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

Leonardo A. Z. Brum[1] **, Eduardo A. L. Meneses**[2] **, Edward D. Moreno**[3]

[1]LaSiD - Laboratório de Sistemas Distribuídos – Universidade Federal da Bahia
Av. Adhemar de Barros, s/n, Instituto de Matemática e Estatística – 40170-110 Salvador, BA

[2]Metalab – Société des arts technologiques
1201 St Laurent Blvd – H2X 2S6 Montréal, QC, Canada

[3]Programa de Pós-Graduação em Ciência da Computação – Universidade Federal de Sergipe
Av.Marcelo Deda Chagas, s/n – 49107-230 São Cristóvão, SE

leonardo.brum@dcomp.ufs.br, emeneses@sat.qc.ca, edward@dcomp.ufs.br

Abstract. *This paper describes PATRICIA, a system that performs real-time singing voice synthesis (SVS) for the Brazilian Portuguese language. A technological mapping and a systematic review were conducted to study the latest developments in real-time SVS and to give directions for PATRICIA design and implementation. Sample-based concatenative synthesis with text files providing the song lyrics in advance was the approach chosen to perform the task. The most recent implemented functionalities are presented and future enhancements are pointed out to overcome the system's current limitations.*

Resumo. *Este artigo descreve PATRICIA, um sistema que realiza a síntese de voz cantada em tempo real para o idioma português do Brasil. Um mapeamento tecnológico e uma revisão sistemática foram conduzidos a fim de estudar os mais recentes desenvolvimentos na área de síntese de canto em tempo real e fornecer diretrizes para o projeto e implementação de PATRICIA. Síntese concatenativa baseada em samples com arquivos texto provendo antecipadamente a letra da canção foi a abordagem escolhida para desempenhar a tarefa. As funcionalidades mais recentemente implementadas são apresentadas, além de futuras melhorias, apontadas no intuito de que o sintetizador supere suas atuais limitações em versões futuras.*

1. Introdução

A finalidade da síntese de voz cantada é a geração de canto por meio de métodos computacionais. Alguns autores consideram que a síntese de canto seja um ramo da tecnologia de conversão de texto em fala, designada *text-to-speech* (TTS) [Alivizatou-Barakou et al. 2017, Tan 2023], 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 este tipo de dado pode ser fornecido por outros meios, como um sinal de áudio de voz [Locqueville et al. 2020, Dong et al. 2014], por exemplo. O outro tipo de entrada de dados é o musical, que provê as qualidades do som, como altura, duração e intensidade, à voz sintetizada.

As possibilidades de uso da tecnologia de síntese de canto têm sido estendidas pelo desenvolvimento de sintetizadores em tempo real. Em tais sistemas, o canto é gerado no mesmo instante em que a entrada de dados — musical, fonética, ou ambas — é fornecida pelo usuário, permitindo a este a execução de uma performance instrumental. Entretanto, a quantidade de sintetizadores em tempo real desenvolvidos nos últimos anos mostrou-se bastante pequena em comparação com o universo da síntese de canto em geral [Brum and Moreno 2019].

Além disso, verificava-se no mercado e na academia a falta de ferramentas que fornecessem a síntese de canto para o português brasileiro. O sistema descrito neste trabalho, chamado PATRICIA, é o primeiro sintetizador especificamente projetado para gerar o canto em tal idioma. Um esboço da arquitetura do sistema foi proposto ainda em 2020 [Brum and Moreno 2020] a título de pesquisa futura, e um protótipo inicial implementado foi apresentado em [Brum et al. 2023].

O presente trabalho está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados, buscados por meio de métodos sistemáticos; por sua vez, a Seção 3 descreve, em linhas gerais, o funcionamento do sintetizador PATRICIA; já a Seção 4 discute as mais recentes funcionalidades implementadas no sistema; finalmente, a Seção 5 traz uma breve conclusão e aponta para os trabalhos futuros.

2. Trabalhos relacionados

No intuito de estabelecer o estado da técnica e o estado da arte na área de síntese de canto em tempo real, um mapeamento tecnológico e uma revisão de literatura baseados em métodos sistemáticos [Kitchenham 2004, Petersen et al. 2008, Moher et al. 2009] foram conduzidos, tendo seus resultados publicados inicialmente em artigos anteriores [Brum and Moreno 2019, Brum and Moreno 2020]. Essas mesmas pesquisas foram reconduzidas em 2022 a fim de se encontrar trabalhos ainda mais recentes, inquirir a respeito das abordagens técnicas empregadas e fornecer diretrizes para o projeto e implementação de PATRICIA. Detalhes a respeito da condução desta nova revisão, incluindo bases de dados, termos de busca e critérios da seleção podem ser encontrados em trabalho mais recente [Brum 2023].

