Tecnologias de Reconhecimento de Fala: uma revisão sistemática de trabalhos no Brasil

Alexandre Henrique Kavalerski Teixeira[1], Ian Macedo Maiwald Santos[1], Jhemeson Silva Mota[1], Jackson Gomes de Souza[1]

[1]Centro Universitário Luterano de Palmas (CEULP/ULBRA) – Palmas, TO – Brasil

{ianmaiscedo,kavalerskialexandre,jhemesonmotta,jackson.souza}@gmail.com

Resumo. *Tecnologias de reconhecimento de fala - também denominada como reconhecimento de voz - possibilitam que aparelhos eletrônicos equipados com microfones possam interpretar a fala humana. Combinando as áreas de conhecimento da linguística e ciência da computação tem-se uma alternativa para a comunicação entre homem e máquina. O presente trabalho apresenta uma revisão sistemática de trabalhos brasileiros relacionados ao reconhecimento de fala.*

1. Introdução

A evolução no desempenho dos computadores por conta de processadores mais velozes acarretou em uma melhora gradual em técnicas e tecnologias relacionadas ao uso da voz. Segundo Cardoso et al. (2010), a utilização de tecnologias de reconhecimento de voz tornaram-se uma opção significativamente vantajosas. Entre diversas tecnologias de voz, Oliveira et al. (2014) listam síntese de voz (ou TTS de *text-to-speech*) e reconhecimento automático de voz (ou ASR, de *automatic speech recognition*) como os mais relevantes.

Ainda segundo os autores, um sistema TTS é composto por partes que transformam textos - em linguagem natural - em voz sintetizada, enquanto um sistema ASR realiza o processo inverso, ou seja, o sinal digitalizado de voz é convertido em texto. Este trabalho tem como foco principal os sistemas de reconhecimento automático de voz.

Yu e Deng (2014) afirmam que ASR é uma tecnologia de grande importância na área do reconhecimento de fala, pois permite e auxilia as interações homem-homem e homem-computador. Ainda para os autores, no passado, a fala não era uma modalidade viável na comunicação homem-computador e isso se deve parcialmente ao fato de que a tecnologia da época não era adequada o suficiente para uma situação de uso real. Outra parte é devido à superioridade de outros meios como *mouse* e teclado que eram mais eficientes e precisos.

Cardoso et al. (2010), dizem que as técnicas mais utilizadas para o reconhecimento de voz atualmente são as redes neurais artificiais, o modelo oculto de Markov (HMM), o modelo híbrido e por audiovisual.

Segundo Cardoso et al. (2010), os sistemas ASR têm várias aplicações práticas, como acionamento de dispositivos em automóveis, atendimento telefônico para solicitação de serviços, programas utilitários em computadores, brinquedos e celulares. O autor reforça que o objetivo desses sistemas é reconhecer a mensagem contida na fala humana e realizar uma ação previamente programada - que pode ser a transcrição do texto - em resposta a entrada.

De acordo com Torres et al. (2016), diversas aplicações que possuem uma interface adaptada para o uso de um reconhecedor de voz começaram a influenciar o modo de interação das pessoas com dispositivos. Neto et al. (2005) reforçam acerca das possibilidades de integração de um sistema ASR com outros sistemas ou módulos, como o de tutores inteligentes, processamento de linguagem natural (PLN), gerenciamento de diálogos, entre outros. Tais sistemas, ainda de acordo com os autores, recebem o resultado obtido pelo ASR, e o uso ou não deles depende da finalidade da aplicação. Assistentes virtuais, tradutores e

ISSN: 2447-0767

ferramentas de busca são algumas das aplicações que integram essa área (TORRES et al., 2016).

Portanto, entende-se que o levantamento de estudos, trabalhos e artigos científicos relacionados às tecnologias de reconhecimento de fala em português é de suma importância para o entendimento do estado da arte de tal área de estudo.

