CADERNOS
de_LINGUÍSTICA

cadernos.abralin.org

REVISÃO DE LITERATURA

CARACTERÍSTICAS PROSÓDICAS ASSOCIADAS AOS SINAIS DE PONTUAÇÃO: **UMA REVISÃO DE ESCOPO**

Julio Cesar GALDINO
Universidade Federal de Alagoas (UFAL)

Kyvia Fernanda Tenório da SILVA
Universidade Federal de Alagoas (UFAL)

Miguel OLIVEIRA JR.
Universidade Federal de Alagoas (UFAL)

OPEN ACCESS

EDITORES
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFS)

AVALIADORES
– Marcus Martins (USP)
– Vera Pacheco (UESB)
– Flaviane Svartman (USP)

SOBRE OS AUTORES
– Julio Cesar Galdino
Investigação, Escrita – Rascunho Original E Escrita – Análise e Edição.
– Kyvia Fernanda Tenório da Silva
Investigação, Metodologia, Escrita – Rascunho Original, Escrita – Análise e Edição.
– Miguel José Alves de Oliveira Júnior
Conceptualização, Administração do Projeto e Supervisão, Escrita - Análise e Edição.

DATAS
– Recebido: 07/08/2021
– Aceito: 03/09/2021
– Publicado: 25/11/2021

COMO CITAR
GALDINO, J.C.; SILVA, K.F.T.; OLIVEIRA JR., M. (2021). Características prosódicas associadas aos sinais de pontuação: uma revisão de escopo. *Cadernos de Linguística*, v. 2, n. 4, e468.

RESUMO

O objetivo deste artigo é apresentar uma revisão de literatura sobre as características prosódicas associadas aos sinais de pontuação. Foi realizado um levantamento bibliográfico a partir da pesquisa de descritores em inglês e português, organizados de acordo com a seguinte sintaxe: prosódia AND acústica AND discurso AND estrutura AND ("sinais de pontuação" OR "pontuação gráfica" OR "sinal de pontuação"), entre os anos 2015 e 2020, sem incluir citações e patentes nas bases de dados: *OvidMedlin*, *Public Medicine Library* (*PubMed*), *Scopus* (*Elsevier*), *Ebscohost* (*Academic Search Premier*), *Gale Academic Online* e *Google Scholar*. Observamos que existe uma diversidade de métodos empregados para analisar a correlação entre os sinais de pontuação e as características prosódicas. Os estudos desta revisão confirmaram nossa pergunta de pesquisa “Há relação entre a intensidade, a duração, o *pitch* e a pausa com os sinais de pontuação?”, evidenciando a relação entre os sinais de pontuação e os aspectos prosódicos.

ABSTRACT

The purpose of this article is to present the literature review of the prosodic features associated with punctuation marks. A bibliographic survey was carried out based on the search for descriptors in English and Portuguese, organized according to the following syntax: prosody AND acoustic AND speaks AND structure AND ("punctuation marks'' OR "graphic punctuation" OR "punctuation mark"), between 2015 and 2020, not including citations and patents in the databases: OvidMedlin, Public Medicine Library (PubMed), Scopus (Elsevier), Ebscohost (Academic Search Premier), Gale Academic Online, and Google Scholar. We note that there are a variety of methods employed to analyze the correlation between punctuation marks and prosodic features. The studies in this review confirmed our research question "Is there a relationship between intensity, duration, pitch and pause with punctuation marks?", highlighting the relationship between punctuation marks and prosodic aspects.

PALAVRAS-CHAVE

Prosódia; Acústica; Sinal de Pontuação; Discurso.

KEYWORDS

Prosody; Acoustics; Punctuation Mark; Speech.

INTRODUÇÃO

A pontuação é um sistema de recursos gráficos que serve para facilitar a leitura e a compreensão de textos, deixando clara a estrutura da língua escrita (BECHARA, 2009; CAGLIARI, 1989; PACHECO, 2003; ALMEIDA & FONSECA, 2018). Na escrita, é relativamente simples identificar a estrutura de um texto, mas, geralmente, na fala isso é problemático (CRYSTAL, 2008). De modo geral, os sinais de pontuação apresentam correspondentes prosódicos que fazem transparecer a intenção subjacente de marcar a estruturação promovida pelas marcas gráficas (ALMEIDA; FONSECA, 2018).

Pacheco (2003), por exemplo, mostra que os sinais de pontuação (dois pontos, ponto de exclamação, de interrogação, ponto final, ponto e vírgula, reticências e vírgula) se diferenciam entre si pela intensidade, pela curva de F0 (Frequência Fundamental) e pela duração ou não da sílaba tônica do componente tônico, delimitado a partir da tônica saliente. Esses sinais de pontuação apresentam intervalos médios de pausa que se distinguem entre si, sendo a pausa mais longa delimitada pelo ponto final, o ponto de interrogação e a vírgula, pela pausa breve.

A presença de sinais de pontuação e de parágrafos concede ao discurso escrito uma estrutura clara, enquanto a estrutura do discurso falado é sinalizada pelos recursos prosódicos, como pausas, duração e tons de fronteira (SILVA, 2017).

> Tom, intensidade e duração ou pitch, loudness and tempo e, ainda, rythm ou ritmo – eis as propriedades, ou traços prosódicos. [...] De um ponto de vista fonológico, pode dizer-se, genericamente, que as línguas utilizam essas propriedades com objectivos diversos: (i) para marcar os *limites das unidades* (o acento pode indicar o fim ou início da palavra; a curva de entonação pode igualmente marcar os limites das unidades prosódicas); (ii) para criar *oposições distintivas* (nas línguas tonais como p.ex. o chinês, o tom de uma sílaba, por contraste com os tons das que a rodeiam, pode opor significados entre duas palavras cujos segmentos são iguais tendo, assim, uma função distintiva; da mesma forma, a duração de uma sílaba pode ter valor distintivo como p. ex. em latim ou em inglês); (iii) para distinguir *significados globais* de construções frásicas (a entonação é usada frequentemente para diferenciar uma interrogação de uma afirmação, por exemplo; neste caso pode dizer-se que a entonação tem valor distintivo (MATEUS, 2004, p. 6 e 7, grifos da autora).

Há vários estudos em diferentes línguas que procuram identificar as relações entre sinais de pontuação presentes na escrita e os fenômenos prosódicos presentes na cadeia da fala, como a duração da pausa e o nível do *pitch* do inglês (YOUNIS, 2017) e os valores de f0 nas variações melódicas do português (SANTOS; PACHECO; OLIVEIRA, 2019). Neste artigo, apresentaremos resultados de alguns desses estudos.

Com o avanço da tecnologia, pesquisas experimentais com foco em sistemas de síntese e reconhecimento de fala desenvolvidas pela área da tecnologia da informação têm se preocupado em analisar essa correspondência entre as características prosódicas e os sinais de pontuação. Para tal, são elaborados modelos neurais, que são treinados, empregando dados contendo informações prosódicas, com o propósito de prever os sinais de pontuação para assegurar a compreensão do que está sendo dito (RAJESWARI;

MAHESWARI, 2012). Os sistemas de síntese e reconhecimento de fala usam os elementos prosódicos com o intuito de proporcionar naturalidade à produção (PEIRÓ-LILJA; FARRÚS, 2020) e garantir níveis fiáveis de inteligibilidade, expressividade e adequação aos sistemas de reconhecimento (MARTIN *et. al*, 2020). O presente artigo também apresentará resultados de estudos que têm implementado, em sistemas de síntese e reconhecimento de fala, as informações prosódicas.

Trabalhos acerca da pontuação geralmente envolvem aspectos sintáticos ou semânticos, deixando pouco clara a relação entre os elementos gráficos da pontuação e a prosódia. Diante disso, o objetivo deste artigo é apresentar uma revisão de literatura sobre as características prosódicas associadas aos sinais de pontuação, nos últimos cinco anos, e, assim, encontrar lacunas, principalmente no contexto da Computação e da Linguística.

O interesse nessas áreas está relacionado ao fato de tratarem da já referida temática do ponto de vista da produção e da percepção. Com o intuito de produzir nos sistemas de conversão de texto para fala (*Text-to-speech* - TTS) e no reconhecimento automático de voz (*Automatic Speech Recognition – ASR*) semelhantes à fala humana, a área da Computação, nos últimos anos, está voltada em desenvolver *softwares* de síntese e reconhecimento de fala, que ainda apresentam erros quanto ao emprego da pontuação. Tendo vista que não há uma relação direta entre as características prosódicas e os sinais de pontuação, faz-se necessário um olhar crítico para essa correlação.

1. METODOLOGIA

Para atingir nosso objetivo, foram percorridas quatro etapas para a construção da revisão: definição da pergunta norteadora, elaboração dos critérios de inclusão e exclusão para a busca da literatura, estabelecimento das informações a serem colhidas dos estudos e apresentação da revisão.

Para realização do levantamento bibliográfico e da discussão da presente pesquisa, a seguinte pergunta norteadora foi elaborada: Há relação entre a intensidade, a duração, a altura (*pitch*) e a pausa com os sinais de pontuação?