A técnicas de síntese de canto podem ser classificadas de acordo com três abordagens principais: abordagem baseada em regras, abordagem concatenativa e abordagem dirigida a dados [Kim 2008]. A abordagem baseada em regras considera como o som é produzido, gerando voz por meio da análise de características físicas do sinal de áudio, como os formantes. Uma das iniciativas pioneiras na área foi a do IRCAM, que desenvolveu na Europa um sistema chamado CHANT [Rodet et al. 1984]. Entre os sintetizadores de canto em tempo real baseado em regras mais recentes, pode-se citar o igualmente europeu Cantor Digitalis [Feugère et al. 2017], que gera vogais a partir de um método de entrada gestual denominado quironomia (*chironomy*) descrito da seguinte maneira pelos autores: 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. A integração de várias instâncias do Cantor Digitalis com o objetivo de formar um coral virtual recebeu o nome de Chorus Digitalis [Le Beux et al. 2011].

Na abordagem concatenativa, um conjunto de amostras de vozes pré-gravadas

é manipulado de acordo com as entradas fonéticas e musicais do sistema, como ocorre no sintetizador de canto comercial Vocaloid [Kenmochi and Ohshita 2007], da Yamaha. Uma versão em tempo real deste sistema, chamada Vocaloid Keyboard [Kagami et al. 2012], foi implementada na forma de um teclado musical para o qual o sintetizador Vocaloid funcionava como sistema embarcado. Outros sintetizadores concatenativos em tempo real encontrados foram o SERAPHIM [Chan et al. 2016], que sintetiza canto em japonês e mandarim a partir de uma entrada gestual, e o VOKinesiS [Delalez and d'Alessandro 2017], que integrou a interface de entrada do já mencionado Cantor Digitalis a pedais, fornecendo melodia a amostras de voz pré-gravadas via protocolo MIDI.

Por fim, na abordagem dirigida a dados, modelos estatísticos paramétricos como HMMs (*Hidden Markov Models*) são usados para aplicar os comportamentos de um determinado sinal à voz sintetizada num processo de treinamento para aprendizagem de máquina. Um dos sistemas mais conhecidos dentre os que utilizam HMMs para realizar a síntese de canto é o SinSy [Oura et al. 2010], desenvolvido pelo Nagoya Institute of Technology. Entre os sintetizadores de canto em tempo real pesquisados, utilizam HMMs o MAGE/pHTS [Veaux et al. 2013] e do I²R Speech2Singing [Dong et al. 2014]. Em sistemas mais recentes, redes neurais profundas têm sido aplicadas para perfazer o treinamento. É o caso do MLP Singer [Tae et al. 2021], projetado para o idioma coreano, e do Full-Band LPCNet [Matsubara et al. 2021], para o japonês.

A Tabela 1 exibe o resultado da revisão de literatura em relação às abordagens técnicas empregadas pelos sintetizadores de canto em tempo real selecionados. Verificou-se que a maior parte deles — isto é, quatro — vale-se da abordagem dirigida a dados, seguidos pelos sintetizadores concatenativos, com três exemplares. A abordagem baseada em regras é utilizada por apenas um dos sintetizadores estudados, o que pode ser explicado pelo fato de que tal abordagem é a mais antiga dentre as três.

Tabela 1. Abordagens técnicas utilizadas pelos sintetizadores estudados.

Sintetizador	Baseado em regras	Baseado em samples	Dirigido a dados
MLP Singer [Tae et al. 2021]			✓
Full-Band LPCNet [Matsubara et al. 2021]			✓
Cantor Digitalis [Feugère et al. 2017]	✓		
VOKinesiS [Delalez and 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]		✓	

