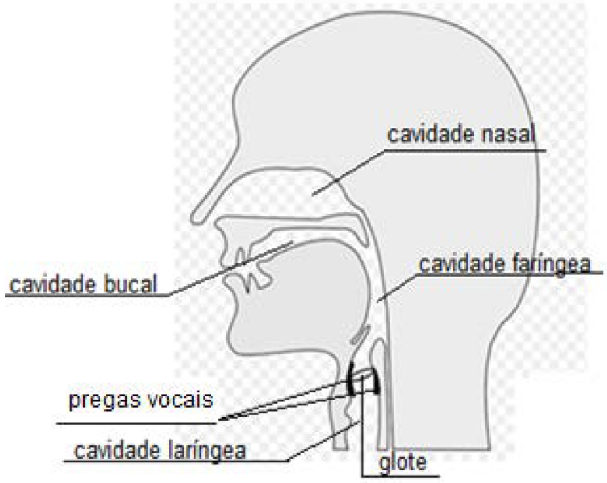
2.3 A voz humana

Uma vez estudadas, na seção anterior, a produção e a percepção do som de maneira geral, a presente seção tratará do mesmo assunto, porém em um campo específico: o da voz humana. Assim, são apresentados desde o mecanismo fisiológico que possibilita a fonação e os elementos acústicos dela resultantes, discorrendo-se também sobre algumas noções de fonética.

2.3.1 A acústica da fonação

O sistema fonador é descrito por Henrique (2002) como sendo constituído por três subsistemas: o aparelho respiratório, formado pela traqueia, brônquios e pulmões, estando estes últimos situados no tórax; as cordas vocais, localizadas acima da traqueia, na laringe, constituídas de cartilagens e músculos; e, finalmente, o trato vocal, conjunto das cavidades laríngea, faríngea, bucal e nasal e serve como estrutura de ressonância para a voz. O espaço entre as cordas vocais é denominado glote. A Figura 11 exibe um esquema do trato vocal.

Figura 11 – Esquema básico do trato vocal.



Fonte: elaboração própria.

Sabe-se que a produção do som em geral depende da existência de uma fonte de energia e de um elemento vibratório. Sendo assim, o mecanismo da fonação em particular pode ser basicamente descrito como tendo por fonte de energia o ar vindo dos pulmões que, por sua vez, entra em vibração por meio das cordas vocais (HENRIQUE, 2002, p. 674).

Passando a considerar as características do som produzido pelo sistema fonador, pode-se dizer que "uma vez que a glote pode ajustar sua frequência e amplitude, nós podemos falar com entonação e cantar" (LOY; CHOWNING, 2011b, p. 409) [5] . Porém, a frequência de que trata Loy é simplesmente aquela dita "fundamental", pois sabe-se que a voz humana não se constitui de ondas sonoras senoidais, de frequência única, mas por sons complexos, compostos pela fundamental e por outras frequências parciais. Sendo assim, convém estudar de que maneira se dá tal composição.

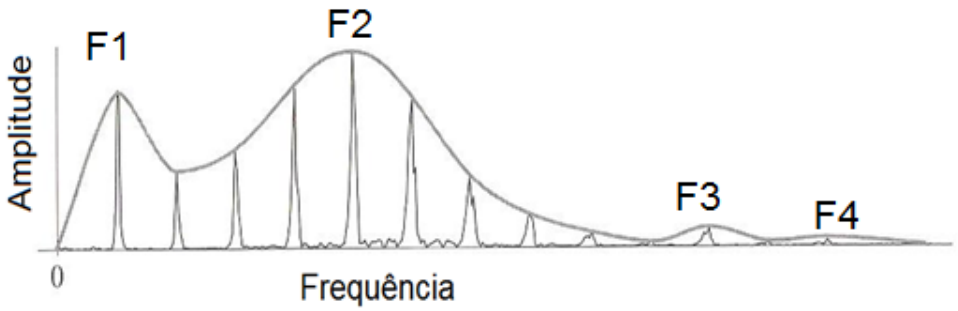
Loy e Chowning (2011a, p. 36) afirma que "quando fazemos diferentes sons vocálicos com nossas bocas, estamos amplificando alguns parciais [...] e atenuando outros"[6] . Para que se verifique adequadamente o papel de cada parcial em um som complexo, inclusive no caso

[5] Tradução nossa do original, em inglês: "Because the glottis can adjust its frequency and amplitude, we can speak with inflection and sing"

[6] Tradução nossa do original, em inglês: "When we make different vowel sounds with our mouths, we are amplifying certain partials [...] and attenuating others"

da voz humana, faz-se necessário estabelecer uma representação do sinal acústico diferente da representação temporal, vista na Seção 2.1.1 . Essa outra representação é chamada *espectro* e "exibe a *distribuição de energia* de uma forma de onda na frequência"[7] (LOY; CHOWNING, 2011a, p. 30). Desse modo, é possível perceber quais parciais num determinado som complexo têm maior amplitude em relação aos demais. A Figura 12 mostra um exemplo de representação espectral de um sinal acústico de voz.

Figura 12 – Espectro de um sinal acústico de voz.



Fonte: adaptado de Loy e Chowning (2011b, p. 410).

Note-se que a distribuição de frequências exibida pela Figura 12 é delineada por uma curva à qual se pode chamar de **envoltória espectral**[8] (FILHO, 2004, p. 40), que não se deve confundir com a curva envoltória descrita na Seção 2.1.3. Nessa envoltória espectral presente na figura acima em particular, observa-se a presença de quatro picos de energia, designados **formantes**, numerados como F1, F2, F3 e F4. Os formantes são conceituados e descritos por Henrique (2002, p. 176 e 177) nos seguintes termos:

> Formantes são zonas do espectro de grande amplitude, e que são independentes das frequências [fundamentais] das notas tocadas. Os formantes correspondem a zonas em que há grande concentração de energia acústica, e surgem como resultado de ressonâncias do sistema amplificador e radiante. A teoria dos formantes foi desenvolvida para explicar e caracterizar os sons vocálicos da fala humana

De fato, se considerarmos as quatro qualidades do som apresentadas na Seção 2.2, pode-se perceber que duas vogais distintas podem ter a mesma altura, intensidade, timbre — dado pela voz do falante — e duração. No entanto, continuam diferenciando-se entre si. Os formantes são o elemento acústico pelo qual existe essa diferenciação.

Em relação ao domínio do tempo do sinal acústico de voz, Taylor e Campbell (2001, p. 769) tecem algumas considerações acerca da importância da curva envoltória para a fala humana, o que, consequentemente, também se aplica ao caso da voz cantada:

[7] Tradução nossa do original, em inglês: "A spectrum shows the *energy distribution* of a waveform in frequency". O itálico é do autor.

[8] Menezes emprega o termo "envelope", que já assinalamos, em nota anterior, ser um sinônimo para envoltória. Preferimos nos valer dessa última palavra para manter a coerência terminológica do trabalho.

> As formas de envoltória desempenham um papel essencial na fala humana. As consoantes são geralmente alterações razoavelmente drásticas na forma envoltória. Uma oclusiva, como o "p", provoca um início de ruído aleatório (ar escapando quando os lábios são abertos) razoavelmente rápido, conduzindo a uma vogal, uma nota estável. Se se permite que o ruído cresça em amplitude mais vagarosamente, o resultado é um "f"[9]

Note-se que os autores associam as vogais a uma "nota estável", ou seja, um som musical, enquanto as consoantes são associadas a ruídos. Aqui já saímos do domínio propriamente acústico para o da percepção. O que interessa ressaltar por agora é que, no âmbito da envoltória, a estabilidade da vogal pode relacionar-se ao regime estacionário da onda, ou seja, à fase de sustentação da curva envoltória, enquanto o caráter ruidoso, em termos de percepção — ou aperiódico, em termos acústicos — das consoantes, as deixam relacionadas às demais fases, ou regimes transitórios. A descrição feita na última citação da transição da consoante "p" para a vogal, coloca esse mesmo "p" na fase de ataque, por exemplo, sendo que uma atenuação no pico do ataque acabaria por soar como "f".

As características das vogais e consoantes, bem como o lugar que ocupam na estrutura conhecida como "sílaba" e suas relações com os elementos acústicos até agora estudados são abordados na subseção a seguir, que tratará de algumas noções de fonética.

2.3.2 Noções de fonética

Compete à fonética "estudar e descrever os sons da linguagem humana" (CAVALIERE, 2011, p. 31). De acordo com essa ciência, as menores unidades distinguíveis na fala humana, denominadas **fonemas**, classificam-se em dois grandes grupos: as **consoantes**, que são vibrações aperiódicas ou ruídos e as **vogais**, que se distinguem das primeiras tanto por serem sons periódicos complexos, podendo incidir sobre elas acento de tom e/ou intensidade, quanto por constituírem núcleo de sílaba (CALLOU; LEITE, 2005, p. 26).

Nougué (2015, p. 93) define a sílaba como "o grupo fonético que se pronuncia de um só golpe ou esforço de voz". O modo como os sons da fala são agrupados em sua estrutura pode aparecer como parâmetro que possibilita a distinção entre os dois grandes grupos de fonemas, conforme explicação de Callou e Leite (2005, p. 29):

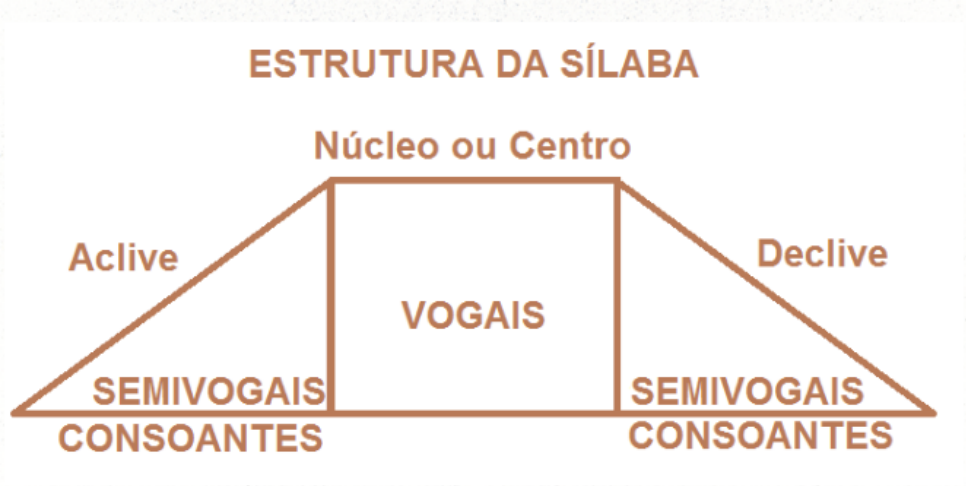
> Do ponto de vista da percepção, considera-se a cadeia sonora como composta de aclives, ápices e declives de sonoridade, cada sílaba sendo constituída de um ápice, que é seu núcleo ou centro ocupado por sons de alta sonoridade, como, por exemplo,

[9] Tradução nossa, do original em inglês: "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'"

> as vogais. Os aclives e declives constituem "vales" de sonoridade que determinam as fronteiras silábicas, suas margens, lugar preferencial das consoantes.

Para além das consoantes e vogais, é certo que há sons vocálicos que podem aparecer à margem da sílaba. Tais sons são denominados **semivogais**, sendo que Cavaliere (2011, p. 72) assinala que "inequivocamente, é esse o elemento que caracteriza a ocorrência de um ditongo na sílaba". Um exemplo de semivogal é o fonema representado pela letra "i" na palavra "pai". Uma esquematização da estrutura da sílaba tal como descrita até aqui é exibida pela Figura 13.

Figura 13 – Esquema básico da estrutura da sílaba.



Fonte: elaboração própria.

Outro fator de diferenciação entre vogais e consoantes consiste no papel que cada um desses tipos de fonema desempenha dentro da sílaba,. De acordo com a teoria quadro/conteúdo (*frame/content theory*) (MACNEILAGE, 1998), a fala é organizada em quadros silábicos, que consistem em ciclos de abertura e fechamento da boca, dentro dos quais há um conteúdo segmental, os fonemas. O quadro silábico é descrito como sendo constituído de três estruturas: o ataque, o núcleo e a coda. Enquanto no ataque e na coda se situam os fonemas consonantais, uma vogal é que constitui o núcleo da sílaba.

Além das vogais e consoantes, há fonemas designados semivogais, que caracterizam os ditongos (CAVALIERE, 2011, p. 72), que podem aparecer no ataque ou coda. De uma perspectiva acústica, a sílaba é uma forma de onda na qual as consoantes e semivogais ocupam as fases de ataque e relaxamento (coda), enquanto as vogais se situam na fase de sustentação (núcleo). A partir dos elementos até agora apresentados é possível inferir que é na vogal que se determina a altura do som cantado, é nela que se concentra a nota musical no canto. A Figura 14 apresenta um esquema de relacionamento entre a estrutura da sílaba "pai", no português brasileiro com as fases da curva envoltória.

No exemplo que acabou de ser dado, a letra "i" na palavra "pai" representava um fonema classificado como semivogal. Porém, em palavras como "igreja", a mesma letra representa uma

Figura 14 – Relação entre a estrututa da sílaba e as fases da envoltória.

Fonte: elaboração própria.

vogal que, por ser oral, tem som distinto da vogal inicial da palavra “importante”, que é nasal. Não sendo, portanto, a relação de representação entre letras e fonemas unívoca, faz-se necessário um sistema que sirva para representar especificamente os fonemas. Duas alternativas de notação para esse fim são brevemente descritas a seguir.

2.3.3 Alfabetos fonéticos

A fim de representar inequivocamente os diversos sons da fala, evitando as ambiguidades presentes nos sistemas ortográficos dos vários idiomas existentes no mundo, foram criados certos conjuntos de símbolos onde cada um deles representa univocamente um fonema. A tais conjuntos de símbolos dá-se o nome de alfabetos fonéticos (HENRIQUE, 2002, p. 704).

O alfabeto fonético que mais atualmente se tende a utilizar é o Alfabeto Fonético Internacional — conhecido também pela sigla IPA, do ingês, *International Phonetic Alphabet* — proposto pela Sociedade Fonética Internacional (CALLOU; LEITE, 2005, p. 34). Tal alfabeto, porém, é constituído por diversos caracteres especiais que não constam do alfabeto português, tais como ʃ, que representa o fonema correspondente à letra “x” na palavra “xícara”, ou ɲ, representante do som do dígrafo “nh” da palavra “senhor”. O emprego de tais caracteres pode constituir uma dificuldade para a usabilidade de sistemas computacionais nos quais a entrada de dados seja fonética, como é o caso dos sintetizadores de canto.

Diante da dificuldade em se lidar com caracteres especiais, principalmente nos computadores mais antigos, a Comunidade Econômica Europeia desenvolveu, no final da década de 1980, uma notação fonética à qual chamou SAMPA (acrônimo para *Speech Assessment Methods*

Phonetic Alphabet) (WELLS et al., 1997). Os símbolos da notação SAMPA são provenientes da tabela computacional padrão ASCII, estando presentes em qualquer teclado comum de computador que contenha o alfabeto latino. Assim, por exemplo, os fonemas representados no IPA pelos símbolos ʃ e ŋ, figuram, na notação SAMPA como S e J, respectivamente.

Uma vez estudadas as características físicas e sensoriais do som em geral e da voz humana em particular, a próxima seção tratará de algumas técnicas pelas quais os sons dos diversos instrumentos musicais podem ser reconstituídos em suas características com o auxílio da tecnologia, especialmente com o emprego de dispositivos digitais. Trata-se das técnicas de síntese sonora.

2.4 Processamento de sinais de som

A partir do momento em que os primeiros instrumentos musicais foram fabricados, a tecnologia está a serviço da música. Porém, durante o século XX, as relações entre música e tecnologia foram consideravelmente ampliadas. A possibilidade de se converter o som em um sinal elétrico analógico e vice-versa abriu caminho não apenas para a invenção de novos instrumentos musicais, mas também para a criação de formas cada vez mais sofisticadas de transmitir, processar e armazenar os sons, causando repercussões na economia, na ciência, na comunicação social e, obviamente, nas artes.

O surgimento e desenvolvimento da eletrônica digital possibilitaram uma melhoria na qualidade do processamento de som e fez com que a tecnologia empregada para isto se tornasse mais acessível, financeiramente, em relação aos equipamentos analógicos. Neste contexto, destaca-se o emprego de computadores integrados a outros dispositivos musicais eletrônicos com o auxílio do protocolo MIDI.

Esta seção apresenta alguns aspectos do processamento analógico e digital de som, explorando alguns conceitos definidos nas seções anteriores e introduzindo o estudo do protocolo MIDI.

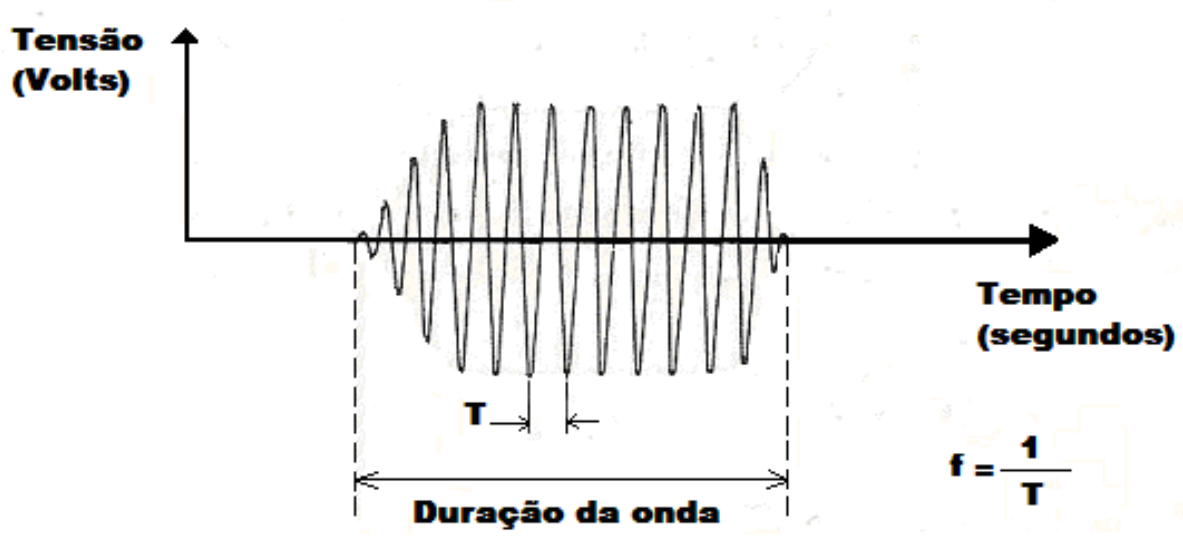
2.4.1 O sinal analógico do som

O termo “analógico” diz respeito à semelhança ou analogia entre um sinal elétrico e o fenômeno que ele pretende representar. As grandezas analógicas “podem variar continuamente, isto é, entre dois quaisquer valores, podem sempre assumir um valor intermédio” (HENRIQUE, 2002, p. 409). O presente trabalho tratará da analogia entre sinais elétricos e formas de ondas sonoras. A conversão do som em um sinal elétrico possibilitou a manipulação de suas características — altura, intensidade, timbre e duração — por meio de dispositivos e a progressiva sofisticação da tecnologia tornou a qualidade deste processamento cada vez melhor, chegando finalmente à representação digital do sinal. A seguir, alguns aspectos do processamento analógico de sons e seu emprego no ambiente musical são abordados.

A transformação do som em um sinal elétrico pode ser conseguida por meio de um dispositivo chamado transduto. As variações de pressão atmosférica do som são convertidas pelo transdutor em variações de nível de tensão elétrica, resultando num sinal com a mesma duração, forma de onda e composição de frequências, podendo este ser transformado novamente em som pela excitação de um alto-falante. Um exemplo de transdutor é o microfone (CAMILO; YABU-UTI; YANO, 1986, p. 388).

A Figura 15 exibe o aspecto de um sinal elétrico correspondente em tensão a uma onda sonora. Pode-se perceber que há certa semelhança entre a Figura 2 e a Figura 15. A variação da tensão elétrica no sinal analógico possui um período T e uma freqüência $f = 1/T$ que corresponde à freqüência da onda sonora, determinando a altura do som. A forma de onda descrita pelo sinal corresponde ao timbre. O intervalo de tempo em que a oscilação da tensão ocorre fornece a duração do som. Por fim, quando o sinal elétrico for convertido novamente em som, a potência de saída nos alto-falantes, regulada pelos controles de volume dos dispositivos, estabelecerá a intensidade sonora. Esta é, em termos gerais, a analogia que se pode fazer entre as qualidades do som e as características do sinal elétrico analógico que o representa.

Figura 15 – Sinal elétrico correspondente em tensão a uma onda sonora.



Fonte: adaptado de Camilo, Yabu-Uti e Yano (1986, p. 388).

A forma de onda do sinal elétrico de som descreverá também uma curva envoltória, sendo que, de acordo com Camilo, Yabu-Uti e Yano (1986, p. 390), a fase de ataque corresponde, em termos elétricos, ao tempo de subida da onda; as fases de decaimento e relaxamento são equivalentes aos tempos de descida; já a fase de sustentação do sinal acústico tem sua contraparte no sinal elétrico no tempo de regime estacionário. Essa forma de onda determina o timbre do som quando de sua reconversão num sinal acústico, enquanto a frequência do sinal elétrico dá a altura e a potência de saída no alto-falante resulta numa maior ou menor sensação de intensidade.

Se a curva envoltória é fator determinante para o timbre do som, pode-se concluir que ela

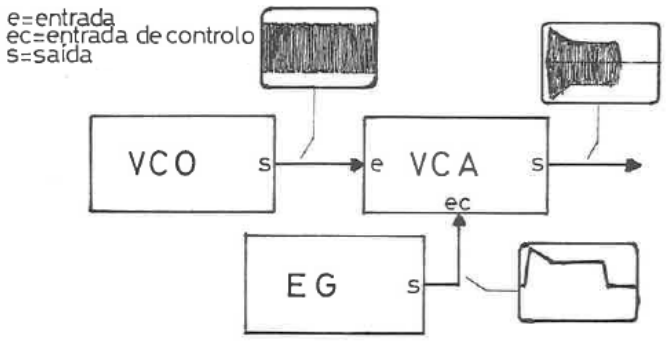
varia de acordo com o instrumento musical. Fisicamente, é possível dizer que o músico é uma fonte de energia que é aplicada ao instrumento e dissipada na forma de som e calor. A forma de dissipação dessa energia é dependente do tipo de instrumento e se reflete na duração de cada uma das fases da envoltória (CAMILO; YABU-UTI; YANO, 1986, p. 391). Por exemplo, o som dos instrumentos de percussão em geral praticamente não possuem tempo de sustentação.

O que acabou de se descrever possibilitou a Robert Moog, em 1964, desenvolver um sintetizador analógico de sons, composto por uma série de módulos (HENRIQUE, 2004, p. 404), sendo cada um deles responsável por dar ao som uma determinada característica:

- VCO (*Voltage Controlled Oscilator*), acionado pelo teclado musical e responsável por gerar uma onda com determinada frequência, conforme a nota tocada, sendo responsável, portanto, por dar altura ao som;
- VCA (*Voltage Controlled Amplifier*), que se conectava a saída do VCO, amplificando seu sinal e conferindo-lhe, desse modo, intensidade;
- EG (*Envelope generator*), que tinha por função modificar o sinal do VCA, conforme parâmetros ADSR da envoltória controlados por um painel, perfazendo uma modulação em amplitude a fim de estabelecer o timbre do som.

Do ponto de vista da percepção sonora, pode-se afirmar que o VCO é responsável por dar ao som sua altura, o VCA, sua intensidade e o EG seu timbre, o que possibilita a um único instrumento, o sintetizador, imitar o som de diversos outros instrumentos musicais. Um esquema básico do sintetizador modular de Moog pode ser visto na Figura 16.

Figura 16 – Diagrama de módulos de um sintetizador analógico.



