REVISTA DA **ABRALIN**

REVISÃO DE LITERATURA

Prosódia e síntese da fala: uma revisão integrativa da literatura

Julio Cesar GALDINO
Universidade Federal de Alagoas (UFAL)

Miguel OLIVEIRA JR.
Universidade Federal de Alagoas (UFAL)

OPEN ACCESS

EDITADO POR
- Raquel Freitag (UFS)

AVALIADO POR
- Camila Leite (UFU)
- Sandra Madureira (PUC-SP)

SOBRE OS AUTORES
- Julio Cesar Galdino
  Investigação, Metodologia, Escrita – rascunho original, Escrita – análise e edição.
- Miguel Oliveira Jr.
  Conceptualização, Administração do Projeto e Supervisão, Escrita – análise e edição.

DATAS
- Recebido: 13/12/2022
- Aceito: 06/04/2023
- Publicado: 22/05/2022

COMO CITAR
Galdino, J. C.; Oliveira Jr., M. (2023). Prosódia e síntese da fala: uma revisão integrativa da literatura. *Revista da Abralin*, v. 22, n. 1, p. 1-15, 2023.

RESUMO

Este é um trabalho de revisão integrativa acerca de estudos produzidos por pesquisadores no Brasil, a partir das relações entre a prosódia e a síntese de fala. A partir da pergunta de pesquisa “Como a prosódia tem sido considerada em trabalhos que visam o aprimoramento da síntese de fala?”, realizamos uma busca no *Google Scholar* com a sintaxe (prosódia OR entoação OR “frequência fundamental”) AND ("text-to-speech" OR TTS OR "síntese de fala" OR “síntese da fala”). Avaliamos os títulos e os resumos dos estudos e, mediante a observação de critérios de inclusão e de exclusão, encontramos 10 estudos, entre 2010 e 2021, que dissertam sobre prosódia e síntese de fala. Os trabalhos selecionados indicam que a frequência fundamental (ou *pitch*) é o recurso mais expressivo para o aprimoramento da fala sintética, embora os sistemas de conversão de texto para a fala utilizem outras características prosódicas para aprimorar seu desempenho. Além disso, os resultados desta revisão mostraram que há ainda pouco estudo no Brasil sobre a relação entre a prosódia e a síntese de fala e que é importante a pesquisa conjunta entre pesquisadores de áreas da linguística e das engenharias, a fim de se obter melhores resultados em sistemas de síntese de fala.

ABSTRACT

This article aims to present an integrative review of studies produced by researchers in Brazil, based on the relationship between prosody and speech synthesis. To achieve this objective, we elaborated the research

question “Which prosodic characteristics are most involved in the improvement of speech synthesis?” and we performed a search on Google Scholar, based on the syntax (prosódia OR entoação OR “frequência fundamental”) AND ("text-to-speech" OR TTS OR "síntese de fala" OR "síntese da fala"). We included 10 studies between 2010 and 2021, which showed that fundamental frequency and pitch are the most expressive features, although text-to-speech systems use other prosodic features to generate synthetic voice intonation or to improve their performance. Furthermore, the results of this review showed that there are still few studies in Brazil on the relationship between prosody and speech synthesis and that joint research between researchers in the fields of linguistics and engineering is important in order to improve the speech synthesis.

PALAVRAS-CHAVE
Síntese de fala. Prosódia. Frequência fundamental.

KEYWORDS
Speech Synthesis. Prosody. Fundamental Frequency.

RESUMO PARA NÃO ESPECIALISTAS
Este é um trabalho que resume e avalia um conjunto de estudos científicos sobre as características entoacionais de vozes produzidas por máquinas. Para atingir esse objetivo, nós realizamos uma busca online e encontramos 10 trabalhos que abordaram esse tema. Verificamos, a partir desta revisão, que existem determinadas características relacionadas à voz humana que podem resultar em uma fala mais natural nessas máquinas. No entanto, ainda há poucas pesquisas no Brasil que investigam essa relação. Este estudo é importante, pois promove uma melhor compreensão dos avanços que estão sendo atingidos na área da tecnologia da fala, uma ferramenta que está, cada vez mais, presente no nosso cotidiano.

Introdução

A síntese de fala é a produção de voz por máquinas, a partir da fonetização automática de frases (DUTOIT, 1997). Ao contrário da simples reprodução de voz, essa síntese objetiva um resultado equivalente à produção da fala humana, com informações fonéticas e prosódicas correspondentes (SAGISAKA, 1990).

A síntese de fala se divide em duas classes e são diferenciadas a partir do tamanho do vocabulário e do campo de aplicação, conforme Egashira (1992). Segundo o autor, na primeira categoria, estão os sistemas de resposta vocal, usados em serviços telefônicos, sistemas de saldo bancário, por exemplo, com frases introdutórias, como "bom dia", "digite sua senha", em que há pouca interação com o usuário. Nesses casos, o vocabulário é limitado, e sua realização é resultado de gravação e armazenamento de fala, a fim de gerar possibilidades de combinações para uma posterior reprodução.

A segunda categoria dos sistemas de síntese de fala são os chamados conversores de texto em fala (TTS – *Text-To-Speech*). Eles contêm uma gama enorme de aplicações, facilitando a interação humano-computador para cegos, lendo notícias, boletins meteorológicos e, principalmente, atuando na automação de *call center* (TAYLOR, 2009). Assim, essa classe de sistema possui um vocabulário irrestrito, tem um custo computacional mais elevado e precisa fazer análises do texto escrito, identificação dos sons equivalentes, associações dos parâmetros de entoação e ritmo e processamento de sinais, o que não traz, muitas vezes, a naturalidade que a fala humana possui (PACHECO, 2010).

Os modernos sistemas de TTS possuem diversas arquiteturas, mas existem pelo menos três blocos que são comuns: *front-end* (pré-processamento de texto), *back-end* (motor de síntese) e *voice font* (base de dados de voz) (BRAGA, 2007). Um resumo dessa arquitetura é mostrado na Figura 1.