3. O sintetizador PATRICIA

O sistema aqui descrito é, até onde se pôde pesquisar, o primeiro sintetizador de canto projetado especificamente para sintetizar o canto no idioma português brasileiro. Ele foi inicialmente concebido como um sistema embarcado que perfaz a síntese de canto em um teclado MIDI para fornecer uma performance em tempo real. Tal concepção

inicial encontrou maior correspondência, dentre os sintetizadores selecionados na revisão sistemática descrita na Seção 2, no Vocaloid Keyboard. Nosso sistema recupera dados fonéticos de um arquivo de texto e parâmetros musicais oriundos de um teclado MIDI para realizar uma síntese concatenativa com auxílio do sintetizador de fala MBROLA [Dutoit et al. 1996], retornando um arquivo de áudio. Este arquivo é reproduzido em *loop* enquanto a respectiva nota MIDI é mantida ativa, o que proporciona o controle da voz em tempo real. O sintetizador se chama PATRICIA, acrônimo para *Programa que Articula em Tempo Real o Idioma Cantado Inscrito em Arquivo*, e foi implementado em SuperCollider (McCartney, 1996), um ambiente e linguagem de programação para síntese de áudio em tempo real. GitHub foi usado para controle de versão. Os mecanismos de entrada, processamento e saída de dados do sintetizador serão descritos em detalhes nas próximas subseções.

3.1. Entrada de dados

De acordo com o estabelecido na Seção 1, um sintetizador de canto lida com dois tipos básicos de dados de entrada: o fonético, que provém da letra da canção a ser sintetizada, e os parâmetros musicais, como altura e duração das notas. Cada sílaba da letra da canção corresponde a uma nota musical.

No protótipo do Vocaloid Keyboard, mencionado na Seção 2, ambos os tipos de entrada eram dados em tempo real. No caso da entrada fonética, a tarefa se mostrou viável por conta da estrutura mais simples das sílabas 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 entrada necessariamente como sendo uma sílaba "TA" uma vez que uma eventual estrutura "AT" não existe na língua japonesa [Kubozono 1989]. Em contraste, o português brasileiro possui sílabas como a última da palavra "magistrais", cuja estrutura fonética complexa aumenta os desafios para o uso de controladores manuais para mapear uma entrada fonética em tempo real.

Por tais razões, PATRICIA recupera dados fonéticos de um arquivo de texto que deve ser preparado antes do início da apresentação musical. A própria versão comercial do Vocaloid Keyboard, o *keytar* VKB-100, vale-se igualmente de tal estratégia [Kashiwase 2017]. Em PATRICIA, a primeira linha do arquivo texto 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 hifens. Para representar os fonemas, PATRICIA utiliza a adaptação de caracteres SAMPA [Wells et al. 1997] definida para os inventários de voz em português do Brasil desenvolvidos para o sintetizador MBROLA.

Por sua vez, os parâmetros musicais são fornecidos por um teclado MIDI conectado ao computador hospedeiro no qual o sistema é executado. A cada vez que uma mensagem MIDI *Note on* é recebida, PATRICIA lê uma linha do arquivo de texto onde se encontra a letra da canção, contendo uma sílaba. Esta sílaba e o número de nota MIDI são utilizados como parâmetros para gerar um arquivo de áudio conforme a descrição a seguir.

3.2. Processamento da síntese

Conforme mencionado, PATRICIA se vale do mecanismo de síntese de um programa externo, o sintetizador de fala MBROLA, que realiza uma síntese concatenativa baseada em

seleção de unidades. As unidades concatenadas são difonos, ou seja, uma conjunção de dois fonemas. Um conjunto de difonos de um determinado idioma é gravado e armazenado em um arquivo que serve como inventário de voz para o sistema. Por sua vez, a fala a ser sintetizada pode ser armazenada em um arquivo de texto com a extensão *.PHO*. Cada linha deste arquivo contém um fonema escrito em notação fonética, 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. Em PATRICIA, uma instrução de linha de comando chama o MBROLA, que realiza a seleção das amostras no inventário de voz indicado, concatenando os difonos de acordo com a descrição dentro do arquivo *.PHO* e, finalmente, gerando um arquivo de áudio.

Para cada nota MIDI e sua sílaba correspondente recuperada por PATRICIA, é gerado um arquivo *.PHO*. Assim, cada linha deste arquivo contém os quatro parâmetros a seguir:

- Um fonema proveniente da sílaba recuperada (o arquivo *.PHO* terá tantas linhas quanto houver fonemas nesta sílaba).
- A duração do fonema, definida arbitrariamente como 25 ms para semivogais e consoantes e 200 ms para vogais.
- Um marco de altura que é sempre de 100%, o que significa que a altura dos fonemas será a mesma durante toda a sua duração.
- Frequência fundamental para fornecer a altura, calculada de acordo com o número da nota MIDI.