Para o levantamento bibliográfico, executou-se uma busca no ano de 2021 nas seguintes bases de dados: *OvidMedlin, Public Medicine Library (PubMed), Scopus (Elsevier), Ebscohost (Academic Search Premier), Gale Academic Online* e *Google Scholar*, disponíveis pelo Portal de Periódicos Capes, que pudessem trazer resultados de textos publicados nestas áreas de pesquisa: ciência, tecnologia, medicina e humanas. Foi feita a pesquisa de descritores em inglês e português, usando a seguinte sintaxe: prosódia AND acústica AND discurso AND estrutura AND ("sinais de pontuação" OR "pontuação gráfica" OR "sinal de pontuação"), sem incluir citações e patentes nas bases de dados.

Foram selecionados artigos e dissertações publicados em português e inglês realizados nos últimos 5 anos (2015 a 2020) que tratavam do tema: características prosódicas associadas aos sinais de pontuação. Foram excluídos estudos com crianças, idosos, pessoas com patologias de voz, alterações cognitivas e auditivas, além de estudos duplicados e de revisão.

Realizou-se a busca e a identificação dos artigos, avaliação dos títulos e resumos, depois a leitura na íntegra e a seleção final para esta revisão. Foram encontrados 1887 arquivos, dos quais 1823 arquivos foram descartados no processo de identificação a partir do título e resumo, restando 64 arquivos.

Ao confrontarmos os arquivos baixados, foram encontrados oito arquivos duplicados que também foram excluídos, sobrando 56 arquivos para serem lidos na íntegra. Foram eliminados 29 arquivos por não abordarem de forma substancial a associação entre as características prosódicas e a pontuação, com foco experimental. A figura 1, baseada na recomendação *Prisma*[1], representa o fluxograma que descreve a seleção das publicações desta revisão.

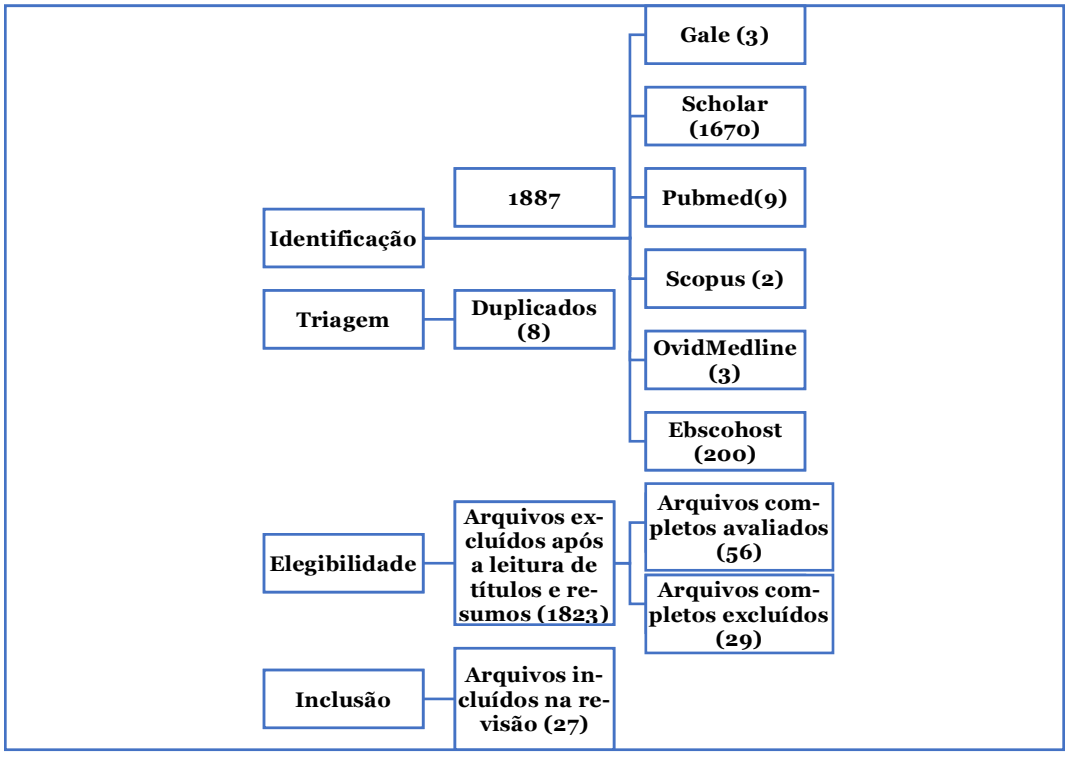

Figura 1. Fluxograma da seleção das publicações.

1 A lista de verificação possui 20 itens essenciais e 2 itens opcionais a serem considerados para realizar uma revisão de escopo. Disponível no site: http://www.prisma-statement.org/Extensions/ScopingReviews.

Para a sumarização dos 27 arquivos incluídos, utilizamos uma análise descritiva para examinar cada estudo. Nos resultados desta revisão, organizamos e compilamos esses trabalhos em duas categorias, elaborando resumos narrativos.

2. RESULTADOS

Conforme apresentado nos Quadros 1 e 2, observa-se a distribuição de 27 estudos, divididos em duas categorias, a saber: nove estão relacionados à associação entre sinais de pontuação e prosódia a partir da leitura e/ou texto escrito e dezoito ligados ao desempenho do software a partir do texto escrito e/ou recursos acústicos.

Associação entre sinais de pontuação e prosódia a partir da leitura e/ou texto escrito					
Autor, Ano e País	**Objetivo**	**Tipo de Estudo**	**Participantes/Banco de dados**	**Variáveis**	**Principais achados**
Santos, Pacheco e Oliveira (2019) Brasil	Avaliar a fluência de leitura dos participantes em diferentes níveis de escolaridade.	Experimental	Banco de dados: três textos distintos, cada um deles com 10 versões, compondo 30 textos, sendo 15 marcadores prosódicos lexicais "perguntou", marcadores prosódicos gráficos (?), (!), (...), (.), frases sem pontuação, e 15 marcadores prosódicos lexicais "disse", marcadores prosódicos gráficos (?), (!), (...), (.). Ao total são 120 ocorrências para cada informante. 12 participantes: grupo I (7 anos); grupo II (entre 15 e 17 anos); Grupo III (entre 30 e 40 anos), excluídos com nível de leitura lentificado e/ou silabado.	Leitores fluentes e não fluentes; marcadores prosódicos e gráficos.	Os leitores mais escolarizados apresentaram variação melódica quanto às marcações prosódicas gráficas e lexicais em detrimento dos que tinham menor escolaridade que apresentaram um padrão recorrente (ascendente) na produção de F0 independentemente do marcador lexical ou gráfico.
Costa (2019) Brasil	Analisar a influência da oralidade no ato de pontuar graficamente os textos escritos através da análise do uso e da presença de sinais de pontuação como marcadores prosódicos, e não sintáticos.	Experimental	Do total de 200 redações, 100 foram produzidas por alunos do 1° ano do ensino médio e 100 por alunos do 2° ano do ensino médio.	Marcação rítmica na fala; uso não convencional de sinais de pontuação na escrita.	Além de analisar as redações, o estudo fez uma análise com dados de fala, solicitando que sujeitos fizessem leitura de enunciados, constatando que existe um contorno entoacional do tipo L*HL entre o sujeito e o predicado associado à cabeça de φ na enunciação de todos os quatro falantes, ainda que em alguns de forma mais evidente e em outros de maneira mais breve. O sujeito nessa frase parece-nos topicalizado. A vírgula, não estando de acordo com as normas padrões de pontuação, pode estar funcionando como marcador prosódico das noções rítmicas dos sujeitos que produziram esses textos.
Almeida; Fonseca (2018) Brasil	Observar como se dá a produção de sinais de pontuação na leitura em voz alta e na escrita, a fim de verificar a relação entre essa produção e a compreensão do texto.	Experimental	22 sujeitos, sendo 20 alunos do 3° ano do ensino médio e 2 professoras da Educação Básica com título de Doutorado em Linguística.	Ausência/presença de pontuação; a quantidade de marcações gráficas) e marcações prosódicas na leitura e a quantidade de erros no teste de compreensão.	Os sinais de pontuação sinalizaram as características prosódicas e não influenciaram na compreensão. Pode influenciar a produção a ausência ou a presença de uma marca inadequada.
Paixão; Serra (2018) Brasil	Investigar de que forma se dá o fraseamento prosódico de estruturas parentéticas (parênteses) no português do Brasil (PB) – fala carioca, em	Experimental	568 sentenças lidas por informantes cariocas do sexo feminino estudantes de pós-graduação da UFRJ e com idades entre 22 e 30 anos (42 sentenças x 15 falantes, excluídos 62 dados),	O cruzamento das variáveis: tipo de parentética, movimento tonal no primeiro IP e	As pausas são mais recorrentes no final dos parênteses. Os contornos descendentes aconteceram no primeiro e no segundo IPs, diferindo do Português Europeu, variedade na qual