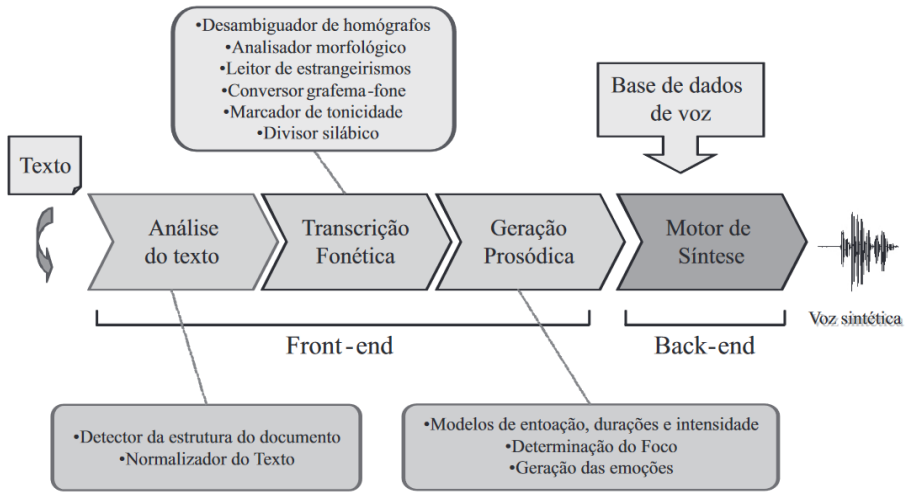

FIGURA 1 – Arquitetura tradicional de um sistema de conversão texto-fala
Fonte: Braga (2007).

As etapas de análise do texto e de transcrição fonética do *front-end* não apresentam grandes dificuldades. Entretanto, a última etapa, a geração prosódica, ainda apresenta desafios, pois os sistemas de síntese carecem de informações prosódicas mais acuradas, resultando em uma artificialidade da voz sintética (SILVA, 2004). Isso significa que é necessário investir no melhoramento da anotação automática de prosódia, para que essa fala sintética se aproxime da naturalidade da fala humana (KLIMKOV *et al.*, 2017).

A prosódia é a organização de várias unidades linguísticas em um ou mais enunciados no processo de produção da fala e sua realização contém características segmentais e suprassegmentais, com o objetivo de transmitir informações linguísticas, paralinguísticas ou não linguísticas (FUJISAKI, 1997). Muitos sistemas TTS predizem representações prosódicas diretamente do texto, mas há risco de o processo de análise cometer erros, suscitando o desafio de gerar conteúdo prosódico, porque o texto codifica principalmente o componente verbal, ignorando a prosódia (TAYLOR, 2009).

Essa preocupação com a prosódia é uma constante em estudos sobre TTS. Trabalhos iniciais sobre o português brasileiro, por exemplo, descrevem os procedimentos desses sistemas, abordando, também, diferentes conceitos relativos à produção da fala, como fones, duração, ritmo, frequência fundamental, intensidade etc. (EGASHIRA, 1992; CHBANE, 1994; MADUREIRA *et al.*, 1995; SILVA & VIOLARO, 1995; OLIVEIRA, 1996; AQUINO, 1998; GOMES, 1998; BARBOSA, 1999; BARBOSA *et al.*, 1999; SIMÕES, 1999). Esses sistemas são investigados, em sua maioria, por profissionais da Engenharia da Computação. No entanto, Simões *et al.* (2000) propõem um sistema TTS para o português brasileiro com colaboração de linguistas, o que evidencia a necessidade de formação interdisciplinar de linguistas neste campo de investigação (BRAGA, 2007). Pelo exposto, a Linguística é uma área fundamental para a geração da fala sintética, uma vez que a descrição das línguas, especificamente no nível prosódico, fornece informações que podem aperfeiçoar a naturalidade da fala sintética.

Levando-se em conta a importante relação entre prosódia e síntese de fala, este artigo busca dissertar sobre como a prosódia tem sido considerada para o aprimoramento da síntese de fala. Pacheco (2010) realiza uma revisão de literatura cerca da síntese de fala, a partir de um resgate histórico dos sistemas mecânicos do século XVIII até os atuais sistemas de geração de síntese de fala, além de fazer um detalhamento das aplicações requeridas para uma boa conversão. Entretanto, falta uma discussão acerca do papel da prosódia nestes sistemas. Assim, este trabalho preenche essa lacuna, apresentando-se como uma contribuição a estudos de interface entre as áreas da Linguística e da Computação.

1. Metodologia

A revisão aqui reportada foi construída a partir de quatro passos, seguindo a recomendação PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*): definição da pergunta norteadora, elaboração dos critérios de inclusão e exclusão para a busca da literatura, síntese das informações dos estudos incluídos e apresentação da revisão. Utilizamos os passos indicados pelo PRISMA, porque eles permitem que, a partir da pergunta de pesquisa, se possa decidir quais as palavras-chave que estarão na busca e selecionar, de forma criteriosa, os estudos que serão incluídos na revisão.

Para realização do levantamento bibliográfico sistemático e da discussão da presente pesquisa, a seguinte pergunta norteadora foi elaborada: "Como a prosódia tem sido considerada em trabalhos que visam o aprimoramento da síntese de fala?". Foi realizada uma busca no ano de 2022 no *Google Scholar*, uma vez que é uma base de dados que reúne trabalhos de diversos tipos, além de ser um dos indexadores mais utilizado por periódicos acadêmicos. A busca com os descritores foi sobre

trabalhos do português brasileiro, produzidos no Brasil, usando a seguinte sintaxe: (prosódia OR entoação OR “frequência fundamental”) AND ("text-to-speech" OR TTS OR "síntese de fala" OR “síntese da fala”), sem incluir citações nas bases de dados.