A Equação 1 mostra a fórmula usada para calcular a frequência fundamental f_0 de cada sílaba, dado seu número de nota MIDI correspondente *n*:

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

Depois que o arquivo *.PHO* é criado, PATRICIA usa o mecanismo do SuperCollider para enviar ao sistema operacional do computador hospedeiro uma linha de comando MBROLA. Esta instrução faz referência ao arquivo *.PHO* e a um dentre três inventários de voz do MBROLA em português do Brasil, denominados *br1*, *br2* e *br3*, para criar o arquivo de áudio correspondente no formato *.AU*. A próxima subseção descreve como PATRICIA manipula esses arquivos de áudio para gerar a voz cantada sintetizada.

3.3. Saída de áudio

Como cada arquivo de áudio gerado pelo MBROLA tem uma duração constante, PATRICIA precisa estendê-la enquanto a respectiva tecla estiver pressionada no instrumento MIDI. Para alcançar este resultado, para cada nota MIDI ativa, o arquivo *.AU* correspondente é reproduzido em um *loop*. A posição inicial do *loop* é estabelecida no centro do arquivo no domínio do tempo, onde se encontra a vogal, que é a parte periódica da forma de onda da sílaba. A posição final é calculada adicionando-se à posição inicial o período em segundos, ou seja, o inverso da frequência calculada. 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 a duração da sílaba. Quando uma mensagem MIDI *Note off* é recebida, o *loop* que corresponde a esta nota é liberado, e o resto do arquivo *.AU* é reproduzido.

A Figura 1 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.

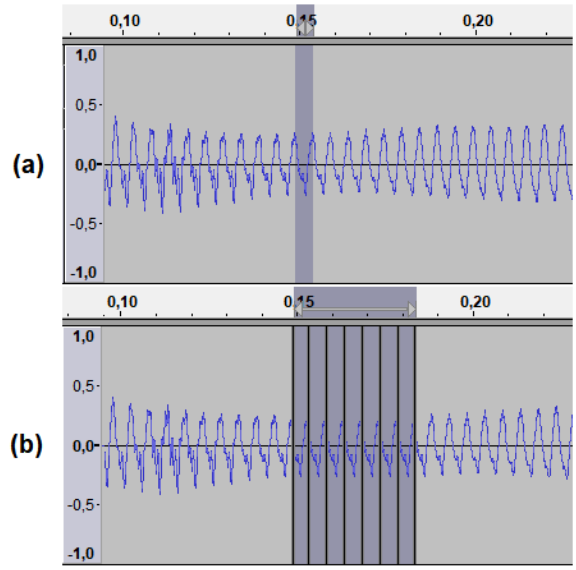

Figura 1. (a) Segmento de *loop* calculado por PATRICIA. (b) Repetição do segmento, estendendo a duração da amostra.

As diferenças entre as taxas de amostragem do MBROLA e do SuperCollider foram levadas em consideração para que o áudio fosse reproduzido corretamente. A repetição sucessiva do processo descrito acima resulta na geração de uma voz cantada sintetizada na saída de áudio do computador.

4. Novas funcionalidades

A descrição de PATRICIA dada na Seção 3 corresponde, em linhas gerais, à versão do sistema apresentada em artigo anterior [Brum et al. 2023]. Tal versão contava apenas com controles de altura e duração, aos quais foram acrescentados os controles de intensidade e timbre, além de um mecanismo de alternância entre as letras de canção para síntese. Essas novas funcionalidades serão apresentadas ao longo desta seção.

4.1. Controle de intensidade

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 se tocar o instrumento, 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

[MIDI-Manufacturers-Association et al. 1996] estabelecem a seguinte fórmula para mapear uma velocidade *V*, que varia de 1 a 127, em uma intensidade *L*, medida em decibéis:

$$L = 40 \cdot log(V/127) \tag{2}$$

A fórmula apresentada pela Equação 2 foi implementada no sistema, 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.

4.2. Controle do timbre e seleção de letras

A qualidade do timbre permite ao ouvinte 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 ao ouvinte a identificação de 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. Uma vez que tal 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.

Adicionalmente, é possível alternar entre diferentes arquivos de texto a fim de que o sintetizador execute canções diversas. A seleção dos arquivos fonéticos se dá por meio da mensagem MIDI *Program Change*, sendo possível disponibilizar até 128 letras de canções diferentes para o sistema.

O diagrama da Figura 2 fornece uma visão geral da arquitetura do sistema PATRICIA, desde a recepção e processamento das entradas musical e fonética até a saída de áudio com o canto sintetizado, passando pela interação com o sintetizador de fala MBROLA.