Fonte: adaptado de Henrique (2004, p. 406).

Considerando-se o esquema apresentado, suponha-se que se quer fazer soar a nota $lá_5$ com som de flauta. Quando do acionamento da tecla correspondente a essa nota, o VCO deverá gerar uma frequência correspondente a 880 Hz — conforme descrito na subseção 2.2.1 — transmitindo-a ao VCA; enquanto isso, os parâmetros de ataque, decaimento, sustentação e

relaxamento devem ser ajustados de modo ao EG conseguir gerar uma envoltória similar à da flauta, modulando o sinal já presente no VCA, que, por fim, deverá amplificar o sinal conforme a intensidade com que foi tocada a tecla e os controles de volume do instrumento, encaminhando-o ao alto-falante, onde se converterá num sinal acústico audível semelhante ao da flauta.

2.4.2 Processamento digital de som

O sinal sonoro analógico é contínuo, requerendo uma discretização a fim de que possa ser tratado computacionalmente. Tal processamento é feito por meio de um dispositivo designado conversor analógico/digital, que está presente, por exemplo, na placa de som dos computadores. Tal dispositivo fará medições periódicas das variações do sinal elétrico contínuo, gerando um sinal discreto aproximado a partir das amostras coletadas (MANNING et al., 2001, p. 203). A fim de que as perdas provocadas pela discretização não sejam significativas, os conversores a perfazem baseando-se no teorema da amostragem, desenvolvido por Shannon e Nyquist, segundo o qual a taxa de amostragem deve ter pelo menos o dobro da frequência mais alta presente no sinal original. Como o limite máximo da audição humana está na faixa de frequências em torno de 20 kHz, uma taxa de amostragem por volta de 40000 amostras por segundo, ou 40 kHz, é suficiente para que um arquivo digital represente fielmente qualquer som audível. Uma das taxas de amostragem padrão tipicamente utilizadas é a de 44100 Hz (COLLINS, 2010, p. 18). Pode-se, então, definir o sinal digital de som como uma "representação de funções acústicas ou ondas de pressão como uma sucessão regular de aproximações numéricas discretas"[10] (MANNING et al., 2001, p. 203). A Figura 17 apresenta o processo de digitalização de uma forma de onda sonora contínua, convertida em uma série de amostras discretas.

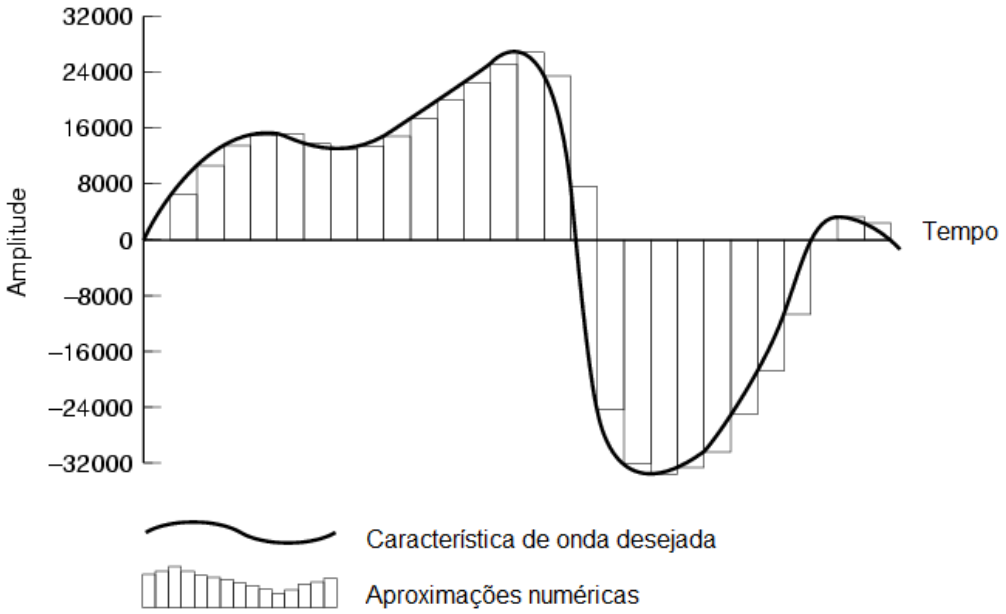
Uma vez convertido para o formado digital, o sinal de som pode ser armazenado em arquivos tais como os de extensão WAV, AU ou AIFF, entre outras. Porém, para que um sinal digital de som se torne audível, é preciso que ele passe por um conversor digital/analógico de modo que, após convertido num sinal elétrico, seja transmitido a um alto-falante do qual sairá o sinal acústico. As placas de som dos atuais computadores suportam tanto microfones quanto alto-falantes como periféricos de entrada e saída.

O advento da digitalização dos sons ampliou as possibilidades de manipulação desse tipo de sinal, inclusive no que diz respeito às técnicas de síntese. Os próprios teclados musicais sintetizadores passaram a incorporar circuitos digitais em contraposição aos antigos módulos analógicos (HENRIQUE, 2002, p. 723). Entretanto, a síntese sonora digital não é realizada apenas de maneira direta por dispositivos de hardware. Diversos tipos de software têm sido desenvolvidos ao longo das últimas décadas com essa finalidade.

Um campo da síntese sonora digital com uma grande variedade de aplicações, desde softwares de tradução até interfaces de acessibilidade, é o da síntese de fala, ou *text-to-speech*

[10] Tradução nossa do original em inglês: "representation of acoustical functions or pressure waves as a regular succession of discrete numerical approximations" (MANNING et al., 2001, p. 203)

Figura 17 – Forma de onda analógica com representação digital.



Fonte: Manning et al. (2001, p. 203).

(TTS), na qual uma entrada textual é convertida em um arquivo digital de som que simula a voz humana. Uma maneira de perfazer esse tipo de síntese é trabalhar com amostras reais de voz pré-gravadas. O maior desafio dessa abordagem é que quanto maior for o tamanho das amostras, mais natural o resultado final soará, porém, ou a gama de expressões possíveis será menor, ou o número de gravações terá de crescer consideravelmente (O'SULLIVAN; IGOE, 2004, p. 360). Assim, por exemplo, caso se queira desenvolver um sintetizador de fala para o português brasileiro, tomar como amostras palavra por palavra demandará centenas de milhares de gravações. Se, por outro lado, as amostras estiverem no nível dos fonemas, serão menos de cem e poderão reproduzir qualquer palavra, mas o resultado final de sua concatenação invariavelmente soará "robótico".

Como exemplo de sintetizador de fala pode-se citar o sistema MBROLA (DUTOIT et al., 1996), que toma como entrada uma lista de fonemas associados a parâmetros de altura e duração que, no contexto da voz falada, têm caráter antes prosódico do que musical. Tal entrada fornece as diretrizes para a manipulação de amostras de voz armazenadas num inventário. No MBROLA, essas amostras são difones, ou seja, uma conjunção de dois fonemas. Um exemplo de inventário MBROLA para o português brasileiro é o *br4*, construído para dar suporte a um sistema chamado Liane TTS[11], desenvolvido pela SERPRO em parceria com a UFRJ. Para gerar a fala em português, foi necessário realizar a gravação de 1025 difones no inventário *br4*. A

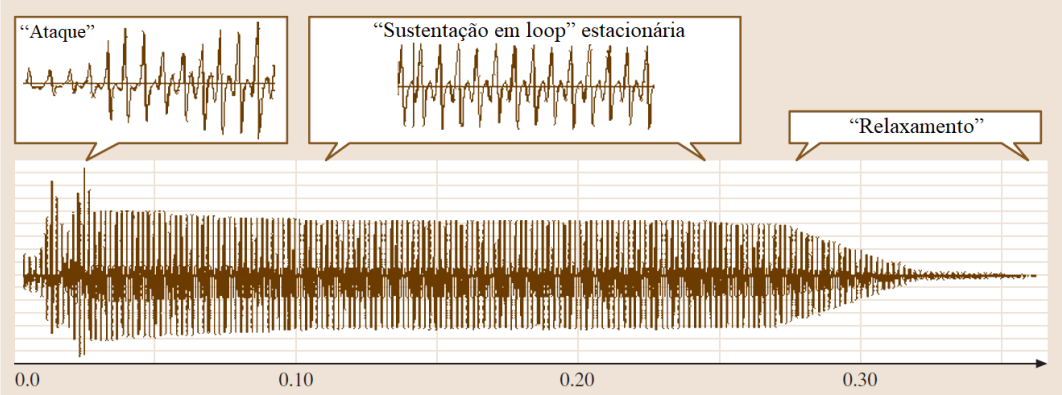
[11] <http://intervox.nce.ufrj.br/~serpro/home.htm>

sucessiva concatenação de tais unidades. gera um arquivo de áudio que possui a mesma taxa de amostragem dos sons presentes no inventário.

Uma técnica similar à que acabou de ser descrita é utilizada para fins musicais. Ela é conhecida como síntese por sons sampleados (do inglês *sample*, "amostra"). Enquanto as técnicas de síntese anteriormente desenvolvidas consistiam numa espécie de imitação, aplicando num sinal de som gerado as caraterísticas do timbre original, o que se tem agora são gravações reais de instrumentos musicais armazenadas e manipuladas conforme as necessidades do músico.

A partir das amostras gravadas, são geradas outras notas musicais de altura próxima, enquanto a duração é tratada da seguinte maneira: se a execução da nota tiver duração menor que a da amostra, a reprodução desta é interrompida; em caso contrário, há duas possibilidades, a depender do instrumento cujo som se deseja reproduzir. Para alguns instrumentos, como o piano, a gravação é executada até o final, extinguindo-se o som; no caso de outros, como os instrumentos de sopro, é desejável que sua reprodução seja prolongada indefinidamente, coincidindo com a duração do acionamento da nota, seja pelo teclado, seja por meio de algum software. Esse prolongamento indefinido é resultante da aplicação da técnica de *looping*, onde uma mesma parte da fase de sustentação da forma de onda é reproduzida continuamente. O fim do acionamento da nota faz a reprodução sair do *loop* em direção à fase de relaxamento (COOK, 2007, p. 752). Os pontos de início e término do *loop* na parte estacionária da forma de onda devem ser bem definidos, de modo que não haja para o ouvinte uma sensação de descontinuidade quando uma execução prolongada ocorrer (BRUM, 2012, p. 4). A Figura 18 apresenta uma forma de onda oriunda de um *sample* de trumpete, onde as fases de ataque, sustentação — em *loop* — e relaxamento são destacadas.

Figura 18 – Técnica de *looping* aplicada à fase de sustentação de uma forma de onda de trumpete.



Fonte: Cook (2007, p. 752).

Uma das tecnologias amplamente utilizadas para a realização da síntese por sons sampleados é o protocolo MIDI, descrito em maiores detalhes a seguir.

2.4.3 O protocolo MIDI

Musical Instrument Digital Interface, também conhecido pelo acrônimo MIDI, é um protocolo desenvolvido década de 1980 "através dos grandes construtores de eletrônica japoneses e americanos" (HENRIQUE, 2002, p. 730), entre os quais estava a Yamaha. Tal protocolo permite a comunicação entre instrumentos musicais eletrônicos e computadores (MOOG, 1986). Sua ampla aceitação fez dele um verdadeiro padrão universal para a computação musical.

As especificações do protocolo (MIDI-MANUFACTURERS-ASSOCIATION et al., 1996) estabelecem que a comunicação entre os equipamentos dá-se pela troca de mensagens constituídas de bytes estruturados de acordo com as especificações do protocolo. Tais mensagens não portam informações de áudio, mas apenas dados digitais de controle que indicam, por exemplo, que tecla foi tocada no instrumento eletrônico e a pressão exercida sobre esta tecla, indicando a altura e a intensidade do som musical. Isso significa que não há um sinal de som digital armazenado nas mensagens MIDI, ou no arquivo cujo formato é especificado pelo protocolo, mas apenas parâmetros que indicam quais *samples* devem ser acessados e como eles devem ser manipulados quando de sua reprodução. Os *samples* em questão podem estar armazenados no computador ou no próprio instrumento eletrônico. O protocolo MIDI é também constituído por especificações de hardware, que determinam como se dão as conexões físicas entre os diversos dispositivos e as transmissões de dados entre eles (O'SULLIVAN; IGOE, 2004, p. 303).

Cada uma das mensagens MIDI é iniciada com um byte de estado e pode possuir um ou mais bytes de dados. O bit mais significativo do byte de estado possui valor binário igual a 1, enquanto o byte de dados tem, no mesmo bit, valor 0. Entre os tipos de mensagem MIDI, interessa ao presente trabalho apresentar as chamadas mensagens de canal de voz (*channel voice messages*), que são identificadas pelos quatro bits mais significativos do seu byte de estado. Os outros quatro bits do mesmo byte indicam o canal de áudio para o qual a mensagem será enviada. Os bytes de dados fornecem os valores de controle que serão alterados pela mensagem. O Quadro 1 apresenta três tipos de mensagens de canal de voz do protocolo MIDI: *Note off* (nota desativada), *Note on* (nota ativada) e *Program change* (mudança de programa).

No Quadro 1 tem-se que:

- $\$nnnn = x$-1, sendo x o número do canal MIDI, isto é, \$0000 corresponde ao canal 1, \$0001 ao canal 2 e assim por diante, até o código \$1111, correspondente ao canal 16.
- $\$kkkkkkk = x$, sendo x o número da nota musical ativada ou desativada, tal que $0 \leq x \leq 127$ e, para a nota dó$_4$, $x = 60$; para dó$\sharp_4$, $x = 61$ e assim sucessivamente. Este valor é inteiro e determina a altura do som.
- $\$vvvvvvv = x$, onde x é a velocidade com que a tecla foi pressionada, o que indica a intensidade do som. A faixa de valores possíveis de velocidade também obedece à variação $0 \leq x \leq 127$, sendo que o valor zero indica a desativação da nota.

Quadro 1 – Mensagens MIDI de canal de voz

BYTE DE ESTADO	BYTES DE DADOS	DESCRIÇÃO
\$1000*nnnn*	\$0*kkkkkkk* \$0*vvvvvvv*	Nota desativada (*Note off*) \$0*kkkkkkk* = código numérico da nota musical \$00000000 = valor nulo que indica a desativação da nota
\$1001*nnnn*	\$0*kkkkkkk* \$0*vvvvvvv*	Nota ativada (*Note on*) \$0*kkkkkkk* = código numérico da nota musical \$0*vvvvvvv* = velocidade com que a tecla é abaixada
\$1100*nnnn*	\$0*ppppppp*	Mudança de programa (*Program change*) \$0*ppppppp* = número do programa

- \$*ppppppp* = *x-1*, sendo *x* o número do programa MIDI, tal que $1 \leq x \leq 128$. Este valor é comumente utilizado para selecionar um entre os diversos sons de instrumentos disponíveis num teclado musical.

Para toda mensagem de nota ativada, deve haver uma mensagem de desativação para a mesma nota no mesmo canal em algum momento. As mensagens podem ser enviadas em tempo real pelos dispositivos ou armazenadas em um arquivo para posterior execução.

Considerando a visão geral das técnicas de síntese sonora que acabou de ser dada, o presente trabalho passará a tratar especificamente do caso da síntese de voz cantada, de modo que a próxima seção apresentará as principais abordagens técnicas empregadas para tal finalidade.

2.5 Técnicas de síntese de voz cantada

Nas seções anteriores, apresentou-se basicamente a natureza do som em geral e da voz humana em particular, tanto do ponto de vista de sua produção, ao abordar os elementos acústicos que os constituem, como do ponto de vista da sua percepção, teorizada seja pela arte musical, seja pela fonética. A Seção 2.4, especificamente, tratou das técnicas de síntese sonora em geral, seja por meio da reconstituição artificial dos elementos constitutivos do som , seja através de uma abordagem orientada à percepção, caso da síntese por *samples*.

A presente seção tem por objetivo partir do caso geral da síntese sonora abordada na seção anterior, para adentrar no caso particular da síntese de voz cantada. Ao longo dos anos, várias técnicas foram desenvolvidas para realizar tal síntese. Essas técnicas podem ser classificadas de acordo com três abordagens principais (KIM, 2008): abordagem baseada em regras, abordagem baseada em *samples* e abordagem dirigida a dados.

Essas três abordagens são apresentandas em linhas gerais a seguir, buscando-se estabelecer

relações com as técnicas anteriormente apresentadas, uma vez que a primeira abordagem orienta-se à produção do som, ao gerar artificialmente constituintes acúticos, como os formantes, enquanto a segunda é orientada à percepção do som, já que se fundamenta na manipulação de amostras pré-gravadas. Já a terceira, desenvolvida mais recentemente, utiliza modelos estatísticos, não-determinísticos.Todas essas abordagens são empregadas para desenvolver sintetizadores de canto em tempo real, o que é apresentado em maiores detalhes ao final desta seção.

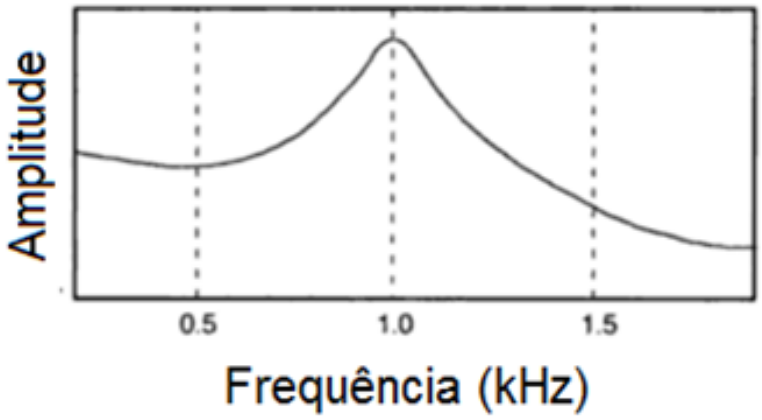
2.5.1 Abordagem baseada em regras

A síntese de canto baseada em regras, similarmente à técnica de modulação em amplitude anteriormente descrita, fundamenta-se no modo como o som é produzido, partindo da análise de suas características físicas e aplicando-as ao sinal artificialmente gerado.

Um exemplo desse tipo de abordagem é a síntese por formantes, desenvolvida em 1984 por Xavier Rodet, do IRCAM[12], resultando num sistema chamado CHANT (RODET; POTARD; BARRIERE, 1984), pioneiro em relação ao uso de voz sintetizadas no domínio artístico e capaz de sintetizar vogais realísticas ao custo de um grande esforço de análise em estúdio.

Viu-se, na Subseção 2.3.1 que os formantes figuram como picos no espectro da onda sonora, ou seja, no domínio da frequência, sendo eles os elementos acústicos que permitem que se distingam as vogais umas das outras. Cada um desses picos da envoltória espectral costuma ser identificado como F1, F2, F3 e assim por diante, conforme corresponderem uma frequência dita "central" no espectro. A numeração dos formantes se dá sucessivamente a partir das frequências centrais mais baixas (KENT; READ, 2015, p. 48). Na Figura 19, vê-se um formante aparecendo como um pico na envoltória espectral de um sinal acústico, tendo sua frequência central o valor de 1kHz.

Figura 19 – Formante com frequência central de 1kHz numa envoltória espectral.



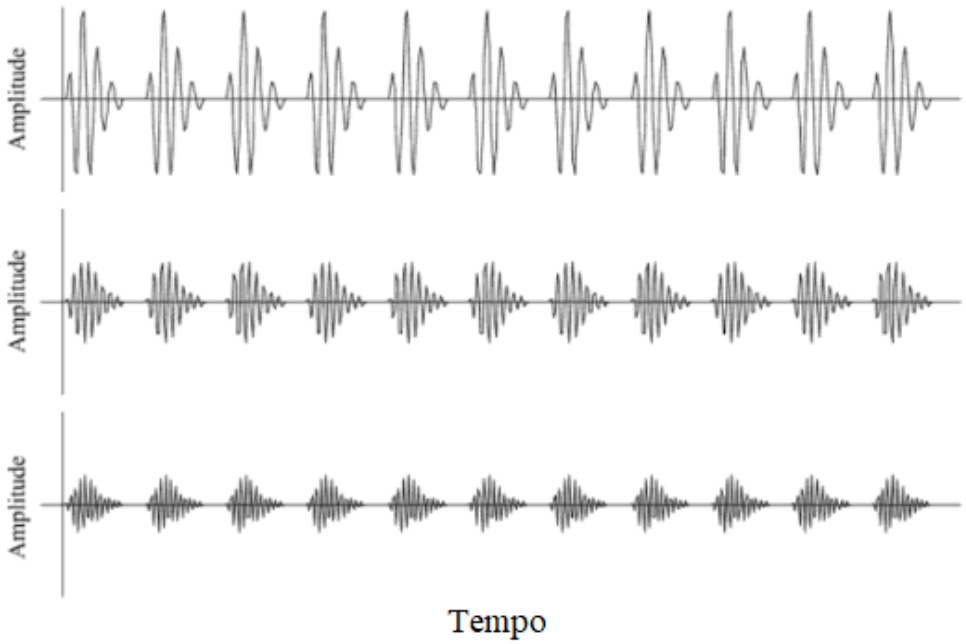
Fonte: adaptado de Roads (1996, p. 297)

[12] Acrônimo para *Institut de Recherche et Coordination Acoustique/Musique*, centro de pesquisa francês dedicado à criação de música contemporânea

A técnica de síntese de formantes caracteriza-se por estabelecer unidades designadas forma de onda formante (FOF, do francês *Forme d'Onde Formant*), sendo que cada uma dessas unidades consiste numa onda senoidal limitada cuja frequência corresponde à frequência central de um dos formantes presentes no fonema que se deseja sintetizar. Para cada fonema, são gerados tantos FOFs quanto houver formantes em sua envoltória espectral. Estabelecidas as unidades, para que se produza um tom contínuo, é necessário gerar uma sequência de FOFs de mesma frequência com uma periodicidade correspondente à frequência fundamental da nota musical desejada. Isso é feito com os demais FOFs do mesmo fonema. Essas sequências, que variam em número entre três e cinco sequências desse tipo são somadas para simular a voz (LOY; CHOWNING, 2011b, p. 418).

Como exemplo, considere-se que se quer sintetizar uma vogal cantando a nota lá$_4$. Supondo que a vogal a ser sintetizada possui três formantes, F1, F2 e F3, cujas frequências centrais tenham o valor de 1, 2 e 3 kHz, respectivamente. Deve-se gerar, então, três FOFs de curtíssima duração, cada um deles com uma das frequências mencionadas. Em seguida, cada um desses FOFs é repetido em sequência, numa frequência de 440 Hz, correspondente à nota musical, o que gerará três sequências distintas de FOFs que se repetem em uma mesma frequência, conforme exibe a Figura 20. Essas três sequencias, ao serem somadas, resultarão na voz cantada sintetizada.

Figura 20 – Sequências de FOFs geradas numa determinada frequência.

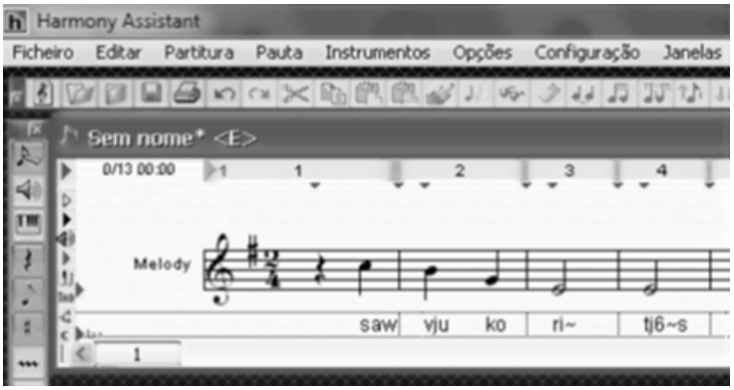


Fonte: adaptado de Loy e Chowning (2011b, p. 418).

