INOVAÇÕES EM TÉCNICAS DE RECONHECIMENTO AUTOMÁTICO DE FALA APLICADAS A SISTEMAS ELÉTRICOS DE POTÊNCIA

Data de submissão: 09/06/2023 *Data de aceite: 03/07/2023*

Ivan Nunes da Silva

Universidade de São Paulo (USP/EESC/SEL)
São Carlos – SP
http://lattes.cnpq.br/0448891472280429

Sofia Moreira de Andrade Lopes

Universidade de São Paulo (USP/EESC/SEL)
São Carlos – SP
http://lattes.cnpq.br/1277390036356439

Victor Hideki Yoshizumi

Universidade de São Paulo (USP/EESC/SEL)
São Carlos – SP
http://lattes.cnpq.br/8378904835843389

Rogério Andrade Flauzino

Universidade de São Paulo (USP/EESC/SEL)
São Carlos – SP
http://lattes.cnpq.br/ 4487681434814567

Danilo Hernane Spatti

Universidade de São Paulo (USP/EESC/SEL)
São Carlos – SP
http://lattes.cnpq.br/7371885828178292

Ivan Gidaro Ricci

ARGO Transmissão de Energia S/A
São Paulo – SP
http://lattes.cnpq.br/6397604500437134

Alexandre Gerber Choupina Latorre

ARGO Transmissão de Energia S/A
São Paulo – SP
http://lattes.cnpq.br/1515823301911325

Ana Cláudia Carvalho Barquete

ARGO Transmissão de Energia S/A
São Paulo – SP
http://lattes.cnpq.br/0533920838682298

Rafael de Oliveira Fernandes

ARGO Transmissão de Energia S/A
São Paulo – SP
http://lattes.cnpq.br/8600035078363689

Pedro Hamilton de Sousa

ARGO Transmissão de Energia S/A
São Paulo – SP
http://lattes.cnpq.br/0293405061704436

RESUMO: Sistemas de reconhecimento automático de fala têm grande importância na literatura, sendo aplicados em diversas áreas, como medicina e cibersegurança. As técnicas utilizadas para compor tais sistemas têm sido também aprimoradas,

de forma que o uso de técnicas de *deep learning* tem se tornado cada vez mais popular. A massiva aplicação deste tipo de sistema para a solução de problemas em diversas áreas tem relação com a sua capacidade de identificar, reconhecer e extrair informações da fala, sendo este o principal modo de comunicação do ser humano. Todavia, a aplicação deste tipo de sistema e de todas as suas potencialidades ainda é tímida no contexto de sistemas elétricos de potência. Assim, este trabalho tem como objetivo reunir e analisar os estudos inovadores e proeminentes na área de sistemas elétricos de potência que utilizam técnicas de reconhecimento automático de fala. Os trabalhos foram analisados de forma comparativa e suas principais características foram detalhadas. Espera-se que este trabalho de investigação sirva de ferramenta para estimular o desenvolvimento de novos estudos na área.
PALAVRAS-CHAVE: Reconhecimento de fala, processamento de fala, sistemas elétricos de potência, aprendizagem de máquinas.

INNOVATIONS IN AUTOMATIC SPEECH RECOGNITION TECHNIQUES APPLIED TO ELECTRICAL POWER SYSTEMS

ABSTRACT: Automatic speech recognition systems have great importance in the literature, being applied in several areas, such as medicine and cybersecurity. The techniques used to compose such systems have also been improved, so that the use of deep learning techniques has become increasingly popular. The massive application of this type of system for solving problems in several areas is related to its ability to identify, recognize and extract information from speech, the main mode of communication of human beings. However, the application of this type of system and all its potential is still timid in the context of electrical power systems. Thus, this work aims to gather and analyze innovative and prominent studies in the area of electrical power systems using automatic speech recognition techniques. The works were analyzed in a comparative way and their main characteristics were detailed. It is expected that this investigation work will serve as a tool to stimulate the development of new studies in the area.
KEYWORDS: Speech recognition, speech processing, electric power systems, machine learning.

1 | INTRODUÇÃO

Sistemas de Reconhecimento Automático de Fala (*Automatic Speech Recognition* – ASR) se baseiam no processo de captação de um sinal de áudio, proveniente da fala do interlocutor, na identificação e reconhecimento das palavras enunciadas e na extração de informações para o uso posterior, a fim de realizar alguma ação ou atividade. Este tipo de sistema é popular em diversas aplicações, pois a fala é a forma de comunicação mais natural e eficiente realizada entre os seres humanos. Usualmente, o processo de identificação e reconhecimento de palavras realizado pelos sistemas se baseia na tecnologia de transcrição de fala para texto (*speech-to-text*), sendo este outro tema frequente de estudos (MALIK *et al.*, 2021).

Devido à sua importância, diversos trabalhos foram desenvolvidos propondo a aplicação de sistemas de ASR nas mais diversas áreas de conhecimento. Artigos de

revisão sobre o tema também são comuns, tais como Malik *et al*. (2021), Vadwala *et al*. (2017) e Oshikawa *et al*. (2018). Durante as buscas nas principais plataformas científicas disponíveis, foram então encontrados diversos artigos na área de reconhecimento de voz utilizando as palavras chave "*speech recognition*" e "*speech-to-text*". Usualmente, tais artigos estão relacionados com aplicações em medicina (NEDJAH *et al*., 2022; ALNASSER e AL-GHOWINEM, 2019), tradução (VENKATASUBRAMANIAN e MOHANKUMAR, 2022; ZHANG, 2022), jogos online (LELARDEUX *et al*., 2017), segurança cibernética (PÉREZ *et al*, 2021), entre outros. Todavia, nota-se uma carência de artigos que utilizam a tecnologia de reconhecimento de voz em aplicações de Sistemas Elétricos de Potência (SEP).

Embora existam diversos problemas que podem ser abordados utilizando sistemas de ASR no contexto de SEP, poucos estudos apresentam destaque na área, sendo a maioria desenvolvida por grupos de pesquisa chineses (JIANGPING *et al*., 2021; ZHANG *et al*., 2021; YU *et al*., 2020; LI *et al*., 2020; LI *et al*., 2019; ZHANG *et al*., 2019). No Brasil a permissão para o uso de teleassistência no contexto do sistema interligado nacional foi permitida a partir de 2019, pela resolução normativa nº 864 publicada pela Agência Nacional de Energia Elétrica (ANEEL); contudo, tal normativa foi revogada e substituída pela resolução Nº 1.005, de 15 de fevereiro de 2022 (ANEEL, 2022). Com esta mudança regulatória, algumas pesquisas foram então desenvolvidas, mas a maior parte dos desenvolvimentos no país ainda se encontra em estágios iniciais de pesquisa (SOUZA *et al*., 2021; JORGE *et al*., 2010).

No contexto de SEP, inúmeras aplicações para ASR podem ser verificadas. As atividades que devem ser realizadas pelos operadores nos centros de operação são normalmente repassadas para eles por meio de comandos de voz via chamadas telefônicas, como a comunicação dos comandos de operação e despacho de energia que ocorre entre o operador e a agência reguladora, por exemplo. Desta forma, os operadores precisam lidar com altas cargas de trabalho em forma de instruções faladas. Este processo está sujeito a diversas falhas, tais como erros de pronúncia e linhas telefônicas ocupadas. Além disso, as operações do dia a dia de um sistema de operação são complexas e exaustivas, sendo que podem afetar negativamente o desempenho dos funcionários, o que pode então ocasionar falhas, levando-se assim a problemas de segurança (XIANG *et al*., 2021). Ademais, quando há a ocorrência de alguma falha, o número de instruções recebidas via comandos de voz aumenta exponencialmente, o que eleva então a possibilidade de erros humanos, colocando-se também em risco a segurança do sistema (YU *et al*., 2020).

Quando há a ocorrência de algum incidente ou falha na operação da planta ou na comunicação, torna-se necessário realizar um processo de auditoria das informações repassadas nas chamadas telefônicas. Por este motivo, os engenheiros responsáveis devem então ouvir uma grande quantidade de horas de gravação para determinar e identificar o que aconteceu e ocasionou a falha. Porém, durante este processo, muito material irrelevante, no formato de arquivos de áudio, é analisado (ZHANG *et al*., 2021).