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

	dados de leitura de frases, a partir de uma abordagem fonológica.			movimento tonal no segundo IP.	a predominância das pausas ocorre no início da parentética e os contornos são ascendentes.
Heggie; Wade-Woolley (2018) Canadá	Investigar a hipótese de que a consciência prosódica de leitores adultos, ou sua capacidade de atender e manipular o ritmo da fala na linguagem oral, está relacionada à capacidade de aplicar corretamente a pontuação.	Experimental	Participantes adultos falantes de inglês com idades entre 18 e 51 anos.	Identificação de sílaba tônica; Manipulação de sílaba tônica; Marcação de sinal de pontuação.	Há correlação entre o desempenho de pontuação dos participantes e a consciência prosódica (r D .436, .392 respectivamente, p < .001). A consciência prosódica surgiu como um preditor robusto da capacidade dos adultos de pontuar, além da influência de seu conhecimento de pontuação, compreensão de leitura e memória de trabalho.
Soncin; Tenani (2017) Brasil	Fornecer uma análise linguística de padrões prosódicos (especialmente o complexo de traços fonológicos suprassegmentais que inclui entonação, pausas e acentos) subjacentes ao uso de vírgulas em textos escritos por estudantes brasileiros.	Experimental	Analisaram apenas 284 dos textos do Banco de Dados de Produções Escritas do Ensino Fundamental, que contém 5.513 textos, escritos por 71 alunos de 13 a 14 anos que, em 2008, cursavam o último ano do ensino fundamental brasileiro.	Coincidência e não coincidência de usos não convencionais de vírgula e limite prosódico	93,4% das vírgulas não convencionais encontram-se nos limites prosódicos. 92% dos usos não convencionais de vírgulas definem limites prosódicos, que delimitam as frases entoacionais. Na modalidade escrita, a linguagem é organizada prosodicamente.
Younis (2017) Egito	Examinar como a prosódia é reproduzida ao traduzir *as* marcas *al-waqf* (pausa) na leitura do Alcorão.	Experimental	Ferramenta paralela do Corpus do Alcorão; Recitações autorizadas dos versos coletados, usando a seção de áudio do Corpus Alcorão.	مـ (pausa obrigatória), لا (não é permitido pausar), صلے (pausa é permitida; continuação é preferível), قلے (continuação é permitida; pausa é preferível). A duração do tempo nas junções; o nível do tom nas pausas e posteriores.	Os recursos prosódicos desempenham um papel na solução de problemas de tradução de pontuação. Por exemplo, os versos (2: 146) e (6: 20) que pareciam ser controversos para os tradutores, assim como as marcas *al-waqf*, pois eles começam de forma semelhante, mas contêm duas pausas diferentes na junção da cláusula, parecem se comportar de maneira diferente no nível acústico. As duas marcas prosódicas examinadas neste estudo são significativas. A duração da pausa nas marcas *al-waqf* (pausa), bem como o nível de *pitch* da linha de base nas pausas e o nível de *pitch* da linha superior após elas, lançam luz sobre a marca de pontuação potencial que deve ser usada na tradução para o inglês.
Barhate *et al.* (2016) Índia	Analisar a prosódia da variação do estilo de leitura de notícias Marathi.	Experimental	Seis ouvintes nativos Marathi.	Marcar todos os limites percebidos e palavras proeminentes na transcrição impressa de notícias de Marathi, que continha vírgulas e ponto final.	Ao realizar testes de percepção, a proeminência depende mais do aumento de F0 na palavra do que da duração. Os limites entre frases foram sinalizados por pausas e alongamento pré-limite da sílaba final. Os limites do tópico exibiram as pausas mais longas.
Fuchs; Krivokapi (2016) Estados Unidos	Investigar a duração entre pressionamentos sucessivos de tecla durante a digitação, a fim de examinar se os limites prosódicos são expressos no processo de escrita. Verificar durações entre teclas que ocorrem ao lado de marcas de pontuação, uma vez que essas marcas de pontuação são muitas vezes realizadas com	Experimental	Quatorze participantes (sete mulheres e sete homens) entre 21-43 anos (média 30,2 anos) foram registrados.	Durações entre teclas que ocorrem ao lado de marcas de pontuação; durações de pausa silenciosa durante a leitura aberta.	A comparação das durações para escrita e leitura (durações entre teclas e pausas dentro e entre as sentenças) produziu um resultado significativo (r 2 ajustado = 0,20, F = 7,53, gl = 51, p = 0,00149). Em relação ao alongamento inicial, foram encontradas diferenças nas durações entre teclas para os casos iniciais de sentença [β = 6,31]. Na produção da fala, na frase, em termos de fronteira prosódica, inicialmente não há diferença entre os limites menores e as palavras. O alongamento final ocorre dentro e entre o nível da

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

	limites prosódicos maiores ou menores durante a leitura aberta.				frase em comparação com o nível da palavra, ou seja, nos limites maior e menor em comparação com o limite da palavra.

Quadro 1. Quadro de síntese dos estudos sobre a associação entre sinais de pontuação e prosódia a partir da leitura e/ou texto escrito.

Desempenho do software a partir do texto escrito e ou recursos acústicos					
Autor, Ano e País	**Objetivo**	**Tipo de Estudo**	**Participantes/Banco de dados**	**Variáveis**	**Principais achados**
Yi e Tao (2019) China	Propor o uso de um modelo baseado em autoatenção para prever sinais de pontuação para sequências de palavras.	Experimental	Conjuntos de dados IWSLT em inglês que contêm palestras TED (Tecnologia, Entretenimento, Design).	Desempenho da pontuação; reconhecimento automático	A fala carrega informações acústicas (pausa e entonações) que não estão presentes no texto simples, possibilitando prever a pontuação. Esse modelo comparado a outros anteriores revelou-se eficaz, pois pode aprender o léxico e as informações acústicas usando qualquer tipo de dados de textos sem áudio e voz correspondente.
Öktem (2019) Espanha	Desenvolver ferramentas abertas que permitem a criação, anotação prosódica, manuseio e visualização de dados da linguagem falada; Compilar e publicar corpora monolíngue e bilíngue adequados para o desenvolvimento baseado em aprendizado de máquina que envolve modelagem prosódico-linguística. Desenvolver um framework para restauração automática de pontuação em transcrições de fala geradas manualmente ou automaticamente, usando recursos lexicais e prosódicos. Avaliar efeito de recursos prosódicos-acústicos na qualidade da restauração de pontuação e processos subsequentes, como análise de dependência e tradução automática. Desenvolver uma estrutura de tradução automática para o domínio do filme que permite a entrada e saída de recursos prosódicos para auxiliar na	Experimental	Conjunto de dados: corpus da popular série de TV de ficção científica de 2000, Heroes16. A série completa consiste em 4 temporadas e 77 episódios e é dublada em vários idiomas, incluindo espanhol, português, francês e catalão. Cada episódio dura uma duração de 42 minutos em média. As palestras TED (Tecnologia, Entretenimento, Design) são um conjunto de palestras com duração média de 15 minutos cada, realizadas no mundo todo em mais de 100 idiomas. Eles incluem uma grande variedade de tópicos, desde tecnologia e design até ciência, cultura e academia. Levando em consideração que o número de palavras por frase no corpus de palestras TED (Tecnologia, Entretenimento, Design) é de 15 a 20 em média, os dados são amostrados em sequências de 50 palavras. As amostras são extraídas sequencialmente das palestras. Cada amostra começa com uma nova frase.	A transcrição automática de fala e a tradução da linguagem falada.	Um *Proscript*[2] foi desenvolvido para anotação acústico-prosódica de dados de fala transcritos. Empregou-se no *Prosograph*[3] dados de fala, sendo possível mostrar todos os enunciados alinhados com suas características prosódicas correspondentes. A combinação dos recursos palavra, tags de classes gramaticais, pausa interlexical e *pitch* funcionou melhor para a restauração precisa dos três principais sinais de pontuação: ponto final, vírgula e ponto de interrogação. A configuração com melhor desempenho entre as três marcas de pontuação obteve uma pontuação F1 de 70,3%, que mostrou uma melhoria de 3,5% em comparação com a abordagem de linha de base, quando apenas dados falados foram usados. Além disso, para marcas de pontuação individuais, pontuações F1 de 83%, 71,8% e 55,2% foram relatadas, empregando várias outras combinações de recursos. O período tende a se beneficiar do uso de recursos de intensidade. As vírgulas foram detectadas geralmente com baixa precisão, mas mostraram que o *pitch* ajudou na detecção. A pontuação baixa das vírgulas foi explicada com a possível variação na anotação de pontuação nas transcrições de referência.

2 Um pacote de códigos que serve para acomodar a criação e o armazenamento de dados prosódicos e para facilitar o processamento com aplicativos de aprendizado de máquina (ÖKTEM, 2019).

3 Um *software* desenvolvido por Öktem (2019), assemelhando-se a uma ferramenta digital de análise musical, para ajudar na visualização de grandes porções de dados falados.