2. Fundamentação Teórica

2.1. *Large Vocabulary Continuous Speech Recognition (*LVCSR)

Bahdanau et al. (2016) diz que trabalhos recentes relacionados ao tema de sistemas avançados de Reconhecimento de Voz e Grande Vocabulário (LVCSR, de *large-vocabulary continuous speech recognition*) têm mostrado resultados promissores. Como exemplo, podemos citar o de Graves et al. (2006) que projetou um sistema de LVCSR que utiliza um modelo de rede neural treinado com Classificação Temporal Conexionista, Roque (2018) afirma que esta classificação é um campo que se apóia nas chamadas redes neurais artificiais como veículo para a expressão matemática dos seus conceitos e teorias.

Segundo Monaco (2017) a maioria dos sistemas LVCSR atuais são híbridos de Redes Neurais com modelos ocultos de Markov e contém componentes modularizados que lidam com diferentes áreas como: modelagem acústica, modelagem de linguagem e decodificação de sequência.

Bahdanau et al. (2016) constata que há uma nova direção de pesquisa relacionada a redes neurais que lida com modelos que aprendem a concentrar sua atenção em partes específicas de sua contribuição. O autor completa que tais sistemas mostram resultados promissores em uma vasta gama de tarefas, incluindo tradução automática, geração de legendas, síntese de caligrafia, classificação de objetos visuais e reconhecimento de fonemas.

2.2. Modelo de Linguagem

Segundo Silva et al. (2004), o modelo de linguagem é uma parte fundamental em muitos sistemas computacionais, por exemplo para o reconhecimento de voz. Ainda para os autores, um modelo estatístico de linguagem fornece a probabilidade de uma cadeia de palavras, comumente chamada de sentença.

Neto et al. (2005) afirmam que não é viável estimar as probabilidades P de uma sequência de palavras, mesmo para uma quantidade moderada de palavras. Por isso utiliza-se o chamado modelo N-grama, que assume uma palavra w_i a partir das N-1 prévias palavras. No caso de N = 3 tem-se um trigrama $P(w_i \mid w_{i-2}, w_{i-1})$, caso N = 2 são os bigramas $P(w_i \mid w_{i-1})$ e N = 1 é um unigrama $P(w_i)$.

A probabilidade para o último símbolo da sentença será considerada ao final da sentença como outra palavra, enquanto o começo desta sentença é tratado somente como uma informação do contexto (SILVA et al., 2004). Os autores ainda ressaltam que para a criar o modelo de linguagem é importante encontrar boas estimativas para probabilidades vinculadas a cada contexto.

2.3. Modelos Ocultos de Markov

Uma cadeia de Markov é um cenário específico de um processo estocástico em um espaço discreto (LEITE, 2008). O autor ressalta que um processo estocástico é uma série de variáveis aleatórias que representam o comportamento de um sistema ao longo do tempo e - mesmo que

o estado inicial seja conhecido - não há como saber de que modo o processo vai evoluir. A Figura 1 representa uma cadeia de Markov com dois estados.

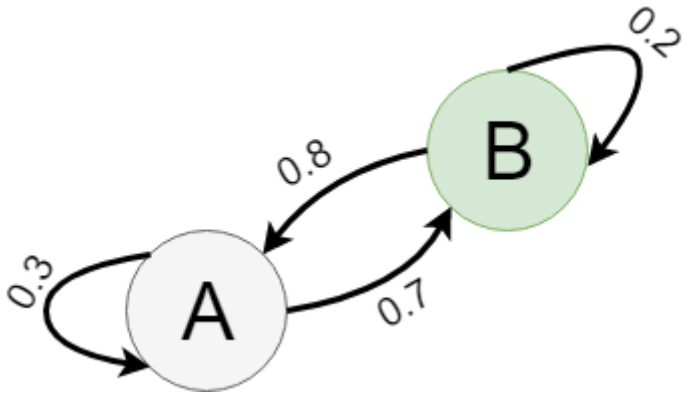

Figura 1. Cadeia de Markov. Os números indicam a probabilidade de alteração de estado.