5. Conclusão e trabalhos futuros

O sistema aqui apresentado, PATRICIA, é um sintetizador de canto em tempo real, sendo o primeiro a ser especificamente projetado para gerar canto no idioma português brasileiro. A pesquisa está em andamento e a atual versão do sistema ainda está em fase de prova de conceito, com um conjunto limitado de requisitos funcionais implementados na linguagem e ambiente SuperCollider. Essas funcionalidades foram testadas e alguns dos testes conduzidos podem ser vistos no canal do projeto no YouTube [1]. Além disso, o sistema foi validado por meio de uma avaliação subjetiva feita por cinco educadores

[1] https://www.youtube.com/@ProjetoPATRICIA

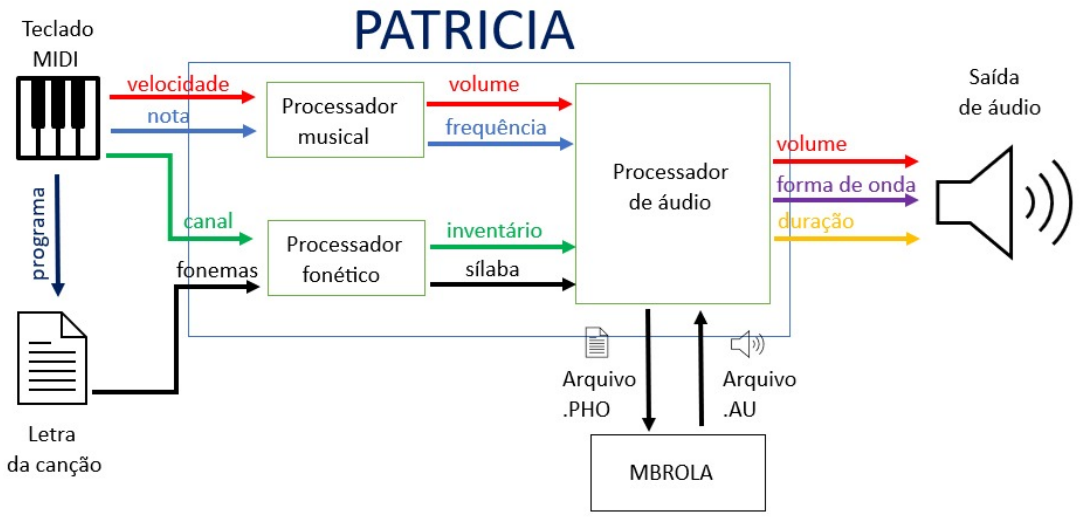

Figura 2. Arquitetura do sistema PATRICIA

musicais atuantes no estado de Sergipe e de uma análise de desempenho cujos resultados encontram-se disponíveis [Brum 2023].

Síntese concatenativa baseada em *samples* e letras de canção fornecidas antecipadamente em um arquivo de texto foram, respectivamente, a abordagem técnica e o método de entrada fonética escolhidos para o sistema. A entrada musical é dada em tempo real via MIDI, e a síntese é realizada pelo MBROLA, um sintetizador de fala adaptado. Esta abordagem mostrou-se mais viável que a dirigida a dados pela ausência de bases de canto em português do Brasil que servissem para a aplicação de algoritimos de treinamento.

Entre as modificações previstas para as futuras versões de PATRICIA, encontram-se as seguintes:

- Criação de um inventário de vozes próprio, com maior qualidade de áudio e com amostras de voz de uma cantora profissional em lugar de *samples* falados, visando melhorar a naturalidade e inteligibilidade do canto sintetizado, dispensando-se o uso dos inventários do MBROLA e estabelecendo-se uma base para treinamento.
- Incorporação do algoritmo de síntese de voz à implementação de PATRICIA, que será outro fator de eliminação da dependência em relação ao MBROLA. Além disso, a aplicação de métodos e técnicas de inteligência artificial ao algoritmo ajudará o sistema a se ajustar ao atual estado da arte.
- Integração do sintetizador a um teclado musical eletrônico, de modo que ele funcione como sistema embarcado para o instrumento. Na versão atual, o teclado atua como um periférico de entrada externo ao sintetizador.

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.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Alivizatou-Barakou et al. (2017). *Intangible Cultural Heritage and New Technologies: Challenges and Opportunities for Cultural Preservation and Development*, pages 129–158. Springer International Publishing, Cham.