Foram incluídos artigos, monografias, dissertações e teses publicados em português, produzidos no Brasil, realizados nos últimos 11 anos (2010 a 2021) que tratassem do tema “prosódia e síntese de fala”. Foram excluídos livros, resenhas, estudos duplicados e de revisão. Essa seleção dos trabalhos foi feita mediante uma avaliação inicial dos títulos e dos resumos. Posteriormente, houve a leitura na íntegra dos trabalhos e a inclusão dos estudos que de fato tratavam do tema desta pesquisa.

Para a síntese das informações dos trabalhos, foi feita uma adaptação de um instrumento proposto por Ursi (2005), em que se identificam os objetivos, os aspectos metodológicos e os resultados.

2. Resultados

A seleção dos estudos para esta revisão está descrita na Figura 2. O fluxograma mostra uma quantidade de 799 trabalhos, em que 780 deles foram excluídos já na fase de análise do título e do resumo. Após a leitura na íntegra dos 19 estudos baixados, incluímos 10 estudos nesta revisão.

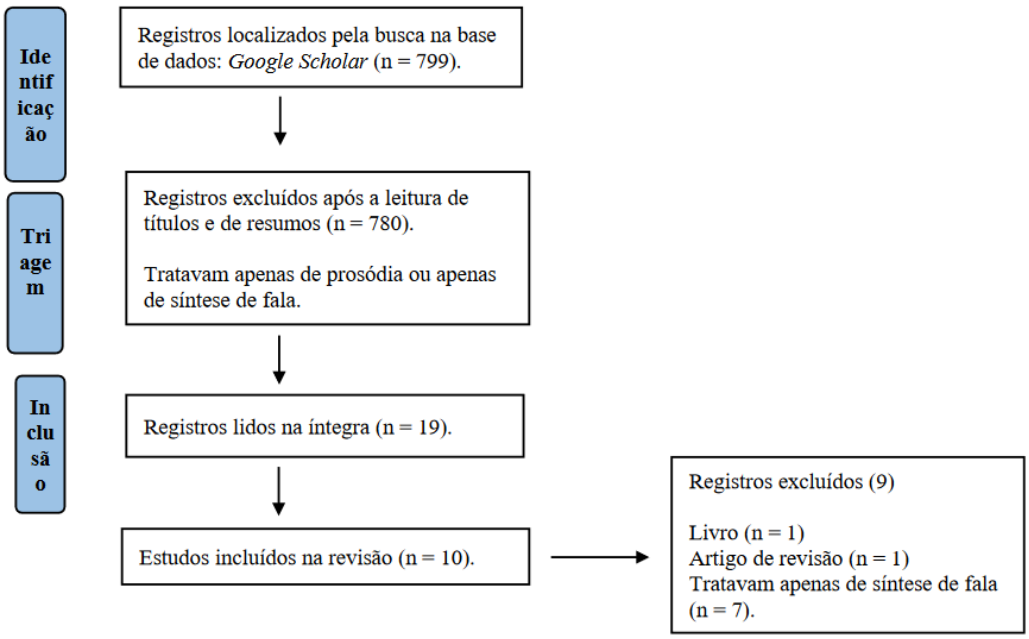

FIGURA 2 – Fluxograma PRISMA para seleção de registros para esta revisão, 2022.
Fonte: elaborada pelos autores.

A seguir, apresentamos, no Quadro 1, a autoria, o ano e o título dos trabalhos incluídos nesta revisão.

Autoria, Ano e Título	Objetivo	Desenho do estudo/ Procedimentos metodológicos	Resultados
Barbosa, 2016. Análise e proposição de modelos de síntese de fala para integração ao framework FIVE	Avaliar um conjunto de mecanismos de síntese de voz, e integrá-los ao Framework FIVE, a fim de obter uma melhor naturalidade e inteligibilidade das vozes geradas para o Português falado no Brasil.	Levantamento sobre os mecanismos de síntese de voz, sobre a arquitetura do Framework FIVE e sobre a plataforma *MaryTTS*; Construção de um conjunto de vozes utilizando a plataforma *MaryTTS* e integradas ao Framework FIVE; Experimentos para avaliação da qualidade das vozes.	Após a avaliação dos mecanismos de síntese de voz, foi possível verificar que a naturalidade e a inteligibilidade das vozes com a técnica de concatenação de unidades são melhores do que a HMMs (*Hidden Markov Models*). Enquanto isso, na perspectiva auditiva acontece o contrário. Além disso, os resultados da perspectiva audiovisual foram melhores do que a perspectiva puramente auditiva.
Latsch, 2011. Desenvolvimento de um sistema de conversão texto-fala com modelagem de prosódia	Apresentar um sistema de apoio à pesquisa e desenvolvimento de um sistema de conversão texto-fala e abordar as etapas, incluindo a modelagem da prosódia.	Parametrização das variáveis prosódicas com base em diferentes atitudes; Descrição de um sistema de conversão de texto-fala, com manipulação prosódica.	O sistema de conversão texto-fala apresentado demonstra uma melhor combinação entre as etapas de alinhamento temporal e mapeamento de *pitch* de um sinal de análise para o sinal de síntese. Em relação à parametrização das variáveis no domínio da sílaba, há uma vantagem, ao oferecer um meio simples de observar e caracterizar as novas atitudes prosódicas. Os resultados demonstram uma descrição das ferramentas de desenvolvimento do sistema de conversão texto-fala, as ferramentas de análise e de síntese da prosódia.
Maia; Seara, 2017. Um sistema TTS baseado em redes neurais profundas usando parâmetros síncronos de pitch	Apresentar formas de usar parâmetros acústicos obtidos de forma síncrona com o *pitch* em sistemas de síntese de fala.	Uso de sentenças na base de dados do projeto FalaBrasil; Implementação da estrutura DNN (*Deep Neural Networks*) com parâmetros síncronos com o *pitch*.	Os resultados experimentais mostraram que o uso de atributos linguísticos obtidos de quadros de tamanhos fixos, juntamente com parâmetros acústicos extraídos de forma síncrona com o *pitch*, produzem melhores resultados em termos de medidas objetivas de qualidade.
Manfio, 2012. Como funcionam alguns fonemas no aplicativo *Balabolka*	Dissertar, à luz de algumas teorias envolvidas com a Sociolinguística, Geografia Linguística e Dialetologia entre outras, sobre ao menos a síntese de voz acerca de um dos vários aplicativos disponíveis: o *Balabolka*.	Descrição sobre características do *Balabolka* e de sua prosódia; Descrição das realizações de fala e de registro; Análise dos fonemas no *Balabolka*.	O aplicativo realiza uma prosódia próxima da fala se comparado a outros softwares de mesma natureza, embora tenha apresente problemas em frases interrogativas. Além disso, o *Balabolka* produz de forma artificial vogais frouxas em ditongos.
Moreira, 2015. Proposta de um *fron-end* em java para sintetizador de voz baseado no MBROLA (*Multi Band Resynthesis OverLap Add*)	Desenvolver um sistema para inclusão digital de deficientes visuais.	Comparação entre frases realizadas por um locutor humano e uma voz sintetizada; Teste de naturalidade da voz, teste de inteligibilidade e teste de usabilidade do software com uma deficiente visual de 40 anos.	Os testes realizados provaram que o resultado sobre as vozes é muito inteligível e causa menos cansaço aos usuários. Essa inteligibilidade também é comprovada nas comparações entre a voz humana e a voz sintética, no domínio do tempo e da frequência, levando-se em conta o depoimento da usuária.