Espindola (2009) afirma que os primeiros artigos relacionados aos Modelos Ocultos de Markov foram publicados ao final da década de 1960 e as primeiras aplicações para essa modelagem estão associadas ao reconhecimento de fala. Posteriormente, na década de 80, HMM foi utilizado para sequenciamento de DNA.

Em um modelo de Markov cada estado representa um evento observável e isso acaba restringindo as possibilidades de uso para este modelo (LEITE, 2008). Desta forma, Cappé, Moulines e Ryden (2006) destacam que no HMM o modelo de Markov está escondido, isto é, os estados não podem ser observados.

A principal diferença entre HMM e outros formalismos Markovianos está na forma no qual o sistema é observado. Uma vez que a maioria dos processos Markovianos são observados diretamente, já que os próprios estados podem ser observados, em HMM a observação é feita por inferência, indiretamente, pois os observáveis são funções probabilísticas dos estados ou das transições entre estados (ESPINDOLA, 2009).

O Modelo Oculto de Markov também pode ser encontrado com os mais diversos nomes, sendo eles: Cadeias Ocultas de Markov (*Hidden Markov Chains*), Processos Ocultos de Markov (*Hidden Markov Processes*), Fontes Markovianas (*Markov Sources*), Funções Probabilísticas de Cadeias de Markov (*Probabilistic Functions of Markov Chains*).

3. Metodologia

Visando reunir evidências e estudos relacionados às tecnologias elencadas como fonte de observação, bem como identificar possíveis sugestões de investigação e/ou lacunas na presente pesquisa, foi proposta uma revisão sistemática do tipo mapeamento sistemático como objetivo deste trabalho. A revisão foi produzida pelo três autores do trabalho sob supervisão do orientador.

3.1 Revisão Sistemática

A revisão sistemática da literatura, do tipo mapeamento sistemático neste trabalho, tem o objetivo de agrupar os principais estudos e tecnologias de reconhecimento de voz produzidos em português. É válido ressaltar que apenas o idioma em que o trabalho foi escrito foi considerado.

3.2 Questão de pesquisa

A pesquisa objetiva responder aos seguintes questionamentos:

ISSN: 2447-0767

- Quais os estudos e tecnologias de reconhecimento automático de voz produzidos em língua portuguesa?
- Quais as técnicas de Inteligência Artificial têm sido utilizadas para reconhecimento automático de voz?
- Qual o estado da arte do reconhecimento automático de voz no Brasil?

3.3 Estratégia de busca e bases de dados

Pesquisas científicas foram realizadas entre março e abril de 2018, nas quais foram realizadas consultas a periódicos encontrados através dos agregadores de bases de dados Scielo, Capes e Google Acadêmico. Em uma busca preliminar realizada nestes repositórios o trabalho de Duarte (2014) foi encontrado

As buscas nos bancos de dados foram realizadas utilizando os seguintes termos como palavras-chave: "reconhecimento de voz", "reconhecimento de fala", "*automatic speech recognition*", "ASR", "Inteligência Artificial", "IA", "revisão sistemática", "estado da arte" e "computação".

A partir dos termos listados anteriormente, foi produzida a seguinte string de busca, com as seguintes combinações das palavras-chave:

("reconhecimento de voz" AND "Inteligência Artificial") OR

("reconhecimento de voz" AND "IA") OR

("reconhecimento de voz" AND "Computação") OR

("reconhecimento de fala" AND "Inteligência Artificial") OR

("reconhecimento de fala" AND "IA") OR

("reconhecimento de fala" AND "Computação") OR

("*automatic speech recognition"*) OR

("revisão sistemática" AND "reconhecimento de voz") OR

("revisão sistemática" AND "reconhecimento de fala") OR

("estado da arte" AND "reconhecimento de voz") OR

("estado da arte" AND "reconhecimento de fala")

3.5 Critérios de seleção de trabalhos

Com o objetivo de permitir a seleção sistematizada de artigos, alguns critérios de inclusão e exclusão foram definidos. O objetivo destes critérios é selecionar somente trabalhos que empregam tecnologias de reconhecimento de fala para o idioma português brasileiro. Além disto, foram selecionados apenas trabalhos produzidos em língua portuguesa que tratem de assuntos relacionados à Ciência da Computação.