Com efeito, sistemas de ASR poderiam ser utilizados para realizar a análise inteligente das informações contidas nos arquivos de áudio, visando-se determinar quais são relevantes de serem monitoradas a fim de indicar o funcionamento da rede. (XIANG *et al.*, 2021).

Ainda em relação à análise sistemática das informações presentes nas chamadas telefônicas realizadas nos centros de operação, é possível formar um banco de informações com dados históricos sobre a segurança da rede contendo as falhas e problemas identificados em sua operação. O uso de tecnologias como as de transcrição de fala para texto permite analisar sistematicamente este tipo de dados, padronizando, rotulando e classificando as informações para uso posterior (YU *et al.*, 2020).

Dentre os estudos analisados na área, verificou-se também a aplicação de sistemas de ASR para auxiliar nos procedimentos de notificação de segurança em subestações. Tais procedimentos são relativos ao processo de leitura de um código de conduta segura, por parte do supervisor, para os funcionários antes de alguma operação na subestação, a fim de alertar e assim evitar possíveis riscos. Dada a importância deste procedimento, é vital que o supervisor leia de forma completa os procedimentos de segurança. Todavia, atualmente não há uma forma automática de assegurar que os procedimentos foram explicados completa e corretamente, pois esta supervisão depende da análise humana, a qual pode também levar a erros (LI *et al.*, 2020).

Além de verificar as principais aplicações de sistemas de ASR no contexto do SEP, a análise dos artigos presentes na literatura também permitiu identificar os principais desafios presentes neste tipo de aplicação. Dentre tais desafios destaca-se a presença de ruído sonoro característico neste tipo de ambiente, como subestações. Além disso, a garantia no nível de precisão no processo de reconhecimento de fala, dentro de um contexto de operação, exige alto grau de acurácia nos processos (JORGE *et al.*, 2010).

Outro problema relatado relaciona-se com o fato das instruções sobre despacho de energia usualmente conterem nomes de entidades do sistema elétrico, o que agrega uma maior dificuldade de reconhecimento por parte dos sistemas de identificação de fala, devido à falta de base de dados. Desta forma, devido à alta presença de nomes próprios de entidades em instruções sobre despacho de energia, normalmente é então necessário realizar o pré-treinamento do modelo de linguagem construído para as instruções de despacho de energia (JIANGPING *et a*l., 2021).

Tais desafios vêm sendo abordados na literatura correlata. Apesar disso, ainda não se verificou estudos que apliquem um sistema completo de ASR no contexto de SEP. Tendo em vista a escassez de trabalhos sobre o tema, este capítulo tem como objetivo destacar e analisar comparativamente os principais trabalhos desenvolvidos na área de SEP utilizando técnicas de ASR. Serão avaliados os métodos utilizados por cada autor e os níveis de acurácia obtidos. Além disso, serão destacados os principais desafios sobre o tema e abordagens para contorná-los. Por fim, serão propostas sugestões para trabalhos futuros. Assim, espera-se que este estudo sirva como motivação e guia para o desenvolvimento

desta área de pesquisa, tendo-se em vista que não foram verificados na literatura artigos de revisão que tratem deste contexto de aplicação, apesar de sua importância.

Este trabalho está dividido em 5 seções. A primeira seção é a presente, a qual contém a introdução ao tema e ao estudo. A segunda seção apresenta os principais conceitos sobre as técnicas de reconhecimento automático de fala. A Seção 3 fornece uma análise sobre o estado da arte de técnicas de ASR aplicadas no contexto de SEP. Análises comparativas sobre os trabalhos de destaque na literatura são descritas na Seção 4. Por fim, a Seção 5 apresenta as conclusões do trabalho, sintetizando o estado da arte do tema de ASR aplicado a SEP, além de discutir também sobre motivos para a ainda tímida aplicação de tais técnicas neste contexto e sugerir ainda abordagens para incentivar mais estudos.

2 | TÉCNICAS DE RECONHECIMENTO AUTOMÁTICO DE FALA

Classificada como uma característica exclusiva do ser humano, a voz concretiza a necessidade de verbalizar pensamentos e permite que o indivíduo compartilhe informações a partir da comunicação oral (CUERVO, 2010; SANTOS, 2015). Dada a importância de tal habilidade no cotidiano das pessoas, tem-se buscado desenvolver máquinas capazes de produzir e entender a voz humana, fazendo-se com que a comunicação vocal homem-máquina ganhe destaque no campo da comunicação por voz.

A interação com sistemas automáticos a partir da utilização da voz possui diversas aplicações que serão exploradas ao longo deste estudo e se mostra vantajosa em diversos cenários, tais como nos casos em que o operador não pode usar suas mãos livremente, quando não é possível ou conveniente a adição de um teclado no ambiente, quando é necessário ter maior mobilidade durante o processo de entrada de dados em um sistema, em situações nas quais o processo deve ser controlado em tempo real, quando o operador deve manter o olhar em um ponto fixo, e quando a robustez e segurança operativa podem ser aprimoradas a partir da garantia de redundância no processo.

O campo da comunicação vocal homem-máquina inclui três tipos de comunicação: sistemas de resposta vocal, sistemas de reconhecimento de locutor e sistemas de reconhecimento de fala. O primeiro tipo consiste em sistemas com capacidade de responder a solicitações de informações com mensagens de voz. Esses sistemas, portanto, transmitem o som em uma única direção, da máquina para o usuário.

Para problemas de reconhecimento de locutor, a tarefa do sistema é verificar a identidade de um locutor específico ou identificar um locutor particular de um determinado conjunto de falantes candidatos. No primeiro caso, o locutor sempre quer ser reconhecido pelo sistema e, portanto, é definido como cooperativo. No segundo caso, o locutor é definido como não cooperativo porque geralmente não quer ser reconhecido pelo sistema, como em aplicações voltadas para a área de criminalística.

A tecnologia de verificação de locutor é semelhante à tecnologia de reconhecimento de fala, o que torna atraente combinar as duas tecnologias no mesmo hardware para aplicações específicas (COSTA, 1994). A terceira abordagem, relacionada com sistemas de reconhecimento de fala, será detalhada a seguir.

2.1 Sistemas de Reconhecimento de Fala

O reconhecimento de fala é um processo importante para simplificar a comunicação homem-máquina, sendo um procedimento que permite aos usuários usar comandos de voz que são reconhecidos e interpretados por sistemas de reconhecimento automático de fala.

A tarefa básica de um sistema de reconhecimento de fala é reconhecer com precisão uma frase falada completa, ou "entender" uma expressão falada (ou seja, responder de forma correta ao que foi falado). O conceito de compreensão, ao invés de reconhecimento, é muito importante para sistemas que lidam com entrada de fala contínua de grande vocabulário, enquanto o conceito de reconhecimento preciso é crítico para sistemas com palavras únicas, vocabulários limitados e baixo número de usuários (RABINER, 1976).

O problema do reconhecimento de fala pela máquina está relacionado à estrutura complexa da voz humana e depende de diversos fatores, como características vocais, entonação, velocidade de fala, estado emocional do usuário, entre outros. Os métodos de reconhecimento de fala podem ser classificados de acordo com o tamanho e a flexibilidade do vocabulário, número de usuários, condições de fala e assim por diante (COSTA, 1994). Em geral, todos os tipos de sistemas de reconhecimento automático de voz podem ser atribuídos a uma das seguintes categorias: sistemas de reconhecimento de palavras isoladas, sistemas de reconhecimento de palavras conectadas, sistemas de reconhecimento dependente do locutor, e sistemas de reconhecimento independente do locutor.

Os sistemas de reconhecimento de palavras isoladas podem ser definidos como aqueles que necessitam de uma pequena pausa antes e depois da sentença que deve ser reconhecida (GU *et al.*, 1991). A duração mínima das pausas separando amostras independentes deve ser da ordem de 100 milissegundos. Intervalos de tempo menores que 100 milissegundos podem ser confundidos com pequenas pausas criadas pela presença de uma consoante oclusiva no meio de uma palavra (MARTIN, 1976).