Entre os sintetizadores comerciais que fazem uso da técnica de síntese de formantes, pode-se citar o Virtual Singer, módulo do editor de partituras Harmony Assistant, da empresa

francesa Myriad. A entrada de dados para o Virtual Singer é uma pauta com letra de canção, na qual sílabas são associadas às notas musicais. A Figura 21 apresenta o primeiro verso do "Hino do Corinthians", composto por Lauro D'Ávila, escrito em notação fonética SAMPA e associado à pauta musical correspondente à melodia do verso.

Figura 21 – Notas musicais associadas a sílabas na interface do Harmony Assistant.



Fonte: elaboração própria.

O sistema dá suporte ao canto em mais de dez idiomas e dialetos e a uma grande variedade de vozes masculinas e femininas, inclusive para canto lírico e gregoriano. A letra da canção pode ser associada às notas musicais tanto conforme a ortografia de cada idioma, quanto segundo a notação fonética SAMPA. Além de ter um formato de arquivo próprio, o Harmony Assistant também importa e exporta arquivos musicais com letras de canção nos formatos MIDI Karaoke e MusicXML.

2.5.2 Abordagens baseadas em *samples*

A síntese de fala ou de canto concatenativa é um tipo de síntese por *samples*, ou seja, é uma técnica que se vale de arquivos de som pré-gravados. No caso dos sintetizadores de canto, amostras de voz vão sendo concatenadas umas às outras à medida que as entradas fonética e musical vão sendo percorridas pelo sistema. Conforme estabelecido na Subseção 2.3.2 é na vogal propriamente que se concentra a nota musical e é ela que basicamente corresponde à fase de sustentação da sílaba. Portanto é na vogal que deve ser aplicada a técnica de *looping* descrita na Suseção 2.4.2, a fim de se prolongar sua duração conforme os parâmetros musicais passados. As consoantes e semivogais aparecerão concatenadas às margens da vogal.

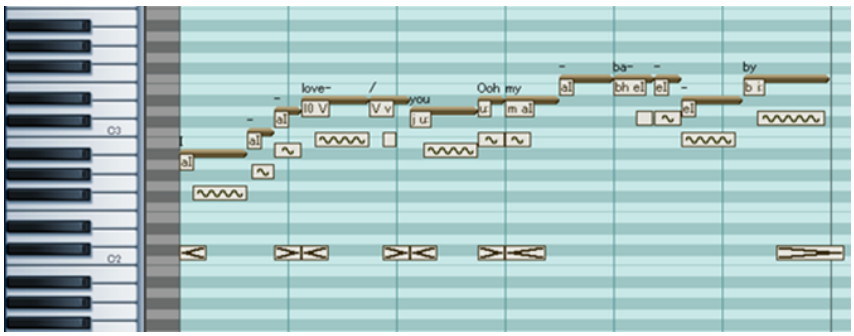
As amostras pré-gravadas ficam armazenadas num inventário de canto constituído por unidades que podem ser modeladas de modo a conter um ou mais fonemas. O fato de que a variação de altura das vogais cantadas é bem menor do que nas faladas, uma vez que as primeiras precisam de certa forma corresponder às notas musicais da canção, não exclui de todo a dificuldade em se obter um resultado tido por "realístico" a partir de amostras. Além disso, se as unidades do inventário forem difones, viu-se, na Subseção 2.4.2, que pouco mais de mil

gravações são necessárias para gerar a fala no português brasileiro. No desenvolvimento de um inventário de canto ainda inédito para este idioma, este número de gravações teria de ser multiplicado de acordo com o intervalo de alturas pretendido para o sintetizador de canto, pois os mesmos difones precisariam ser gravados tanto para regiões mais graves, quanto para as mais agudas do canto.

O sistema LYRICOS é citado como exemplo de sintetizador de canto baseado em concatenação por Kim (2008). Tal sistema, pioneiro na sintese de canto baseada em *samples*, é descrito no trabalho de Macon et al. (1997). Esse sintetizador utilizava o protocolo MIDI para controlar o acesso aos samples e sua manipulação. Já entre os sequenciadores comerciais atuais que produzem canto artificialmente por meio da síntese baseada em concatenação, está o Vocaloid, desenvolvido pela Yamaha (KENMOCHI; OHSHITA, 2007), que também se vale do protocolo MIDI e cujo funcionamento é apresentado, em linhas gerais, a seguir.

O Vocaloid dispõe de uma interface do tipo *piano roll*, composta por um teclado virtual agregado a uma tabela cujo preenchimento corresponde às notas musicais escolhidas. A entrada de dados pode ser feita por periféricos convencionais, como um mouse, ou por meio de instrumentos musicais eletrônicos via protocolo MIDI. A letra da canção é associada às notas à medida que é digitada no piano roll. Efeitos de expressão, como *vibrato*, e acentos agógicos podem ser também atribuídos às notas. A Figura 22 apresenta uma visão parcial do *piano roll* de uma das versões do Vocaloid.

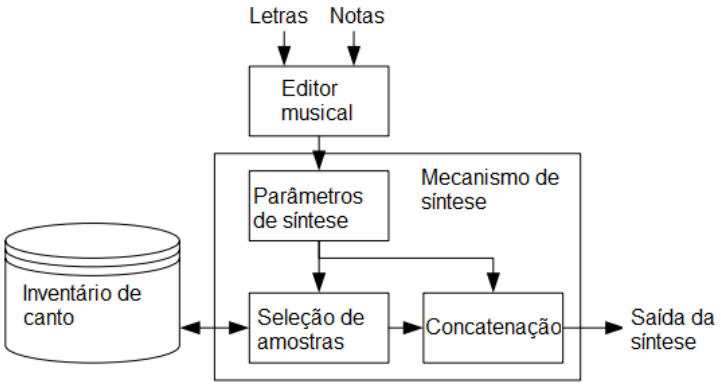
Figura 22 – Visão parcial do *piano roll* do Vocaloid.



Fonte: elaboração própria.

A partir desse editor musical, que serve de entrada para a letra da canção e as notas a ela associadas, os parâmetros musicais e fonéticos são transmitidos ao mecanismo de síntese, onde servirão de critério para a seleção de amostras proveniente do inventário de canto do sistema. Da concatenação dessas mesmas amostras, resultará o canto sintetizado como saída do processamento de dados. O diagrama da Figura 23 apresenta o funcionamento básico do Vocaloid.

Figura 23 – Diagrama de funcionamento do sistema Vocaloid.



Fonte: adaptado de Kenmochi e Ohshita (2007, p. 4009).

2.5.3 Abordagens dirigidas a dados

Nos últimos anos, alguns sintetizadores de fala (TTS) e canto vêm sendo desenvolvidos com base em modelos probabilísticos (TABET; BOUGHAZI, 2011), em contraposição ao caráter determinístico da abordagem baseada em regras. Entre eles estão o HNM (*Harmonic plus Noise Model*) e o HMM (*Hidden Markov Model*), que são descritos brevemente a partir de agora.

O modelo HNM pressupõe que o som possui uma parte harmônica e uma parte ruidosa. A parte harmônica é modelada a partir de senoides que representam múltiplos de uma determinada frequência, chamada frequência fundamental. Já na parte ruidosa, como não há frequência fundamental que gere múltiplos, as diversas frequências são geradas a partir de um filtro aplicado numa onda especificada arbitrariamente com 100 Hz. A parte harmônica representa os sons do idioma chamados "sonoros" em fonética, como as vogais e algumas consoantes que fazem vibrar a laringe. Já a parte ruidosa, representa os sons da fala foneticamente designados como "surdos". Entre os sintetizadores de canto que empregam o modelo HNM pode-se citar o sistema descrito no trabalho de Gu e Liao (2008).

Por sua vez, o modelo HMM, têm caráter estatístico e seu uso consistem em duas fases. Na fase de treinamento, certos parâmetros são extraídos a parrtir de uma gravação de áudio natural, que não precisa ficar armazenada para o sistema final. Na segunda fase, a de síntese, o áudio é gerado conforme a entrada textual a partir das inferências estatísticas dos parâmetros obtidos na fase de treinamento. A qualidade do áudio gerado não é tão boa quanto a da síntese concatenativa, mas os custos de armazenamento são muito menores.

O modelo HMM, que vem sido utilizado com sucesso em sistemas TTS (LI; CHEN; REN, 2014; SUGIURA; ZETTSU, 2016), mostou-se útil para, por exemplo, aplicar em amostras contendo um único fonema o comportamento advindo da análise estatística da voz de um cantor (KHAN; LEE, 2015; FREIXES; SOCORÓ; ALÍAS, 2016), o que diminui o tamanho do inventário de canto ao mesmo tempo em que minimiza de modo mais eficaz que a abordagem concatenativa a questão da "não naturalidade" da voz sintetizada.

O primeiro sintetizador de canto desenvolvido com base em HMMs foi o SinSy (OURA et al., 2010), do Instituto de Tecnologia de Nagoya. O sistema está disponível num site, no qual é possível fazer o upload de um arquivo no formato MusicXML, que pode ser gerado pela maior parte dos editores de partitura disponíveis, devolvendo o canto sintetizado em formato WAV. Os idiomas disponíveis no SinSy são o inglês e o japonês.

Mais recentemente, outros modelos, como as redes neurais, e métodos computacionais como *deep learning* têm sido utilizados para desenvolver sistemas sintetizadores de voz cantada.

2.5.4 Síntese de canto em tempo real

Em softwares como o Vocaloid, o usuário descreve as entradas (letra e notas musicais) para que o sistema num momento posterior gere o canto, de modo análogo ao que acontece geralmente nos ambientes de programação, onde o tempo de projeto e o tempo de execução são distintos.

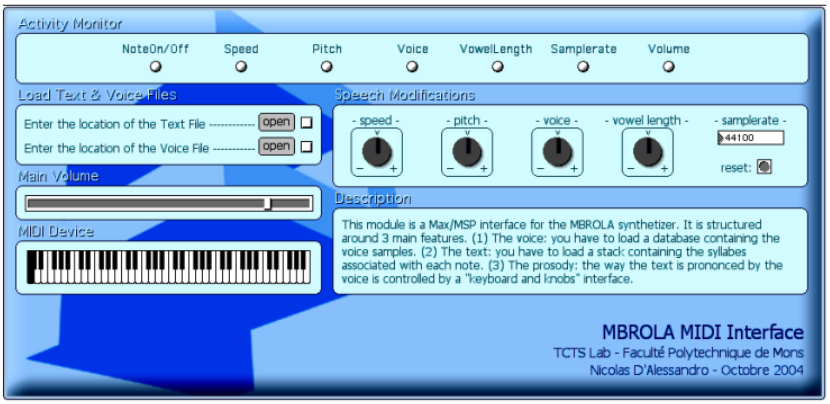
Essa limitação tem sido superada pela produção de sintetizadores de canto em tempo real, sistemas embarcados que geram o canto artificialmente no mesmo instante em que as entradas são indicadas pelo usuário, o que permite a ele o uso do sintetizador de canto como se fosse um instrumento musical (CHAN et al., 2016).

Há diversas ferramentas disponíveis para perfazer o processamento digital de áudio em tempo real, entre as quais destacam-se as desenvolvidas por Miller Puckette. Este pesquisador criou na década de 1980, em seu trabalho no IRCAM, uma linguagem de programação chamada Max, cuja instanciação e configuração de objetos era controlada por um módulo externo, um ambiente gráfico voltado para a computação musical denominado The Patcher (PUCKETTE, 1988). Alguns anos mais tarde, Puckette desenvolveu de maneira independente uma outra linguagem de programação no intuito de superar as limitações de Max. Esta nova ferramenta chama-se Pure Data (PUCKETTE, 1997) e foi concebida em si mesma como uma linguagem de programação visual com capacidades de processamento de áudio em tempo real. Uma versão comercial da linguagem Max, chamada Max/MSP (ZICARELLI, 1997), foi então lançada pela companhia Cycling ’74, fundada por David Zicarelli. Max/MSP incorporou as características mencionadas de Pure Data a Max.

Um exemplo de sintetizador de voz cantada em tempo real é a ferramenta MaxMBROLA (D’ALESSANDRO et al., 2005), que por meio dos mecanismos da linguagem Max/MSP, controla por meio do protocolo MIDI o áudio gerado pelo sintetizador de fala MBROLA, mencionado na Subseção 2.4.2, transformando-o, em tempo real, numa voz cantada. A inferface do sistema MaxMBROLA pode ser vista na Figura 24.

Outra ferramenta que possibilita o processamento de áudio em tempo real é um software livre chamado Supercollider (MCCARTNEY, 1996). A arquitetura de uma aplicação SuperCollider é composta, basicamente, pelos seguintes componentes:

Figura 24 – Interface do sintetizador MaxMBROLA.

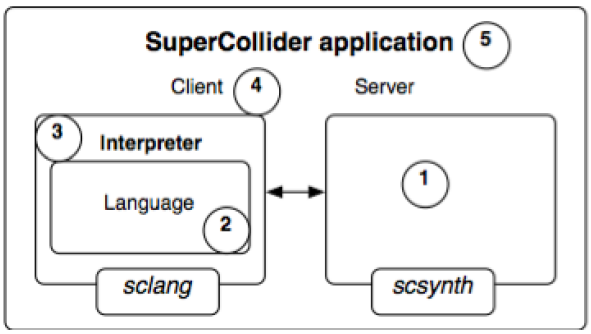


Fonte: D'Alessandro et al. (2005, p. 4)

1. Um processo servidor de áudio (*scsynth*) que consegue tratar entradas e saídas desse tipo de mídia em tempo real, além de lidar com mensagens MIDI, entre outras funções;
2. A linguagem de programação de áudio SuperCollider, que é dinamicamente tipada e orientada a objetos;
3. Um interpretador para a referida linguagem de programação (*sclang*);
4. O programa interpretado que serve como cliente para o servidor de áudio.
5. A aplicação propriamente dita, que consiste na integração entre os componentes anteriores.

Estes cinco componentes e seus relacionamentos são apresentados na Figura 25, extraída da documentação do SuperCollider[13], conforme a numeração dada acima. O SuperCollider ainda conta com um ambiente integrado de desenvolvimento (*scide*), composto por um editor de texto para a linguagem e um sistema de ajuda.

Figura 25 – Arquitetura de uma aplicação SuperCollider.



Fonte: documentação do SuperCollider.

[13] Disponível em <https://doc.sccode.org/Guides/ClientVsServer.html>.

O sintetizador de canto em tempo real PATRICIA, objeto de estudo da presente dissertação, controla o áudio oriundo do sintetizador de fala MBROLA, de maneira similar ao que faz o sistema MaxMBROLA. Entretanto, ao contrário deste último, PATRICA foi implementado em SuperCollider. Os detalhes desta implementação são expostos na Subseção 4.1.3

A fim de melhor compreender o ramo da síntese de canto em tempo real, um mapeamento científico foi conduzido para estabelecer seu estado da arte e da técnica. Os estudos selecionados são apresentados em maiores detalhes no próximo capítulo.

3

Trabalhos Relacionados

A fim de realizar a busca por trabalhos relacionados, optou-se pelo emprego de métodos sistemáticos, pois eles garantem a reprodutibilidade da pesquisa, além de torná-la menos suscetível ao enviesamento pelas preferências pessoais do autor, uma vez que este deve se comprometer a analisar todos os trabalhos advindos da busca definida. É certo que uma tal maneira de conduzir a pesquisa pode acabar por não levar em consideração trabalhos relevantes que não se encaixarem nos critérios estabelecidos, mas entendeu-se que, para fins científicos, as vantagens da utilização dos métodos sistemáticos aqui expostas superam as eventuais desvantagens.

O presente trabalho realizou um mapeamento sistemático da literatura que atualiza os resultados de um mapeamento anterior conduzido em 2019 (BRUM; MORENO, 2019; BRUM; MORENO, 2020). Segundo (PETERSEN et al., 2008), este método consiste em um protocolo sistemático para busca e seleção de estudos relevantes com o objetivo de extrair dados e mapear resultados para um problema de pesquisa específico. O processo foi conduzido com o auxílio de outros dois métodos complementares: um mapeamento tecnológico (ORJUELA et al., 2019) e uma revisão sistemática da literatura (KITCHENHAM, 2004).

No mapeamento tecnológico, buscou-se patentes relacionadas à síntese de canto em tempo real. A revisão sistemática da literatura realiza a busca em bases de dados científicas. Esses dois métodos fornecem, respectivamente, uma melhor compreensão do estado da técnica e o estado da arte da síntese da voz cantada em tempo real. Ambos os processos são descritos nas duas seções a seguir.

3.1 Mapeamento Tecnológico

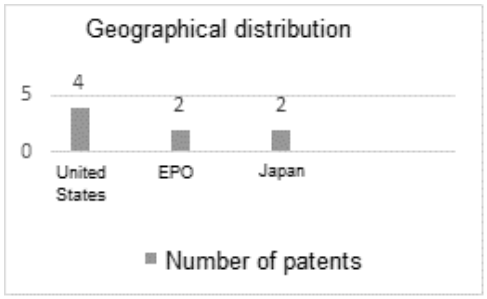
A busca de patentes relacionadas a sintetizadores de voz cantada em tempo real foi realizada na base de dados da *OMPI* (*Organização Mundial da Propriedade Intelectual*), denominada

Patentscope[1] A seguinte string de pesquisa foi utilizada:

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

Esta busca resultou em oito depósitos de patentes. Todos eram de propriedade da Yamaha Corporation, do Japão, e tinham como objeto um aparato, método e meio de armazenamento para síntese de canto em tempo real. No entanto, a maioria das patentes foi registrada fora do Japão, provavelmente para garantir proteção legal internacional. A figura 26 apresenta a distribuição geográfica das patentes.

Figura 26 – Número de patentes encontradas em cada região geográfica.



Fonte: Patentscope.

O produto patenteado foi apresentado como um protótipo em um artigo científico de 2012 (KAGAMI et al., 2012). Tal protótipo foi denominado Vocaloid Keyboard e consistia em um teclado musical com sintetizador de canto Vocaloid como sistema embarcado, permitindo ao seu usuário realizar uma performance instrumental. A presença do Vocaloid como software do instrumento indica que a abordagem técnica empregada foi a de síntese baseada em *samples*. Em se tratando do hardware, uma placa Arduino foi utilizada, ao menos na fase de prototipação.

O protótipo do Vocaloid Keyboard possuía, em seu painel de controle, botões alfabéticos à esquerda, organizados da seguinte forma: duas fileiras horizontais com sinais diacríticos e consoantes e, abaixo delas, cinco botões com as vogais, organizadas em forma de cruz. Com a mão esquerda, o usuário pode pressionar esses botões para gerar sílabas; enquanto isso, o teclado musical pode ser tocado pela mão direita para fornecer as notas musicais. Um display com caracteres japoneses *katakana* mostra as sílabas geradas. A figura 27 apresenta o protótipo descrito.

No entanto, a versão comercial do Vocaloid Keyboard, o VKB-100 keytar, abandonou os botões fonéticos presentes no protótipo em favor de fornecer a letra da música antecipadamente, e não mais em tempo real (KASHIWASE, 2017). Os botões fonéticos foram, assim, substituídos por outros que dão expressão à música, controlados pela mão esquerda do usuário.

[1] Disponível em <https://patentscope.wipo.int/>.

Figura 27 – Protótipo do Vocaloid Keyboard.

Fonte: Kagami et al. (2012, p. 839).

3.2 Revisão Sistemática da Literatura

Nesta seção, descreve-se como foi realizado o processo de busca e seleção dos estudos primários. Para tanto, foi necessário definir as questões de pesquisa, a estratégia de busca e seleção e os critérios de seleção.

3.2.1 Questões de pesquisa

O mapeamento sistemático proposto procurou identificar e analisar as ferramentas existentes para realizar a síntese de canto em tempo real. Para alcançar tal objetivo, foram elaboradas duas questões a serem respondidas por meio da análise das obras selecionadas:

1. Qual é a abordagem de síntese de canto empregada pela maioria dos sistemas em tempo real?
2. Como tais sistemas recuperam os dados de entrada fonética?

3.2.2 Estratégia de busca e seleção

As seguintes bases de dados de Ciência da Computação foram selecionadas para a realização da busca: Scopus, IEEE Xplore e Web of Science. O serviço CAFE[2], disponível no portal de periódicos da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), foi utilizado para acessar todos as trabalhos presentes nas três bases de dados.

Uma string foi desenvolvida de acordo com os termos e palavras-chave do objetivo da pesquisa, o que padroniza a busca nas bases citadas. A string utilizada foi a seguinte:

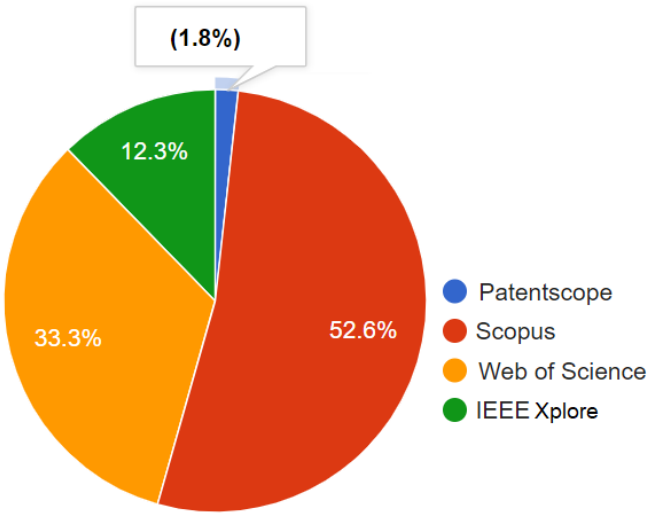
```
( ( "singing") AND ( "voice synthesis"OR "vocal synthesis"OR
"singing synthesis") AND ( "real time"OR "real-time"OR "performative") ) .
```

Essas buscas foram realizadas em novembro de 2022, com o auxílio da ferramenta Parsifal[3], e retornaram um total de 56 trabalhos encontrados, conforme a seguinte distribuição: Scopus (30),

[2] Disponível em <https://www.rnp.br/servicos/servicos-avancados/cafe>.
[3] Disponível em <https://parsif.al/>.

Web of Science (19) e IEEE Xplore (7). Em tais obras foi iniciada a fase de seleção, obedecendo aos critérios e procedimentos apresentados na subseção a seguir. O artigo que apresenta o Vocaloid Keyboard (KAGAMI et al., 2012), encontrado no mapeamento tecnológico descrito na subseção anterior, foi adicionado manualmente ao mapeamento sistemático, totalizando 57 trabalhos. A Figura 28 apresenta o percentual de artigos identificados por base bibliográfica.

Figura 28 – Percentual de artigos identificados por base bibliográfica.



Fonte: Parsifal.

3.2.3 Critérios de seleção

A fim de filtrar os artigos relevantes para efeito deste mapeamento sistemático, foram definidos os critérios de inclusão (IC) e exclusão (EC) dos artigos. Eles são apresentados no Quadro 2.