Autoria, Ano e Título	Objetivo	Desenho do estudo/ Procedimentos metodológicos	Resultados
Neto, 2011. Ferramentas e recursos livres para reconhecimento e síntese de voz em português brasileiro	Descrever o desenvolvimento de recursos e ferramentas livres para reconhecimento e síntese de voz em PB (Português Brasileiro) e apresentar um novo método para reavaliar o resultado do reconhecimento baseado em HMMs.	Descrição de recursos para síntese e reconhecimento de voz a partir de um dicionário fonético para o PB; Avaliação de conversores, locutores, sistemas; Avaliação de modelos de linguagem de modelos acústicos de locutores.	Houve melhoria dos recursos para os conversores em PB, em especial para conversão grafema-fone e para sílaba, além de melhoras na utilização de técnicas para adaptação ao locutor para minimização de efeitos negativos entre os dados. Os resultados da avaliação apresentaram uma estratégia inovadora para aprimorar os resultados provenientes de um sistema baseado em HMMs, a partir da extração de frequência fundamental e de outros parâmetros referentes ao espectro da voz e à excitação, como os coeficientes MFCCs (*Mel-frequency cepstral coefficients*).
Reis *et al.*, 2011. Síntese prosódica da fala em português do Brasil	Apresentar um sistema TTS (*text-to-speech*), capaz de reproduzir a fala com nuances de emoção.	Descrição de um modelo prosódico, a partir da identificação de fonemas, sílabas, palavra prosódica, sintagma entoacional; Descrição de modelo emocional para estados neutro, feliz, triste e bravo. Utilização do software MBROLA com dados do br4, banco de dados do Serviço Federal de Processamento de Dados e da UFRJ (Universidade Federal do Rio de Janeiro).	O modelo prosódico é capaz de gerar falas próximas à fala natural, possibilitando a adição de nuances emotivas ao discurso computacional. O modelo prosódico mostrou-se eficaz para sentenças afirmativas simples, isto é, para um único tipo de curva entoacional.
Sá, 2018. Geração de prosódia para o português brasileiro em sistemas *text-to-speech*	Investigar sistemas *text-to-speech* existentes através do estudo de seus algoritmos para síntese de voz e geração de prosódia para diversas línguas, com foco no PB.	Levantamento de sistemas TTS desenvolvidos para o PB; Criação de um módulo de prosódia que permite fazer anotações prosódicas manuais, a partir de um programa para o *front*-end, do programa MBROLA para converter a saída e do INTSINT (*International Transciption System for Intonation*) para análise e síntese de contornos de f0.	Observou-se uma carência de suporte à síntese expressiva. Linguagens computacionais já começaram a ser integradas a sistemas comerciais de TTS, mas há trabalhos escassos para o PB. Em relação aos modelos de anotação entoacional, ainda não há uma solução considerada a mais apropriada para analisar o português brasileiro, mas há uma grande quantidade de trabalhos de contornos melódicos que podem ser adaptados e convertidos em parâmetros para sistemas TTS.
Souza, 2010. Síntese de fala em português brasileiro baseada em modelos ocultos de Markov	Abordar a construção de um algoritmo de determinação da sílaba tônica de palavras, um algoritmo de conversão de grafemas em fonemas, e um algoritmo de separação silábica de palavras transcritas foneticamente.	Descrição das etapas de sistemas de conversão de texto em fala; Apresentação dos fundamentos necessários acerca dos Modelos Ocultos de Markov; Descrição dos detalhes da implementação realizada.	A construção do projeto foi bem-sucedida, com taxa de acerto muito significativa, em relação ao mecanismo integrado de determinação de sílaba tônica de palavras, conversão de grafemas para fonemas e divisão silábica da palavra foneticamente. Com uma sequência de logaritmos, a a f0 foi obtida inicialmente e depois com uma sequência de vetores, a partir de uma escala mel (melodia).

Autoria, Ano e Título	Objetivo	Desenho do estudo/ Procedimentos metodológicos	Resultados
Thomaz, 2012. Modelagem de prosódia para conversores texto-fala	Ampliar a funcionalidade de manipulação prosódica das atitudes do sistema SASPRO (Sistema de Análise e Síntese da Prosódia).	Classificação de atitudes prosódicas com base em estudos linguísticos já investigados; Classificação de atitudes em estruturas silábicas diferentes com base em estudos linguísticos já investigados; Aplicação do modelo e aplicação de teste subjetivo com 20 voluntários.	O modelo de atitudes prosódicas baseia-se em três aspectos do sinal de voz (duração, intensidade e *pitch*), para 14 atitudes. Em relação à avaliação dos resultados do trabalho a partir dos testes com voluntários, 9 das 14 atitudes tiveram uma nota maior que 3, enquanto as outras 5 foram consideradas inaceitáveis no modelo de prosódia proposto no sistema.