Para sistemas de reconhecimento de palavras conectadas, o modo de entrada de palavras é mais conveniente para os usuários, pois se aproxima da forma natural de falar do ser humano. Contudo, devido ao nível atual da tecnologia de reconhecimento de fala, esse método de comunicação tem algumas limitações. Além disso, a entrada de palavras conectadas não equivale à fala contínua com um vocabulário de milhares de palavras, de modo que os usuários não se comunicam com o sistema da mesma forma que se comunica com outros indivíduos. Este modelo de sistema pode ser utilizado de forma eficaz em aplicações nas quais um grupo restrito de usuários pode ser instruído sobre como operar

o sistema (UNNIKRISHNAN *et al.*, 1991; BROWN *et al.*, 1991). Para intervalos curtos de palavras conectadas, taxas de fala de mais de 300 palavras por minuto podem ser então alcançadas (MARTIN, 1976).

Os sistemas dependentes do locutor são caracterizados por serem treinados para seguir as características de fala específicas de um usuário. Uma vantagem desta abordagem é seu vocabulário de centenas de palavras, que é definido e criado pelo usuário, além da facilidade com a qual esse vocabulário pode ser modificado e atualizado. Todavia, a principal desvantagem desses sistemas é a necessidade de treiná-lo para cada usuário, o que pode tornar a abordagem proibitiva se houver a necessidade de um grande vocabulário e um elevado número de usuários. Além disso, o desempenho é sensível às mudanças na voz do usuário devido ao estresse, fadiga, rouquidão, etc. O desempenho de um sistema de reconhecimento de fala de alta qualidade (fala contínua) é de cerca de 90% em condições laboratoriais bem definidas (HUANG, 1992). No entanto, a avaliação de dispositivos reais de reconhecimento de fala depende da aparência de palavras semelhantes, do estado físico e emocional do locutor, do tipo e posição do microfone e do ruído ambiente.

Por fim, sistemas de reconhecimento independentes do locutor podem ser definidos como aqueles que não estão vinculados a nenhuma característica particular da fala de um usuário. Sua principal característica é oferecer desempenho de reconhecimento aceitável para um grande número de usuários. No entanto, este tipo de sistema não funciona de maneira uniforme para todos os locutores, especialmente quando estes possuem diferentes sotaques, gêneros, dialetos, ou até mesmo comportamentos distintos de fala. Usualmente, o vocabulário desses sistemas é fixo e muito menor do que o dos sistemas que são dependentes do locutor. Além disso, como o vocabulário deve ser gerado pelo fabricante, a sua atualização torna-se então um tanto custosa. Assim, sistemas de reconhecimento dependentes do locutor devem ser geralmente usados quando a identificação do locutor e a modificação do vocabulário de aplicação são necessárias (SAMBUR e RABINER, 1975).

2.2 Arquitetura de Sistemas de Reconhecimento Automático de Fala

Os atuais sistemas de reconhecimento contínuo de fala se baseiam no princípio do reconhecimento estatístico de padrões, no qual os sinais acústicos são convertidos em uma série de símbolos, analisados e estruturados em unidades de subpalavras, que sejam capazes de representá-los com a mínima perda de informação possível.

A etapa inicial para um sistema de reconhecimento de fala consiste na aquisição do sinal de voz que servirá de entrada para o sistema. Esse processo pode ser realizado por microfones e amplificadores, os quais fornecem um sinal elétrico analógico. Depois que o sinal de áudio do microfone é amostrado, quantizado e codificado, o pré-processamento é realizado com o intuito de remover o ruído presente no sinal. Deste modo, deve-se realizar a depuração de sinal para aprimorar o processo de codificação e eliminar componentes indesejáveis, de maneira a realizar uma escalada do sinal para reduzir sua faixa dinâmica

e evitar possíveis erros de quantificação (PICONE, 1993).

O estágio de pré-processamento do sinal de voz pode ser dividido nas seguintes etapas: pré-ênfase, segmentação, janelamento e transformada de Fourier. Os aspectos mais importantes de cada um são detalhados a seguir:

- Pré-ênfase – Consiste na aplicação de um filtro passa-alta digital de primeira ordem com o intuito de compensar os efeitos dos impulsos glóticos e enfatizar as frequências dos formantes (MORENO, 1996). Esta etapa evita que dados sejam perdidos durante o processo de segmentação, uma vez que a maior parcela das informações se encontra contidas nas baixas frequências, além também de remover inteiramente o componente DC do sinal a fim de suavizá-lo espectralmente.
- Segmentação – Utilizada para definir precisamente os pontos iniciais e finais das palavras. É preciso diferenciar as partes do sinal que contêm informações de áudio daquelas que não contêm, a fim de reduzir o custo computacional exigido. Assim, o sinal de voz é dividido em quadros relativamente pequenos, de forma que eles assumam características para que possam ser considerados quase estacionários (SEPÚLVEDA, 2004). O tamanho do quadro é normalmente de 20 a 30 milissegundos, com um deslocamento típico de 10 milissegundos entre os quadros, o qual evita a perda de representação do segmento. Uma vez que o sinal é segmentado, os quadros são salvos como vetores de atributos para processamento posterior.
- Janelamento – A segmentação de sinais de voz introduz problemas de descontinuidade no início e no final de cada quadro, pois cada um deles começa e termina abruptamente. Sendo assim, precisa-se mitigar esse efeito multiplicando cada quadro por uma janela apropriada para suavizar as bordas do quadro até chegar a zero e, em seguida, enfatizar a parte central para destacar as propriedades características do segmento, conforme mostrado na Figura 1. Existem muitos tipos de janelas que podem ser utilizadas no reconhecimento de fala, no entanto, a mais utilizada é a janela de Hamming (MITRA, 2010).
- Transformada de Fourier – Levando em consideração a tarefa de reconhecimento de fala, quando o sinal é transformado em suas componentes de frequência, é então possível distinguir as vozes de locutores distintos e definir as palavras faladas (SIQUEIRA, 2011). Como o sinal de voz é não estacionário, o espectro de potência de cada quadro janelado pode ser extraído usando a Transformada Discreta de Fourier (OPPENHEIM e SCHAFER, 1999).

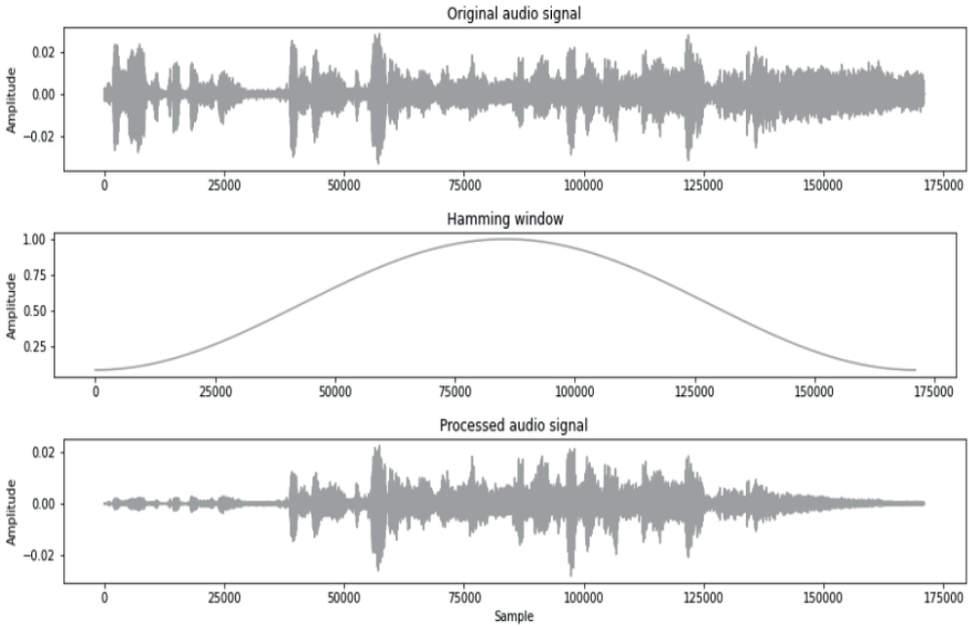

Figura 1. Aplicação da janela de Hamming ao segmento de áudio.