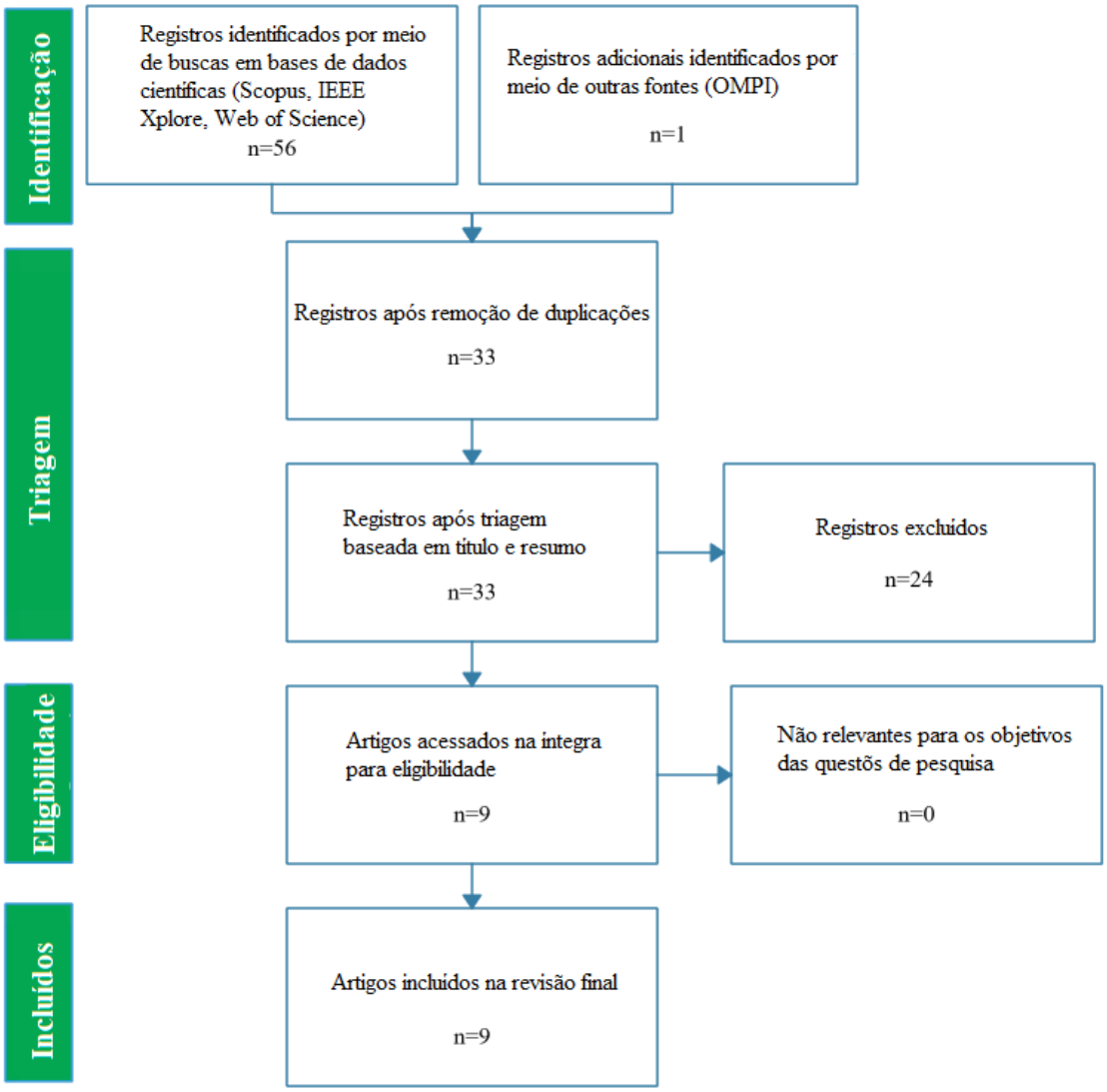
Quadro 2 – Critérios de Inclusão e Exclusão

Critérios de Inclusão (CI)	Critérios de Exclusão (CE)
CI1- Trabalhos desenvolvidos nos últimos 10 anos (2012-2022). CI2- Obras que descrevem um novo sintetizador de canto.	CE1- Obras duplicadas. CE2- Assunto duplicado (mesmo sintetizador descrito). CE3 - Trabalhos que não estão totalmente disponíveis.

Para ilustrar a aplicação do Quadro 2 neste estudo, a Figura 29 mostra o processo de seleção dos trabalhos. Após a busca nas bases de dados científicas, os artigos do resultado

da busca tiveram seus títulos e resumos lidos e aplicados os critérios de inclusão e exclusão. Um diagrama de fluxo PRISMA — *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (MOHER et al., 2009) é usado para apresentar cada etapa desse processo, o número de estudos elegíveis excluídos e as respectivas justificativas.

Figura 29 – Diagrama PRISMA que descreve o processo de seleção dos artigos.



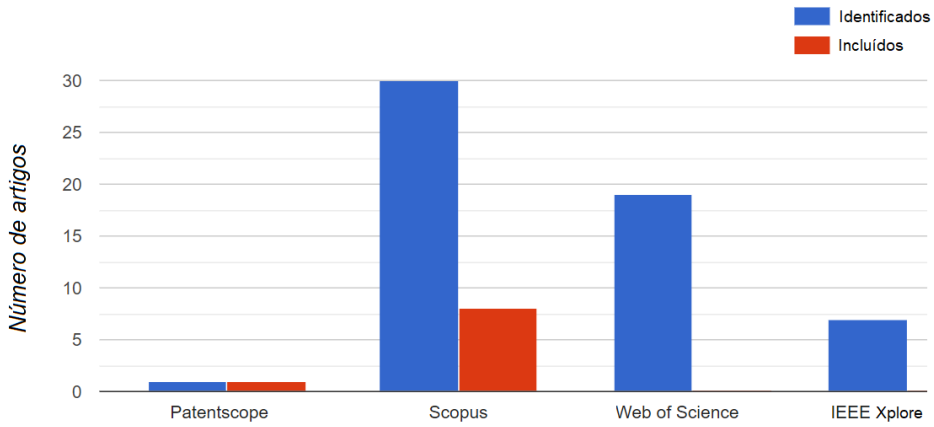
Fonte: elaboração própria.

3.3 Artigos selecionados

Ao final do processo, nove trabalhos foram incluídos na revisão final. Uma relação entre os artigos identificados e os artigos incluídos no estudo por base bibliográfica é exibida pela Figura 30. Como a base Scopus exibe resultados oriundos de outras bases, houve a tendência

de incluir os artigos na revisão final via Scopus e rejeitar os registros das bases originais como trabalhos duplicados.

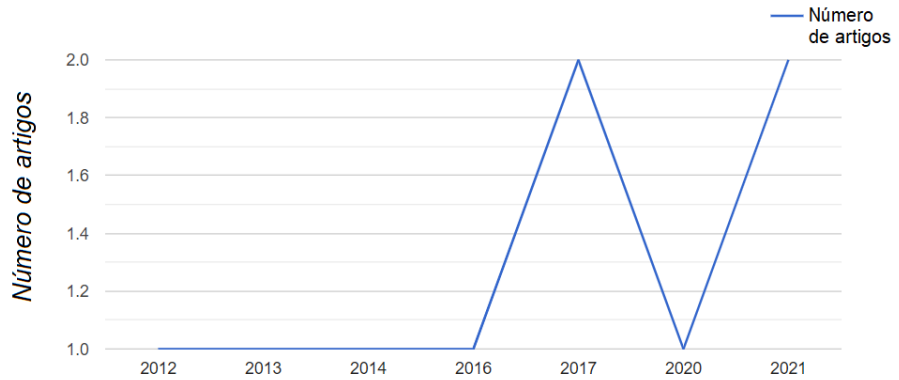
Figura 30 – Relação entre artigos identificados e incluídos por base bibliográfica.



Fonte: Parsifal.

Outro dado extraído do processo de revisão sistemática foi o da quantidade de artigos incluídos por ano de publicação, conforme mostrado pela Figura 31. O que se verifica é que, embora a síntese de voz cantada em geral seja uma área bem estabelecida na computação musical, com quatro décadas de trabalhos desenvolvidos, a síntese de canto em tempo real ainda é um campo de pesquisa restrito, com poucas publicações, embora venha contando com novos desenvolvimentos nos últimos anos. Em todo caso, procurou-se restringir o escopo da pesquisa a fim de analisar soluções mais próximas de PATRICIA. Uma busca por sintetizadores de canto em geral, ou mesmo de voz, tornaria a pesquisa demasiado abrangente, de modo que muito do que seria descrito teria pouca relação com o objeto de estudo proposto.

Figura 31 – Quantidade de artigos incluídos na revisão por ano de publicação.



Fonte: Parsifal.

Os sintetizadores de canto em tempo real descritos pelos trabalhos pesquisados estão apresentados a seguir, ordenados do mais recente para o mais antigo.

MLP Singer (TAE; KIM; LEE, 2021). Sintetizador que se vale de técnicas de *deep learning*. Utiliza uma rede neural perceptron multicamadas (*multi-layer perceptron*, MLP) para prover a síntese de canto no idioma coreano, o que representa uma abordagem dirigida a dados. A rede neural foi treinada com 45 canções interpretadas por uma cantora profissional. Cada canção foi cantada duas vezes em diferentes tonalidades, sendo acompanhadas por dados MIDI e anotações textuais, que são, respectivamente, os formatos de entrada musical e fonético do sistema. As sílabas são dividas em três partes: ataque, núcleo e coda, sendo o núcleo ocupado pela vogal, recebendo a maior parcela do tempo. As entradas percorrem as camadas da rede neural, que aplicam a elas diversas transformações, resultando numa saída de áudio projetada em um espectrograma.

Full-Band LPCNet (MATSUBARA et al., 2021). Sintetizador de fala e canto que funciona como um *vocoder*, isto é, modula o sinal de voz de acordo com outro sinal que lhe fornece a frequência, denominado onda portadora (*carrier*). Utiliza uma rede neural treinada com 48 canções executadas *acapella* por uma cantora japonesa.

Voks (LOCQUEVILLE et al., 2020). Família de instrumentos musicais digitais que, via protocolo MIDI, fornecem melodia para sinais de voz pré-gravados, transformando-os em canto em tempo real. Seu método de entrada gestual é denominado quironomia (*chironomy*), que consiste numa analogia entre movimentos manuais e os parâmetros fonéticos e musicais requeridos pela síntese de canto. Um dos instrumentos da família é o T-Voks (XIAO et al., 2019), um theremin com funcionalidades aumentadas.

Cantor Digitalis (FEUGÈRE et al., 2017). Sintetizador que também se vale da quironomia para perfazer a síntese por formantes, que se limita a produzir apenas vogais. Com uma das mãos, o usuário aciona um tablet com um *stylus*, dispositivo similar a uma caneta, no intuito de informar o contorno melódico desejado; ao mesmo tempo, com os dedos da outra mão, indica gestualmente no tablet a vogal a ser sintetizada.

VOKinesiS (DELALEZ; D'ALESSANDRO, 2017). Sintetizador que aproveitou a interface do Cantor Digitalis, integrarando-a a pedais para se constituir outro sistema. Neste, amostras de voz pré-gravadas são transformadas por meio do controle de altura provido pelo Cantor Digitalis, enquanto os pedais fornecem parâmetros temporais que alteram o ritmo das amostras originais.

SERAPHIM (CHAN et al., 2016). Propondo-se a superar certas limitações do Cantor Digitalis — que só sintetiza vogais — e do Vocaloid Keyboard, cujas capacidades de síntese em tempo real estão no nível do quadro (sílaba) e não do conteúdo (fonemas)(MACNEILAGE, 1998), o sistema SERAPHIM fornece uma entrada gestual que possibilita a síntese fonema-a-fonema em tempo real, tanto de vogais quanto de consoantes. A técnica utilizada é a de síntese

concatenativa, baseada em samples, estando o inventário de canto armazenado em estruturas indexadas denominadas *wavetables*. O sistema provê síntese de fala e canto para o idioma mandarim e apenas de canto para o japonês.

I²R Speech2Singing (DONG et al., 2014). Sistema desenvolvido em Singapura que converte instantaneamente uma entrada de voz em canto, aplicando à voz do usuário as características das vozes de cantores profissionais presentes em sua base de dados. Trata-se por tanto, de um exemplo de sistema que se vale da abordagem dirigida a dados, onde se extraem os parâmetros de um sinal para aplicá-los a outro.

MAGE/pHTS (VEAUX et al., 2013). Sistema distribuído que utiliza modelos HMM treinados para perfazer a síntese de voz cantada em tempo real por meio de um controle gestual, valendo-se do dispositivo Kinect. Utiliza o ambiente PureData de programação em tempo real. Permite o controle interativo e simultâneo das entradas do sistema por múltiplos usuários.

Vocaloid Keyboard (KAGAMI et al., 2012) Produto desenvolvido pela Yamaha que consiste num teclado musical com um sintetizador de canto embarcado, possibilitando ao usuário a realização de uma performance instrumental com seu cantor virtual. A entrada fonética é feita pela mão esquerda, que aciona botões com caracteres alfabéticos. Descrito em maiores detalhes na Seção 3.1.

3.4 Resultados do mapeamento

A primeira das questões de pesquisa estabelecidas na revisão sistemática foi "Qual é a abordagem de síntese de canto empregada pela maioria dos sistemas em tempo real?".

Para respondê-la, o Quadro 3 apresenta as abordagens técnicas utilizadas pelos sintetizadores de voz cantada em tempo real descritos por cada um dos artigos selecionados. Os trabalhos aparecem em ordem cronológica.

Quadro 3 – Abordagens técnicas utilizadas pelos sintetizadores estudados.

Sintetizador	Baseado em regras	Baseado em samples	Dirigido a dados
MLP Singer (TAE; KIM; LEE, 2021)			✓
Full-Band LPCNet (MATSUBARA et al., 2021)			✓
Voks (LOCQUEVILLE et al., 2020)		✓	
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)		✓	

Constatou-se que a maior parte dos sintetizadores de canto em tempo real pesquisados

utiliza a abordagem baseada em samples ou a abordagem dirigida a dados, totalizando quatro sintetizadores cada uma delas, sendo que apenas um utiliza a abordagem baseada em regras. O fato de a abordagem baseada em regras ser a mais antiga pode ser a explicação para que ela esteja menos presente nos trabalhos mais recentes. A Subseção 4.2.1 discute a abordagem de síntese escolhida para o sintetizador PATRICIA tendo em vista os resultados apresentados pela trabela 3.

A outra questão de pesquisa proposta por este trabalho indagava o seguinte: "Como os sintetizadores em tempo real obtêm os parâmetos fonéticos?", um elemento crítico em se tratando de sistemas que se propõem a prover uma performance em tempo real. Os artigos analisados apontaram para três formas de entrada: arquivos texto, sinais de áudio e controles manuais. O Quadro 4 apresenta a comparação entre os artigos sob este último aspecto.

Quadro 4 – Métodos de entrada fonética dos sintetizadores de canto descritos nos artigos analisados.

Sintetizador	**Arquivos de texto**	**Sinais de áudio**	**Controles manuais**
MLP Singer (TAE; KIM; LEE, 2021)	✓		
Full-Band LPCNet (MATSUBARA et al., 2021)		✓	
Voks (LOCQUEVILLE et al., 2020)		✓	
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)			✓

Verificou-se que a maior parte dos sintetizadores pesquisados, a saber, cinco deles, emprega um sinal de áudio para fornecer a entrada fonética. Três se valem de algum tipo de controle manual, enquanto os arquivos de texto são utilizados por somente um dos sintetizadores para este fim. Uma hipótese que pode explicar tal resultado é a de que os arquivos texto precisam ser preparados previamente, tirando o caráter de tempo real da entrada fonética, enquanto os controles manuais acabam sendo bastante limitados para representar toda a gama de fonemas das diversas linguagens existentes. A contribuição de tal resultado para a implementação do sintetizador é abordada na Subseção 4.2.2 .

O mapeamento científico descrito acima forneceu as diretrizes para o projeto e implementação do protótipo de sintetizador de canto proposto por este trabalho, definindo sua abordagem técnica e método de entrada fonética. O desenvolvimento do protótipo e sua relação com os resultados do mapeamento são descritos em maiores detalhes no próximo capítulo.

4

O sintetizador PATRICIA

O sistema aqui proposto é, até onde se pôde pesquisar, o primeiro sintetizador de canto projetado especificamente para cantar no idioma português brasileiro. O sintetizador foi denominado PATRICIA, acrônimo para *Programa que Articula em Tempo Real o Idioma Cantado Inscrito em Arquivo*.

O processo de desenvolvimento do sistema, os aspectos técnicos que guardam relação com o mapeamento sistemático descrito no Capítulo 3, o controle das qualidades do som na saída de áudio do sintetizador e os experimentos conduzidos para validar os requisitos funcionais e avaliar o desempenho de PATRICIA são discutidos nas seções seguintes.

4.1 Desenvolvimento do sistema

A presente seção descreve o modelo de ciclo de vida de software empregado para realizar o desenvolvimento do sintetizador, a articulação entre os diversos componentes que constituem a arquitetura do sistema e a tecnologia e algoritmo utilizados em sua implementação.

4.1.1 Modelo de desenvolvimento

O desenvolvimento do sintetizador PATRICIA buscou seguir o modelo de prototipação evolucionária de software, na qual se começam a implementar os requisitos funcionais mais bem entendidos, em contraposição à prototipação rápida, ou descartável, na qual os requisitos menos entendidos são implementados primeiro (SHERRELL, 2013).

Levando-se em conta que o sistema é um sintetizador de canto, o que se tem é que os requisitos mais claros são aqueles que correspondem às propriedades objetivas da canção, constituída de letra e melodia, que representam, respectivamente, seus aspectos fonéticos e musicais.

Deste modo, espera-se que o sintetizador, do ponto de vista fonético, selecione corretamente as sílabas que compõem a letra da canção. Musicalmente, o que se deseja é que o sistema controle de modo adequado as qualidades do som definidas na Seção 2.2: altura, duração, intensidade e timbre. Tais qualidades podem ser mapeadas quantitativamente em termos de grandezas físicas, como frequência e amplitude.

O primeiro protótipo do sistema, descrito no presente trabalho, procurou servir de prova de conceito, ou seja, demonstrou ser factível a construção de um sintetizador de canto em tempo real para o português brasileiro, levando em consideração os requisitos funcionais que acabaram de ser descritos. Esta primeira implementação é apresentada em maiores detalhes nas próximas seções.

Porém, com as sucessivas refinações do sistema, pretende-se contemplar, em suas futuras versões, uma outra gama de requisitos funcionais. Tais requisitos guardam maior relação com as impressões subjetivas do ouvinte: inteligibilidade (BENOîT; GRICE; HAZAN, 1996), ou seja, não apenas a correta seleção das sílabas de um ponto de vista técnico, mas sua efetiva compreensão por parte de um sujeito avaliador; naturalidade (TERNSTRÖM, 2002), que seria a apreciação de quanto a voz sintetizada estaria próxima de uma voz humana natural; por fim, expressividade (UMBERT et al., 2015), relacionada às capacidades de se realizar uma performance artística única, em contraposição à reprodução meramente "mecânica"e sempre igual daquilo que está definido numa partitura musical.

4.1.2 Arquitetura do sistema

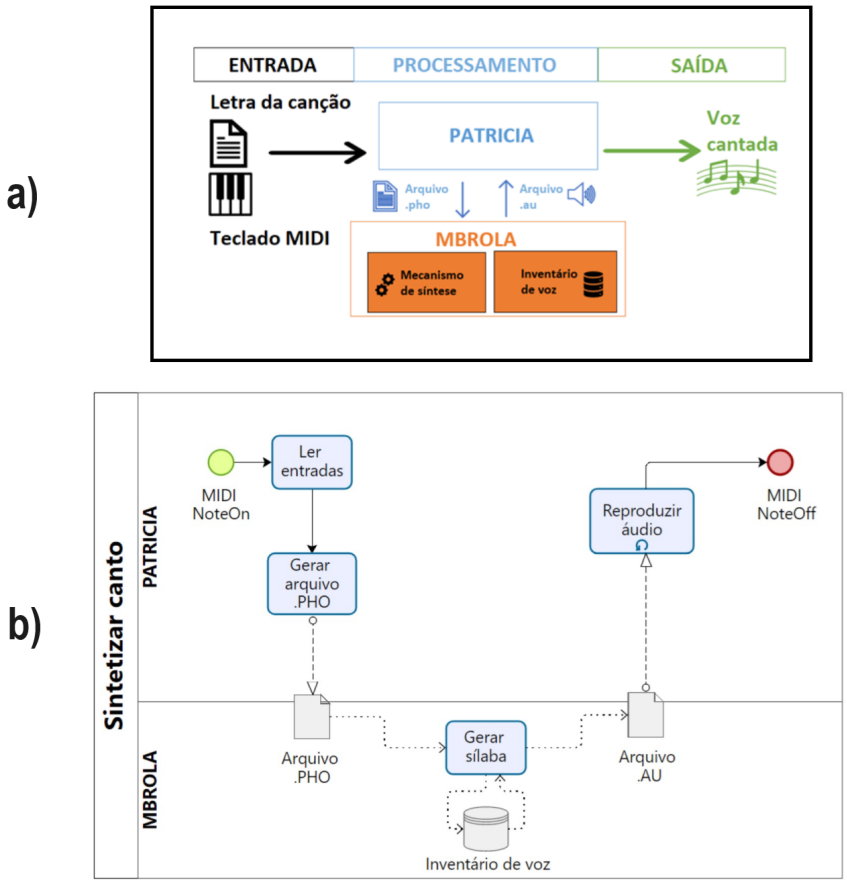
Inicialmente, o sintetizador PATRICIA foi concebido como um sistema embarcado que perfaz a síntese de canto em um teclado MIDI para fornecer uma performance em tempo real. No mapeamento sistemático, constatou-se que esta solução já estava implementada e patenteada, resultando no produto Vocaloid Keyboard (KAGAMI et al., 2012).

Entretanto, na atual versão do protótipo, PATRICIA ainda não funciona como um sintetizador embarcado, mas como um sistema executado em um computador hospedeiro para o qual o teclado musical é um dispositivo externo, servindo como periférico de entrada. Enquanto o teclado MIDI fornece a entrada musical para o sistema, um arquivo de texto que contém a letra da canção, presente na memória secundária do computador, provê a entrada fonética.

Além disso, a síntese sonora propriamente dita é feita por outro software, o sintetizador de fala MBROLA. PATRICIA processa as entradas musical e fonética, transformando-as num formato de texto *.pho*, que o MBROLA reconhece. O mecanismo de síntese do MBROLA, por sua vez, consulta seu inventário de voz de acordo com os dados do arquivo *.pho* e devolve um arquivo de áudio *.au* a PATRICIA, que o manipula para finalmente gerar, como saída, o canto sintetizado. A Figura 32 exibe um diagrama da arquitetura do sistema e o processo de síntese aqui descrito em notação BPMN. Os aspectos técnicos do processamento de dados são detalhados na

Seção 4.2.

Figura 32 – (a) Diagrama de arquitetura de PATRICIA. (b) Processo de síntese de canto entre PATRICIA e MBROLA.



Fonte: elaboração própria.

Uma arquitetura semelhante ao do sistema PATRICIA pode ser encontrada no sintetizador MaxMBROLA (D'ALESSANDRO et al., 2005), mencionado na Subseção 2.5.4, mas o artigo que o descreve foi excluído da revisão sistemática por estar fora do intervalo de tempo proposto de dez anos. Ademais, o MaxMBROLA foi projetado para ser executado em ambiente Max/MSP, que é um software proprietário, enquanto para a implementação de PATRICIA foi escolhido um software livre. Os detalhes de tal implementação são abordados na subseção a seguir.

4.1.3 Implementação

O sintetizador PATRICIA foi implementado em SuperCollider (MCCARTNEY, 1996), um ambiente de programação para síntese de áudio em tempo real cujo funcionamento foi descrito, em linhas gerais, na Subseção 2.5.4. GitHub foi usado para controle de versão.

Embora a linguagem SuperCollider seja orientada a objetos, pode-se dizer que a programação de PATRICIA foi fundamentalmente orientada a eventos. Tais eventos constituem-se basicamente do acionamento e liberação das teclas do instrumento MIDI que fornece a entrada musical do sistema. O Algoritmo 1 descreve, em linhas gerais, a implementação do sintetizador, apresentada em pseudocódigo. O código fonte SuperCollider de PATRICIA pode ser encontrado no Apêndice A.