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

	tradução e gerar resultados para síntese.				
Szaszák (2019) Hungria	Avaliar uma abordagem inspirada em prosódia e um modelo de sequência de frases implementado como uma rede neural recorrente para prever as marcas de pontuação do áudio. Fazer uma avaliação subjetiva para quantificar os benefícios da pontuação na legibilidade humana, ou seja, mostrar se quando uma precisão de pontuação é alcançada, os humanos são capazes de distinguir a pontuação automática da produzida pelo homem, mesmo que a primeira possa conter erros de pontuação.	Experimental	Bancos de dados: Húngaro BABEL, um corpus de fala lida gravado a partir de falantes nativos não profissionais; e um corpus Broadcast News (BN). O corpus BABEL é dividido para treinar, validar e testar conjuntos (80%, 10%, 10% dos enunciados, respectivamente). O corpus BN é usado para testes. Participantes: 35 sujeitos, 28 homens e 7 mulheres com média de idade de 29,6 anos. A maioria era estudantes universitários ou funcionários do setor terciário.	Previsão da pontuação pelo software; julgamento da pontuação automática e humana quanto à precisão.	O software conseguiu prever a pontuação automática a partir dos recursos prosódicos. Os participantes não conseguiram distinguir a pontuação automática da produzida pelo homem, mesmo a primeira contendo erros de pontuação, sendo os indivíduos mais sensíveis a erros do reconhecimento automático da fala do que a erros de pontuação.
Szaszák; Tündik (2019) Hungria	Propor uma abordagem que se adapte bem a dois tipos de idiomas – inglês, a ordem de palavras menos restrita, e húngaro, altamente aglutinante.	Experimental	O banco de dados húngaro é derivado de serviços públicos e transmissões comerciais de TV, que contém vários gêneros (previsões do tempo, notícias, conversas, revistas, esportes). O conjunto de dados em inglês consiste em transcrições IWSLT TED Talk, que é um conjunto de dados de referência comumente usado para modelos de pontuação em inglês, contendo 2,1M, 296K e 13K palavras, respectivamente, e lidando com três tipos de pontuação (vírgula, ponto e ponto de interrogação). Os dados de áudio são traduções de palestras IWSLT2011 contendo 6 horas de fala.	Previsão de pontuação combinando características de caractere, palavra e prosódia em dois idiomas: húngaro e inglês	Para o húngaro, de ordem de palavras altamente aglutinante e relativamente livre, houve uma melhora significativa na pontuação geral, em relação à linha de base, ou seja, ao modelo de nível de palavra, adicionando os modelos de caracteres e prosódia (significativo por $p < 0,01$). Para o inglês, adicionar prosódia levou a uma melhoria de 4,4% em relação aos modelos de caracteres nas transcrições ASR (ainda significativo por $p < 0,05$).
Cenceschi *et al.* (2019) Itália	Investigar se as pistas derivadas do texto poderiam melhorar o reconhecimento de formas prosódicas simples em unidades de informação, constituídas de realização textual (frases escritas - declaração, perguntas e exclamações) e uma realização acústica (a gravação da produção da tal frase).	Experimental	Banco de dados para o *PESInet*[4] ebooks, audiobooks EPUB3 (contém texto e gravação de áudio, alinhado ao tempo no nível da frase), e o corpus LIT / DIA-LIT (contém gravações de áudio de programas de TV italianos, com transcrições). Participantes: Confrontar o *PESInet* com ouvintes: 302 falantes de italiano que	Julgamento do *PESInet* e dos participantes italianos quanto às formas prosódicas baseadas em pistas textuais e acústicas	O *PESInet* foi treinado com cerca de 1,5 milhão de frases de texto e 60.000 declarações de fala recitada e espontânea, conseguindo alcançar precisão de 80% em três classes e 91% em pergunta versus não pergunta. Em comparação com os ouvintes humanos, o *PESInet* atingiu uma melhor precisão de 89% em detrimento a 80% para os ouvintes humanos.

4 PESInet é um sistema de Reconhecimento Automático de Prosódia com o objetivo de classificar as Unidades de Informação em Afirmação, Pergunta ou Exclamação.

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

			selecionavam qual unidade de informação carregava a forma prosódica esperada.		
Kocharov; Kachkovskaia; Skrelin (2019) Rússia	Prever todos os limites prosódicos potenciais com base na sintaxe e entre esses limites potenciais, escolher aqueles que são marcados acusticamente.	Experimental	Banco de dados: CORPRES - um corpus desenvolvido em 2009 no Departamento de Fonética da Universidade Estadual de São Petersburgo, contendo gravações de textos fictícios lidos por oito falantes nativos do russo padrão, 4 homens e 4 mulheres.	Desempenho da detecção automática de fronteira prosódica	A detecção automática de fronteira prosódica mostrou alta eficiência. Sem dados sintáticos, a acústica sozinha fornece a eficiência de 0,86. O pré-processamento sintático permite eliminar da análise posterior uma parte substancial das junções de palavras onde um limite é extremamente improvável. Isso levou a uma eficiência (em termos de medida F 1) de mais de 0,912, precisão acima de 0,93 e recuperação de 0,90. Os autores constataram que não era necessário medir parâmetros acústicos, como amplitude média e intervalo de F0, em todo o grupo de clíticos: cálculos sobre a sílaba tônica forneceram quase os mesmos resultados, revelando que a detecção de limite prosódico deve ter um bom desempenho, independentemente de a segmentação em nível de palavra estar disponível ou não.
Dominguez *et al.* (2019) Espanha	Prever o estudo da interface IS-prosódia[5] partir de um perspectiva metodológica baseada em uma configuração de implementação de síntese de fala.	Experimental	Banco de dados: Um corpus criado a partir de texto obtido na web em alemão sobre conselhos para rotinas de sono e notícias locais. O corpus contém oito textos com um total de 1.418 palavras.	Desempenho de síntese de fala manual e automática	O estudo possibilita a transição do trabalho teórico na interface IS-prosódia para a integração do enriquecimento da prosódia baseado em temática para alcançar um discurso sintetizado mais expressivo. Um teste t mostra que os resultados gerais (média) para as modificações automáticas de prosódia (AUT = 3,30) alcançam significância estatística em p <0,05 em comparação com a pontuação padrão (DEF = 3,01).
Liu, Li e Zhao (2018) China	Discutir a viabilidade do reconhecimento de pontuação na oralidade por meio da detecção de fala e dos caracteres da fala não sonora (sinais de pontuação).	Experimental	Banco de dados: em dois períodos, os professores de inglês conseguiram 2100 frases. 1000 frases foram para fins de treinamento, enquanto 1100 frases usadas para reconhecimento. Participantes: 10 pós-graduados (6 homens, 4 mulheres) em aulas de inglês no Centro de Aprendizagem de Línguas Estrangeiras leram 210 frases de seu livro didático.	Reconhecimento da pontuação automática por meio de parâmetros acústicos (duração, variação melódica, intensidade).	O software apresentou reconhecimento médio de 97,2% na detecção sonora e não sonora. Ao dividir um sintagma nominal, vírgula e ponto, em frases mais curtas, tiveram maior chance de quebrar entidades, principalmente se envolverem frases longas.
Orosanu; Jouvet (2018) França	Analisar a detecção da modalidade (perguntas e afirmações) "sentença" em francês quando aplicada em transcrições de fala para texto automático.	Experimental	10 mil perguntas e 98 mil declarações extraídas dos ESTER2[6] e ETAPE[7] e do projeto EPAC ESTER2 e EPAC[8] que são notícias transmitidas em francês coletadas de vários canais de rádio (discurso preparado e entrevistas). Os	Frases declarativas e interrogativas; segmentação automática e manual	Ao avaliar as sequências de palavras e partes do discurso (linguísticas) e as informações acústicas (prosódica), o classificador linguístico forneceu resultados muito melhores: 72% quando é aplicado em transcrições de fala para texto automático (com limites perfeitos de frase) contra 74% quando

5 A interface de Estrutura de Informação (IS)-prosódia tem uma perspectiva metodológica baseada em uma configuração de implementação de síntese de fala.

6 É um banco de dados provenientes da extensão da campanha de avaliação ESTER inicial de 2003–2005, visando uma ampla variedade de estilos de fala e sotaques.