Após o pré-processamento do sinal de voz, realiza-se o processo de extração de características. Esta etapa objetiva representar o sinal de fala de uma maneira conveniente para o reconhecedor a partir de um conjunto de vetores de *n* componentes, capazes de representar o espectro de cada segmento de fala. Desta forma, é possível realizar a compressão do sinal por meio deste conjunto de vetores e suprimir informações irrelevantes para análise fonética futura dos dados pré-processados. Esse conjunto de vetores pode ser retratado de diversas maneiras com parâmetros que representam aspectos distintos do sinal. Ainda se faz necessário observar que quanto mais parâmetros um vetor tiver, os resultados da implementação se tornam menos confiáveis, além de demandar um maior custo computacional. Portanto, o número de parâmetros deve ser o menor possível, a fim de se evitar que a base de dados seja sobrecarregada.

Outra etapa essencial consiste na definição de um modelo de linguagem (gramática). Um modelo de linguagem utiliza o contexto das palavras e informações sobre a frequência com que as palavras são pronunciadas a fim de encontrar opções possíveis que apontem quais palavras têm mais probabilidade de vir antes ou depois de alguma outra.

As restrições impostas pelo modelo de linguagem melhoram significativamente o desempenho do sistema e reduzem o espaço de busca por frases corretas. De modo geral, o modelo da linguagem possui a função de calcular a probabilidade de uma palavra $P(W)$ em uma determinada sentença, dada todas as palavras que a precedem W_1, W_2,..., W_n. Desta forma, é possível expressar $P(W)$ da seguinte maneira:

$$P(W) = \prod_{i=1}^{n} P(W_i \mid W_1, W_2, \ldots, W_{i-1})$$

onde: $P\ (W_i \mid W_1,\ W_2,\ldots,\ W_{i-1})$. é a probabilidade de que W_i seja selecionada após a sequência de palavras (W_1, W_2,..., W_{i-1}).

O bloco final do sistema de reconhecimento, definido como bloco de classificação, consiste em três subestruturas básicas destinadas a mesclar e comparar os vetores de características obtidos com padrões de referência. Essas referências representam os diversos objetos a serem identificados como sílabas, fonemas ou palavras, dependendo da arquitetura e do modelo de linguagem do sistema reconhecedor. Após a obtenção do vetor de características e do padrão de referência, é feita uma comparação entre a referência e a frase reconhecida. Essa forma de comparação está relacionada com o projeto do sistema de reconhecimento, exigindo-se a criação de modelos eficientes para identificar uma palavra entre muitas possíveis (RODRÍGUEZ *et al*., 2004). Diferentes técnicas podem ser utilizadas para esse propósito, permitindo-se realizar a representação e construção de modelos de classificação.

Desta forma, podem-se representar as principais etapas de um sistema automático de reconhecimento de fala através do fluxograma apresentado na Figura 2.

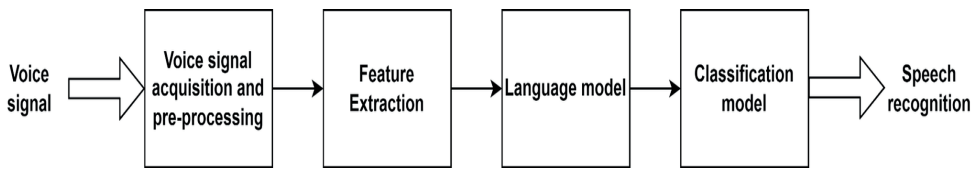

Figura 2. Estrutura de um sistema de reconhecimento automático de fala.

3 | O ESTADO DA ARTE EM SISTEMAS DE RECONHECIMENTO AUTOMÁTICO DE FALA PARA SEP

Esta seção tem como objetivo apresentar os artigos proeminentes na literatura que utilizam técnicas de reconhecimento de fala aplicadas a problemas de sistemas elétricos de potência. Para a seleção dos artigos discutidos aqui, foram então realizadas buscas padronizadas nas mais diversas bases técnico-científicas. Este procedimento de busca segue um critério de filtro a partir das palavras chaves definidas, tendo-se em vista a grande quantidade de trabalhos correlatos encontrados. A principal palavra chave utilizada foi "*speech recognition*", sendo esta combinada com os termos tradicionais encontrados em estudos de aplicações no SEP, tais como "*power system*", "*power plant*" e "*power grid*". Além disso, aplicaram-se filtros de citações e idade da pesquisa, limitando-se os trabalhos a resultados entre 2016 e 2022.

No contexto de SEP, muitos estudos visam tratar de questões relacionadas com o despacho de energia. Em Jiangping *et al*. (2021), por exemplo, desenvolveu-se um modelo voltado para identificação do nome de entidades presentes em instruções relativas ao

despacho de energia. Neste trabalho, para compor a biblioteca de vocabulários, foram utilizados 600 termos relacionados com nomes de entidades citadas em instruções para despacho de energia, incluindo-se termos de especificação de contato de despacho, nomes de equipamentos do sistema de energia, ordens de operação típicas, entre outros.

Os autores utilizaram um modelo BERT-BIGRU-CRF em sua aplicação. O algoritmo BERT (*Bidirectional Encoder Representations from Transformers*) consiste em um modelo de linguagem baseado na extração de características. Ele utiliza uma grande quantidade de dados não rotulados para aprender o modelo de linguagem. No trabalho, o modelo BERT é utilizado para realizar o reconhecimento de nomes de entidades do sistema elétrico. Ainda referente a esta aplicação, os autores utilizam a rede neural BIGRU (*Bidirectional Gated Recurrent Unit*) para extrair padrões das instruções de despacho e construir um contexto de dependência com tais padrões. Por fim, o método CRF (*Conditional Random Fields*) foi utilizado para considerar a dependência entre vetores de palavras adjacentes, a fim de aumentar a acurácia do modelo de reconhecimento. Para o teste do método completo foram utilizadas 2000 instruções de despacho rotuladas. Os resultados indicaram que o modelo proposto com pré-treinamento do nome das entidades apresentou os melhores resultados do que os modelos tradicionais, atingindo-se em torno de 94% de acurácia (JIANGPING *et al*., 2021).

Instruções relacionadas com despacho de energia também foram analisadas em Zhang *et al*. (2021). Neste estudo, os autores desenvolveram um sistema inteligente para reconhecimento de fala. O diferencial deste sistema é sua adaptação para dados de voz relacionados com despacho de energia. Os autores criaram uma base de conhecimento sobre despacho de energia, construindo um dicionário acústico de conversas relacionadas a este contexto. Foram utilizadas técnicas de pré-processamento para reduzir ruídos e aumentar a robustez do modelo acústico. Para a construção do dicionário customizado, foram então selecionadas palavras comuns no dia a dia da operação, mas que não estão no vocabulário comum, e que por isso poderiam ser interpretadas erroneamente, tais como subestação, linha, ou nomes próprios, referentes aos nomes das subestações. Desta forma, o dicionário foi formado por dois conjuntos principais: nomes comuns para a operação, como nomes de subestações, por exemplo, e verbos comumente utilizados pelos operadores para dar e receber comandos.

O modelo acústico utilizado em Zhang *et al*. (2021) foi composto pela rede LSTM (*Long Short Term Memory*) e pela técnica CRF. Além disso, o método *N*-Gram foi utilizado para compor o modelo de linguagem. Como fonte de dados, os autores utilizaram ao todo um milhão de diálogos sobre o despacho de energia. Estes dados foram coletados em uma província na China. Dentre os principais resultados, verificou-se que o uso do dicionário personalizado ao invés do dicionário universal promove uma melhora na acurácia do reconhecimento das falas no contexto de despacho de energia.

Em Souza *et al*. (2021), o foco do trabalho é no sistema de classificação. Neste

artigo os autores buscaram desenvolver um sistema inteligente para classificar em seis categorias as chamadas de voz realizadas entre os centros de operação das subestações de uma empresa de energia elétrica (ENGIE) e a operadora nacional do sistema brasileiro (ONS). O principal foco do estudo é o de facilitar o processo de auditoria das ações internas dos operadores e aumentar a eficiência operacional. Para compor o sistema de classificação, foram então utilizadas ferramentas de *machine learning*. A primeira etapa utiliza a ferramenta *Amazon Web Service* para realizar a conversão das chamadas de *speech-to-text*. Os autores também propuseram análises para aumentar a precisão da ferramenta utilizando-se um vocabulário customizado voltado para palavras encontradas no dia a dia de uma planta operativa.