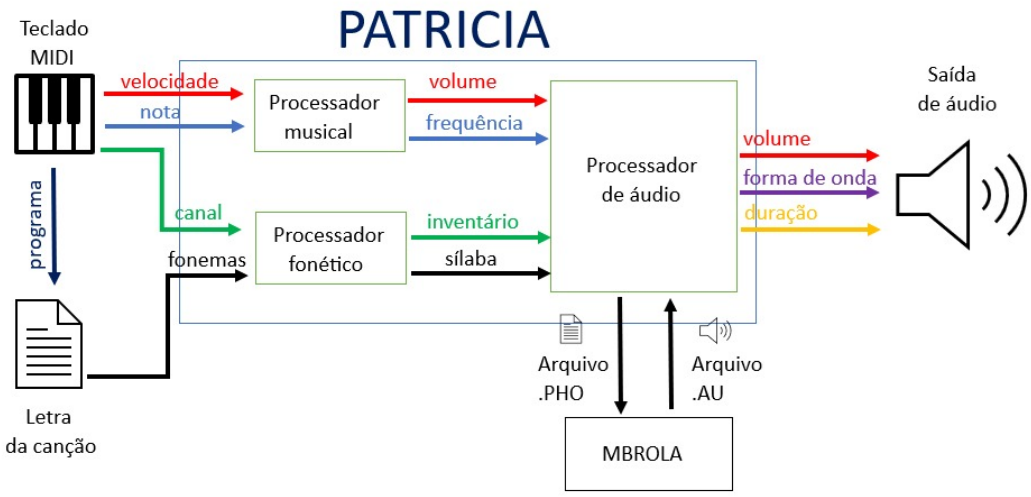
Algoritmo 1 – Algoritmo utilizado pelo sintetizador PATRICIA.

```
1  início
2      servidorAudio.inicializar();
3      # laço infinito de eventos:
4      repita
5          mensagem ← servidorAudio.escutarMensagensMIDI();
6          se mensagem = NoteOn então
7              # entrada fonética:
8              silaba ← obtenhaProximaLinha(arqTextoLetra);
9              # entrada musical:
10             nota ← mensagem.MidiNote;
11             volume ← mensagem.MidiVelocity;
12             # controla o timbre:
13             inventario ← mensagem.MidiChannel;
14             # controla a altura:
15             frequencia ← 440 · ( ∛2)^(nota−81);
16             # interface PATRICIA-MBROLA:
17             arquivoPHO ← gerarArquivoPHO(silaba, frequencia);
18             samples[nota] ← gerarAudioMBROLA(arquivoPHO, inventario);
19             # controla a intensidade:
20             servidorAudio.volumeSaida ← 40 · log(volume/127);
21             # controla a duração:
22             inicioLoop ← samples[nota].duracao / 2;
23             fimLoop ← inicioLoop + 1/frequencia;
24             servidorAudio.reproduzirLoop(samples[nota], inicioLoop, fimLoop);
25         se mensagem = NoteOff então
26             # sai do loop e reproduz o restante do audio:
27             servidorAudio.liberarLoop(samples[nota]);
28         se mensagem = ProgramChange então
29             # escolhe o arquivo fonético conforme o programa:
30             arqTextoLetra ← "lyrics"+ mensagem.MidiProgram ;
31     até falso;
32 fim
```

Dado o algoritmo acima, podem-se distinguir, na arquitetura interna de PATRICIA, um processador musical e um processador fonético, que lidam com ambos os tipos de entrada do sistema e cujas saídas alimentam o processador de áudio, que se comunica com o sintetizador MBROLA e manipula o canto gerado na saída do sistema. A Figura 33 apresenta a arquitetura

interna de PATRICIA por meio de um diagrama de blocos.

Figura 33 – Arquitetura interna de PATRICIA.



Fonte: elaboração própria.

No intuito de fornecer uma melhor compreensão acerca da implementação de PATRICIA, as Seções 4.2 e 4.3 apresentarão, respectivamente, os aspectos técnicos concernentes às entradas musical e fonética do sistema e o processamento do controle de áudio que gera, como saída, o canto sintetizado.

4.2 Aspectos técnicos

O mapeamento sistemático descrito no Capítulo 3 indagou acerca de dois aspectos técnicos empregados nos mais recentes desenvolvimentos relativos à síntese de voz cantada em tempo real: as abordagens de síntese e métodos de entrada fonética . Esta seção discute como ambos os aspectos, tais como implementados em PATRICIA, relacionam-se com os resultados do referido mapeamento.

4.2.1 Abordagem de síntese

A concepção inicial de PATRICIA, enquanto sistema embarcado que provê a síntese de canto para um teclado musical, encontrou maior correspondência no produto encontrado no mapeamento tecnológico apresentado na Seção 3.1, o Vocaloid Keyboard. Sendo assim, uma abordagem baseada em *samples*, usada tanto em teclados MIDI em geral quanto no sistema Vocaloid original, descrito na Subseção 2.5.2, foi escolhida para a implementação de PATRICIA. Além disso, a revisão sistemática, conduzida de acordo com a Seção 3.2 apontou para o fato de que tal abordagem técnica permanece relevante nos últimos desenvolvimentos de síntese de canto em tempo real, conforme os resultados do mapeamento mostrados no Quadro 3.

No entanto, na primeira implementação de PATRICIA, a complexidade da técnica de síntese concatenativa e a construção do banco de dados de voz foram transferidas para um sistema já existente, o sintetizador de fala MBROLA, que realiza uma síntese concatenativa por seleção de unidades. As unidades em questão são difones, ou seja, uma conjunção de dois fonemas. Um conjunto de difones de um determinado idioma é gravado e armazenado em um arquivo que serve como inventário de voz para o sistema. MBROLA foi escolhido, pois, além de ser um software livre com uma gama variada de inventários de voz à disposição, funciona por meio de linhas de comando, respondendo de modo satisfatório às chamadas em tempo real feitas pelo servidor de áudio do SuperCollider.

Incialmente, o inventário de voz MBROLA[1] *br4*, mencionado na Subseção 2.4.2, foi utilizado para dar a PATRICIA uma voz feminina. Entretanto, a fim de implementar o controle de timbre descrito na Subseção 4.3.4, substituiu-se *br4* por três inventários de voz masculina denominados *br1*, *br2* e *br3*, desenvolvidos por Denis R. Costa. Isso permitiu ao sistema o uso de três vozes diferentes, cujos inventários compartilham a mesma notação para a representação dos difones.

MBROLA toma como entrada uma lista de fonemas associados a parâmetros de altura e duração que, num sintetizador de fala, têm caráter antes prosódico do que musical. Tal entrada fornece as diretrizes para a concatenação dos difones, gerando um arquivo de áudio que possui a mesma taxa de amostragem dos *samples* armazenados no inventário. O modo como os fonemas são fornecidos ao MBROLA por meio do sintetizador PATRICIA é descrito a seguir.

4.2.2 Método de entrada fonética

Conforme mencionado na Introdução, um sintetizador de canto lida com dois tipos básicos de dados de entrada: o fonético, que vem da letra da música a ser sintetizada, e os parâmetros musicais, como altura e duração das notas. Cada sílaba da letra da canção corresponde a uma nota musical.

A técnica de síntese concatenativa de difones escolhida forneceu diretrizes para selecionar o método de entrada fonética. Como PATRICIA conta com inventários MBROLA de amostras de vozes pré-gravadas, não pretende funcionar como um vocoder, nem como um sistema de treinamento que transforma um sinal de acordo com as características de outra voz como em (DONG et al., 2014; MATSUBARA et al., 2021; TAE; KIM; LEE, 2021; VEAUX et al., 2013). Portanto, um sinal de áudio como entrada fonética, utilizado pela maioria das obras selecionadas na revisão sistemática, conforme o Quadro 4, acabou por ser descartado.

Por outro lado, um controle manual de difones usando controladores gestuais tem suas próprias limitações, conforme discutido no artigo que descreve o theremin aumentado T-Voks (XIAO et al., 2019). De acordo com este, "a síntese de canto concatenativa baseada em difones

[1] Os inventários de voz MBROLA aqui apresentados estão disponíveis em <https://github.com/numediart/MBROLA-voices>.

soa muito natural, mas dificilmente pode ser aplicada à síntese performativa em tempo real porque não parece possível selecionar qualquer sequência arbitrária de difones na hora". De fato, se essa tarefa se mostrou possível pelo protótipo do Teclado Vocaloid, que apresenta botões para introdução de letras em tempo real (KAGAMI et al., 2012), isso aconteceu por conta da estrutura mais simples das moras, ou unidades prosódicas, do idioma japonês. A maioria delas é composta por uma consoante e uma vogal, nesta ordem. Assim, quando o usuário pressiona, por exemplo, os botões "T" e "A" simultaneamente no painel Vocaloid Keyboard, o mecanismo de síntese interpreta esta mora necessariamente como sendo uma sílaba "TA" uma vez que uma eventual mora "AT" não existe na língua japonesa (LABRUNE, 2012). Em contraste, o português brasileiro possui sílabas como a última da palavra *"magistrais"*, cuja estrutura é consoante-consoante-vogal-semivogal-consoante, complexidade que aumenta os desafios para o uso de controladores manuais para mapear uma entrada fonética em tempo real.

Entretanto, de acordo com o que foi mencionado na Seção 3.1 até mesmo a versão comercial do Vocaloid Keyboard renuncia a este método de entrada fonética para fornecer a letra da música antecipadamente (KASHIWASE, 2017). Assim, prover uma entrada fonética previamente preparada em um arquivo de texto parece ser o método mais adequado para a implementação atual de PATRICIA, embora seja o menos utilizado pelos sintetizadores de canto em tempo real mais recentes de acordo com o resultado do mapeamento sistemático apresentado no Quadro 4. Uma vez determinado o método de entrada fonética, é necessário discutir o formato de tal entrada. As sílabas poderiam ser escritas de acordo com a ortografia oficial portuguesa, mas preferiu-se uma notação fonética para dar ao usuário dois tipos de controle sobre a síntese da voz:

1. A letra da música não é um texto em prosa, mas em versos, então certos fenômenos que ocorrem na versificação como a elisão, onde duas sílabas se fundem em uma, devem ser levados em consideração. As sílabas dos versos não coincidem necessariamente com as gramaticais (BANDEIRA, 1997) e são melhor controladas em nível fonético.

2. O português brasileiro possui diversas variantes regionais. Por exemplo, a primeira sílaba da palavra "tia" pode ser pronunciada como [ti] ou [tʃi][2], dependendo da região do falante brasileiro (SILVA; LEITE, 2015), de modo que uma sílaba ortográfica "ti" na entrada do sistema seria transformada pelo mecanismo de síntese em uma variante fonética específica escolhida arbitrariamente, excluindo outras possibilidades.

Deste modo, a solução é fornecer a entrada fonética diretamente em uma notação fonética. A notação SAMPA (WELLS et al., 1997) mostrou-se adequada para a tarefa, por ser baseada em caracteres ASCII, presentes em qualquer teclado padrão de computador, enquanto o Alfabeto Fonético Internacional (IPA) possui muitos caracteres especiais cuja entrada é mais difícil de ser feita. Além disso, a notação SAMPA foi empregada nos primeiros inventários MBROLA, como

[2] A notação do Alfabeto Fonético Internacional (IPA) foi utilizada nos exemplos dados.

o *fr1*, criado para o idioma francês (DUTOIT et al., 1996). Contudo, a representação fonética empregada pelos inventários MBROLA para o português brasileiro é apenas uma adaptação da notação SAMPA, divergindo desta em alguns pontos. O fonema cuja notação IPA é [ʃ], por exemplo, em SAMPA é representado por [S] e, nos inventários brasileiros, por [x]. Deste modo, os inventários MBROLA aproximam-se da ortografia do idioma português, já que o fonema em questão é o primeiro em palavras como "xícara".

PATRICIA, portanto, recupera dados fonéticos de um arquivo de texto que deve ser preparado antes do início da apresentação musical. A primeira linha do arquivo é reservada ao título da canção. Quanto às demais linhas, cada uma delas deve conter uma sílaba escrita em notação fonética, tendo seus fonemas separados por hífens. PATRICIA utiliza a adaptação de caracteres SAMPA para o português do Brasil que foi definida igualmente para os inventários MBROLA *br1*, *br2* e *br3*. A relação dos fonemas disponíveis e suas respectivas representações são apresentadas no Apêndice B.

É possível alternar entre diferentes arquivos a fim de que o sintetizador execute canções diversas. A seleção dos arquivos fonéticos se dá por meio da mensagem MIDI *Progam Change*, descrita na Subseção 2.4.3, o que significa que é possível disponibilizar até 128 letras de canção diferentes para o sistema.

O canto em outros idiomas por meio de PATRICIA seria possível por meio do uso de outros inventários de voz MBROLA disponíveis, mas o sistema teria de ser adaptado para tratar dos símbolos fonéticos utilizados por cada um. Como a intenção era projetar um sintetizador de canto especificamente para o português brasileiro, tal adaptação não foi implementada.

4.3 Controle de áudio

De acordo com o que foi estabelecido nas Subseções 4.1.2 e 4.2.1, no sistema MBROLA, a fala a ser sintetizada pode ser armazenada em um arquivo de texto com a extensão *.pho*. Cada linha deste arquivo contém um fonema, dado conforme a notação estebelecida pelo inventário de voz, uma duração em milissegundos e uma série de marcos de altura compostos por dois números: a posição do marco dentro do fonema, que é uma porcentagem de sua duração total e o valor da frequência em Hertz em tal posição. Uma instrução de linha de comando chama o MBROLA, que realiza a seleção da unidade no inventário de voz indicado, concatenando os difones de acordo com a descrição dentro do arquivo *.pho* e finalmente gerando um arquivo de áudio.

Para cada mensagem MIDI *Note on* e sua sílaba correspondente recuperada pela PATRICIA, é gerado um arquivo *.pho*. Assim, cada linha deste arquivo contém estes quatro parâmetros:

1. Um fonema correspondente à sílaba recuperada (o arquivo *.pho* terá tantas linhas quanto

houver fonemas nesta sílaba);

2. Duração de um fonema, definida como 25 ms para semivogais e consoantes e 200 ms para vogais;

3. Um marco de altura que é sempre de 100%, o que significa que a altura dos fonemas é a mesma durante toda a sua duração.

4. Frequência fundamental para fornecer a altura, calculada de acordo com o código numérico MIDI da nota musical.

No arquivo.*pho*, uma duração maior foi dada às vogais por corresponderem acusticamente à fase de sustentação da sílaba e, musicalmente, à nota cantada, conforme a Subseção 2.3.2. As consoantes corresponderiam basicamente a ruídos, cujo prolongamento seria indesejado. Os valores atribuídos em milissegundos a cada tipo de fonema possibilitaram a aplicação da técnica de *loop*, empregada para controlar o áudio de saída do sistema, de acordo com a explicação a ser dada na Subseção 4.3.2. Quanto ao cálculo da frequência fundamental em função do código de nota MIDI, este é descrito em maiores detalhes na Subseção 4.3.1.

Depois que o arquivo.*pho* é criado, PATRICIA usa o mecanismo do SuperCollider para enviar ao sistema operacional[3] do computador uma linha de comando MBROLA. Esta instrução faz referência ao arquivo *.pho* e a um dos três inventários de voz do MBROLA em português do Brasil para criar o arquivo de áudio correspondente no formato *Au*.

As próximas subseções descrevem como PATRICIA faz o controle das qualidades de som dos arquivos *Au* mencionados para gerar a voz cantada sintetizada.

4.3.1 Controle da altura

Esta subseção explica como se dá o cálculo da freqência fundamental de cada nota musical indicada pelo protocolo MIDI, registrada no arquivo *.pho* a ser processado pelo MBROLA.

Já foi estabelecido, na Subseção 2.2.1 que o intervalo musical de uma oitava separa uma nota daquela que possui o dobro de sua freqüência fundamental. Assim, obtém-se a Equação 4.1:

$$\frac{f(x_{n+1})}{f(x_n)} = 2 \tag{4.1}$$

Onde x é uma nota musical qualquer, n indica a oitava à qual x pertence e f é uma função que relaciona x_n à sua frequência fundamental. Viu-se também que, no sistema de afinação temperado, o intervalo acústico entre duas notas consecutivas quaisquer da escala cromática, separadas pelo intervalo musical de um semitom, é uma constante. Como o intervalo de oitava

[3] PATRICIA funciona nos sistemas operacionais Windows e Linux, desde que o MBROLA e o SuperCollider estejam instalados corretamente.

corresponde a doze semitons, conclui-se que elevar uma nota uma oitava acima é o mesmo que multiplicar sua frequência doze vezes pela constante mencionada, obtida pela razão entre as freqüências de duas notas distantes um semitom uma da outra. Assim, é possível escrever a Equação 4.2 da seguinte forma:

$$f(x_n) \cdot \left(\frac{f(x\sharp_n)}{f(x_n)} \right)^{12} = f(x_{n+1}) \tag{4.2}$$

Onde o sinal "♯"indica a alteração sustenido, ou seja,$x\sharp_n$ é a nota que se encontra um semitom acima de x_n. Das Equações 2.1 e 2.2 tem-se que:

$$\left(\frac{f(x\sharp_n)}{f(x_n)} \right)^{12} = \left(\frac{f(x_{n+1})}{f(x_n)} \right) = 2$$

Logo,

$$\frac{f(x\sharp_n)}{f(x_n)} = \sqrt[12]{2} \tag{4.3}$$

Portanto, ao se multiplicar a freqüência de uma determinada nota musical por $\sqrt[12]{2}$, obtém-se a freqüência da nota que se encontra um semitom acima dela na escala cromática (CAMILO; YABU-UTI; YANO, 1986). Daí o valor $\sqrt[12]{2}$ ser chamado **constante do semitom cromático**.

Weber (2003), por sua vez, apresenta a fórmula para se calcular a frequência fundamental f de qualquer nota musical em Hertz, segundo a Equação 4.4:

$$f = 440 \times 2^{oitava-4+(nota-10)/12} \tag{4.4}$$

Onde a *oitava* é um número inteiro que varia de 0 a 8 e a *nota* é associada a um número inteiro entre 1 e 12, inclusive, de acordo com a ordem de aparição na escala cromática: dó = 1, dó♯ = 2, ré = 3 e assim por diante. A Equação 4.5 exibe um outro arranjo matemático para a mesma fórmula.

$$f = 440 \cdot (\sqrt[12]{2})^{12 \cdot (oitava-4)+(nota-10)} \tag{4.5}$$

Parar reescrever a Equação 4.5, adequando-a ao padrão das equações anteriores, pode-se definir uma função $f(x_n)$, que nos dá a frequência fundamental da nota x na oitava n e uma outra função $g(x)$, que indica a posição da nota x na escala cromática, variando de 1 a 12. Assim, o termo *oitava* da Equação 4.5 é substituído pela variável n e o termo *nota*, por $g(x)$, resultando na Equação 4.6:

$$f(x_n) = 440 \cdot (\sqrt[12]{2})^{12\cdot(n-4)+(g(x)-10)} \tag{4.6}$$

Na fórmula apresentada pela Equação 4.6, é possível distinguir três elementos: o primeiro é a frequência fundamental da nota lá$_4$, no valor de 440 Hz, que serve referência no sistema temperado, conforme a Subseção 2.2.1; o segundo é a constante do semitom cromático, cujo valor é $\sqrt[12]{2}$ e serve de base para o terceiro elemento, um expoente que indica a distância, em semitons, entre a nota variável x_n e a nota de referência lá$_4$.

No cálculo da frequência da própria nota lá$_4$, por exemplo, a distância em semitons seria nula, pois n seria igual a 4 e g(lá) teria valor 10. Assim, o cálculo do expoente ficaria $12 \cdot (4 - 4) + (10 - 10) = 0$ e o resultado da exponenciação seria 1, elemento neutro da multiplicação, restando o próprio valor de 440 Hz. Para notas musicais anteriores a lá$_4$, o expoente negativo resultará em sucessivas divisões da frequência de referência pela constante do semitom cromático, em direção a frequências — e alturas musicais — mais baixas. O contrário se dá em notas posteriores a lá$_4$: expoente positivo, multiplicações sucessivas, frequências e alturas mais elevadas.

O protocolo MIDI, porém, não representa os sons musicais no padrão $nota_{oitava}$ e sim, como visto na Subseção 2.4.3, por um número inteiro que varia entre 0 e 127, percorrendo, semitom por semitom, sucessivas escalas cromáticas entre diversas oitavas. Deste modo, cada som musical nas mensagens MIDI é representado por um código numérico único, independente da oitava em que se encontre. Para o caso específico da nota lá$_4$, o código numérico MIDI correspondente é 69. Assim, pode-se reescrever a Equação 4.6 da seguinte maneira:

$$f = 440 \cdot (\sqrt[12]{2})^{m-69} \tag{4.7}$$

Onde f é a frequência fundamental da nota musical cujo código numérico MIDI é m.

Entretanto, os experimentos realizados nesta dissertação mostraram que, ao implementar a fórmula da Equação 4.7 no sintetizador PATRICIA, a voz sintetizada soava uma oitava abaixo do esperado por alguma contingência do MBROLA ou do inventário de voz. Para compensar tal deficiência, a fórmula efetivamente implementada no sistema indica uma frequência elevada uma oitava acima. Para tanto, preferiu-se alterar o código MIDI no expoente para 81, corresponde a nota lá$_5$, em lugar de dobrar a frequência de referência para 880 Hz. Acredita-se que a manutenção do valor de referência 440 na fórmula serviu para que a alteração necessária não comprometesse sua inteligibilidade. Deste modo, o controle de altura em PATRICIA se dá pelo registro, em um arquivo *.pho* da frequência fundamental f de cada nota musical de código MIDI m tocada pelo usuário de acordo com o cálculo [4] apresentado pela Equação 4.8.

[4] A implementação desta fórmula matemática pode ser vista na linha 32 do código fonte de PATRICIA, disponível no Apêndice B desta dissertação.

$$f = 440 \cdot (\sqrt[12]{2})^{m-81} \tag{4.8}$$

4.3.2 Controle da duração

Como cada arquivo de áudio gerado pelo MBROLA tem uma duração constante, PATRICIA precisa estendê-la em conformidade com a intenção musical do usuário. Tal intenção, entretanto, não é expressa em tempos e compassos musicias (vide Subseção 2.2.2) a serem registrados *a priori* numa interface como a dos sintetizadores de canto convencionais apresentados nas Subseções 2.5.1 e 2.5.2. À medida que PATRICIA é um sintetizador em tempo real, a duração de cada nota precisa coincidir simplesmente com o tempo em que a respectiva tecla estiver pressionada no instrumento MIDI.

Para alcançar este resultado, para cada mensagem MIDI *Note on*, o arquivo *.au* correspondente é reproduzido em um *loop*, de acordo com a técnica descrita na Subseção 2.5.2. A posição inicial do *loop* está no centro do arquivo no domínio do tempo, onde se encontra a vogal, que é a parte periódica da forma de onda da sílaba, conforme estabelecido a Subseção 2.3.2. A posição final é calculada adicionando-se à posição inicial o período em segundos, ou seja, o inverso da frequência fundamental, cujo cálculo é feito de acordo com a Equação 4.8. O intervalo entre as posições inicial e final corresponde a um ciclo ou período de oscilação da forma de onda. Este ciclo é repetido enquanto a nota MIDI correspondente for mantida ativa, estendendo indefinidamente a duração da sílaba. Quando uma mensagem MIDI *Note off* é recebida, o *loop* que corresponde a esta nota é liberado, e o restante do arquivo *.au* é reproduzido.

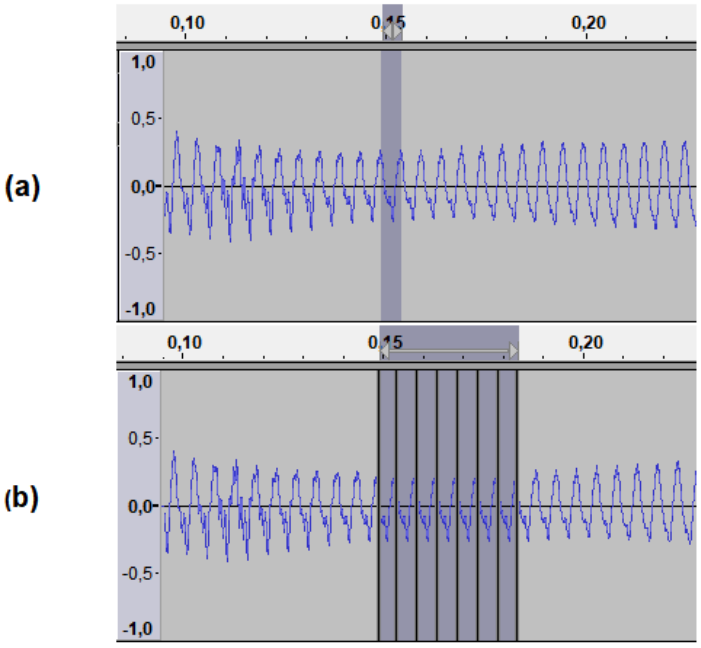
A figura 34 mostra uma visão parcial de uma forma de onda gerada pelo MBROLA para a primeira sílaba da palavra *"quando"* cantando uma nota sol_4. A região selecionada em (a) corresponde a um período da forma de onda calculado por PATRICIA, partindo-se do centro da amostra. A repetição de tal segmento, que aparece sete vezes em (b), prolonga a duração do áudio.