QUADRO 1 – Síntese dos trabalhos avaliados.

Fonte: elaborado pelos autores.

3. Discussão

Nesta revisão, encontramos estudos que pertencem a diversas áreas. Eles estão distribuídos na Engenharia Elétrica (n=3), na Engenharia de Computação (n=1), na Engenharia de Teleinformática (n=1), na Engenharia Mecatrônica (n=1), na Engenharia Eletrônica e de Computação (n=1), na Ciência da Computação (n=2) e na área da Linguística (n=1).

Essa distribuição de áreas demonstra que a maioria das pesquisas foi desenvolvida em âmbitos do conhecimento de linhas mais exatas e técnicas. Entretanto, a fim de aprimorar o desempenho de sistemas de conversão de texto, esses trabalhos precisam utilizar informações que permitam entender as características próprias da fala. Por isso, os estudos linguísticos são muito importantes no aprimoramento da fala sintética, pois é a Linguística o campo do conhecimento capaz de fornecer informações que descrevam os fenômenos que o caráter multissistêmico da língua apresenta, sobretudo os aspectos prosódicos.

Esta revisão apresenta 10 estudos que estabelecem a relação entre a prosódia e a síntese de voz. Os objetivos destes trabalhos são diferentes, mas todos se preocupam em aprimorar o desempenho dos sistemas de conversão de texto para a fala, seja para testar um novo modelo, desenvolver um algoritmo ou para aprimorar determinadas características, como a emoção, por exemplo.

Em relação à metodologia utilizada nos estudos, apesar de estarem inseridos em áreas mais exatas, há uma preocupação em usar informações linguísticas que possam tornar a fala sintética mais natural. Thomaz (2012), por exemplo, a fim de aprimorar a sintetização da fala, faz uso da Linguística ao modelar uma voz neutra em 14 atitudes prosódicas, entre elas, ironia, pedido e sugestão, descritas por Moraes (2008). O trabalho alcançou seu objetivo, ao fazer um teste de percepção, em que 9 dessas atitudes modeladas foram bem aceitas pelos juízes, porém, 5 delas não tiveram boa avaliação. Pode-se inferir, a partir disso, que é interessante investigar, em trabalhos futuros, quais informações prosódicas podem ser consideradas para melhorias e qual a diferença entre os contornos entoacionais da fala humana e da fala sintética nessas atitudes, por exemplo.

Reis *et al.* (2011) fazem uso de teoria prosódica para determinar o contorno da fala nos sistemas de conversão. O modelo proposto pelos autores foi eficaz para um tipo apenas de curva entoacional, nas sentenças afirmativas simples, o que demonstra a necessidade de se trabalhar outras naturezas de curva e levar em conta outros parâmetros. Dessa forma, esses exemplos mostram que é importante o trabalho conjunto de linguistas e pesquisadores das engenharias, pois, se esses estudos envolvessem profissionais de ambas as áreas, haveria uma melhora na constituição dos blocos dos sistemas de síntese e no entendimento de características prosódicas que pudessem aprimorar esses sistemas.

Entre as características prosódicas investigadas nos 10 estudos, a frequência fundamental é a mais expressiva, como em Sá (2018), que propõe um modelo para síntese de contornos de f0. Esses parâmetros da fala são gerados e manipulados mediante algoritmos, logaritmos e outros recursos matemáticos e computacionais. Em Souza (2010) e Barbosa (2016), a f0 é obtida inicialmente com uma sequência de logaritmos e depois com uma sequência de vetores, a partir de uma escala de melodia, em que, para cada tom com uma certa frequência em Hz, é associado um valor.

A f0 também pode ser utilizada com extração de uma base de voz, modelada através de Modelos Ocultos de Markov (NETO, 2011). Zen *et al*. (2009) explicam que a síntese baseada nesses modelos unifica os blocos *front-end* e *back-end*, gerando uma nova estrutura, o que se torna uma vantagem para o desempenho geral de um sistema TTS, pois, segundo os autores, com os dois blocos em conjunto, é possível obter eficácia em usar análise de texto e análise acústica em um único bloco.

Entre os estudos revisados, a f0 aparece como equivalente ao *pitch* (MAIA & SEARA, 2017; THOMAZ, 2012; LATSCH, 2011; MOREIRA, 2015). É importante, no entanto, distinguir os dois conceitos, pois a f0 é mensurável e pertence à esfera da produção, enquanto o *pitch* pertence à esfera da percepção (BARBOSA, 2019).

Outras características prosódicas também podem participar do processo de geração da entoação da síntese de fala. Manfio (2012), por exemplo, investiga a síntese de voz em um aplicativo chamado Balabolka, que utiliza partes gravadas da fala humana e possui uma prosódia próxima à da fala real, embora o software não consiga atingir um bom desempenho em frases interrogativas e na produção dos ditongos. O autor ressalta boa realização de características prosódicas no aplicativo, como a delimitação de pausas feitas pelas vírgulas, bem como o resultado satisfatório de volume e tonalidade nas frases que são formadas por enumeração, por exemplo.

Embora haja predominância de certos recursos prosódicos para a geração e o aprimoramento de TTS, para gerar nuances de emoção, por exemplo, o estudo de Reis *et al*. (2011) faz uso de várias características prosódicas, como duração, *pitch*, velocidade e contorno de frequência fundamental. Recentemente, essa tentativa de inserir expressões emocionais para aprimorar a expressividade dos sistemas tem sido constante (INOUE *et al*., 2017; ROBINSON *et al*., 2019; TAHON *et al*., 2018). Esse processo de geração envolve geralmente a conversão de uma voz neutra para uma voz emocional, sendo uma característica que os ouvintes esperam, devido ao contexto em que a frase está sendo emitida (ROBINSON *et al*, 2019).