Em seguida, as chamadas de voz transcritas foram classificadas em seis categorias: mudança de estado, modulação de geração, mudança de tensão, nível do reservatório, chuva, entre outros. Para compor o conjunto de treinamento foram utilizados 650 áudios transcritos, que foram previamente rotulados por especialistas. Em relação à transcrição de texto, a taxa de reconhecimento das palavras ficou em torno de 30%. Em relação à classificação, foram avaliados diferentes modelos de classificação (*random forest*, *extra trees*, *gradient boosting*, LightGBM, XGBoost, *decision tree*), sendo que todos obtiveram uma acurácia de aproximadamente 93%. No contexto geral, os autores identificaram maiores erros de classificação para classes cujas palavras-chave tinham menor taxa de reconhecimento pelo sistema de transcrição de texto (SOUZA *et al*., 2021).

O foco na etapa de classificação da técnica de reconhecimento de fala também é verificado em Yu *et al*. (2020), em que o estudo propõe um modelo inteligente para classificar e rotular conjuntos de texto relacionados com a ocorrência de violações na operação de um sistema elétrico. Os autores utilizaram o método *Jieba Cutter* para dividir os textos das violações em palavras. As palavras chave sobre cada violação foram selecionadas como amostras para o treinamento do modelo de classificação. Para realizar a classificação foi utilizada a rede neural recorrente LSTM (*Long Short Term Memory Network*). O conjunto de treinamento foi composto por 1660 amostras, incluindo-se todos os tipos de violação que ocorrem na rede elétrica, sendo os dados obtidos de um sistema localizado ao sul da China. Neste estudo, os resultados mostram um nível de acurácia, em torno de 99%, para o modelo de classificação LSTM (YU *et al*., 2020).

Em Li *et al*. (2020), o foco do trabalho foi desenvolver um sistema inteligente de reconhecimento de fala em tempo real. O foco deste sistema é auxiliar nos procedimentos de notificação de segurança em campo, garantindo que todas as instruções de segurança sejam lidas e explicadas corretamente pelo responsável técnico. De acordo com os autores, os arquivos de voz obtidos foram pré-processados para a retirada do ruído presente no ambiente de subestação. Esse processamento do sinal foi realizado utilizando o método de subtração espectral, com determinação adaptativa de parâmetros. Após a remoção de ruídos do sinal de áudio, os autores utilizaram a rede LSTM para realizar o

reconhecimento de fala. Os métodos foram testados com três tipos de dados: arquivos de áudio gravados em ambientes silenciosos, arquivos de áudio com a adição artificial de ruídos, e arquivos de áudio obtidos em ambientes de subestação com o ruído característico destes locais. Os resultados obtidos pelos autores indicam que o nível de precisão no processo de reconhecimento de fala é incrementado pelo uso do pré-processamento dos arquivos de áudio para a remoção de ruídos. Além disso, de acordo com os autores, após o reconhecimento de fala realizado pela LSTM, também é realizada a comparação com as instruções de segurança originais, a fim de verificar se todas as etapas foram descritas. Todavia, os autores não indicam no trabalho como é realizada a comparação.

Aplicações em campo também são o foco em Li *et al*. (2019), em que os autores desenvolveram um sistema inteligente de operação móvel para ser utilizado durante o processo de manutenção de sistemas de energia. O funcionamento do sistema proposto é baseado na tecnologia de reconhecimento de fala. De acordo com os autores, o método segue as seguintes etapas: é realizada a captação do arquivo de áudio por meio de um microfone; é realizado um processo de pré-processamento de dados para remoção de ruídos do ambiente; em seguida, aplica-se um processo de extração de características do arquivo de áudio processado, a fim de filtrar as informações que mais representam o conteúdo do áudio; os dados extraídos são agrupados em categorias com o uso de métodos de clusterização; finalmente, o reconhecimento da fala é então realizado por meio do processo de reconhecimento de padrões. O processo de tratamento do áudio com ruído é realizado utilizando o Modelo oculto de Markov. O equipamento móvel desenvolvido foi testado em diferentes ambientes de manutenção e os testes foram realizados com diversos números de palavras e avaliados sistematicamente para cada ambiente. Os resultados indicam uma taxa de reconhecimento de fala de aproximadamente 98% para ambientes com alta presença de ruídos (LI *et al*., 2019).

As técnicas de reconhecimento de fala também foram utilizadas para auxiliar no processo de treinamento dos operadores. Em Jorge *et al*. (2010) foi desenvolvida uma interface homem-máquina baseada no reconhecimento de fala automático. Tal interface foi desenvolvida para ser integrada a uma mesa de controle virtual de uma usina nuclear. Esta mesa virtual visa treinar operadores e avaliar a ergonomia do sistema ao simular o funcionamento de uma mesa de controle de uma usina nuclear. O uso de tecnologia de reconhecimento de voz integrado com a mesa virtual tem como finalidade trazer maior naturalidade para o treinamento, retirando a necessidade de se utilizar teclados e mouses devido às limitações físicas que tais equipamentos trazem. O reconhecimento de fala automático proposto no trabalho recebe comandos de voz compostos por palavras isoladas. Estas devem ser reconhecidas para que, em seguida, o comando correspondente possa ser executado (JORGE *et al*., 2010).

O método de reconhecimento de fala proposto pelos autores é dividido em duas etapas. A primeira etapa consiste no pré-processamento dos sinais de fala utilizando

uma análise “cepstral”, que extrai parâmetros de sinais de voz para serem utilizados em processos de reconhecimento. Em seguida, os autores utilizaram Redes Neurais Artificiais (RNA) para identificar o comando correspondente aos padrões de voz obtidos da primeira análise. Duas RNA foram avaliadas, uma rede perceptron multicamada treinada com *backpropagation* e uma *General Regression Neural Network* (GRNN). Os autores também testaram a utilização das redes em paralelo para verificar se haveria aumento da robustez do método de classificação. Os resultados apresentaram uma alta taxa de reconhecimento dos comandos de voz, com níveis acima de 90% (JORGE *et al.*, 2010).

Por fim, destacam-se alguns trabalhos que visam utilizar técnicas de reconhecimento de fala para substituir o operador em algumas atividades. Este tipo de aplicação propõe a redução da intensidade de trabalho operacional do agente e visa liberar os funcionários da execução de trabalho repetitivo. Em Zhang *et al.* (2019), por exemplo, os autores buscaram desenvolver um modelo do comportamento do operador de uma planta de energia, a fim de construir um operador virtual capaz de realizar as funções de operação. Para atingir tais objetivos, os autores aplicaram modelos de processamento de linguagem natural e modelos de operação do comportamento humanoide. Para aumentar a precisão do modelo de linguagem natural, os autores estabeleceram um vocabulário comum ao ambiente de operação de uma planta termoelétrica, sendo este o ambiente de aplicação do trabalho. Desta forma, foram definidas bibliotecas de palavras contendo termos relativos às atividades de operação, a nomes de equipamentos e de estados do sistema. Com base nos resultados, os autores indicaram que o robô de operação virtual corretamente reconhece a instrução de fala, é capaz de executar a operação especificada no simulador da usina, e retorna o feedback de informações para o atendente. Todavia, não foram especificados níveis de acurácia para o método proposto (ZHANG *et al.*, 2019).

A substituição de ações do operador também é abordada em Xiang *et al.* (2021), em que os autores levantam a hipótese da construção de um sistema inteligente para realizar o despacho de energia, utilizando-se um módulo de agente inteligente, que representaria as ações do operador. O projeto propõe que o processo de reconhecimento de fala seja realizado pelo modelo WaveNet. Além disso, os autores também indicam a possibilidade de utilizar técnicas de aprendizado de máquinas para o reconhecimento de nomes de entidades do sistema elétrico, assim como técnicas para tornar o processo de reconhecimento de fala adaptável para diferentes interlocutores. Entretanto, o trabalho apenas apresenta a proposta do modelo, sem, no entanto, apresentar resultados de sua validação.