4. Resultados

Seguindo os critérios e procedimentos descritos na seção anterior, foram encontrados oitocentos e três trabalhos, dos quais apenas três atenderam aos parâmetros estabelecidos. Um diagrama exemplificando os totais de trabalhos, bem como as respectivas fontes pode ser visto na Figura 2.

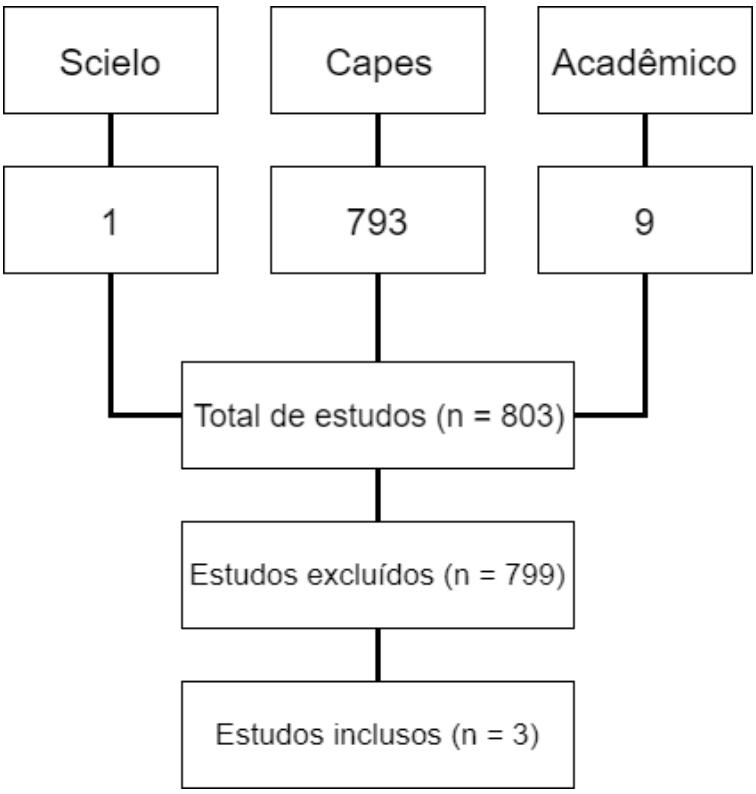

Figura 2. Diagrama com resultados de busca obtidos

Destaca-se a discrepância nos totalizadores de trabalhos encontrados e incluídos devido ao objetivo de filtrar somente os que fossem produzidos com o idioma português, conforme descrito na seção anterior. Na plataforma Capes, por exemplo, ao buscar por "*automatic speech recognition*" sem adicionar filtros, 11.063 resultados foram retornados, no entanto, quase todos estes estavam escritos na língua inglesa e após filtrar para português, 21 estudos foram retornados.

A partir dos trabalhos selecionados, foi feito um estudo destes visando identificar as técnicas citadas pelos autores. O resultado deste estudo foi elencado e pode ser visto na Tabela 1, logo a seguir.

Tabela 1. Verificações das técnicas citadas nos trabalhos

Trabalhos	**Técnica / Software**		
	Hidden Markov Model Toolkit (HTK)	Microsoft Internet Explorer add-in	*Mel Frequency Cepstral Coeficients* (MFCC)
Alencar e Alcaim (2008)	X		
Nguessan, Pardini e Martini (2017)		X	
Silva, Fernandes e Castro (2015)			X

ISSN: 2447-0767

Como pode ser observado, o artigo de Nguessan, Pardini e Martini (2017) não afirma diretamente qual(is) técnica(s) foi(foram) utilizada(s) para o reconhecimento de voz. No entanto, os autores dizem que, no sistema desenvolvido e descrito em seu estudo, o *software* Microsoft Internet Explorer add-in foi utilizado na etapa do reconhecimento de fala.