Brum, L. A. Z. (2023). Patricia: um sintetizador de canto em tempo real para o português brasileiro. Master's thesis, Universidade Federal de Sergipe.

Brum, L. A. Z., Meneses, E. A. L., and Moreno, E. D. (2023). Patricia: a real-time singing synthesizer prototype for the brazilian portuguese language. In *Proceedings of the International Computer Music Conference 2023*, page 176 – 180.

Brum, L. A. Z. and Moreno, E. D. (2019). State of art of real-time singing voice synthesis. In *Anais do XVII Simpósio Brasileiro de Computação Musical*, pages 50–57, Porto Alegre, RS, Brasil. SBC.

Brum, L. A. Z. and Moreno, E. D. (2020). Challenges and perspectives on real-time singing voice synthesis. *Revista de Informática Teórica e Aplicada*, 27(4):118–126.

Chan, P. Y., Dong, M., Ho, G. X. H., and Li, H. (2016). SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms. In *Proc. Interspeech 2016*, pages 1225–1229.

Delalez, S. and d'Alessandro, C. (2017). Adjusting the Frame: Biphasic Performative Control of Speech Rhythm. In *Proceedings of Interspeech 2017*, pages 864–868, Stockholm, Sweden.

Dong, M., Lee, S. W., Li, H., Chan, P., Peng, X., Ehnes, J. W., and Huang, D. (2014). I2r speech2singing perfects everyone's singing. In *Proc. Interspeech 2014*, pages 2148–2149.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (1996). 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*, volume 3, pages 1393–1396 vol.3.

Feugère, L., d'Alessandro, C., Doval, B., and Perrotin, O. (2017). Cantor digitalis: chironomic parametric synthesis of singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):1–19.

Kagami, S., Hamano, K., Kashiwaze, K., and Yamamoto, K. (2012). Development of realtime japanese vocal keyboard. *Information Processing Society of Japan INTERACTION*, pages 837–842.

Kashiwase, K. (2017). An over-the-shoulder keyboard that extends the potential for vocaloid performance. *Yamaha Corporation*. Accessed: 2023-01-29.

Kenmochi, H. and Ohshita, H. (2007). VOCALOID - commercial singing synthesizer based on sample concatenation. In *Proc. Interspeech 2007*, pages 4009–4010.

Kim, Y. E. (2008). *Singing Voice Analysis, Synthesis, and Modeling*, pages 359–374. Springer New York, New York, NY.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kubozono, H. (1989). The mora and syllable structure in japanese: Evidence from speech errors. *Language and Speech*, 32(3):249–278.

Le Beux, S., Feugère, L., and d'Alessandro, C. (2011). Chorus digitalis: experiment in chironomic choir singing. In *Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 2005–2008. ISCA.

Locqueville, G., d'Alessandro, C., Delalez, S., Doval, B., and Xiao, X. (2020). Voks: Digital instruments for chironomic control of voice samples. *Speech Communication*, 125:97–113.

Matsubara, K., Okamoto, T., Takashima, R., Takiguchi, T., Toda, T., Shiga, Y., and Kawai, H. (2021). Full-band lpcnet: A real-time neural vocoder for 48 khz audio with a cpu. *IEEE Access*, 9:94923–94933.

MIDI-Manufacturers-Association et al. (1996). The complete midi 1.0 detailed specification. *Los Angeles, CA, The MIDI Manufacturers Association*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLOS Medicine*, 6(7):1–6.

Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., and Tokuda, K. (2010). Recent development of the hmm-based singing voice synthesis system—sinsy. In *Seventh ISCA Workshop on Speech Synthesis*.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, EASE'08, page 68–77, Swindon, GBR. BCS Learning Development Ltd.

Rodet, X., Potard, Y., and Barriere, J.-B. (1984). The chant project: From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31.

Tae, J., Kim, H., and Lee, Y. (2021). Mlp singer: Towards rapid parallel korean singing voice synthesis. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Tan, X. (2023). *Beyond Text-to-Speech Synthesis*, pages 175–179. Springer Nature Singapore, Singapore.

Veaux, C., Astrinaki, M., Oura, K., Clark, R. A. J., and Yamagishi, J. (2013). Gesture control of hmm-based singing voice synthesis. In *Proc. 8th ISCA Workshop on Speech Synthesis (SSW 8)*, pages 247–248.

Wells, J. C. et al. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4:684–732.