As Tabelas 1 e 2 reúnem as principais características dos trabalhos avaliados. A Tabela 1 foca na estrutura do sistema de reconhecimento de fala proposto, tendo como base o esquema apresentado na Figura 2. Enquanto a Tabela 2 considera outros aspectos relevantes da pesquisa, tais como a presença ou não de análises comparativas dos resultados e a utilização de técnicas de otimização para os métodos utilizados. As análises sobre os artigos apresentados e as tabelas serão desenvolvidas na próxima seção.

Ref.	Type of ASR	Automatic Speech Recognition System			
		Voice signal acquisition and pre-processing	Feature Extraction	Language model	Classification model
Jiangping *et al.* (2021)	Connected speech recognition.	Continuous Bag of Words Model was used to segment the instructions.	BiGRU model.	BERT model.	Conditional Random Fields (CRF).
Zhang *et al.* (2021)	Connected speech recognition.	–	Kaldi method to extract multidimensional Mel-Frequency Ceptral Coefficients (MFCC).	N-Gram language model.	Long Short-Term Memory (LSTM) network and Conditional Random Fields (CRF).
Yu *et al.* (2020)	Isolated words recognition (Text classification).	Jieba Cutter to cut each sentence into words. Then, keywords are selected as learning samples, and the word2vector model is used to transform them into eigenvectors.	Recurrent Neural Network (RNN) extracts features from the data.	–	Recurrent Neural Network (RNN).
Li *et al.* (2020)	Not specified.	Improved Parameter Adaptive Spectral Subtraction (IPASS).	–	–	Long Short-Term Memory (LSTM).
Li *et al.* (2019)	Isolated words recognition.	Hidden Markov Model (HMM).	Not specified.	Not specified.	Not specified.
Zhang *et al.* (2019)	Connected speech recognition.	Forward maximum matching algorithm for sentence segmentation.	–	N-Gram language model.	–
Souza *et al.* (2021)	Isolated words recognition (Text classification).	–	Not specified.	–	Decision tree, random forest, extra trees, gradient boosting, XGBoost and LightGBM.
Jorge *et al.* (2010)	Isolated words recognition.	Blank space elimination using short-time energy method.	Cepstral analysis.	–	Artificial neural networks (feedforward) trained with backpropagation algorithm and General Regression Neural Network (GRNN).
Xiang *et al.* (2021)	Not specified.	Applied frame cutting.	MFCC features were extracted using natural language processing technology.	Bidirectional Encoder Representations from Transformers (BERT).	–

Tabela 1. Estrutura dos Sistemas de Reconhecimento de Fala Presentes na Literatura Avaliada.

Ref.	Main method	Full model accuracy	Custom vocabulary	Input data	Comparative analysis	Optimization method
Jiangping *et al.* (2021)	BERT-BIGRU-CRF.	≈94%	Power dispatch instructions.	3 generic datasets were analyzed: clue2020, MsrA and people's daily datasets. In addition, 2000 power dispatch instructions were used.	Different datasets and different models were comparatively evaluated.	–
Zhang *et al.* (2021)	LSTM-CRF.	Word error and sentence error rates, respectively, 1,38% e 9,28%.	Built a customized dictionary using energy dispatch commands.	1 million samples related to energy dispatch speech were collected in an energy dispatch center in China.	Different feature extraction methods were evaluated. In addition, the proposed model is compared with other traditional techniques and the use or not of a customized dictionary is also evaluated.	Cepstral Mean and Variance Normalization (CMVN) method is used after feature extraction to reduce channel and noise effects and improve the robustness of the acoustic model.
Yu *et al.* (2020)	LSTM.	~99%	–	1660 samples of text collected from South China Grid.	The proposed method was compared with Naive Bayes Classifier and Logistic Regression.	-
Li *et al.* (2020)	IPASS-LSTM.	19.5 (Word Error Rrate - WER).	–	Both simulated and on-site data were analyzed and used as input for the methods. Different levels of ambient noise were also tested. The number of samples is not specified.	The proposed IPASS method for noise reduction is compared with traditional spectral subtraction and multiband spectral subtraction.	Proposes an optimal parameter selection for IPASS based on signal-to-noise ratio to allow noise robustness.
Li *et al.* (2019)	Not specified.	98.09%	–	Is indicated that three datasets of 10 words, 30 words and 50 words were tested in different noisy environments.	Different types and levels of environmental noises were tested.	–
Zhang *et al.* (2019)	Microsoft SAPI voice development platform.	–	Customized vocabulary based on thermal power plants terms-operation terms, equipment names and state terms.	–	–	–

Souza *et al.* (2021)	Amazon transcribe (AWS) for speech-to-text.	The highest classification accuracy for one label was 99,5%, while the lowest was 89% for random forest method.	Customized vocabulary with proper nouns, technical terms and commonly used acronyms to improve speech-to-text performance.	For the classification phase, 650 labeled transcribed audios were used for training.	300 combinations were analyzed - different classification models and different combinations of input features.	–
Jorge *et al.* (2010)	ANN.	~97%	–	20 samples for each command (class) were used as data. The datasets were created using both random and sequential speech.	There was a comparative analysis between the ANN used and between an approach using both ANNs in parallel.	–
Xiang *et al.* (2021)	WaveNet is used for the speech-to-text process.	–	Power dispatching vocabulary.	Power grid dispatching rules and power grid historical data.	–	To create an ASR system that is adaptive, the paper uses memory sensing network to extract dispatcher's voice auxiliary features.

Tabela 2. Características dos Estudos Avaliados.

4 I ANÁLISE DAS PRINCIPAIS CARACTERÍSTICAS DO ESTADO DA ARTE

Inicialmente, pode-se então avaliar a motivação dos estudos analisados na seção anterior. Verifica-se que a maioria das abordagens presentes na literatura para aplicações de ASR em SEP tem como finalidade agregar maior segurança aos processos de operação, a fim de auxiliar no processo de verificação de instruções de despacho (JIANGPING *et al.*, 2021), garantindo notificações de segurança (LI *et al.*, 2020) e auxiliando no treinamento de pessoal (JORGE *et al.*, 2010). Alguns estudos como em Zhang *et al.* (2019) e Xiang *et al.* (2021) também indicam o aumento da eficiência operacional ao substituir parcialmente por um sistema inteligente as ações realizadas pelo operador. Também se observou a necessidade de facilitar os processos de auditoria (SOUZA *et al.*, 2021), que atualmente são executados por meio de processos custosos, demandando aos operadores ouvirem horas de arquivos de áudio. Todavia, ainda existem diversas abordagens que podem ser exploradas neste contexto, como por exemplo, a avaliação de sentimentos para determinar o nível de segurança das atividades de cada operador. Este tipo de análise é muito presente na literatura, mas ainda não foi abordado no contexto de SEP. Um artigo de revisão sobre o tópico é encontrado em Koolagudi e Rao (2012).

Com base na Tabela 1 observa-se que a maioria dos autores utiliza como

técnicas de pré-processamento os métodos de segmentação para dividir sentenças em palavras, facilitando-se assim o processo de seleção de palavras-chave que representem aquela sentença e também o processo de classificação posteriormente realizado para o reconhecimento da palavra. Não obstante, não foi verificada a predileção por nenhum método específico para realizar o processo de segmentação. O processo de extração de características também foi avaliado. Verificou-se que alguns artigos como Souza *et al*. (2021), Yu *et al*. (2020) e Li *el al*. (2019) indicam a extração de características dos dados sem no entanto explicitar o método utilizado. Dentre os trabalhos que apresentaram os métodos utilizados, foi possível observar uma predileção pelo uso dos *Mel-Frequency Cepstral Coefficients* (MFCC), que consiste em um tipo de análise muito comum na literatura de sistemas de reconhecimento de fala. Os MFCC são utilizados para evidenciar as componentes de baixa frequência da fala frente às de alta frequência, pois as primeiras possuem maior nível de informação (KOOLAGUDI *et al*., 2012).