De forma geral, os estudos descritos nesta revisão mostram que a prosódia tem sido considerada como essencial para o desenvolvimento da síntese de fala, a partir de informações linguísticas aliadas ao campo das áreas das engenharias. A presença da Linguística nesses trabalhos demonstra que ela é uma área importante, uma vez que contribui para uma voz sintética mais expressiva e aceita pelos usuários, especificamente nos níveis prosódicos.

Além disso, os resultados desta revisão mostraram que há ainda pouco estudo no Brasil sobre a relação entre a prosódia e a síntese de fala e que é importante a pesquisa conjunta entre pesquisadores de áreas da linguística e das engenharias, a fim de se obter melhores resultados em sistemas de síntese de fala.

4. Considerações finais

Em nossa revisão, observamos que há 10 estudos que estabelecem relação entre prosódia e síntese de fala, encontrados na base de dados *Google Scholar*, apenas em português, o que demonstra que há escassez de pesquisas realizadas no Brasil. São trabalhos que estão inseridos em áreas mais exatas, porém, também fazem uso de aporte teórico da Linguística, a fim de aprimorar o desempenho de sistemas de conversão de texto para a fala. De forma geral, os sistemas utilizam várias características prosódicas para esse aprimoramento, mas a frequência fundamental (*pitch*) é o recurso mais expressivo.

Estudos em outras línguas, como o inglês, o chinês, o tailandês, entre outros, contêm uma variedade maior de análises entre prosódia e fala sintética, com o objetivo principal de propor modelos para geração de contornos de frequência fundamental (KAMEOKA *et al.*, 2015; KORIYAMA & KOBAYASHI, 2015; THOMAS *et al.*, 2015; MOUNGSRI *et al.*, 2017; CHEN *et al.*, 2018; RAO, 2017). Em línguas como o mandarim, a F0 baseia-se em "tons" lexicais que diferem em significado, havendo necessidade de um bom resultado dos padrões entoacionais (CHEN *et. al.*, 2018). Assim, definir esses contornos é importante, pois uma prosódia que se distancia das características da fala natural pode prejudicar a inteligibilidade dos sistemas.

Além dos sistemas TTS, outro tipo de aplicação tecnológica que se ampara de informações prosódicas da fala são os sistemas ASR (*Automatic Speech Recognition*), em que, ao contrário da síntese de fala, a voz é inserida no sistema e é convertida em texto. A prosódia pode ajudar a melhorar esses sistemas, uma vez que recursos, como a pausa, a intensidade, o *pitch* e a frequência fundamental, podem possibilitar o reconhecimento de voz por meio de redes neurais (BALLESTEROS & WANNER, 2016; LIU, LIU & SONG, 2018; SZASZÁK & TÜNDIK, 2019; YI & TAO, 2019). A inteligência artificial já apresenta bons resultados no reconhecimento dessa fala, mas há escassez de trabalhos em língua portuguesa (TEIXEIRA *et al.*, 2016). Assim como o reconhecimento de fala, os resultados desta revisão mostraram que há ainda pouco estudo no Brasil sobre a relação entre a prosódia e a síntese de fala.

Informações complementares

Avaliação e resposta dos autores

Avaliação: https://doi.org/10.25189/rabralin.v22i1.2130.R

Conflito de Interesse

Os autores não têm conflitos de interesse a declarar.

REFERÊNCIAS

AQUINO, P. A. O papel das vogais reduzidas pós-tônicas na construção de um sistema de síntese concatenativa para o português do Brasil. *Revista Sínteses do Instituto de Estudos da Linguagem IEL*, Unicamp, Campinas, v. 3, p. 9-18, 1998. Disponível em: https://revistas.iel.unicamp.br/index.php/sinteses/article/view/6078. Acesso em: 07 abr. 2023.

BALLESTEROS, M.; WANNER, L. A Neural Network Architecture for Multilingual Punctuation Generation. *Association for Computational Linguistics*, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, p. 1048-1053, 2016. Disponível em: http://dx.doi.org/10.18653/v1/D16-1111. Acesso em: 8 fev. 2022.

BARBOSA, D. S. *Análise e proposição de modelos de síntese de fala para integração ao framework FIVE.* Dissertação (Mestrado em Engenharia de Computação), Universidade de Pernambuco, Recife, 2016.

BARBOSA, P. A. Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala. *In: Estudos de prosódia.* SCARPA, E. M. (org.). Campinas, SP: Editora da Unicamp, 1999.

BARBOSA, P. A. *Prosódia*. São Paulo: Parábola, 2019.

BARBOSA, P. A.; VIOLARO, F.; ALBANO, E. C.; SIMÕES, F.; AQUINO, P. A.; MADUREIRA, S.; FRANÇOSO, E. Aiuruetê: a high-qulity concatenative text-to-speech system for brazilian portuguese with demisyllabic analysis-based units and a hierarchical model of rhythm production. *Eurospeech*, Budapeste. Proceedings do Eurospeech'99, 1999. v. 5. p. 2059-2062. Disponível em: https://www.semanticscholar.org/paper/Aiuruete%3A-a-high-quality-concatenative-system-for-a-Barbosa-Violaro/6fe3d550425ba35042b41c59b79c11f35dd59e3d. Acesso em: 07 abr. 2023.

BRAGA, D. Máquinas falantes: Novos paradigmas da língua e da linguística. *Colóquio Política Linguística*, 2007. Disponível em: http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/ColoquioPoliticaLinguistica_2007.pdf. Acesso em: 07 ago. 2022.

CHBANE, D. T. *Desenvolvimento de sistema para conversão de textos em fonemas no idioma português.* Dissertação (Mestrado em Engenharia). Universidade de São Paulo, Escola Politécnica, São Paulo, 1994.