7 O *corpus* ETAPE visa uma variedade mais ampla de qualidade de fala espontânea a partir do material advindo de TV.

8 O corpus EPAC é constituído por um conjunto de 100 horas de fala coloquial transcrita manualmente e pelas saídas de ferramentas automáticas (segmentação automática, transcrição etc.) aplicadas em todo o corpus de áudio ESTER 1 na língua francesa, que se trata de cerca de 1700 horas de gravações de áudio de programas radiofônicos.

			dados da ETAPE correspondem a debates recolhidos em várias estações de rádio e TV (fala espontânea).		é aplicado em transcrições manuais de referência. A combinação das duas características nas transcrições automáticas de fala chegou a 72%.
Tündik e Szaszák (2018) Hungria	Abordar a pontuação em nível de caractere e mostrar que usando informações em nível de caractere pode ajudar a superar não apenas a linha de base Entropia Máxima[9], mas também as linhas de base da rede neural recorrente baseadas em nível de palavra.	Experimental	O banco de dados húngaro é proveniente de transmissões públicas do Fundo de Gerenciamento de Ativos e Apoio de Serviços de Mídia (MTVA), Hungria. Os dados contêm vários gêneros de TV, como previsões do tempo, notícias e conversas transmitidas (BC), revistas, notícias esportivas e revistas esportivas. O conjunto de dados usado para esses experimentos é um subconjunto de transcrição manual, incluindo pontuação. Além disso, usou-se o conjunto de dados *IWSLT* que contém transcrições de palestras TED em inglês. O *IWSLT* foi usado várias vezes para avaliar modelos de pontuação em inglês e recuperação de pontuação.	Desempenho da pontuação; reconhecimento automático e manual.	Devido às particularidades de cada língua, na língua húngara, houve predição por vírgula e, no inglês, ocorreu para ponto final e interrogação. O modelo de pontuação possui poucas entradas de vocabulário, mas com desempenho próximo ao modelo de palavra. A combinação da pontuação e da palavra (modelo híbrido) mostrou desempenho superior em comparação com a linha de base. Além disso, o modelo de pontuação foi bem-sucedido.
Liu; Liu; Song (2018) China	Treinar o modelo de recuperação de pontuação em três fases: modelo com recursos de texto, modelo com recursos de texto e tempo de pausa e modelo com três características (texto, tempo de pausa e *pitch*).	Experimental	Banco de dados: os dados do texto são compostos principalmente pelo Sogou-News e pelo Diário do Povo. Como dados de desenvolvimento, dez horas de conteúdos de notícias são usadas nesse jornal. O conjunto de teste contém uma hora de conteúdos de notícias.	Desempenho do modelo de reconhecimento automático de fala	O modelo pode reter as vantagens do LSTM (memória de curto prazo longa) e o tom introduzido pode se adaptar ao idioma chinês, que é uma língua tonal. Além disso, o modelo consiste em três partes independentes umas das outras, podendo ser usado o primeiro modelo para prever a posição e o tipo de pontuação sem tempo de pausa e tom. Da mesma forma, nas condições correspondentes, podem usar o segundo ou terceiro modelo separadamente.
Su; Thida (2018) Myanmar	Desenvolver um sintetizador de fala em Myanmar que possa produzir uma qualidade aceitável de saída sintetizada em quase tempo real em um dispositivo móvel usando a técnica de síntese de fala concatenada.	Experimental	Banco de dados: 133 fonemas gravados, são registrados com taxa de amostragem de 44100 Hz. A qualidade da voz do sistema MTTS (Myanmar Text-To-Speech) é avaliada por seis avaliadores, sendo 5 mulheres e um homem.	Desempenho de síntese de fala concatenada	Ao calcular a taxa de erro de palavra para cada pessoa, a taxa de erro de palavra para o teste de inteligibilidade foi de apenas 7% e a pontuação média da pontuação de opinião para inteligibilidade é 3,7. Para o teste de naturalidade, os avaliadores foram questionados se o discurso de Mianmar gerado pelo sistema é natural ou não. De acordo com suas respostas, 2% dos ouvintes consideram a saída de voz muito natural, 21% consideram a fala natural e 51% dos ouvintes identificam a voz como aceitável. Cerca de 23% assumiram que a saída de voz precisa obter mais naturalidade. A pontuação média da pontuação de opinião para 400 sentenças é 2,96. Assim, a saída de voz do sistema atingiu o nível aceitável de naturalidade.
Moró; Szaszák (2017a) Hungria	Propor um modelo leve de pontuação inspirado na prosódia,	Experimental	Banco de dados: o Húngaro BABEL, um banco de dados de leitura de voz gravada de falantes nativos não profissionais.	Modelo de segmentação	Obtiveram bons resultados de pontuação em dados de fala lidos (Húngaro BABEL) com um método rápido e facilmente treinável, que além

9 É um sistema com energia interna que se desequilibra por perda de ligação interna, evoluindo até um novo equilíbrio, chamado de entropia máxima.

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

	capaz de trabalhar com baixa latência.			automática para frases fonológicas	disso apresenta pouca latência. Como um modelo baseado em prosódia, é robusto contra erros de ASR, confirmados por experimentos. Descreveram o impacto moderado da densidade de alinhamento de frase fonológica e comprimento de sequência no desempenho de pontuação. Nos dados de notícias transmitidas, a previsão do período foi satisfatória, mas a previsão das vírgulas provou ser mais problemática.
Moró; Szaszák (2017b) Hungria	Propor uma abordagem pura e leve, que captura a estrutura de informação do enunciado com base na prosódia da fala.	Experimental	Banco de dados: Húngaro BABEL, dados de leitura de voz gravada de falantes nativos não profissionais é usado para treinar o modelo de rede neural recorrente de sequência de frases. Há quase 20mil faixas de palavras e 7 a 20 mil faixas de frases fonológicas, o último dependendo da densidade do alinhamento de frase fonológica controlado pelo parâmetro de verossimilhança do logaritmo de inserção de frase. Um subconjunto de 10% do conjunto de treinamento é reservado para validação e outros 10% para teste. Um subconjunto de 10% do conjunto de treinamento é reservado para validação e outros 10% para teste. Também testaram em um corpus Broadcast News (BN) de 50 blocos curtos com 3 mil palavras, 300 vírgulas e 500 períodos. Este material é uma hipótese ASR obtida pela taxa de erro de palavra = 10,5%	Modelo de pontuação automática para frases fonológicas	O método proposto é rápido de treinar e apresenta pouca latência por ser robusto contra erros de reconhecimento de fala. Há leve impacto da densidade de alinhamento de frase fonológica e da janela de sequência na pontuação. Mudando para um corpus de notícias de transmissão, o poder de pontuação foi mantido por períodos usando alguns ciclos de adaptação para o modelo de rede neural recorrente, mas houve uma queda considerável de desempenho para vírgulas. Na verdade, a evidência acústica parece ser limitada em relação aos slots de vírgula em notícias de transmissão, o que pode ser contrabalançado pela inclusão de recursos baseados em texto em uma abordagem híbrida.
Klimkov *et al.* (2017) Suécia	Melhorar a previsão de pausas na leitura da conversão de texto em voz.	Experimental	Dados extraídos de audiolivros, contendo um conjunto de aproximadamente 500 livros de não ficção, de 9 talentos vocais (cerca de 55 livros por locutor).	Desempenho do modelo de previsão de pausas na leitura da conversão de texto em voz	Apesar de haver preferência significativa para o modelo de fraseamento CART *(Classification-Regression Decision Trees)* pelos ouvintes ingênuos e profissionais, no quadro geral, o BiLSTM *(Bi-directional Long Short-Term Memory)* faz grandes contribuições para melhorar o número de quebras inadequadas.
Ballesteros; Wanner (2016) Espanha	Gerar pontuação no discurso escrito pela arquitetura de rede neural sem usar recursos acústicos e sintáticos, apenas o texto.	Experimental	Banco de dados em cinco idiomas: tcheco, inglês, francês, alemão e espanhol.	Texto sem pontuação(entrada); introduz pontuação sem usar nenhum recurso acústico e sintático	A arquitetura adicionou marcas de pontuação sem quaisquer recursos sintáticos ou acústicos, podendo ser usada para reconhecimento de fala automático e correção de erros gramaticais.
Igras-Cybulska; Ziółko; Zelasko; Witkowski (2016) Polônia	Verificar se as informações das pausas podem ser úteis para sistemas de biometria de falantes (experimento 1) e para o reconhecimento de diferentes tipos de fala espontânea (experimento 2), bem como distinguir entre fala lida e espontânea (experimento 3).	Experimental	Fala de 30 falantes (16 homens, 14 mulheres). O primeiro grupo de gravações (30 min) é formado por declarações de discursos ou apresentações públicas. A segunda parte do corpus (30 min) consistiu em gravações da tradução em tempo real das orações durante as sessões do Parlamento Europeu [20]. O terceiro tipo de gravação (60 min) são as emissões de rádio.	Pausas para identificação de falantes (experimento 1); diferentes tipos de fala espontânea (experimento 2); fala lida e fala espontânea (experimento 3).	No experimento com reconhecimento automático de três tipos de fala espontânea, obtiveram 78% de acerto, utilizando o classificador GMM. Os recursos relacionados à pausa silenciosa permitiram distinguir entre a fala lida e espontânea por meio de um aumento de gradiente extremo com 75% de precisão, mostrando a utilidade das pausas para distinguir diferentes contextos situacionais e tarefas cognitivas.