Em relação aos métodos de classificação utilizados, foi possível verificar uma predileção dos autores pelo uso de redes neurais artificiais, principalmente a rede LSTM. Esta rede consiste em um tipo de rede neural recorrente e sua popularidade para este tipo de aplicação pode ser justificado pela sua capacidade de atuar em frases longas devido a sua habilidade de memorização. Além disso, de acordo com os autores, esta rede é capaz de extrair informações de sentenças com comprimento variado, o que é uma vantagem frente a outros modelos (YU *et al*., 2020).

Na Tabela 2 observa-se que diversos trabalhos fazem uso de vocabulário customizado com termos do contexto de sistemas elétricos de potência. Verifica-se ainda que tais trabalhos são os que utilizam modelos de reconhecimento de fala ou de transcrição de texto que já são sedimentados na literatura, como em Jiangping *et al*. (2021) e Xiang *et al*. (2021), em que se usa o modelo BERT; em Zhang *et al*. (2021) e Zhang *et al*. (2019), em que o modelo de linguagem *N*-Gram é aplicado; e em Souza *et al*. (2021), em que se utiliza o AWS. Desta forma, infere-se que o uso do vocabulário customizado atua no processo de *transfer learning* e *fine tuning* dos métodos. A vantagem de utilizar um modelo já consolidado na literatura é poder utilizar o aprendizado deste modelo que já foi treinado, usualmente com um grande número de dados, para atuar em um problema particular.

O processo de transferência de aprendizado se baseia na hipótese de que, se um modelo foi treinado com um conjunto de dados grande o suficiente para certa aplicação, ele pode então ser considerado como um modelo generalista para aplicações similares (TENSORFLOW, 2023). No caso, modelos que já foram treinados com uma grande quantidade de amostras de áudio podem ser considerados modelos gerais para reconhecimento de fala, mesmo que não sejam focados em termos de sistemas elétricos de potência. O processo de ajuste fino funciona para especializar tais modelos generalistas para uma determinada aplicação. Desta forma, podem-se utilizar os vocabulários customizados para termos específicos de SEP, para ensinar ao modelo, que já possui um

conhecimento anterior, sobre esta nova aplicação (TENSORFLOW, 2023). Os processos de transferência de aprendizado e ajuste fino são muito comuns em aplicações de *deep learning*, por facilitarem o processo de treinamento dos modelos e contornarem problemas como a falta de dados de treinamento, por exemplo.

Com base nos resultados dos métodos apresentados na Tabela 2, verifica-se um nível elevado de acurácia, atingindo taxas acima de 94% para a maioria dos trabalhos avaliados. Em relação aos dados de entrada, verificou-se que alguns artigos como Li *et al*. (2020), Li *et al*. (2019) e Xiang *et al*. (2021) não indicaram o número de amostras utilizados, mas informaram apenas o tipo de amostras que foram avaliadas. De fato, Li *et al*. (2020) é o único artigo que indica o uso de amostras simuladas, que foi justificado pelo teste de diferentes níveis de ruído. Em relação aos outros artigos, verificou-se uma predominância do uso de dados coletados de redes elétricas chinesas, como em Jiangping *et al*. (2021), Zhang *et al*. (2021) e Yu *et al*. (2020). Em Jiangping *et al*. (2021) também é verificado o uso de datasets genéricos disponíveis online. Um aspecto interessante a ser observado é a pequena quantidade de amostras utilizadas em Souza *et al*. (2021) para o processo de classificação de texto, assim como em Jorge *et al*. (2010) para o reconhecimento de palavras isoladas. Mesmo com um número reduzido de amostras, ambos os estudos apresentaram bons níveis de acurácia, o que indica a qualidade do dataset selecionado.

Um aspecto em comum entre os artigos é que a maioria apresenta algum tipo de análise comparativa. Alguns, como em JIANGPING *et al*. (2021), Zhang *et al*. (2021), Yu *et al*. (2020) e Souza *et al*. (2021) apresentam comparação entre diferentes métodos de classificação, enquanto outros apresentam comparações entre diferentes tipos de dados (LI *et al*., 2019), diferentes métodos de pré-processamento (LI *et al*., 2020) ou diferentes abordagens de um mesmo modelo (JORGE *et al*., 2010). Ressalta-se a importância de análises comparativas para validar o método proposto no artigo frente a outros métodos existentes, além de auxiliar na expansão do tema de estudo, por avaliar diferentes possibilidades de aplicação.

Em relação aos métodos de otimização, poucos artigos utilizam tais técnicas. Destacam-se aqueles apresentados em Zhang *et al*. (2021) e Li *et al*. (2020), os quais utilizam diferentes métodos de otimização para tratar da presença de ruídos nos arquivos de áudio coletados nos ambientes de subestação. Esta abordagem de tratamento de ruídos é importante, pois os ambientes relacionados com aplicações de SEP geralmente apresentam elevados níveis de ruído característicos, tais como a presença de ruídos de máquinas, de linhas de transmissão em operação e também a presença de outras vozes ao fundo, devido à ocorrência de chamadas telefônicas simultâneas no centro de operação.

Por fim, um aspecto importante a ser considerado é o nível de desenvolvimento dos trabalhos encontrados na literatura para aplicações de ASR em SEP. Observa-se que alguns trabalhos simplesmente tratam do processo de classificação de texto, como em Souza *et al*. (2021) e Yu *et al*. (2020), não realizando a construção completa de um sistema

de reconhecimento de fala. Além disso, um dos trabalhos, apresentado em Xiang *et al*. (2021), apenas levanta hipóteses para a construção de um ASR para este tipo de aplicação, não trazendo simulações ou resultados que validam tais hipóteses. Por fim, verifica-se que trabalhos que realizam a construção de um ASR completo, como em Jorge *et al*. (2010), usualmente o fazem para reconhecimento de palavras isoladas e não de frases, ou então, fornecem pouca informação sobre as características do modelo, como em Li *et al*. (2019) e Zhang *et al*. (2019). Desta forma, verifica-se uma lacuna dentre os estudos da área para trabalhos que descrevam e forneçam detalhes sobre a construção de sistemas de reconhecimento de fala completos aplicados à área de SEP.

5 | CONCLUSÕES

Este capítulo teve como objetivo apresentar uma revisão sobre trabalhos que utilizam técnicas de reconhecimento de fala no contexto de sistemas elétricos de potência. A análise da literatura e a seleção dos trabalhos apresentados foram feitas de forma sistemática seguindo uma metodologia projetada para selecionar os artigos mais recentes e proeminentes do tema. A principal conclusão sobre a revisão de literatura reside na baixa incidência de artigos que tratam sobre ASR no contexto de SEP, de forma que ainda é um tema com muitos tópicos a serem explorados. Tais tópicos foram ressaltados ao longo do texto para auxiliar trabalhos futuros.

A análise dos artigos selecionados foi realizada de forma comparativa, utilizando-se duas tabelas principais contendo as características primordiais de cada trabalho. Diversos aspectos foram avaliados, tais como: métodos utilizados, níveis de acurácia, utilização de métodos de otimização, natureza do conjunto de dados, entre outros. Dentre as análises, o ponto de principal atenção foi a ausência de estudos que apresentem de forma detalhada a construção de um sistema completo de ASR para a identificação de frases completas. Além disso, observou-se a carência de análises sobre problemas tradicionais no contexto de reconhecimento de fala, tais como a presença de diferentes interlocutores, com tons de ritmos de fala distintos. Desta forma, este trabalho visa ressaltar a importância do tema, pois existem inúmeros problemas e processos em SEP que podem ser resolvidos e melhorados com o uso da tecnologia de ASR. Com efeito, espera-se também estimular o desenvolvimento de novos trabalhos, pois há ainda uma grande área a ser explorada neste tema.

REFERÊNCIAS

ALNASSER, A.; AL-GHOWINEM, S. Vocally specified text recognition in natural scenes for the blind and visually impaired. *In*: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Computing. SAI 2018. **Advances in Intelligent Systems and Computing**, v. 858, pp 231–248, 2019.

ANEEL. **Resolução Normativa Nº 1.005** (15 de fevereiro de 2022), 2022.