As diferenças entre as taxas de amostragem dos inventários MBROLA e do SuperCollider, 16000 Hz e 44100 Hz, respectivamente, tiveram que ser levadas em consideração para que o áudio fosse reproduzido corretamente. Se o áudio do MBROLA, cuja taxa de amostragem é mais baixa, fosse reproduzido na taxa padrão do SuperCollider, o resultado seria uma reprodução acelerada e um aumento da frequência fundamental do próprio sinal acústico, que soaria mais agudo. Para que a proporção entre ambas as taxas de amostragem fosse mantida, foi necessário definir uma taxa de reprodução correspondente à divisão entre elas [5], ou seja 16000/44100 Hz.

A repetição sucessiva do processo descrito acima para cada nota tocada no instrumento MIDI resulta na geração de uma voz cantada sintetizada na saída de áudio do computador.

[5] Esta divisão é passada como parâmetro ao método que implementa o *loop*, como pode ser visto na linha 94 do código fonte do sistema, presente no Apêndice A desta dissertação.

Figura 34 – (a) Segmento de *loop* calculado por PATRICIA. (b) Repetição do segmento, estendendo a duração da amostra.



Fonte: elaboração própria.

4.3.3 Controle da intensidade

Na Subseção 2.4.3, viu-se que as mensagens MIDI *Note on* e *Note off* possuem dois bytes de dados: um que indica o código numérico da nota musical e outro que fornece a velocidade com que a tecla foi abaixada no instrumento. Em mensagens *Note on*, a velocidade varia entre 1 e 127. Já nas mensagens *Note off* a velocidade é sempre igual a zero, indicando a soltura da tecla e, consequentemente, a desativação da nota.

Como a intensidade do som depende da força empregada ao tocar o instrumento, conforme exposto na Subseção 2.2.3, o parâmetro de velocidade pode ser mapeado para um volume ou intensidade de som medido em decibéis(dB), que é uma unidade de medida que emprega uma escala logarítmica. As especificações do protocolo MIDI estabelecem a seguinte fórmula para mapear um volume V, que varia de 1 a 127 em uma intensidade L, medida em decibéis:

$$L = 40 \cdot log(V/127) \tag{4.9}$$

A fórmula apresentada pela Equação 4.9 foi implementada no sistema[6], sendo seu resultado atribuído ao controle de volume de saída de áudio do SuperCollider. Assim, para cada nota tocada pelo usuário, a intensidade do som resultante variará de acordo com a velocidade de acionamento das teclas.

[6] Tal implementação pode ser verificada na linha 96 do código fonte do sistema

4.3.4 Controle do timbre

A qualidade do timbre, abordada na Subseção 2.2.4, permite a identificação da fonte sonora, ainda que fontes distintas produzam sons musicais de mesma altura, duração e intensidade. Da mesma forma que cada instrumento musical tem um timbre que lhe é próprio, também em cada voz humana é perceptível tal característica, possibilitando identificar quem está falando ou cantando.

Como cada um dos três inventários MBROLA utilizados por PATRICIA é composto por *samples* oriundos de gravações de uma voz específica, isso significa que alterar o inventário utilizado modificará o timbre do canto resultante.

A seleção dos inventários de voz se dá pelo número do canal MIDI, presente nos quatro bits menos significativos do byte de estado das mensagens MIDI *Note on* e *Note off*, conforme descrito na Subseção 2.4.3. Como este número pode assumir valores entre 1 e 16, isso significa que PATRICIA pode suportar até dezesseis inventários de voz diferentes.

A mudança de canal pode ser feita em tempo real, possibilitando a utilização de até três timbres de voz diferentes ao longo da mesma canção. A viabilidade da mudança de canal durante a performance musical dependerá de como os controles MIDI estão dispostos em cada dispositivo ou instrumento musical utilizado.

4.4 Experimentos e avaliação

Na qualidade de prova de conceito, o sintetizador de canto descrito neste trabalho possui apenas um conjunto mínimo de requisitos funcionais implementados. Tais requisitos são validados pelos experimentos descritos a seguir, que buscaram avaliar as capacidades musicais do sistema e fazer uma análise comparativa entre seu desempenho em um computador pessoal e num computador de placa única.

4.4.1 Demonstração musical

Uma breve demonstração das capacidades musicais de PATRICIA foi gravada e se encontra disponível no seguinte endereço:

<https://youtube.com/playlist?list=PLIJKwH6F5ow-DV9IvaN8xoa0c4CcyEFUl>.

Trata-se de uma playlist contendo três vídeos. O experimento foi realizado em um computador pessoal com processador Intel(R) Core(TM) i5-7300U, 2.60GHz, 8 GB de memória RAM e sistema operacional Windows 11 Pro de 64 bits, ao qual estava conectado um controlador MIDI M-AUDIO Keystation Mini32 MK3.

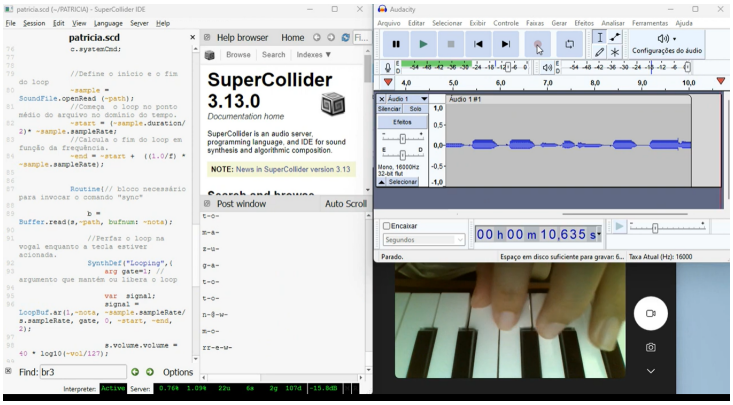
Os vídeos apresentam a execução de PATRICIA no IDE do SuperCollider. Ao mesmo tempo, uma imagem *picture-in-picture* mostra o teclado MIDI sendo tocado, em ângulo espelhado,

enquanto o editor de áudio Audacity grava o sinal de saída. Em cada um dos vídeos um inventário de voz diferente foi utilizado. Um quadro do vídeo gravado é exibido na Figura 35.

As canções escolhidas para o experimento são de domínio público e foram extraídas do cancioneiro popular brasileiro infantil. São elas:

- "O cravo brigou com a rosa", na qual se utilizou o inventário *br1*;
- "Atirei o pau no gato", valendo-se do inventário *br2*;
- "Terezinha de Jesus", para a qual se empregou o inventário *br3*.

Figura 35 – Quadro de um dos vídeos de demonstração musical de PATRICIA.

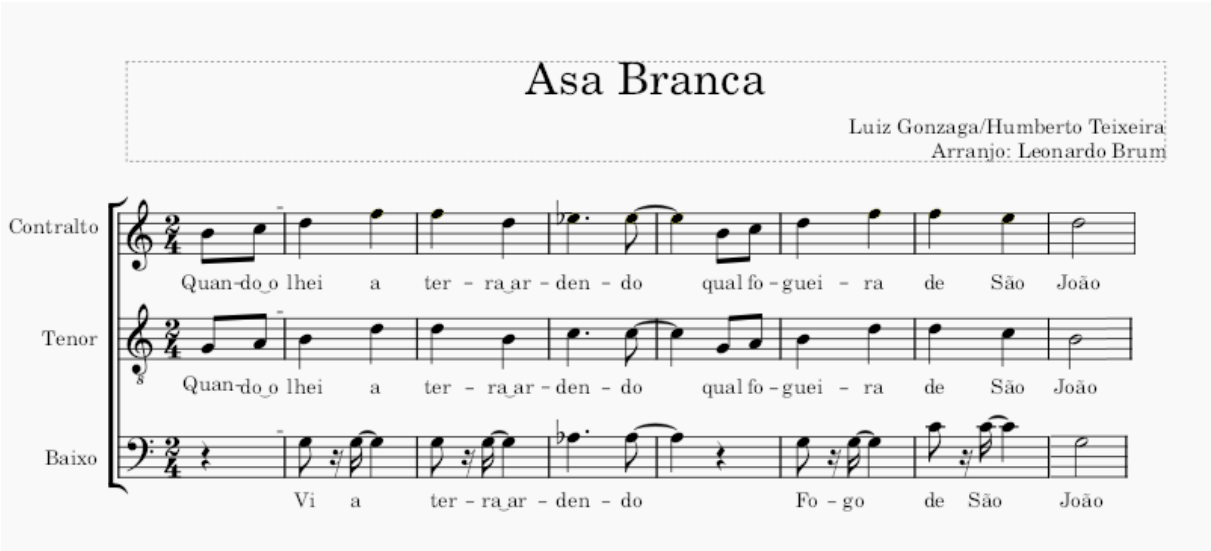


Fonte:elaboração própria.

Em outra gravação foi possível demonstrar um potencial uso pedagógico do sistema PATRICIA. Um arranjo para coro a três vozes dos dois primeiros versos da canção "Asa Branca", de Luiz Gonzaga e Humberto Teixeira, foi feito pelo autor da dissertação. A partitura do arranjo pode ser vista na Figura 36. Utilizou-se PATRICIA para gerar o áudio de cada uma das vozes, empregando-se o inventário *br1* para o tenor, que faz a melodia principal, *br2* para a voz contralto e *br3* para o baixo. Os áudios, no editor Audacity, são executados em separado e, em seguida, conjuntamente no vídeo disponível em <https://youtu.be/HU42aDK8ico>. No ensino do canto coral, é comum que o professor ou regente apresente aos coralistas as linhas melódicas de cada voz valendo-se de algum instrumento musical, como piano ou teclado. Ao se empregar o sintetizador PATRICIA, será possível utilizar um instrumento musical para gerar não apenas a linha melódica, mas o áudio da própria voz a ser cantada.

Essas demonstrações foram objeto de uma avaliação por parte de educadores musicais, o que será descrito em maiores detalhes na próxima subseção.

Figura 36 – Arranjo a três vozes dos dois primeiros versos da canção "Asa Branca".



Fonte: elaboração própria.

4.4.2 Avaliação por educadores musicais

As quatro demonstrações musicais descritas na subseção anterior, em conjunto com uma demonstração adicional, que havia sido preparada quando o sintetizador ainda contava com o inventário *br4*[7], foram avaliadas por cinco educadores musicais que atuam no Estado de Sergipe.

A avaliação foi feita por meio de um formulário contendo vinte questões, sendo quatro para cada demonstração, que o educador deveria assistir antes de responder. Para as demonstrações correspondentes aos inventários *br4*, *br1*, *br2* e *br3*, as quatro questões foram as seguintes:

- Você considera que as sílabas cantadas são compreensíveis?
- Você considera que o canto gerado tem naturalidade?
- Você considera que as alturas musicais geradas correspondem à execução do músico?
- Você considera que a duração das notas geradas correspondem à execução do músico?

Essas quatro questões têm por objetivo avaliar, respectivamente, a inteligibilidade, a naturalidade, o controle de altura e o controle de duração do canto gerado. Os dois primeiros requisitos foram apresentados na Subseção 4.1.1. Já os dois últimos foram objetos das Subseções 4.3.1 e 4.3.2. As respostas possíveis às questões, correspondentes a uma escala de Likert com quatro opções, abrindo-se mão de uma posição neutra, seguem abaixo:

- Concordo totalmente;

[7] A demonstração com o inventário *br4* foi feita valendo-se também da canção "Asa Branca"e encontra-se disponível no seguinte endereço: <https://youtu.be/XK7y-mCw7KI>.

- Concordo parcialmente;
- Discordo parcialmente;
- Discordo totalmente.

Já em relação à demonstração do canto a três vozes, as duas últimas questões foram modificadas. A penúltima indagava "Você considera que o canto gerado corresponde ao arranjo exibido pela partitura acima?", considerando-se a partitura exibida pela Figura 36. Por fim, a última pergunta visou avaliar a utilidade pedagógica do sintetizador, ao propor a seguinte questão aos educadores: "Você considera que o sintetizador PATRICIA seria uma ferramenta útil para o ensino de canto coral?".

As questões foram pontuadas de 1 a 4, da posição mais desfavorável à mais favorável na escala de Likert. Somando-se as pontuações dadas por cada um dos cinco usuários em cada questão e multiplicando-se o resultado por cinco, obtém-se um percentual correspondente à avaliação de cada requisito. As Tabelas de 1 a 5, a seguir, apresentam os resultados obtidos por cada uma das demonstrações. Nelas, os educadores são identificados como E1, E2 ... E5 e as questões como Q1, Q2...Q20.

Tabela 1 – Avaliação da demonstração "Asa Branca", com inventário *br4*.

	E1	E2	E3	E4	E5	TOTAL
Q1(inteligibilidade)	3	4	3	3	3	80%
Q2(naturalidade)	3	3	2	1	1	50%
Q3(controle de altura)	3	4	4	4	4	95%
Q4(controle de duração)	2	4	4	3	3	80%

Tabela 2 – Avaliação da demonstração "O Cravo Brigou com a Rosa", com inventário *br1*.

	E1	E2	E3	E4	E5	TOTAL
Q5(inteligibilidade)	2	4	3	3	3	75%
Q6(naturalidade)	2	3	2	1	1	45%
Q7(controle de altura)	3	4	4	4	3	90%
Q8(controle de duração)	3	4	4	3	3	85%

Tabela 3 – Avaliação da demonstração "Atirei o Pau no Gato", com inventário *br2*.

	E1	E2	E3	E4	E5	TOTAL
Q9(inteligibilidade)	3	4	3	3	3	80%
Q10(naturalidade)	2	3	3	2	1	55%
Q11(controle de altura)	3	4	4	4	4	95%
Q12(controle de duração)	3	4	4	4	4	95%

Tabela 4 – Avaliação da demonstração "Terezinha de Jesus", com inventário *br3*.

	E1	E2	E3	E4	E5	TOTAL
Q13(inteligibilidade)	2	3	2	3	1	55%
Q14(naturalidade)	2	3	2	1	1	45%
Q15(controle de altura)	2	4	4	4	4	90%
Q16(controle de duração)	3	4	4	4	4	95%

Tabela 5 – Avaliação da demonstração "Asa Branca a três vozes", com inventários *br1*, *br2* e *br3*.

	E1	E2	E3	E4	E5	TOTAL
Q17(inteligibilidade)	3	3	3	3	3	75%
Q18(naturalidade)	3	3	3	2	1	60%
Q19(conformidade ao arranjo)	3	4	4	4	4	95%
Q20(utilidade pedagógica)	4	4	4	3	2	75%

O formulário de avaliação aplicado encontra-se, na íntegra, no Apêndice C, enquanto os resultados obtidos, sob forma de gráficos, podem ser vistos no Apêndice D. Espera-se que as limitações aqui apontadas possam ser superadas em versões posteriores do sistema. Eventuais soluções são abordadas no Capítulo 5, no qual se discutem os trabalhos futuros.

4.4.3 Análise de desempenho

Os experimentos que visaram analisar o desempenho do sistema foram conduzidos em dois dispositivos: o mesmo computador pessoal utilizado para fazer a demonstração musical e um computador de placa única Raspberry Pi. O Quadro 5 apresenta, em maiores detalhes, as configurações dos referidos computadores. O teclado musical utilizado nesses experientos também foi o controlador MIDI M-AUDIO Keystation Mini32 MK3.

Quadro 5 – Configurações dos computadores utilizados nos experimentos.

Computador	Notebook Positivo Master N8140 Blackstone	Raspberry Pi 4 Model B Rev 1.5
Sistema Operacional	Windows 11 Pro v.22H2 64 bits	Raspberry Pi OS Debian 11 (bullseye) 64 bits
Processador	Intel(R) Core(TM) i5-7300U CPU @ 2.60GHz	Broadcom BCM2711 SoC @ 1.8GHz
Memória RAM instalada	8 GB	8 GB

A análise comparativa proposta baseia-se na medida, em cada dispositivo, dos seguintes parâmetros relativos ao servidor de áudio do SuperCollider, o processo *scsynth*:

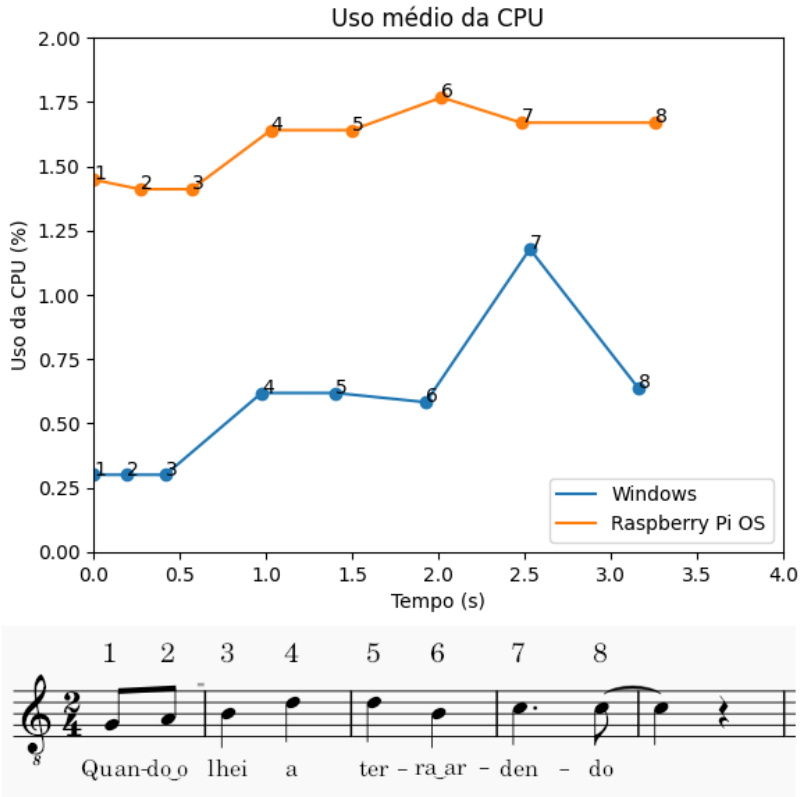
- Uso médio da CPU: representa a carga média da CPU ao longo do tempo durante a execução do servidor de áudio. Calcula-se a média de uso da CPU pelo servidor de áudio em um

determinado intervalo de tempo. Trata-se de uma métrica útil para avaliar o desempenho geral do servidor.

- Picos de uso da CPU: indica o pico mais alto de uso da CPU durante a execução do servidor de áudio, registrando o valor máximo alcançado pela carga da CPU durante a execução de processamentos de áudio. Tal medida pode ser utilizada para identificar momentos em que a carga da CPU atinge níveis críticos ou quando ocorrem picos de demanda de processamento.

A Figura 37 exibe as medidas de uso médio da CPU pelo servidor de áudio em cada um dos dispositivos, indicados no gráfico por seus respectivos sistemas operacionais. A execução do primeiro verso da canção "Asa Branca" foi escolhida para perfazer o experimento. Os diversos pontos representam os instantes em que cada nota foi tocada, sendo numerados de 1 a 8 para que se faça a correspondência com a partitura colocada na parte inferior do gráfico. Um número maior de pontos poderia ser gerado para abranger uma maior duração da execução musical, mas a correspondência nota a nota entre as duas performances ficaria menos clara e a partitura tenderia ou a se tornar ilegível, ou a ocupar demasiado espaço.

Figura 37 – Uso médio de CPU em cada dispositivo



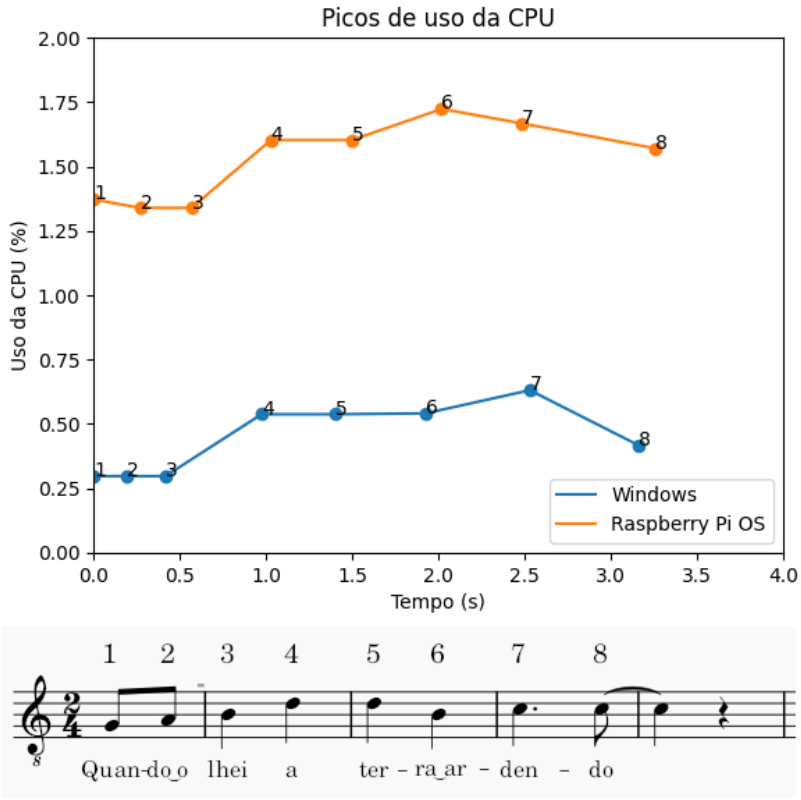
Fonte: elaboração própria.

Observou-se que, embora o uso médio da CPU tenha sido mais intenso no Raspberry Pi do que no computador pessoal, em ambos os dispositivos a variação da medida ao longo do

tempo se comportou de maneira similar, com a exceção de um valor mais elevado de uso da CPU observado quando da ativação da sétima nota no computador pessoal. No Raspberry Pi, o uso médio da CPU pelo servidor de áudio do SuperCollider variou dentro do intervalo de 1,25 até 1,75%, enquanto a variação no computador pessoal se deu dentro do intervalo entre 0,25 e 1,25%.

Por sua vez, a Figura 38 apresenta as medições dos picos de uso da CPU pelo processo *scsynth* em ambos os computadores, relacionando-os às notas musicais presentes na partitura exibida na parte inferior, de modo similar ao padrão adotado na Figura 37.

Figura 38 – Picos de uso da CPU em cada dispositivo



Fonte: elaboração própria.

Aqui o comportamento do gráfico foi similar ao da medida anterior. Os picos de uso da CPU foram mais intensos no Raspberry Pi, mas a variação da medida dentro do patamar apresentado por cada dispositivo foi semelhante. Os picos de uso da CPU variaram entre 1,25 e 1,75% no Raspberry Pi e entre 0,25 e 0,75% no computador pessoal. Acredita-se que exploração dos recursos de parelelismo do SuperCollider possa melhorar o desempenho do sintetizador no Raspberry Pi.

4.5 Resultados e discussão

A demonstração musical descrita na Subseção 4.4.1 pôde validar a versão atual do sintetizador enquanto prova de conceito. O controle em tempo real das qualidades básicas do som — altura, intensidade, duração e timbre — além da correta seleção dos difones, mostraram-se viáveis, pois puderam ser verificados tanto do ponto de vista auditivo quanto visualmente, ao se analisar o comportamento do sinal de áudio gerado no editor Audacity à medida que as canções eram executadas. É possível verificar que até as imprecisões de execução se refletiram na saída de áudio, o que é uma característica desejada em se tratando de um sintetizador em tempo real.