DOI 10.25189/2675-4916.2021.V2.N4.ID468 ISSN: 2675-4916 V. 2, N. 4, 2021

Igras; Ziółko (2016) Polônia	Melhorar os sistemas de reconhecimento automático de voz com detecção automática de limites de frases, explorando informações de pistas acústicas ou prosódicas.	Experimental	Banco de Dados de Fala Audiovisual AGH (Igras *et al.,* 2012) e uma gravação do monólogo preparado. Finalmente, o conjunto de teste continha 159 sinais de pontuação (98 pontos finais e 61 vírgulas).	Correlatos acústicos da pontuação no polonês falado para automatizar a inserção da pontuação nas transcrições ASR.	Foi desenvolvido um algoritmo para detecção dos limites sintáticos que aparecem nos locais dos sinais de pontuação. No primeiro estágio, o algoritmo detecta pausas e divide um sinal de fala em segmentos. Na segunda etapa, verifica a configuração dos traços acústicos e coloca hipóteses das posições dos sinais de pontuação. A classificação é realizada com parâmetros que descrevem a duração do fonema e energia, taxa de fala, contornos de frequência fundamental e bandas de frequência. Os melhores resultados foram obtidos para o classificador Naive Bayes. A eficiência do algoritmo é de 52% de precisão e 98% de recuperação.

Quadro 2. Quadro de síntese dos estudos sobre desempenho do software a partir do texto escrito e/ou recursos acústicos.

3. DISCUSSÃO

Nesta revisão, encontramos pesquisas experimentais feitas em 14 países. Observou-se que 17% das pesquisas foram desenvolvidas no Brasil (n=5) e na Hungria (n=5), 11% na China (n=3) e Espanha (n=3), 4% Polônia (n=2), 4% Canadá (n=1), Egito (n=1), Estados Unidos (n=1), França (n=1), Índia (n=1), Itália (n=1), Myanmar (n=1), Rússia (n=1) e Suécia (n=1).

Observamos a presença, neste levantamento bibliográfico, de nove estudos que, a partir da leitura ou da presença da oralidade na escrita, analisaram a relação entre os sinais de pontuação e os aspectos prosódicos.

Os estudos encontrados comprovaram essa relação quando expuseram que há correlação entre o desempenho de pontuação de falantes e a consciência prosódica (HEGGIE; WADE-WOOLLEY, 2018) e que os sinais podem influenciar a produção da leitura em voz alta (ALMEIDA; FONSECA, 2018). Falantes mais escolarizados apresentam variação melódica em marcações prosódicas gráficas e lexicais, enquanto os leitores menos escolarizados apresentaram um padrão recorrente (ascendente) na produção de frequência fundamental (SANTOS; PACHECO; OLIVEIRA, 2019).

No português brasileiro (PB), a vírgula funciona como marcador prosódico das noções rítmicas dos falantes e é responsável por influenciar a pontuação gráfica do texto (COSTA, 2019). Em relação ao tipo de limites prosódicos definidos por usos não convencionais de vírgulas, foi revelado que em 92% dos casos em que os usos não convencionais de vírgulas definem limites prosódicos, esses limites delimitam frases entonacionais (SONCIN; TENANI, 2017).

Os limites entre as frases são sinalizados, geralmente, por pausas e pelo alongamento da sílaba final (FUCHS & KRIVOKAPI, 2016; BARHATE *et. al.*, 2016; YOUNIS, 2017; PAIXÃO E SERRA, 2018), mostrando que as pistas prosódicas podem sinalizar limites sintáticos (NESPOR; VOGEL, 1986).

Além desses nove estudos mencionados, encontramos 18 trabalhos que envolvem a associação entre sinais de pontuação e a prosódia em aplicações tecnológicas, nos sistemas de conversão de texto para fala e no reconhecimento automático de voz, todos internacionais. Houve predominância de artigos sobre ASR, somando-se 13, enquanto houve apenas quatro artigos sobre TTS e um envolvendo ambos os sistemas.

Em relação aos sistemas ASR, as pesquisas mostraram que as pausas são indicadores mais fortes dos sinais de pontuação, sendo um recurso prosódico bastante útil para distinguir diferentes contextos situacionais e tarefas cognitivas (IGRAS E ZIOŁKO, 2016; IGRAS-CYBULSKA *et. al.* 2016). As pausas, segundo Cagliari (1981), são unidades rítmicas obrigatórias para diferenciar significados que ocorrem dentro do enunciado nas mais diversas posições, sendo marcadas com símbolos na escrita.

Pausa e entonação não fornecem pistas corretas quanto ao limite da frase, diferentemente do que ocorre na escrita (CRYSTAL, 2008). A segmentação da fala não coincide com a segmentação da escrita, pois a fala não apresenta espaços entre as palavras, como ocorre com a partição na escrita, pois são os aspectos prosódicos responsáveis pelo fracionamento da fala em unidades prosódicas (ALCÂNTARA, 2006). Nos casos de segmentação não convencional da escrita, os aspectos prosódicos irão delimitar as fronteiras no fluxo da fala, conduzindo a percepção da segmentação da escrita (CAPRISTANO, 2004).

Algumas pesquisas apontaram pouca identificação de vírgulas nos sistemas ASR - reconhecimento automático de fala - que transformam a fala em texto (MORÓ & SZASZÁK, 2017a; MORÓ; SZASZÁK, 2017b). A vírgula pode indicar três tons (suspensivo, asserção e pergunta), sendo marcadores prosódicos que podem ser expressos de diversas maneiras e servem para orientar o leitor quanto às variações melódicas e entoacionais do discurso (CAGLIARI, 1989).

A não previsão de vírgulas pelos sistemas que traduzem a fala em texto torna-se um problema para a comunicação, pois o reconhecimento de sinais de pontuação favorece a distinção dos constituintes da frase e facilitam a leitura (MOREIRA, 1985). Ao empregar o método de dividir sintagma nominal, um estudo mostrou que vírgula e ponto possuem maior chance de quebrar entidades, principalmente se envolverem frases longas, favorecendo o reconhecimento de 97,2% na detecção sonora e não sonora (LIU, LI E ZHAO, 2018).

Em relação aos sistemas TTS, que convertem texto em fala, os estudos têm mostrado que a prosódia desempenha um papel importante na melhoria da qualidade dessa conversão (RAMU REDDY; SREENIVASA RAO, 2016). O sistema de conversão de texto em fala é constituído de duas partes: processamento de linguagem natural e processamento do sinal digital, que geram uma representação acústica da entoação a partir de outros níveis linguísticos (DUTOIT, 1997). A inserção das informações acústicas – altura (F0), intensidade e pausa - provenientes de dados de fala em um conversor texto-fala permite que a fala produzida pelo sistema tenha maior aceitabilidade quanto aos aspectos da inteligibilidade e da naturalidade por parte dos ouvintes (TEIXEIRA, 1995).

Em nossos achados sobre os sistemas TTS, é possível detectar essa preocupação para um discurso sintetizado mais expressivo (DOMINGUEZ *et. al.* 2019) e aceitável (SU; THIDA, 2018). A utilização de redes neurais parece ser um recurso importante para a previsão de pausas na leitura nesses sistemas (KLIMKOV *et.al.* 2017). Além disso, a detecção de limite prosódico deve ter um bom desempenho, independentemente de a segmentação em nível de palavra estar disponível ou não (KOCHAROV; KACHKOVSKAIA; SKRELIN, 2019).

Encontramos um único trabalho que investiga ambos os sistemas TTS e ASR. Utilizando o *Prosograph*, Öktem (2019) conseguiu prever três principais sinais de pontuação (ponto final, vírgula e ponto de interrogação), a partir dos recursos "palavra", "pausa interlexical" e

"altura(*pitch*)". As vírgulas foram detectadas geralmente com baixa precisão, o que mostra, novamente, a dificuldade de os sistemas ASR conseguirem prever esse sinal. O trabalho de Öktem (2019), no entanto, revelou que o *pitch* ajudou na detecção da vírgula.

De forma geral, para fazer a previsão e análise dos recursos prosódicos na síntese e no reconhecimento de fala, são empregadas redes neurais artificiais, que são sistemas que utilizam algoritmos, sendo definidos como códigos responsáveis pelo processamento e pela manipulação de variáveis ordenadas, podendo reconhecer padrões e estabelecer correlações em dados brutos (CUNHA, 2020).

Nesta revisão, os autores dos estudos desenvolveram as mais variadas redes neurais, sendo elas: (6) artigos que utilizaram RNN – Rede Neural Recorrente, sendo capaz de processar toda a sequência de dados e pontos específicos (SZASZÁK, 2019; LIU, LI e SONG, 2018; TÜNDIK e SZASZÁK, 2018; MORÓ e SZASZÁK, 2017a; SZASZÁK e TÜNDIK, 2017b; BALLESTEROS e WANNER, 2016); (2) *BiLSTM - Bi-directional Long Short-Term Memory*, constituída por duas camadas de rede *LSTM - Long Short-Term Memory* (conjunto de sub-redes ligadas por blocos de memória, contendo células que armazenam o estado temporal da rede), sendo uma camada responsável por analisar os estados passados e outra com os estados futuros (MORÓ e SZASZÁK, 2017b; CENCESCHI *et. al.*, 2019); (1) *Multi-Head Self-attencion* – arquitetura capaz de prever a partir de emprego de múltiplos fatores de entrada para o modelo (YI e TAO, 2019); (1) *Bark Wavelet Transform* – ferramenta para extrair parâmetros estatísticos dos sinais de fala em sub-bandas (LIU, LI e ZHAO, 2018); (1) *Multilayer Perceptron*, arquitetura que, além das camadas de entrada e saída, tem várias camadas ocultas no meio (OROSANU e JOUVET, 2018); (1) com uso de CART - *Classification-Regression Decision Trees*, classificador não paramétrico baseado em regras simples, *multi-class Discriminant Analysis* (Linear – LDA e Quadrática – QDA, *Naive Bayes classifier* (NB), utilizando modelos gaussianos de distribuição de recursos e *k-nearest Neighbors classifier* (kNN), classificador de aprendizado de máquina mais simples (IGRAS e ZIOŁKO, 2016) nos processos de reconhecimento de voz.