BROWN, M. K.; MCGEE, M. A.; RABINER, L. R.; WILPON, J. G. Training set design for connected speech recognition. **IEEE Transactions on Signal Processing**, v. 39, n. 6, p. 1268-1281, 1991.

COSTA, W. C. A. **Reconhecimento de fala utilizando modelos de markov escondidos (HMM's) de densidades contínuas**. Tese (Mestrado em Engenharia Elétrica) – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande. Campina Grande, p. 103, 1994.

CUERVO, L. **Introdução à fisiologia da voz – práticas vocais para a educação musical**. Porto Alegre: Departamento de Música (UFRGS), 2010.

GU, H.; TSENG, C.; LEE, L. Isolated-utterance speech recognition using hidden Markov models with bounded state durations. **IEEE Transactions on signal processing**, v. 39, n. 8, p. 1743-1752, 1991.

HUANG, X. D. Phoneme classification using semicontinuous hidden Markov models. **IEEE Transactions on Signal Processing**, v. 40, n. 5, p. 1062-1067, 1992.

JIANGPING, J. *et al*. Analysis of power grid dispatching instructions based on BERT-BIGRU mode. *In*: IEEE CONFERENCE ON TELECOMMUNICATIONS, OPTICS AND COMPUTER SCIENCE, 2021, Shenyang, China. **Anais** […] Piscataway: IEEE, 2021.

JORGE, C. A. F. *et al*. Human-system interface based on speech recognition: application to a virtual nuclear power plant control desk. **Progress in Nuclear Energy**, v. 52, n. 4, p. 379-386, 2010.

KOOLAGUDI, S. G.; RAO, K. S. Emotion recognition from speech: a review. **International Journal of Speech Technology**, v. 15, p. 99-117, 2012.

KOOLAGUDI, S. G.; RASTOGI, D.; RAO, K. S. Identification of language using mel-frequency cepstral coefficients (MFCC). **Procedia Engineering**, v. 38, p. 3391-3398, 2012.

LELARDEUX, C. P. *et al*. Communication system and team situation awareness in a multiplayer real-time learning environment: application to a virtual operating room. **The Visual Computer**, v. 33, p. 489-515, 2017.

LI, H.; LI, Z.; RAO, Z. Mobile operation platform for power system maintenance based on intelligent speech recognition. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING, AUTOMATION AND SYSTEMS, 2019, Chongqing, China. **Anais** […] Piscataway: IEEE, 2019.

LI, C. *et al*. A novel speech recognition algorithm for substation safety notification based on IPASS-LSTM. *In*: INTERNATIONAL CONFERENCE ON SMART GRIDS AND ENERGY SYSTEMS, 2020, Perth, Australia. **Anais** […] Piscataway: IEEE, 2020.

MALIK, M.; MALIK, M. K.; MEHMOOD, K.; MAKHDOOM, I. Automatic speech recognition: a survey. **Multimedia Tools and Applications**, v. 80, p. 9411-9457, 2021.

MARTIN, T. B. Practical applications of voice input to machines. **Proceedings of the IEEE**, v. 64, n. 4, p. 487-501, 1976.

MITRA, S. K. **Digital signal processing: a computer-based approach**. 4. ed. New York: McGraw-Hill, 2010.

MORENO, P. J. **Speech recognition in noisy environments**. Tese (Doutorado em Engenharia Elétrica e Computação) – Department of Electrical and Computer Engineering, Carnegie Mellon University. Pittsburgh, Pennsylvania, p. 130, 1996.

NEDJAH, N.; SANTOS, I.; MOURELLE, L. M. Sentiment analysis using convolutional neural network via word embeddings. **Evolutionary Intelligence**, v. 15, p. 2295–2319; 2022.

OPPENHEIM, A. V; SCHAFER, R. W. **Discrete-time signal processing**, 2. ed. Upper Saddle River: Prentice-Hall, 1999.

OSHIKAWA, R.; QIAN, J.; WANG, W. Y. A survey on natural language processing for fake news detection. **ArXiv**, v. abs/1811.00770, 2018.

PÉREZ, F.J. *et al.* Multimedia analysis platform for crime prevention and investigation. **Multimedia Tools and Applications**, v. 80, p. 23681-23700, 2021.

PICONE, J. Signal modeling techniques in speech recognition. **Proceedings of the IEEE**, v. 81, n. 9, p. 1215-1247, 1993.

RABINER, L. R. Special issue on man-machine communication by voice. **Proceedings of the IEEE**, v. 64, n. 4, p. 403-404, 1976.

RODRÍGUEZ, J. L. O.; GUERRA, S. S.; FERNÁNDEZ, R. B. **Reconocimiento de voz usando segmentación de energía usando modelos ocultos de Markov de densidad continua**. Cidade do Méximo: CIC-IPN, 2004.

SAMBUR, M. R.; RABINER, L. R. A speaker-independent digit-recognition system". **The Bell System Technical Journal**, v. 54, n. 1, p. 81-102, 1975.

SANTOS, M. O. **Análise acústica de desvios vocais infantis utilizando a transformada wavelet**. Tese (Mestrado em Engenharia Elétrica) – Instituto Federal de Educação, Ciência e Tecnologia da Paraíba. João Pessoa, p. 80, 2015.

SEPÚLVEDA, F. A. **Extracción de parámetros de señales de voz usando técnicas de análisis en tiempo-frecuencia**. Tese (Mestrado em Automação Industrial) – Facultad de Ingenier´ıa y Arquitectura, Universidad Nacional de Colombia. Manizales, Colômbia, p. 115, 2004.

SIQUEIRA, J. K. **Reconhecimento de voz contínua com atributos MFCC, SSCH e PNCC, wavelet denoising e redes neurais**. Tese (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, p. 85, 2011.

SOUZA, F. D. *et al.* Multi-Label classification of voice calls from power plant operation centers. *In*: IEEE PES INNOVATIVE SMART GRID TECHNOLOGIES CONFERENCE – LATIN AMERICA), 2021, Lima, Peru. **Anais** […] Piscataway: IEEE, 2021.

TENSORFLOW. Transferência de aprendizado e ajuste fino. **TensorFlow**, 2023. Disponível em: <https://www.tensorflow.org/tutorials/images/transfer_learning>. Acesso em: 09 de jun. de 2023.

UNNIKRISHNAN, K. P.; HOPFIELD, J. J.; TANK, D. W. Connected-digit speaker-dependent speech recognition usin neural network with time-delayed connections. **IEEE Transactions on Signal Processing**, v. 39, n. 3, p. 698-713, 1991.

VADWALA, A. Y.; SUTHAR, K. A.; KARMAKAR, Y. A.; PANDYA, N. Survey paper on different speech recognition algorithm: challenges and techniques. **International Journal of Computer Applications***, v. 175*, p. 31-36. 2017.

VENKATASUBRAMANIAN, S., MOHANKUMAR, R. A deep convolutional neural network-based speech-to-text conversion for multilingual languages. **Advances in Intelligent Systems and Computing**, v. 1420, p. 617–633, 2022.

XIANG, Z. *et al*. Design of intelligent dispatching system based on human voice adaptive speech recognition. In: INTERNATIONAL CONFERENCE ON POWER SYSTEM TECHNOLOGY, 2021, Haikou, China. **Anais** […] Piscataway: IEEE, 2021.

YU, Y. *et al*. Intelligent classification and automatic annotation of violations based on neural network language model. *In*: *INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS*, 2020, Glasgow, UK. **Anais** […] Piscataway: IEEE, 2020.

ZHANG, H; XIAO, L.; YAN P.; XIAO, Q. Research on speech recognition of power grid dispatching based on big data and deep learning. *In*: INTERNATIONAL CONFERENCE ON POWER SYSTEM TECHNOLOGY, 2021, Haikou, China. **Anais** […] Piscataway: IEEE, 2021.

ZHANG, Q.; MA, J.; WANG, J. Application research of virtual operation robot in smart power plant. *In*: CHINESE AUTOMATION CONGRESS, 2019, Hangzhou, China. **Anais** […] Piscataway: IEEE, 2019.

ZHANG, Y. Design of automatic evaluation of machine English translation based on BP neural network. **Lecture Notes on Data Engineering and Communications Technologies**, v. 123, p. 359-367, 2022.