Em relação à avaliação por educadores musicais, apresentada na Subseção 4.4.2, o requisito que teve avaliação menos favorável foi a naturalidade. A principal causa disso é provavelmente a dependência de PATRICIA em relação ao MBROLA, que não foi projetado para sintetizar a voz cantada. Outras causas são a ausência de um controle de expressão da voz e um algoritmo de extensão de duração que gera um segmento de *loop* muito curto (ROADS, 1996, p. 122), abrangendo apenas um período da forma de onda da voz sintetizada. Além disso, ruídos do tipo "clique"foram perceptíveis em algumas transições silábicas.

Quanto à inteligibilidade das sílabas geradas, a maior parte das avaliações foi positiva, mas isso pode se dever ao fato de que, muito provavelmente, os educadores musicais já conheciam de antemão a letra das canções avaliadas. O inventário de avaliação mais desfavorável nesse quesito foi o *br3*, como se pode ver na Tabela 4. Ademais, o controle esperado em nível fonético do canto sintetizado não foi de todo possível, dadas as configurações dos inventários de voz utilizados, uma vez que, ao menos nos casos de *br1* e *br3*, alguns difones, como [ti] e [di], já tinham sido programados para corresponder a uma variante específica da língua portuguesa brasileira, resultando nas sílabas palatalizadas [tʃi] e [dʒi], respectivamente.

O controle de altura e duração também foram bem avaliados pelos educadores, inclusive em relação à demonstração do canto a três vozes, em que esses parâmetros foram avaliados em conjunto sob o nome de "conformidade ao arranjo". A utilidade pedagógica, por sua vez, dividiu opiniões, embora três dos cinco educadores tenham declarado concordar totalmente que o sintetizador PATRICIA seria uma ferramenta útil para o ensino de canto coral.

Por fim, ambas as medições do desempenho do sistema descritas na Subseção 4.4.3 resultaram em valores muito baixos de uso da CPU e, consequentemente, um desempenho bastante satisfatório. Tal resultado indica que as escolhas, em termos de software, foram acertadas: SuperCollider mostrou-se bastante eficiente no processamento de áudio em tempo real e MBROLA respondeu bem às chamadas feitas pelo servidor. Quanto ao hardware, contudo, os experimentos levam a supor que o sintetizador poderia ser implantado num dispositivo de potência computacional muito menor, tornando a solução mais barata, sobretudo quando do desenvolvimento de sua versão final, como sistema embarcado.

5

Conclusão e trabalhos futuros

O sistema aqui proposto, PATRICIA, é um sintetizador de canto em tempo real, sendo o primeiro a ser especificamente projetado para gerar canto no idioma português brasileiro. Trata-se de um trabalho ainda em andamento, e o sintetizador descrito é uma prova de conceito com um conjunto limitado de requisitos funcionais implementados na linguagem e ambiente SuperCollider. A versão atual do sistema será apresentada na *International Computer Music Conference*, um dos principais eventos acadêmicos do ramo da Computação Musica.

Síntese concatenativa baseada em samples e letras de canção fornecidas antecipadamente em um arquivo de texto mostraram ser, respectivamente, a abordagem técnica e o método de entrada fonética mais adequados para o sistema, conforme discutido. A entrada musical é dada em tempo real via MIDI, e a síntese é realizada pelo MBROLA, um sintetizador de fala adaptado.

Para as futuras versões do sistema, a implementação de uma série de modificações está prevista:

- Criação de um inventário de vozes próprio, com maior qualidade de áudio e com com amostras de voz de uma cantora profissional em lugar de *samples* falados, visando melhorar a naturalidade e inteligibilidade do canto sintetizado. Tal inventário será inicialmente com auxílio do MBROLATOR[1], uma ferramenta para criação de arquivos de bases de voz MBROLA, mas a versão final do sistema contará com um inventário próprio, fazendo sua dependência para com o MBROLA desaparecer.
- Incorporação do algoritmo de síntese concatenativa à implementação de PATRICIA, que será outro fator de eliminação da dependência em relação ao MBROLA.
- Tratamento mais abrangente das mensagens MIDI, a fim de dotar o sistema de controles de expressão artística.

[1] Disponível em <https://github.com/nummediart/MBROLATOR>

- Otimização do desempenho do sinterizador por meio do uso do processo *supernova* em lugar de *scsynth*, o que permitirá explorar os recursos de paralelismo do processador multicore do dispositivo.
- Integração do sintetizador a um instrumento musical eletrônico, de modo que ele funcione como sistema embarcado para o instrumento. Na versão atual, o instrumento atua como um periférico de entrada externo ao sintetizador.
- Transformação de PATRICIA em um instrumento aumentado, ou seja, dotado de controles que vão além da técnica convencional pelo qual seriam tocados. Num tal contexto, controles gestuais que permitam que de alguma forma a entrada fonética também se dê em tempo real poderiam ser implementados em PATRICIA, o que seria benéfico para a seu uso em ambientes mais artísticos.

Já entre os desafios a serem enfrentados nas futuras implementações, podem-se citar os seguintes:

- A dicussão a respeito das patentes do Vocaloid Keyboard, que é um produto similar e propriedade da Yamaha Corporation.
- A construção do inventário de voz a partir das gravações de uma cantora levantará questões éticas em torno de pesquisas envolvendo seres humanos.
- A sistematização de testes do sistema que permitam uma avaliação do produto por parte de artistas que, ao utilizá-lo, validarão os requisitos funcionais tidos como mais subjetivos.
- O estabelecimento de uma plataforma de hardware mais adequada, em termos de custo/benefício, para a implantação do sintetizador.

Esta pesquisa, iniciada em 2019, interrompida pela pandemia de COVID-19 em 2020 e retomada em 2022, teve os resultados do primeiro mapeamento sistemático conduzido apresentados no XVII Simpósio Brasileiro de Computação Musical (BRUM; MORENO, 2019) e publicados na Revista de Informática Teórica e Aplicada da UFRGS (BRUM; MORENO, 2020). Quanto ao sistema, já implementado, este foi apresentado na *International Computer Music Conference* de 2023, realizada em Shenzhen, China (BRUM; MENESES; MORENO, 2023).

Pode-se dizer, por fim, que PATRICIA é um sistema classificado em uma área de pesquisa com poucos desenvolvimentos — a síntese de canto em tempo real — e com uma característica pioneira: ser projetado para o português brasileiro. Suas limitações, portanto, não o impedem de ser uma contribuição relevante para a Computação Musical, além de abrir uma série de possibilidades e desafios para futuras pesquisas.

Referências

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*. Cham: Springer International Publishing, 2017. p. 129–158. ISBN 978-3-319-49607-8. Disponível em: <https://doi.org/10.1007/978-3-319-49607-8_5>. Citado na página 15.

BANDEIRA, M. *A versificação em língua portuguesa*. [S.l.]: Editora Delta, 1997. Citado na página 67.

BENOîT, C.; GRICE, M.; HAZAN, V. The sus test 1: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, v. 18, n. 4, p. 381 – 392, 1996. Cited by: 156. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030166343&doi=10.1016%2f0167-6393%2896%2900026-X&partnerID=40&md5=3bb9fdaaf7c1c54666f2f8a6b0308fd0>. Citado na página 62.

BLATTER, A. *Revisiting music theory: basic principles*. [S.l.]: Taylor & Francis, 2016. Citado na página 26.

BRUM, L. A. Z. Technical aspects of concatenation-based singing voice synthesis. *Scientia Plena*, v. 8, n. 3(a), may 2012. Disponível em: <https://www.scientiaplena.org.br/sp/article/view/919>. Citado na página 41.

BRUM, L. A. Z.; MENESES, E. A. L.; MORENO, E. D. Patricia: a real-time singing synthesizer prototype for the brazilian portuguese language. In: *Proceedings of the International Computer Music Conference 2023*. [S.l.: s.n.], 2023. p. 176 – 180. Citado na página 83.

BRUM, L. A. Z.; MORENO, E. D. State of art of real-time singing voice synthesis. In: *Anais do XVII Simpósio Brasileiro de Computação Musical*. Porto Alegre, RS, Brasil: SBC, 2019. p. 50–57. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbcm/article/view/10422>. Citado 3 vezes nas páginas 16, 52 e 83.

BRUM, L. A. Z.; MORENO, E. D. Challenges and perspectives on real-time singing voice synthesis. *Revista de Informática Teórica e Aplicada*, v. 27, n. 4, p. 118–126, Dec. 2020. Disponível em: <https://seer.ufrgs.br/index.php/rita/article/view/RITA_VOL27_NR4_118>. Citado 2 vezes nas páginas 52 e 83.

CALLOU, D.; LEITE, Y. *Iniciação à fonética e à fonologia*. [S.l.]: Jorge Zahar, 2005. Citado 2 vezes nas páginas 33 e 35.

CAMILO, D.; YABU-UTI, J. B. T.; YANO, Y. Circuitos lógicos. *São Paulo: LCTE*, 1986. Citado 5 vezes nas páginas 22, 30, 37, 38 e 70.

CAVALIERE, R. S. *Pontos essenciais em fonética e fonologia*. [S.l.]: Nova Fronteira, 2011. Citado 2 vezes nas páginas 33 e 34.

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: *Proc. Interspeech 2016*. [S.l.: s.n.], 2016. p. 1225–1229. Citado 5 vezes nas páginas 16, 49, 58, 59 e 60.

COLLINS, N. *Introduction to Computer Music*. [S.l.]: Wiley, 2010. ISBN 9780470714553. Citado na página 39.

COOK, P. R. Computer music. In: ____. *Springer Handbook of Acoustics*. New York, NY: Springer New York, 2007. p. 747–778. ISBN 978-1-4939-0755-7. Disponível em: <https://doi.org/10.1007/978-1-4939-0755-7_17>. Citado na página 41.

D'ALESSANDRO, N. et al. Maxmbrola: A max/msp mbrola-based tool for real-time voice synthesis. In: *2005 13th European Signal Processing Conference*. [S.l.: s.n.], 2005. p. 1–4. Citado 3 vezes nas páginas 49, 50 e 63.

DELALEZ, S.; D'ALESSANDRO, C. Adjusting the Frame: Biphasic Performative Control of Speech Rhythm. In: *Proceedings of Interspeech 2017*. Stockholm, Sweden: [s.n.], 2017. p. 864–868. Disponível em: <https://hal.sorbonne-universite.fr/hal-01672232>. Citado 3 vezes nas páginas 58, 59 e 60.

DONG, M. et al. I2r speech2singing perfects everyone's singing. In: *Proc. Interspeech 2014*. [S.l.: s.n.], 2014. p. 2148–2149. Citado 3 vezes nas páginas 59, 60 e 66.

DUTOIT, T. et al. The mbrola project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. [s.n.], 1996. v. 3, p. 1393–1396 vol.3. Disponível em: <https://doi.org/10.1109/ICSLP.1996.607874>. Citado 2 vezes nas páginas 40 e 68.

FEUGÈRE, L. et al. Cantor digitalis: chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, SpringerOpen, v. 2017, n. 1, p. 1–19, 2017. Disponível em: <https://doi.org/10.1186/s13636-016-0098-5>. Citado 3 vezes nas páginas 58, 59 e 60.

FILHO, F. M. *A acústica musical em palavras e sons*. [S.l.]: Atelie Editorial, 2004. Citado 2 vezes nas páginas 24 e 32.

FREIXES, M.; SOCORÓ, J. C.; ALÍAS, F. Adding singing capabilities to unit selection tts through hnm-based conversion. In: ABAD, A. et al. (Ed.). *Advances in Speech and Language Technologies for Iberian Languages*. Cham: Springer International Publishing, 2016. p. 33–43. ISBN 978-3-319-49169-1. Disponível em: <https://doi.org/10.1007/978-3-319-49169-1_4>. Citado na página 48.

GU, H.-Y.; LIAO, H.-L. Mandarin singing voice synthesis using an hnm based scheme. In: *2008 Congress on Image and Signal Processing*. [S.l.: s.n.], 2008. v. 5, p. 347–351. Citado na página 48.

HENRIQUE, L. L. *Acústica musical*. [S.l.: s.n.], 2002. Citado 15 vezes nas páginas 18, 19, 20, 21, 22, 23, 24, 29, 30, 31, 32, 35, 36, 39 e 42.

HENRIQUE, L. L. *Instrumentos musicais*. [S.l.]: Fundação Calouste Gulbenkian, 2004. (Manuais universitários). Citado na página 38.

KAGAMI, S. et al. Development of realtime japanese vocal keyboard. *Information Processing Society of Japan INTERACTION*, p. 837–842, 2012. Citado 7 vezes nas páginas 53, 54, 55, 59, 60, 62 e 67.

KASHIWASE, K. An over-the-shoulder keyboard that extends the potential for vocaloid performance. *Yamaha Corporation*, 2017. Accessed: 2023-01-29. Disponível em: <https://www.yamaha.com/en/about/design/synapses/id_104>. Citado 2 vezes nas páginas 53 e 67.

KENMOCHI, H. VOCALOID and Hatsune Miku phenomenon in Japan. In: *Proc. First Interdisciplinary Workshop on Singing Voice (InterSinging 2010)*. [S.l.: s.n.], 2010. p. 1–4. Citado na página 15.

KENMOCHI, H. Singing synthesis as a new musical instrument. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2012. p. 5385–5388. Disponível em: <https://doi.org/10.1007/978-981-10-0281-6_20>. Citado na página 15.

KENMOCHI, H.; OHSHITA, H. VOCALOID - commercial singing synthesizer based on sample concatenation. In: *Proc. Interspeech 2007*. [S.l.: s.n.], 2007. p. 4009–4010. Citado 3 vezes nas páginas 15, 47 e 48.

KENT, R.; READ, C. *Análise Acústica da Fala*. [S.l.]: Cortez Editora, 2015. Citado na página 44.

KHAN, N. U.; LEE, J. C. Hmm based duration control for singing tts. In: PARK, D.-S. et al. (Ed.). *Advances in Computer Science and Ubiquitous Computing*. Singapore: Springer Singapore, 2015. p. 137–143. ISBN 978-981-10-0281-6. Disponível em: <https://doi.org/10.1007/978-981-10-0281-6_20>. Citado 2 vezes nas páginas 15 e 48.

KIM, Y. E. Singing voice analysis, synthesis, and modeling. In: ____. *Handbook of Signal Processing in Acoustics*. New York, NY: Springer New York, 2008. p. 359–374. ISBN 978-0-387-30441-0. Disponível em: <https://doi.org/10.1007/978-0-387-30441-0_23>. Citado 2 vezes nas páginas 43 e 47.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado 2 vezes nas páginas 17 e 52.

LABRUNE, L. *The phonology of Japanese*. [S.l.]: Oxford University Press, 2012. Citado na página 67.

LI, M.; CHEN, S.; REN, K. Enabling private and non-intrusive smartphone calls with liptalk. In: *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. [s.n.], 2014. p. 191–192. Disponível em: <https://doi.org/10.1109/INFCOMW.2014.6849220>. Citado na página 48.

LOCQUEVILLE, G. et al. Voks: Digital instruments for chironomic control of voice samples. *Speech Communication*, v. 125, p. 97–113, 2020. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639320302788>. Citado 3 vezes nas páginas 58, 59 e 60.

LOY, G.; CHOWNING, J. *Musimathics, Volume 1: The Mathematical Foundations of Music*. [S.l.]: MIT Press, 2011. (Musimathics). ISBN 9780262516556. Citado 5 vezes nas páginas 21, 22, 26, 31 e 32.

LOY, G.; CHOWNING, J. *Musimathics, Volume 2: The Mathematical Foundations of Music*. MIT Press, 2011. (The MIT Press). ISBN 9780262292764. Disponível em: <https://books.google.com.br/books?id=88fxCwAAQBAJ>. Citado 3 vezes nas páginas 31, 32 e 45.

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. Disponível em: <https://doi.org/10.1017/S0140525X98001265>. Citado 2 vezes nas páginas 34 e 58.

MACON, M. et al. Concatenation-based midi-to-singing voice synthesis. In: AUDIO ENGINEERING SOCIETY. *Audio Engineering Society Convention 103*. [S.l.], 1997. Citado na página 47.

MANNING, P. et al. Computers and music. In: *The New Grove Dictionary of Music and Musicians*. Oxford University Press, 2001. ISBN 9781561592630. Disponível em: <https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040583>. Citado 2 vezes nas páginas 39 e 40.

MATSUBARA, K. et al. Full-band lpcnet: A real-time neural vocoder for 48 khz audio with a cpu. *IEEE Access*, v. 9, p. 94923–94933, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2021.3089565>. Citado 4 vezes nas páginas 58, 59, 60 e 66.

MCCARTNEY, J. Supercollider: a new real time synthesis language. In: *Proceedings of the 1996 International Computer Music Conference*. [s.n.], 1996. p. 257–258. Disponível em: <http://hdl.handle.net/2027/spo.bbp2372.1996.078>. Citado 2 vezes nas páginas 49 e 63.

MED, B. *Teoria da música*. [S.l.]: Brasília: Musimed, 1996. v. 996. Citado 7 vezes nas páginas 23, 24, 25, 26, 27, 28 e 29.

MIDI-MANUFACTURERS-ASSOCIATION et al. The complete midi 1.0 detailed specification. *Los Angeles, CA, The MIDI Manufacturers Association*, 1996. Citado na página 42.

MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLOS Medicine*, Public Library of Science, v. 6, n. 7, p. 1–6, 07 2009. Disponível em: <https://doi.org/10.1371/journal.pmed.1000097>. Citado na página 56.

MOOG, R. A. Midi: Musical instrument digital interface. *Journal of the Audio Engineering Society*, Audio Engineering Society, v. 34, n. 5, p. 394–404, 1986. Disponível em: <http://www.aes.org/e-lib/browse.cfm?elib=5267>. Citado na página 42.

NOUGUÉ, C. A. A. *Suma gramatical da língua portuguesa: gramática geral e avançada*. [S.l.]: E Realizações Editora, 2015. ISBN 9788580332032. Citado na página 33.

ORJUELA, A. et al. Methodological proposal for the identification of incremental innovations in smes. *European Research Studies Journal*, XXII, n. 4, p. 199–214, 2019. Citado na página 52.

O'SULLIVAN, D.; IGOE, T. *Physical computing: sensing and controlling the physical world with computers*. [S.l.]: Course Technology Press, 2004. Citado 2 vezes nas páginas 40 e 42.

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. Citado na página 49.

PABON, P. et al. Future Perspectives. In: *The Oxford Handbook of Singing*. Oxford University Press, 2019. ISBN 9780199660773. Disponível em: <https://doi.org/10.1093/oxfordhb/9780199660773.013.67>. Citado na página 15.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*. Swindon, GBR: BCS Learning Development Ltd., 2008. (EASE'08), p. 68–77. Citado 2 vezes nas páginas 17 e 52.

PRIOLLI, M. L. d. M. *Princípios Básicos da Música para a Juventude*. [S.l.: s.n.], 1989. v. 1. Citado na página 27.

PUCKETTE, M. The patcher. In: INTERNATIONAL COMPUTER MUSIC ASSOCIATION. *Proceedings of the 1988 International Computer Music Conference*. 1988. Disponível em: <http://hdl.handle.net/2027/spo.bbp2372.1988.046>. Citado na página 49.

PUCKETTE, M. S. Pure data. In: INTERNATIONAL COMPUTER MUSIC ASSOCIATION. *Proceedings of the 1997 International Computer Music Conference*. 1997. Disponível em: <http://hdl.handle.net/2027/spo.bbp2372.1997.060>. Citado na página 49.

ROADS, C. *The computer music tutorial*. [S.l.]: MIT press, 1996. Citado 2 vezes nas páginas 44 e 81.

RODET, X.; POTARD, Y.; BARRIERE, J.-B. The chant project: From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, MIT Press, v. 8, n. 3, p. 15–31, 1984. Disponível em: <https://doi.org/10.2307/3679810>. Citado na página 44.

SHERRELL, L. Evolutionary prototyping. In: ____. *Encyclopedia of Sciences and Religions*. Dordrecht: Springer Netherlands, 2013. p. 803–803. ISBN 978-1-4020-8265-8. Disponível em: <https://doi.org/10.1007/978-1-4020-8265-8_201039>. Citado na página 61.

SILVA, T. C.; LEITE, C. Padrões sonoros emergentes:(oclusiva alveolar + sibilante) no português brasileiro. *Caderno de Letras*, n. 24, p. 15–36, 2015. Disponível em: <https://doi.org/10.15210/cdl.v0i24.7270>. Citado na página 67.

SUGIURA, K.; ZETTSU, K. Analysis of long-term and large-scale experiments on robot dialogues using a cloud robotics platform. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. [s.n.], 2016. p. 525–526. Disponível em: <https://doi.org/10.1109/HRI.2016.7451838>. Citado na página 48.

TABET, Y.; BOUGHAZI, M. Speech synthesis techniques. a survey. In: *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*. [s.n.], 2011. p. 67–70. Disponível em: <https://doi.org/10.1109/WOSSPA.2011.5931414>. Citado na página 48.

TAE, J.; KIM, H.; LEE, Y. Mlp singer: Towards rapid parallel korean singing voice synthesis. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. [s.n.], 2021. p. 1–6. Disponível em: <https://doi.org/10.1109/MLSP52302.2021.9596184>. Citado 4 vezes nas páginas 58, 59, 60 e 66.

TAYLOR, C.; CAMPBELL, M. Sound. In: *The New Grove Dictionary of Music and Musicians*. Oxford University Press, 2001. ISBN 9781561592630. Disponível em: <https://www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000026289>. Citado 3 vezes nas páginas 19, 20 e 32.

TERNSTRÖM, S. Session on naturalness in synthesized speech and music. In: *143rd ASA meeting, Pittsburgh*. [S.l.: s.n.], 2002. Citado na página 62.

UMBERT, M. et al. Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. *IEEE Signal Processing Magazine*, v. 32, n. 6, p. 55–73, 2015. Citado na página 62.

VEAUX, C. et al. Gesture control of hmm-based singing voice synthesis. In: *Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8)*. [S.l.: s.n.], 2013. p. 247–248. Citado 3 vezes nas páginas 59, 60 e 66.

WEBER, R. F. *Arquitetura de computadores pessoais*. [S.l.]: Sagra Luzzatto, 2003. Citado 3 vezes nas páginas 21, 30 e 70.

WELLS, J. C. et al. Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, Mouton de Gruyter Berlin, v. 4, p. 684–732, 1997. Citado 2 vezes nas páginas 36 e 67.

WISNIK, J. M. S. Som e o sentido: uma outra história das músicas. 2001. Citado 2 vezes nas páginas 21 e 24.

XIAO, X. et al. T-voks: the singing and speaking theremin. In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. Zenodo, 2019. p. 110–115. Disponível em: <https://doi.org/10.5281/zenodo.3672886>. Citado 2 vezes nas páginas 58 e 66.

YOUNG, H.; FREEDMAN, R. *Sears & Zemansky física II: termodinâmica e ondas*. Pearson Addison Wesley, 2008. ISBN 9788588639331. Disponível em: <https://books.google.com.br/books?id=Z3GZPgAACAAJ>. Citado 2 vezes nas páginas 18 e 20.

ZICARELLI, D. Max/msp software. *San Francisco: Cycling '74*, 1997. Citado na página 49.