Quantos aos estudos sobre síntese de voz, foram empregadas as seguintes redes neurais: (1) estudo utilizou CART - *Classification-Regression Decision Trees* e BiLSTM - *Bi-directional Long Short-Term Memory* (KLIMKOV *et. al.*, 2017); (1) Classificador *Random Forest*, usado para detectar os limites prosódicos reais usando um pequeno conjunto de recursos acústicos (KOCHAROV *et. al.*, 2019); (1) SSML - *Speech Synthesis Markup Language*, que atribui uma variedade de tags de prosódia SSML com base na estrutura de temática de cada frase (DOMINGUEZ *et. al.*, 2019); (1) G2P - *English Grapheme to Phoneme Conversion* , utilizado para converter grafemas (ortografia) em fonemas (pronúncia) (SU e THIDA, 2018). Houve apenas uma pesquisa que realizou experimentos com reconhecimento de voz e síntese de fala, usando a RNN (ÖKTEM, 2019).

A previsão das redes neurais utilizadas em ambos os sistemas ainda não alcançou 100% de eficácia por apresentar limitações na previsão da pontuação e na síntese de fala a partir dos recursos prosódicos, pois nenhuma rede neural pode satisfazer completamente os processos de conversão de fala para texto ou de texto para fala. Isso pode estar relacionado às decisões metodológicas quanto ao tipo de rede neural e ao tipo de *corpus* (fala, texto escrito e lido).

As pistas prosódicas (pausa, intensidade e entonações) possibilitaram o reconhecimento de fala (ASR), a partir das redes neurais mencionadas (BALLESTEROS & WANNER, 2016; MORÓ & SZASZÁK, 2017a; MORÓ; SZASZÁK, 2017b; OROSANU; JOUVET, 2018; LIU, LIU & SONG, 2018; TÜNDIK & SZASZÁK, 2018; CENCESCHI *et. al.* 2019; SZASZÁK, 2019; SZASZÁK & TÜNDIK, 2019; YI & TAO, 2019).

Todos os parâmetros prosódicos descritos nos resultados variam a depender da intenção comunicativa do falante e são determinantes para dar fluidez e continuidade à fala (PACHECO, 2017). De forma geral, os estudos encontrados nas diferentes áreas de pesquisa (Linguística, Linguística Computacional e Tecnologia da Informação) apontaram a correspondência entre os sinais de pontuação e o emprego dos recursos prosódicos, inclusive para a previsão da pontuação no reconhecimento automático de voz e para a melhoria da síntese de fala.

4. CONCLUSÃO

Os estudos desta revisão confirmaram nossa pergunta de pesquisa, evidenciando a relação entre os sinais de pontuação e os aspectos prosódicos. A maioria dos trabalhos relacionados à tecnologia desenvolveu diferentes redes neurais para transformar texto em fala e/ou para converter fala em texto e mostrou que as pausas são apontadas como indicadores mais fortes dos sinais de pontuação.

Identificamos poucas pesquisas entre os sistemas TTS e os sinais de pontuação, fato que indica uma necessidade de pesquisas futuras, a fim de melhorar a qualidade das conversões de texto para fala.

Via de regra, os sistemas ASR são usados para reconhecer declarações faladas e convertê-las em texto (BOHOUTA, 2020), não para produzirem sequências de saída com sinais de pontuação. Entretanto, os sinais comprometem a legibilidade das transcrições de fala, sendo importante prever sinais de pontuação para transcrições de fala (YI *et al.*, 2017; YI; TAO, 2019).

A identificação das características prosódicas junto com o reconhecimento de fonemas ou sílabas são incorporadas pelos sistemas de reconhecimento da fala, sendo as informações prosódicas responsáveis por solucionar os problemas de ambiguidade

da linguagem, visto que é uma informação que revela a expressão (MÜLLER, 2006). O aspecto prosódico deve ser considerado no desenvolvimento desse tipo de sistema, pois ainda há problemas no reconhecimento dos diversos padrões prosódicos de uma língua (PAIXÃO, 2014).

REFERÊNCIAS

ALCÂNTARA, R. G. *Unidades microlinguísticas, relações entre oralidade e escrita e gêneros textuais: possibilidades de abordagens no ensino da língua portuguesa*. Dissertação (Mestrado em Educação). Universidade Federal do Espírito Santo, Vitória. 2006.

ALMEIDA, S. A.; FONSECA, A. A. A relação entre os sinais de pontuação e o processamento de leitura de alunos concluintes do Ensino Médio. *Signo*, Santa Cruz do Sul, v. 43, n. 77, p. 74-86, maio, 2018. ISSN 1982-2014. DOI http://dx.doi.org/10.17058/signo.v43i77.11534. Acesso em: 12 fev. 2021.

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, 1048-1053, 2016. DOI http://dx.doi.org/10.18653/v1/D16-1111. Acesso em: 8 fev. 2021.

BARHATE, S.; KSHIRSAGAR, S.; SANGHVI, N.; SABU, K.; RAO, P.; BONDALE, N. Prosodic features of Marathi news reading style. *IEEE* Region 10 Conference (TENCON), Singapore, pp. 2215-2218, 2016. DOI https://doi.org/10.1109/TENCON.2016.7848421. Acesso em: 12 fev. 2021.

BECHARA, E. *Moderna gramática portuguesa*. 37. ed. rev., ampl. e atual. conforme o novo Acordo Ortográfico. Rio de Janeiro: Nova Fronteira. 2009.

BOHOUTA, G. M. *Improving Wake-Up-Word and General Speech Recognition Systems.* Thesis (Ph.D.) - Florida Institute of Technology, 2020.

CAGLIARI, L. C. *Elementos de fonética do português brasileiro*. Campinas: Unicamp. Tese de (Livre-Docência), 1981.

CAGLIARI, L. C. Marcadores prosódicos na escrita. *In:* SEMINÁRIO DO GRUPO DE ESTUDOS LINGUÍSTICOS, 18, 1989, Lorena. *Anais do XVIII Seminário do Gel*. Lorena: Grupo de Estudos Linguísticos de São Paulo, p. 195-203, 1989.

CAPRISTANO, C. C. A propósito da escrita infantil: uma reflexão sobre as segmentações não convencionais. *Letras de Hoje*. Porto Alegre, v. 39, n. 3, p. 245-260, setembro, 2004.

CENCESCHI, S.; TEDESCO, R.; SBATTELLA, L; LUCHETTI, M. PESInet: Automatic Recognition of Italian Statements, Questions, and Exclamations With Neural Networks. *CLiC-it (2019)*. Disponível em: https://www.researchgate.net/publication/336848037_PESInet_Automatic_Recognition_of_Italian_Statements_Questions_and_Exclamations_With_Neural_Networks. Acesso em: 8 fev. 2021.

COSTA, R. A. S. *Implicações prosódicas da oralidade no uso da vírgula: uma interface entre fonologia e sintaxe*. (Dissertação de Mestrado) – Universidade Estadual de Campinas, Campinas, 2019.

CRYSTAL, D. *A Dictionary of Linguistics and Phonetics*. 6 ed. Oxford: Blackwell, 2008.

CUNHA, L. C. *Redes Neurais Convolucionais e Segmentação de Imagens [manuscrito]: Uma revisão bibliográfica.* Monografia (Bacharelado) - Universidade Federal de Ouro Preto. Escola de Minas, 2020.

DOMINGUEZ, M; BURGA, A.; FARRÚS, M.; WANNER, L. Towards expressive prosody generation in TTS for reading aloud applications. *Proc. IberSPEECH*, p. 40-44, 2018. DOI http://dx.doi.org/10.21437/IberSPEECH.2018-9. Acesso em: 25 maio 2021.

DUTOIT, T. *An introduction to text-to-speech synthesis*. [S.l.: s.n.], 1997.

FUCHS, S.; KRIVOKAPIC, J. Prosodic Boundaries in Writing: Evidence from a Keystroke Analysis. *Frontiers in Psychology.* Volume 7, 2016. DOI https://doi.org/10.3389/fpsyg.2016.01678. Acesso em: 05 fev. 2021.

HEGGIE, L.; WADE-WOOLLEY, L. Prosodic Awareness and punctuation ability in adult readers. *Reading Psychology*, v. 39, p. 188–215, 2018. Disponível em: https://doi.org/10.1080/02702711.2017.1413021. Acesso em: 05 fev. 2021.