CHEN, J.; YANG, H.; WU, X.; MOORE, B. C.J. The effect of F0 contour on the intelligibility of speech in the presence of interfering sounds for Mandarin Chinese. *The Journal of the Acoustical Society of America*, v. 143, n. 2, p. 864-877, 2018. Disponível em: https://doi.org/10.1121/1.5023218. Acesso em: 22 maio 2022.

DUTOIT, T. *An introduction to text-to-speech synthesis*. Kluwer Academic Publishers, 1997.

EGASHIRA, F. *Síntese de voz a partir de texto para a língua portuguesa.* Dissertação (Mestrado em Engenharia Elétrica). Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica, Campinas, 1992.

FUJISAKI, H. Prosody, models, and spontaneous speech. IN: SAGISAKA, Y.; CAMPBELL, N.; HIGUCHI, N. (edits). *Computing Prosody*: Computational Models for Processing Spontaneous Speech. New York, Springer, 1997.

GOMES, L. C. T. *Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras*. Dissertação (Mestrado em Engenharia Elétrica). Unicamp, Faculdade de Engenharia Elétrica e de Computação, 1998.

INOUE, K.; HARA, S.; ABE, M.; HOJO, N.; IJIMA, Y. An investigation to transplant emotional expressions in DNN-based TTS synthesis. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC), 2017, p. 1253-1258, Disponível em: 10.1109/APSIPA.2017.8282231. Acesso em: 23 abr. 2022.

KAMEOKA, H.; YOSHIZATO, K.; ISHIHARA, T.; KADOWAKI, K.; OHISHI, Y.; KASHINO, K. Generative Modeling of Voice Fundamental Frequency Contours. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 23, n. 6, p. 1042-1053, jun. 2015. Disponível em: 10.1109/TASLP.2015.2418576. Acesso em: 18 jul. 2022.

KLIMKOV, V.; NADOLSKI, A.; MOINET, A.; PUTRYCZ, B.; BARRA-CHICOTE, R.; MERRITT, T.; DRUGMAN, T. Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information. *Proc. Interspeech*, p. 1064-1068, 2017. Disponível em: http://dx.doi.org/10.21437/Interspeech.2017-419. Acesso em: 25 maio 2022.

KORIYAMA, T.; KOBAYASHI, T. Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), p. 4929-4933, 2015. Disponível em: https://doi.org/10.1109/ICASSP.2015.7178908. Acesso em: 18 jul. 2022.

LATSCH, V. L. *Desenvolvimento de um sistema de conversão texto-fala com modelagem de prosódia***.** Tese (Doutorado em Engenharia Elétrica). Universidade Federal do Rio de Janeiro, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Rio de Janeiro, 2011.

LIU, X.; LIU, Y.; SONG, X. Investigating for Punctuation Prediction in Chinese Speech Transcriptions. 2018 *International Conference on Asian Language Processing (IALP), IEEE*, p. 74-78, 2018. Disponível em: https://doi.org/10.1109/IALP.2018.8629143. Acesso em: 25 maio 2022.

MADUREIRA, S.; SILVA, C. H.; AQUINO, P. Pitch Patterns and Duration: Analysis and Synthesis. *XIII International Congress of Phonetic Sciences*, Estocolmo. Proceedings of the XIII International Congress of Phonetic Sciences. Stockholm, v. 3. p. 406-410, 1995. Disponível em: https://www.coli.uni-saarland.de/groups/BM/phonetics/icphs/ICPhS1995/13_ICPhS_1995_Vol_2/p13.2_406.pdf. Acesso em: 07 abr. 2023.

MAIA, R.; SEARA, R. Um sistema TTS baseado em redes neurais profundas usando parâmetros síncronos de pitch. *XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais* – São Pedro, SP, 3-6 de set., 2017. Disponível em: https://www.sbrt.org.br/sbrt2017/anais/1570361943.pdf Acesso em: 16 jul. 2022.

MANFIO, E. R. Como funcionam alguns fonemas no aplicativo Balabolka. *Revista de Linguística e Teoria Literária*, *Via Litterae*, Anápolis, v. 4, n. 2, p. 191-204, jul./dez. 2012. Disponível em: www2.unucseh.ueg.br/vialitterae. Acesso em: 08 ago. 2022.

MORAES, J. A. "The Pitch Accents in Brazilian Portuguese: analysis by synthesis". *Proceedings of the Fourth Conference on Speech Prosody*, p. 389– 398, maio, 2008. Disponível em: https://www.isca-speech.org/archive_v0/sp2008/papers/sp08_389.pdf. Acesso em: 08 ago. 2022.

MOREIRA, N. A. M. *Proposta de um front-end em java para sintetizador de voz baseado no* MBROLA. Dissertação (Engenharia de Teleinformática). Universidade Federal do Ceará, Centro de Tecnologia, Departamento de Engenharia de Teleinformática, Fortaleza, 2015.

MOUNGSRI. D.; KORIYAMA, T.; KOBAYASHI, T. Enhanced F0 generation for GPR-based speech synthesis considering syllable-based prosodic features. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC), 2017, p. 1524-1527, Disponível em: https://doi.org/10.1109/APSIPA.2017.8282285. Acesso em: 18 jul. 2022.

NETO, C. S. N. *Ferramentas e recursos livres para reconhecimento e síntese de voz em português brasileiro*. Tese (Doutorado em Engenharia Elétrica com ênfase em Computação Aplicada). Universidade Federal do Pará, Instituto de Tecnologia, Belém, 2011.

OLIVEIRA, L. M. V. V. C. *Síntese de fala a partir de texto*. Dissertação (Mestrado em Engenharia Electrotécnica e de Computadores). Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa, 1996.