Apêndices

APÊNDICE A – Código fonte do sintetizador PATRICIA

Listing A.1 – Código fonte do sintetizador PATRICIA.

```
1  s.reboot;
2  MIDIIn.free;
3  MIDIClient.init;
4  MIDIIn.connectAll;
5  MIDIClient.sources;
6  ~arq1 = true;
7
8
9  // Caminho dos arquivos do projeto:
10 p = Platform.userHomeDir ++ Platform.pathSeparator ++ "PATRICIA" ++
       Platform.pathSeparator;
11
12
13 //Abre o arquivo fonetico para leitura.
14 ~lyrics = File.open (p++"lyrics.txt","r");
15
16 //Tecla pressionada
17
18 m = MIDIFunc.noteOn({arg ...args;
19
20   //Inicializa o contador
21   i = 0;
22   Clock.awake;
23
24   //Obtem a proxima silaba a cada nota MIDI tocada.
25   ~syllable = ~lyrics.getLine;
26
27   ~nota = args[1];
28   ~vol = args[0];
29   ~inv = "br" ++ (args[2]+1)  ;
30
31   //Obtendo a frequencia conforme a nota MIDI.
32   f = 440 * ((2 ** (1.0/12.0)) ** (~nota-81));
33
```

```
34  //Enquanto houver silabas no arquivo
35  if (((~syllable != nil) && (~syllable.asString != "!")),
36    {
37      //Alterna entre dois arquivos .pho do mbrola, para evitar
           conflito.
38      if(~arq1,
39        {~phones = "f1.pho";},
40        {~phones = "f2.pho";}
41      );
42      ~phonetics = File.open(p++~phones,"w");
43
44      //Inicializando a duracao fonetica.
45      t = 50;
46
47      //Converte a linha silabica do arquivo lyrics.txt num arquivo
           .pho do mbrola
48      while({i < (~syllable.size-1)},
49        {
50          if((~syllable[i].asString != "-"),
51            {
52              ~phonetics.write(~syllable[i].asString );
53              if( (Set["a","e","i","o","u","@"].includes
54                        (~syllable[i].asString)),
55                {t = 200; //tempo para vogais
56                },{}
57              );
58            },{
59              ~phonetics.write(" "++t.asString ++ " 100 "++f++" \n");
60              t = 50;//tempo para consoantes e semivogais
61          });
62
63          //Incrementa o contador
64          i = i + 1;
65      });
66      ~phonetics.close;
67
68      //Nomeia o arquivo com o codigo da nota MIDI.
69      ~path = p ++ ~nota ++".au";
70
71
72      //Invoca o mbrola, gerando o audio da silaba.
```

```
73      c = p ++ "mbrola " ++ p ++  ~inv + p ++ ~phones ++ " " ++
            ~path;
74      c.systemCmd;
75
76
77      //Define o inicio e o fim do loop
78      ~sample = SoundFile.openRead (~path);
79      //Come a  o loop no ponto medio do arquivo no dominio do
            tempo.
80      ~start = (~sample.duration/2)* ~sample.sampleRate;
81      //Calcula o fim do loop em funcao da frequencia.
82      ~end = ~start +  ((1.0/f) * ~sample.sampleRate);
83
84
85      Routine{// bloco necessario para invocar o comando "sync"
86
87        b = Buffer.read(s,~path, bufnum: ~nota);
88
89        //Perfaz o loop na vogal enquanto a tecla estiver acionada.
90        SynthDef("Looping",{
91          arg gate=1; //argumento que mantem ou libera o loop
92
93          var  signal;
94          signal = LoopBuf.ar(1,~nota,
                ~sample.sampleRate/s.sampleRate, gate, 0, ~start, ~end,
                2);
95
96          s.volume.volume = 40 * log10(~vol/127);
97
98          Out.ar(0, signal);
99        }).send(s);
100
101       // Sincroniza o servidor para evitar erro ao criar o no.
102       s.sync;
103       s.sendMsg("/error",0);
104       //Libera o no tocado nesta nota pela ultima vez.
105       s.sendMsg("/n_free",  args[1]);
106       postln(~syllable);
107
108       // Cria um novo no nomeado a partir do codigo MIDI da nota.
109       s.sendMsg("/s_new", "Looping", args[1]);
110     }.play;
```

```
111
112
113         //Inverte o booleano que alterna entre os arquivos .pho
114         ~arq1 = ~arq1.not; },
115       {
116         //Se nao achar mais silabas
117         postln("FIM DA CANCAO!");
118         MIDIIn.free;
119       }
120     )
121 });
122
123 //Tecla liberada
124 n = MIDIFunc.noteOff({arg ...args;
125   //Modifica o argumento gate, liberando o loop do no que
        corresponde a respectiva nota MIDI.
126   s.sendMsg("/n_set", args[1], \gate, 0);
127 });
```

APÊNDICE B – Fonemas do português brasileiro e suas representações

Tipo de fonema	Símbolo IPA	Representação no MBROLA	Exemplos
Vogal	a	a	b**a**rco
	ɐ	@	c**â**mera
	ɐ̃	am	amanh**ã**
	e	e	cab**e**lo
	ɛ	ee	qu**e**ro
	ẽ	em	qu**en**te
	i	i	p**i**co
	ĩ	im	br**in**co
	o	o	t**o**lo
	ɔ	oo	b**o**la
	õ	om	**om**bro
	u	u	d**u**ro
	ũ	um	alg**um**
Consoante	b	b	**b**arco
	k	k	**c**om
	d	d	**d**oce
	g	g	**g**rande
	p	p	**p**ai
	t	t	**t**aco
	f	f	**f**ácil
	v	v	**v**inho
	ʒ	j	**j**ato
	s	s	**s**ala
	s	s2	cas**c**a
	ʃ	x	**ch**ave
	z	z	**z**ebra
	m	m	**m**esmo
	n	n	**n**unca
	ɲ	nh	gali**nh**a
	l	l	**l**aranja
	ʎ	lh	a**lh**o
	ɾ	r	pu**r**o
	ɾ	r2	ha**r**pa
	ʁ	rr	to**rr**e
Semivogal	j	y	ma**i**s
	w	w	ma**u**

O Quadro acima apresenta os fonemas do português brasileiro agrupados conforme sua classificação em vogais, consoantes em semivogais. Para cada fonema apresenta-se seu

símbolo no IPA, sua represntação nos inventários *br1*, *br2* e *br3* do MBROLA e uma palavra que exemplifica o valor fonético, indicando-se em negrito as letras da palavra que correspondem a cada fonema apresentado. Observe-se que os sons representados no MBROLA por "s2"e "r2"são foneticamente idênticos aos representados por "s"e "r", respectivamente. As representações s2 e r2 foram necessárias para que o MBROLA articulasse corretamente os difones quando tais fonemas aparecessem ao final da sílaba.

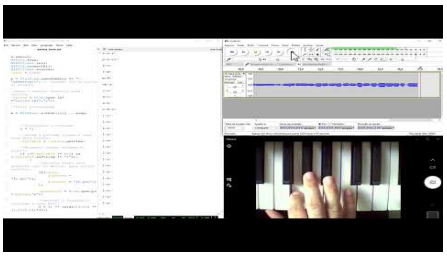
APÊNDICE C – Formulário de avaliação do sintetizador PATRICIA

Avaliação do sintetizador PATRICIA

Canção "Asa Branca"

* Indica uma pergunta obrigatória

Demonstração da canção "Asa Branca". Assista e responda às quatro perguntas a seguir.



http://youtube.com/watch?v=XK7y-mCw7KI

1. Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

2. Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

3. Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

4. Quanto à execução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico? *

Marcar apenas uma oval.

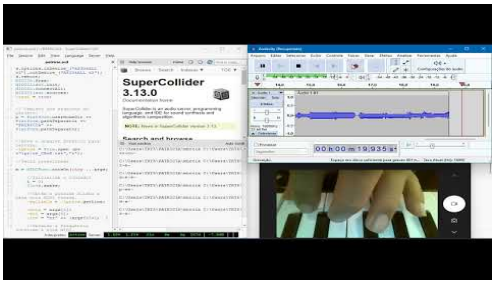
- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

Pular para a pergunta 5

Avaliação do sintetizador PATRICIA

Canção "O Cravo Brigou com a Rosa"

Demonstração da canção "O Cravo Brigou com a Rosa". Assista e responda às quatro perguntas a seguir.



http://youtube.com/watch?v=MVc3eyzrfAg

5. Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

6. Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

7. Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

8. Quanto à execução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico? *

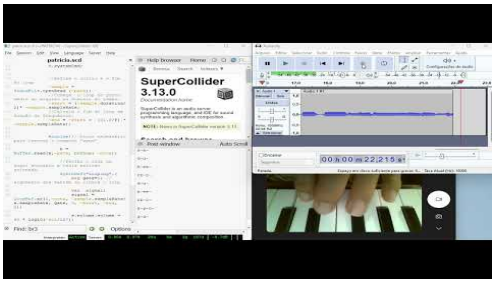
Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

Avaliação do sintetizador PATRICIA

Canção "Atirei o Pau no Gato"

Demonstração da canção "Atirei o Pau no Gato". Assista e responda às quatro perguntas a seguir.



http://youtube.com/watch?v=Gw2P2yRb9cs

9. Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

10. Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

11. Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

12. Quanto à execução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico? *

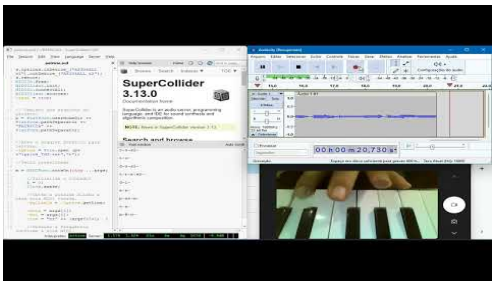
Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

Avaliação do sintetizador PATRICIA

Canção "Terezinha de Jesus"

Demonstração da canção "Terezinha de Jesus". Assista e responda às quatro perguntas a seguir.



http://youtube.com/watch?v=yoFT1YiTzvY

13. Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

14. Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

15. Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

16. Quanto à execução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico? *

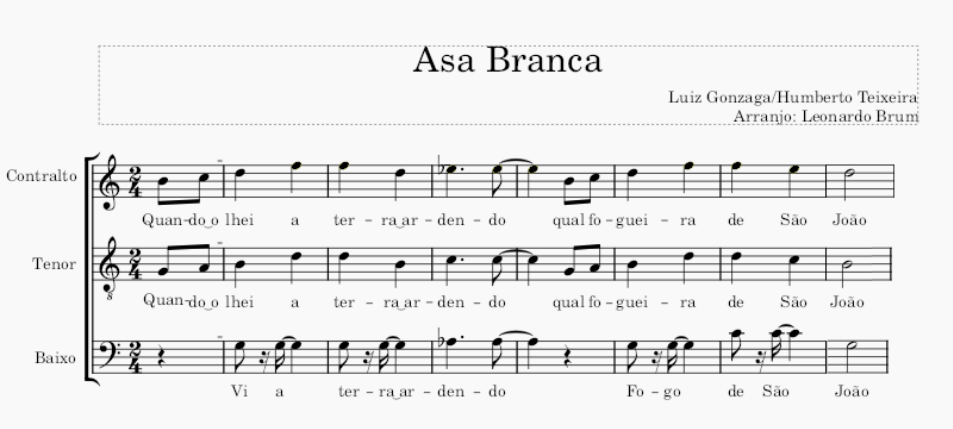
Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

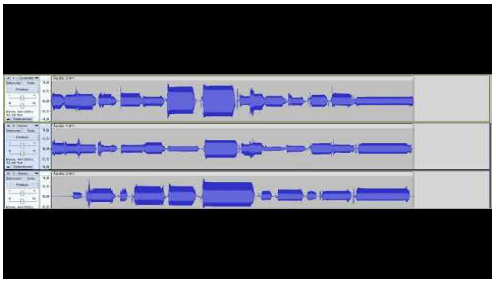
Avaliação do sintetizador PATRICIA

Canção "Asa Branca" a três vozes

Arranjo utilizado para a demonstração



Demonstração da canção "Asa Branca" a três vozes. Assista e responda às quatro perguntas a seguir.


http://youtube.com/watch?v=HU42aDK8ico

17. Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

18. Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

19. Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que o canto gerado corresponde ao arranjo exibido pela partitura acima? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

20. Você considera que o sintetizador PATRICIA seria uma ferramenta útil para o ensino de canto coral? *

Marcar apenas uma oval.

- ◯ Concordo totalmente
- ◯ Concordo parcialmente
- ◯ Discordo parcialmente
- ◯ Discordo totalmente

APÊNDICE D – Respostas ao formulário de avaliação

Avaliação do sintetizador PATRICIA

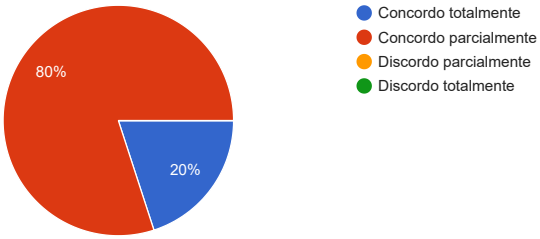
5 respostas

Publicar análise

Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis?

Copiar

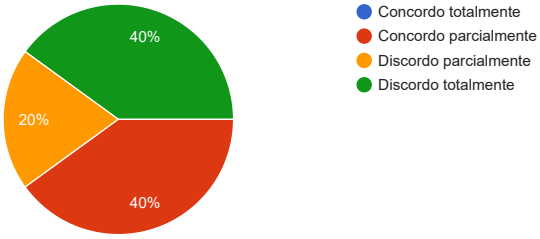
5 respostas



Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade?

Copiar

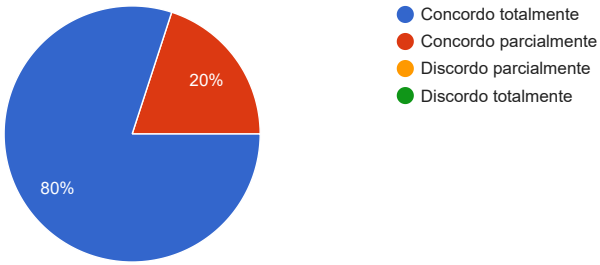
5 respostas



Quanto à reprodução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico?

Copiar

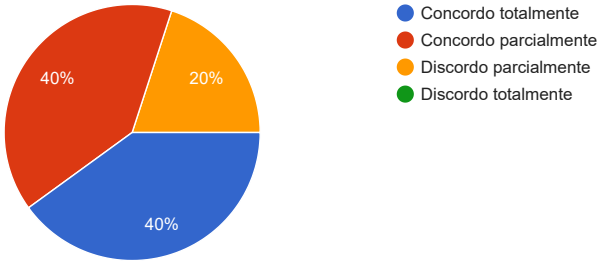
5 respostas



Quanto à execução da canção "Asa Branca" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico?

Copiar

5 respostas

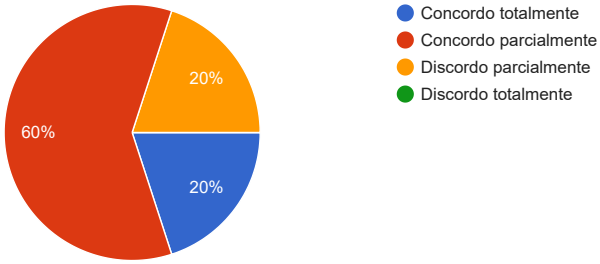


Avaliação do sintetizador PATRICIA

Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis?

Copiar

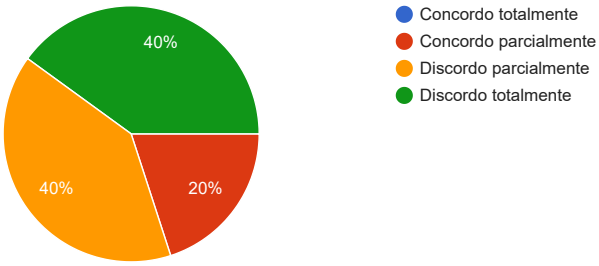
5 respostas



Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade?

Copiar

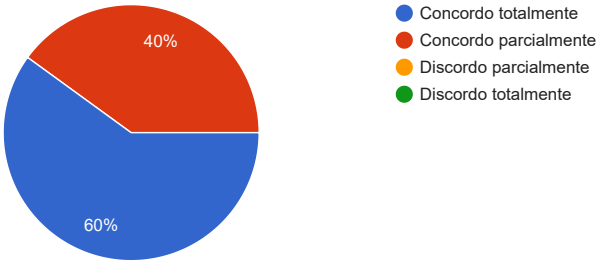
5 respostas



Quanto à reprodução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)?

Copiar

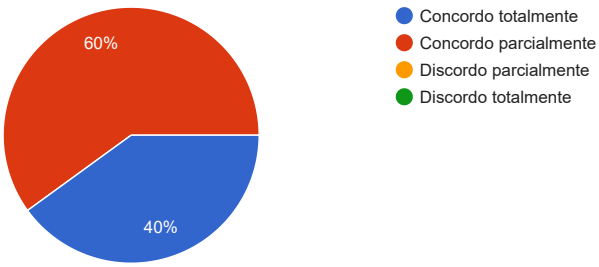
5 respostas



Quanto à execução da canção "O Cravo Brigou com a Rosa" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico?

Copiar

5 respostas

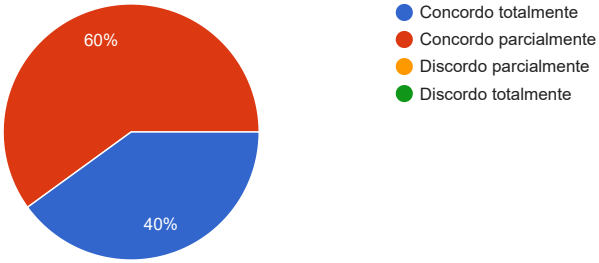


Avaliação do sintetizador PATRICIA

Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis?

Copiar

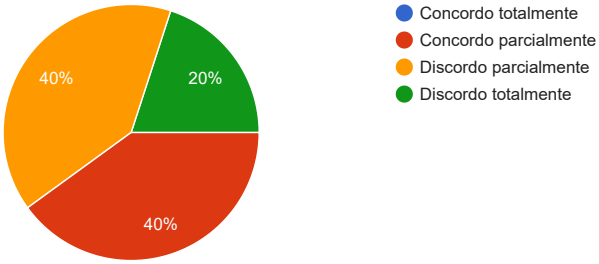
5 respostas



Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade?

Copiar

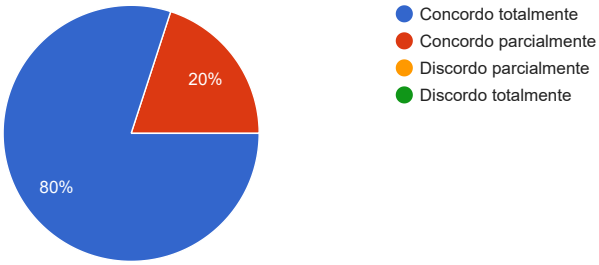
5 respostas



Quanto à reprodução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)?

Copiar

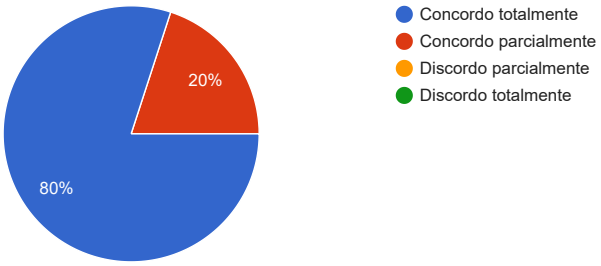
5 respostas



Quanto à execução da canção "Atirei o Pau no Gato" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico?

Copiar

5 respostas

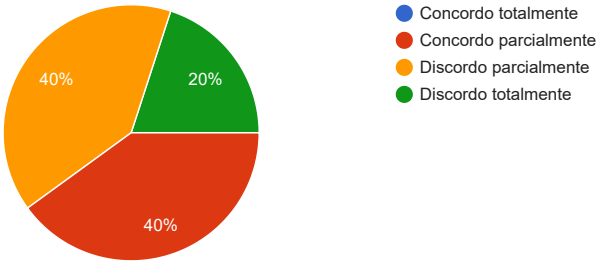


Avaliação do sintetizador PATRICIA

Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis?

Copiar

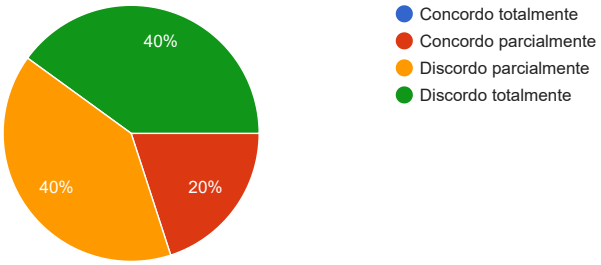
5 respostas



Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade?

Copiar

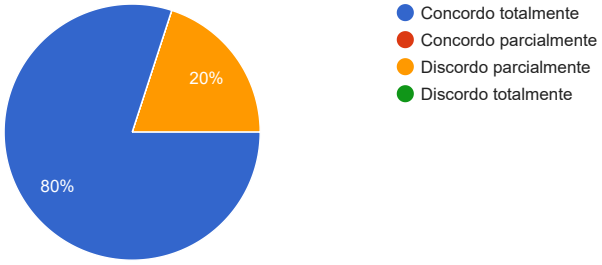
5 respostas



Quanto à reprodução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que as alturas musicais geradas correspondem à execução do músico (leve em conta que a filmagem foi feita em ângulo espelhado)?

Copiar

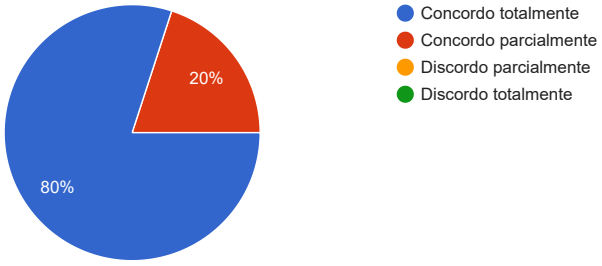
5 respostas



Quanto à execução da canção "Terezinha de Jesus" pelo sintetizador PATRICIA, você considera que a duração das notas geradas correspondem à execução do músico?

Copiar

5 respostas

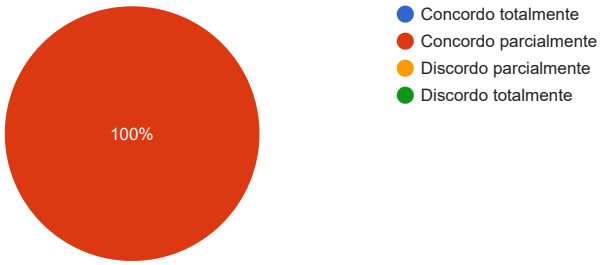


Avaliação do sintetizador PATRICIA

Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que as sílabas cantadas são compreensíveis?

Copiar

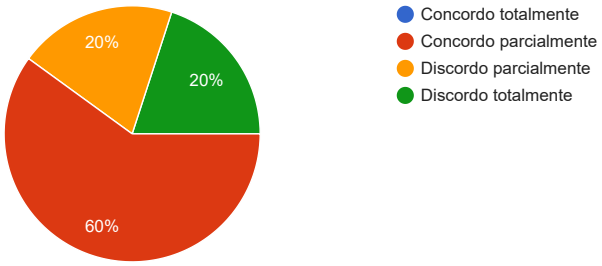
5 respostas



Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que o canto gerado tem naturalidade?

Copiar

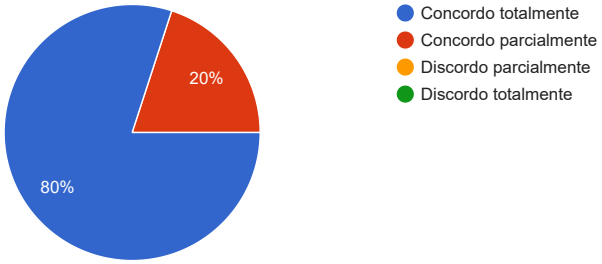
5 respostas



Quanto à reprodução da canção "Asa Branca" a três vozes pelo sintetizador PATRICIA, você considera que o canto gerado corresponde ao arranjo exibido pela partitura acima?

Copiar

5 respostas



Você considera que o sintetizador PATRICIA seria uma ferramenta útil para o ensino de canto coral?

Copiar

5 respostas