IGRAS-CYBULSKA, M.; ZIÓŁKO, B.; ŻELASKO, P.; WITKOWSKI, M. Structure of pauses in speech in the context of speaker verification and classification of speech type. *Journal on Audio, Speech, and Music Processing*. 2016. DOI https://doi.org/10.1186/s13636-016-0096-7. Acesso em: 05 fev. 2021.

IGRAS, M.; ZIÓŁKO, B. Detection of Sentence Boundaries in Polish Based on Acoustic Cues. *Archives of Acoustics*. Volume 41, Number 2, 2016. DOI http://dx.doi.org/10.1515/aoa-2016-0023. Acesso em: 05 fev. 2021.

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. DOI http://dx.doi.org/10.21437/Interspeech.2017-419. Acesso em: 25 maio 2021.

KOCHAROV, D.; KACHKOVSKAIA, T.; SKRELIN, P. Prosodic boundary detection using syntactic and acoustic information. *Computer Speech & Language*, v. 53, p. 231-241, 2019. DOI https://doi.org/10.1016/j.csl.2018.07.001. Acesso em: 26 maio 2021.

LIU, J.; LI, C.; ZHAO, L. Recognition of Punctuation in Voiced And Unvoiced Speech For Ib-CET. *AECT Convention Proceedings*. Volume 2, p. 335-341, 2018. Disponível em: https://members.aect.org/pdf/Proceedings/proceedings18/2018i/18_13.pdf. Acesso em: 08 fev. 2021.

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. DOI https://doi.org/10.1109/IALP.2018.8629143. Acesso em: 25 maio 2021.

MATEUS, M. H. Estudando a melodia da fala - traços prosódicos e constituintes prosódicos. Palavras - *Revista da Associação de Professores de Português*, n. 28, p. 79-98. Estudando a melodia da fala: traços prosódicos e constituintes prosódicos. In: O ENSINO DAS LÍNGUAS E A LINGUÍSTICA. ENCONTRO DA APL E ESE de SETÚBAL, Lisboa, 2004.

MARTIN, F. A.; MALFAZ, M., CASTRO-GONZÁLEZ, A.; CASTILHO, J. C.; SALICHS, M. A. Four-Features Evaluation of Text to Speech Systems for Three Social Robots. *Electronics 2020*, *9* (2), 267 (Madrid), Spain, 2020. DOI https://doi.org/10.3390/electronics9020267. Acesso em: 02 abr. 2021.

MOREIRA, N. C. R. Leitura: Problemas de processamento. *Revista de Letras*, Fortaleza, 8 (1) – jan./jun. 1985. Disponível em: http://www.periodicos.ufc.br/revletras/article/view/19448. Acesso em: 14 jul. 2021.

MORÓ, A.; SZASZÁK, G. A prosody inspired RNN approach for punctuation of machine produced speech transcripts to improve human readability. 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), *IEEE*, p. 000219-000224, 2017a. DOI https://doi.org/10.1109/CogInfoCom.2017.8268246. Acesso em: 26 maio 2021.

MORÓ, A.; SZASZÁK, G. A Phonological Phrase Sequence Modelling Approach for Resource Efficient and Robust Real-Time Punctuation Recovery. *Proc. Interspeech*, p. 558-562, 2017b. DOI http://dx.doi.org/10.21437/Interspeech.2017-204. Acesso em: 26 maio 2021.

MÜLLER, D. N. *COMFALA - Modelo Computacional do Processo de Compreensão da Fala*. Tese (Doutorado em Ciência da Computação). 130 p. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

NESPOR, M.; VOGEL, L. *Prosodic Phonology*. Dordrecht, Netherlands: Foris Publications, 1986.

ÖKTEM, A. *Incorporating prosody into neural speech processing pipelines: applications on automatic speech transcription and spoken language machine translation.* Tese. Universitat Pompeu Fabra. Departament de Tecnologies de la Informació i les Comunicacions, 2019.

OROSANU, L.; JOUVET, D. Detection of Sentence Modality on French Automatic Speech-to-text Transcriptions, *Procedia Computer Science*, v. 128, p. 38-46, 2018. DOI https://doi.org/10.1016/j.procs.2018.03.006. Acesso em: 12 fev. 2021.

PACHECO, V. Escrita, prosódia e leitura. *In:* FREITAG, R. M. KO.; LUCENTE, L. (Orgs.) *Prosódia da Fala: Pesquisa e Ensino*. São Paulo: Blucher, 1ª edição, p. 103-115, 2017.

PACHECO, V. *Investigação fonético-acústico e experimental dos sinais de pontuação enquanto marcadores prosódicos*. Dissertação (Mestrado em Linguística). Instituto de Estudos da Linguagem, Unicamp, Campinas, 2003.

PAIXÃO, V. B. *A prosódia das interrogativas totais na fala carioca: fala espontânea versus leitura*. Dissertação (Mestrado em Letras Vernáculas). Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

PAIXÃO, V. B.; SERRA, C. R. Fraseamento prosódico de estruturas parentéticas em dados de leitura no Português do Rio de Janeiro. *Working Papers em Linguística*, Volume 19 (2): 113-135, Florianópolis, ago./dez., 2018. DOI https://doi.org/10.5007/1984-8420.2018v19n2p113. Acesso em: 8 fev. 2021.

PEIRÓ-LILJA, A.; FARRÚS, M. Naturalness enhancement with linguistic information in end-to-end TTS using unsupervised parallel encoding. In: *Proceedings of Interspeech 2020*, 2020 Oct 25-29; Shanghai, China. DOI http://dx.doi.org/10.21437/Interspeech.2020-1788. Acesso em: 2 abr. 2021.

RAJESWARI, K. C; MAHESWARI, P. U. Prosody Modeling Techniques for Text-to-Speech - Synthesis Systems - A Survey. *International Journal of Computer Applications* (0975 – 8887) Volume 39– No.16, 2012. DOI http://dx.doi.org/10.5120/4902-7399. Acesso em: 02 abr. 2021.

RAMU REDDY, V.; SREENIVASA RAO, K. Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks, *Neurocomputing,* v. 171, p. 1323-1334, 2016.

SANTOS, A. J.; PACHECO, V.; OLIVEIRA, M. S. O papel dos marcadores prosódicos na fluência de leitura / The role of prosodic markers in reading fluency. *Revista de Estudos da Linguagem*, [S.l.], v. 27, n. 3, p. 1417-1457, july 2019. ISSN 2237-2083. DOI http://dx.doi.org/10.17851/2237-2083.27.3.1417-1457. Acesso em: 12 fev. 2021.

SILVA, E. W. R. *A relação entre produção e percepção de pistas prosódicas na segmentação de narrativas espontâneas*. Dissertação (Mestrado em Linguística), Universidade Federal de Alagoas, Faculdade de Letras, Maceió, 2017.

SONCIN, G.; TENANI, L. Evidence of the role of prosody in argumentative writing: Comma use in texts written by Brazilian students aged 11–14. *Writing & Pedagogy*. Volume 9, p. 77-101, 2017. DOI http://dx.doi.org/10.1558/wap.26498. Acesso em: 12 fev. 2021.

SU, C.; THIDA, A. Phoneme based Myanmar text to speech system. International. *Journal of Advanced Computer Research*. v. 8. p. 47-58, 2018. DOI http://dx.doi.org/10.19101/IJACR.2017.733036. Acesso em: 21 maio 2021.

SZASZÁK, G. An Audio-based Sequential Punctuation Model for ASR and its Effect on Human Readability. *Acta Polytechnica*, Hungarica Vol. 16, No. 2, 2019. DOI https://doi.org/10.12700/aph.16.2.2019.2.6. Acesso em: 12 fev. 2021.

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. DOI http://dx.doi.org/10.21437/Interspeech.2019-2132. Acesso em: 25 maio 2021.

TEIXEIRA, J. P. R. *Modelização paramétrica de sinais para aplicação em sistemas de conversão texto-fala*. Dissertação (Mestrado em Engenharia). Universidade do Porto, Porto, 1995.

TÜNDIK, M. Á.; SZASZÁK, G. Joint Word- and Character-level Embedding CNN-RNN Models for Punctuation Restoration, *IEEE* International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 2018, pp. 000135-000140, DOI https://doi.org/10.1109/CogInfoCom.2018.8639876. Acesso em: 08 fev. 2021.

YI, J.; TAO, J.; WEN, Z.; LI, Y. Distilling Knowledge from an Ensemble of Models for Punctuation Prediction. *INTERSPEECH*, 2017.

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, 2019, pp. 7270-7274. DOI https://doi.org/10.1109/ICASSP.2019.8682260. Acesso em: 08 fev. 2021.

YOUNIS, N. When Punctuation Marks are not so 'Punctual': Translating al-waqf Marks at the Prosodic-Orthographic Interface. *Occasional Papers in the Development of English Education*. Volume 63, p. 451-484, 2017. DOI https://dx.doi.org/10.21608/opde.2017.88216. Acesso em 08 fev. 2021.