Além disto, é possível elencar os artigos de acordo com sua finalidade. Uma tabela com a associação entre os artigos e suas respectivas finalidades pode ser vista na Tabela 2.

Tabela 2. Associação entre artigos e finalidades

Artigo	**Finalidade**
Alencar e Alcaim (2008)	Avaliar o desempenho de diversos atributos para reconhecimento automático de voz obtidos de parâmetros LSF e parâmetros LPC
Nguessan, Pardini e Martini (2017)	Analisar um sistema de mobilidade para deficiente visual na sua rota de caminhada até o local onde ele queira
Silva, Fernandes Castro (2015)	Apresentar uma visão geral acerca de Verificação de Locutor Independente de Texto, demonstrando o funcionamento básico de um sistema baseado na aplicação do método da fusão de escores

Como pode ser observado, há uma variação nas finalidades dos trabalhos encontrados, tendo trabalhos que vão desde uma busca de aperfeiçoamento de algoritmos e técnicas, como nos trabalhos de Alencar e Alcaim (2008) e Silva, Fernandes e Castro (2015); até o trabalho de Nguessan, Pardini e Martini (2017), o qual trata-se de uma proposta de sistema que visa auxiliar o processo de mobilidade para deficientes visuais, em que, um dos módulos destes sistema utiliza reconhecimento de voz.

6. Considerações Finais

Ao longo deste trabalho, alguns elementos relacionados à área de reconhecimento de fala foram abordados. Além disto, foi feita uma revisão sistemática - que por sua vez, teve os procedimentos descritos - com o objetivo de identificar o estado da arte dos trabalhos produzidos em português.

A inteligência artificial já vem apresentando melhorias nos métodos de reconhecimento de fala, sendo tal progresso de grande importância a diversos setores, como saúde e acessibilidade, além de auxiliar no processo de automatização de tarefas. Todavia, os conteúdos relacionados ao reconhecimento automático de fala escritos no idioma português são demasiado escassos.

Além disso, os poucos trabalhos encontrados, possuem focos diferentes, demonstrando a gama de possibilidades de implementação das tecnologias descritas. Para isto, foi exemplificada a relação entre os trabalhos relacionados e suas finalidades. Como conclusão, pode ser questionada a escassez de trabalhos nesta área produzidos na língua portuguesa.

Referências

ALENCAR, Vladimir F. S. de; ALCAIM, Abraham. **Atributos eficientes em reconhecimento automático de voz distribuído**. Sba: Controle & Automação Sociedade Brasileira de Automatica, [s.l.], v. 19, n. 2, p.147-154, jun. 2008. FapUNIFESP (SciELO). http://dx.doi.org/10.1590/s0103-17592008000200004.

CAPPÉ, Olivier; MOULINES, Eric; RYDEN, Tobias. **Inference in Hidden Markov Models**. [s.l.]: Springer Science & Business Media, 2006. 653 p. Disponível em:

<https://books.google.com.br/books?id=4d_oEYn8Fl0C&printsec=frontcover&hl=pt-BR#v=onepage&q&f=true>. Acesso em: 30 abr. 2018.

CARDOSO, SERGIO A. et al. Sesame: sistema de reconhecimento de comandos de voz utilizando pds e rna. In: **XVIII Congresso Brasileiro de Automática**. 2010. p. 1316-1323.

ESPINDOLA, Luciana da Silveira. **Um Estudo sobre Modelos Ocultos de Markov HMM:** Hidden Markov Model. Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul, 2009. 33 p. Disponível em: <http://www.inf.pucrs.br/peg/pub/tr/TI1_Luciana.pdf>. Acesso em: 27 abr. 2018.