PACHECO, F. S. Artigo de Revisão: Sistemas de Síntese de Fala. *Revista Ilha Digital*, ISSN 2177-2649, v. 2, p. 3-17, 2010. Disponível em: http://ilhadigital.florianopolis.ifsc.edu.br/index.php/ilhadigital/article/view/17. Acesso em: 07 ago. 2022.

RAO, M.V. A.; GHOSH, P. K. Pitch prediction from Mel-generalized cepstrum – a computationally efficient pitch modeling approach for speech synthesis. *2017 25th European Signal Processing Conference (EUSIPCO)*, p. 1629-1633, 2017. Disponível em: https://doi.org/10.23919/EUSIPCO.2017.8081485. Acesso em: 22 maio 2022.

REIS, B. F.; MARTINS, V. V.; PEREIRA-BARRETTO, M. R.; MOSCATO, L. A. Síntese prosódica da fala em português do Brasil. *XSABAI – Simpósio Brasileiro de Automação Inteligente*, X, 2011. São João del-Rei, Minas Gerais, p. 1185-1188, 2011. Disponível em: https://fei.edu.br/sbai/SBAI2011/86262.pdf. Acesso em: 18 jul. 2022.

ROBINSON, C.; OBIN, N.; ROEBEL, A. Sequence-to-sequence Modelling of F0 for Speech Emotion Conversion. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6830-6834, 2019. Disponível em: https://doi.org/10.1109/ICASSP.2019.8683865. Acesso em: 18 jul. 2022.

SÁ, F. C. *Geração de prosódia para o português brasileiro em sistemas text-to-speech.* Monografia (Bacharelado em Ciência da Computação). Universidade Federal do Rio Grande do Norte, Natal, 2018.

SAGISAKA, Y. Speech synthesis from text. *IEEE Communications Magazine*, v. 28 n. 1, 35–41, 1990. Disponível em: https://doi.org/10.1109/35.46669. Acesso em: 07 ago. 2022.

SILVA, C. H.; VIOLARO, F. Modelamento prosódico para conversão texto-fala do português falado no Brasil. *Revista Brasileira de Telecomunicações*, v. 10, n. 1, 1995. Disponível em: https://jcis.sbrt.org.br/jcis/article/view/179/93. Acesso em: 08 ago. 2022.

SILVA, S. Z. *Um estudo de modelos básicos de prosódia para o Português Brasileiro*. Tese (Mestrado em Engenharia Elétrica), Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, 2004.

SIMÕES, F. O. *Implementação de um sistema de conversão texto-fala para o português do Brasil.* Dissertação (Mestrado em Engenharia Elétrica), Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, 1999.

SIMÕES, F. O.; VIOLARO, F.; BARBOSA, P. A.; ALBANO, E. C. Um sistema de conversão texto-fala para o português falado no Brasil. *Journal of Communication and Information Systems*, v. 15, n. 2, 2000. Disponível em: http://dx.doi.org/10.14209/jcis.2000.8. Acesso em: 08 ago. 2022.

SOUZA, C. F. S. *Síntese de fala em português brasileiro baseada em modelos ocultos de Markov*. Dissertação (Mestrado em Ciência da Computação). Universidade Federal de Pernambuco, Centro de Informática, Recife, 2010.

SZASZÁK, G., TÜNDIK, M. Á. Leveraging a character, word and prosody triplet for an ASR error robust and agglutination friendly punctuation approach. *Proc. Interspeech*, p. 2988-2992, 2019. Disponível em: http://dx.doi.org/10.21437/Interspeech.2019-2132. Acesso em: 25 maio 2022.

TAHON, M.; LECORVÉ, G.; LOLIVE, D. Can We Generate Emotional Pronunciations for Expressive Speech Synthesis? *IEEE Transactions on Affective Computing*, v. 11, n. 4, p. 684-695, 1 Oct.-Dec. 2020. Disponível em: https://doi.org/10.1109/TAFFC.2018.2828429. Acesso em: 18 jul. 2022.

TAYLOR, P. *Text–to–Speech Synthesis*. Cambridge University Press, 2009.

TEIXEIRA, A. H. K.; SANTOS, I. M. M.; MOTA, J. S.; GOMES DE SOUZA, J. Tecnologias de reconhecimento de fala: uma revisão sistemática de trabalhos no Brasil**.** XX *Encoinfo – Congresso de Computação e Tecnologias da Informação*, p. 160-167, 2016. Disponível em: http://ulbra-to.br/encoinfo/wp-content/uploads/2020/03/Tecnologias-de-Reconhecimento-de-Fala-uma-revis%C3%A3o-sistem%C3%A1tica-de-trabalhos-no-Brasil.pdf. Acesso em: 07 ago. 2022.

THOMAS, C.; GOKUL, P.; THOMAS, N.; GOPINATH, D. P. Synthesizing intonation for Malayalam TTS. *International Conference on Control Communication & Computing India* (ICCC)**,** 2015, p. 522-527, Disponível em: https://doi.org/10.1109/ICCC.2015.7432949. Acesso em: 22 maio 2022.

THOMAZ, L. A. *Modelagem de prosódia para conversores texto-fala*. Monografia (Graduação em Eletrônica e Computação). Universidade Federal do Rio de Janeiro, Escola Politécnica, Departamento de Eletrônica e de Computação, Centro de Tecnologia, Rio de Janeiro, 2012.

URSI, E. S. *Prevenção de lesões de pele no perioperatório: revisão integrativa da literatura*. Dissertação (Mestrado em Enfermagem). Universidade de São Paulo, Ribeirão Preto, 2005.

YI, J.; TAO, J. Self-attention Based Model for Punctuation Prediction Using Word and Speech Embeddings. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Brighton, UK, p. 7270-7274, 2019. Disponível em: https://doi.org/10.1109/ICASSP.2019.8682260. Acesso em: 08 fev. 2022.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. *Speech Communication*, Elsevier, v. 51, n. 11, p. 1039-1064, 2009. Disponível em: https://doi.org/10.1016/j.specom.2009.04.004. Acesso em: 16 ago. 2022.