LEITE, Paula Beatriz Cerqueira. **Identificação de Tipos de Culturas Agrícolas a partir de Seqüências de Imagens Multitemporais Utilizando Modelos de Markov Ocultos**. 2008. 79 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008. Disponível em: <https://www.maxwell.vrac.puc-rio.br/12960/12960_4.PDF>. Acesso em: 28 abr. 2018.

NETO, N. et al. **Desenvolvimento de Software Livre Usando Reconhecimento e Síntese de Voz:** O Estado da Arte para o Português Brasileiro. Trilha Nacional de Workshop Software Livre, [s.l.], p. 7, 2005. ISBN: 1595932240, DOI: 10.1145/1111360.1111396.

NGUESSAN, Desire; PARDINI, Bruno; MARTINI, Sidney. Um sistema baseado nas NTICs para auxílio aos deficientes visuais em sua caminhada por locais desconhecidos. **Research, Society And Development**, [s.l.], v. 4, n. 2, p.90-101, fev. 2017

OLIVEIRA, R. et al. **Recursos para desenvolvimento de aplicativos com suporte a reconhecimento de voz para desktop e sistemas embarcados**. 12th International Conference on Computational Processing of the Portuguese Language (PROPOR), [s.l.], no December 2014, p. 6, 2011.

SILVA, Ênio et al. Modelos de Linguagem N-grama para Reconhecimento de Voz com Grande Vocabulário. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E LINGUAGEM HUMANA, 2., 2004, Salvador. **Anais**... Salvador: Til, 2004. p. 82 - 90. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/til/2004/009.pdf>. Acesso em: 30 abr. 2018.

SILVA, Mayara Ferreira da; FERNANDES, Dênis; CASTRO, Maria Cristina Felippeto de. Fonética Forense: o uso da fusão de escores para verificação de locutor independente de texto. **Revista Brasileira de Criminalística**, [s.l.], v. 4, n. 2, p.33-38, 31 ago. 2015. Associacao Brasileira de Criminalistica - ABC. http://dx.doi.org/10.15260/rbc.v4i2.92. Disponível em: <http://rbc.org.br/ojs/index.php/rbc/article/view/92/pdf_34>. Acesso em: 29 abr. 2018.

TEVAH, R. **Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro**. [s.l.], p. 102, 2006.

TORRES, Elaynne da S. et al. Redes Neurais Convolucionais no Reconhecimento de Fala em Português para jogos em plataformas móveis. In: SIMPÓSIO BRASILEIRO DE JOGOS E ENTRETENIMENTO DIGITAL, 15., 2016, São Paulo. **Proceedings of SBGames 2016.** São Paulo: Sbc, 2016. p. 246 - 249. Disponível em: <http://www.sbgames.org/sbgames2016/downloads/anais/157742.pdf>. Acesso em: 3 mar. 2018.

YU, Dong; DENG, Li. **Automatic Speech Recognition:** A Deep Learning Approach. Redmond: Springer, 2014. 321 p. Disponível em: <https://books.google.com.br/books?id=rUBTBQAAQBAJ&printsec=frontcover&hl=pt-BR#v=onepage&q&f=true>. Acesso em: 3 mar. 2018.

ISSN: 2447-0767

WOODLAND, Philip C. et al. Large vocabulary continuous speech recognition using HTK. In: **Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on**. Ieee, 1994. p. II/125-II/128 vol. 2.

MONACO, Juliana. **Deep Learning:** Discurso/Reconhecimento de Fala. 2017. Disponível em: <http://www.semantix.com.br/deep-learning-reconhecimento-de-fala/>. Acesso em: 18 abr. 2018.

ROQUE, Antonio. **Conexionismo e Redes Neurais**. Disponível em: <sisne.org/Disciplinas/PosGrad/PsicoConex/aula1.pdf>. Acesso em: 18 abr. 2018.

Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In ICML-06.