Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Engenharia Elétrica

Trabalho de Conclusão de Curso

Interface homem-máquina controlada pela fala com correção de erro

Nicolas Vieira Magalhães

João Monlevade, MG
2026

Nicolas Vieira Magalhães

Interface homem-máquina controlada pela fala com correção de erro

Trabalho de Conclusão de Curso apresentado à Universidade Federal de Ouro Preto como parte dos requisitos para obtenção do Título de Bacharel em Engenharia Elétrica pelo Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto.
Orientador: Prof. Dr. Glauco Ferreira Gazel Yared

Universidade Federal de Ouro Preto
João Monlevade
2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

M188i Magalhães, Nicolas Vieira.
Interface homem-máquina controlada pela fala com correção de erro. [manuscrito] / Nicolas Vieira Magalhães. - 2026.
31 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Glauco Ferreira Gazel Yared.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia Elétrica .

1. Correção de erros de fala (Computadores). 2. Interação humano-máquina. 3. Processamento de linguagem natural (Computação). 4. Reconhecimento automático de fala. 5. Reconhecimento automático da voz. I. Yared, Glauco Ferreira Gazel. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.5:004.934

Bibliotecário(a) Responsável: Flavia Reis - CRB6/2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA ELETRICA

FOLHA DE APROVAÇÃO

Nicolas Vieira Magalhães

Interface homem-máquina controlada pela fala com correção de erro

Monografia apresentada ao Curso de Engenharia Elétrica da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia Elétrica

Aprovada em 5 de março de 2026.

Membros da banca

Doutor - Glauco Ferreira Gazel Yared - Orientador - Universidade Federal de Ouro Preto
Doutora - Gilda Aparecida de Assis - Universidade Federal de Ouro Preto
Doutor - Marcelo Moreira Tiago - Universidade Federal de Ouro Preto

Glauco Ferreira Gazel Yared, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/04/2026.

Documento assinado eletronicamente por **Glauco Ferreira Gazel Yared**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 14/04/2026, às 21:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1089260** e o código CRC **A3A6A62C**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0818 - www.ufop.br

Dedico este trabalho a Deus, por me conceder força, sabedoria e perseverança ao longo desta jornada. Aos meus pais, pelo amor, apoio e pelos valores que me ensinaram, que foram fundamentais para que eu chegasse até aqui. E à minha esposa, Camila, que me lembra todos os dias do que realmente importa, sendo meu apoio, inspiração e porto seguro.

Resumo

O objetivo deste trabalho é propor um sistema de correção de erros em interfaces de comunicação homem-máquina controladas pela fala, capaz de aprimorar a acurácia do reconhecimento e reduzir falhas de interação. A metodologia emprega algoritmos que processam palavras reconhecidas pelo sistema, realizam comparações com um banco de dados de comandos e aplicam técnicas de aproximação e substituição para corrigir eventuais discrepâncias. A aplicação dessa abordagem busca uma interface mais precisa e eficiente, elevando a taxa de reconhecimento correto e proporcionando uma interação mais satisfatória ao usuário. Os resultados experimentais demonstraram que é possível atingir altas taxas de acerto, mesmo com frases contendo erros inseridos artificialmente. Entre os métodos avaliados, o método de espaço multi-dimensional destacou-se como o mais robusto, apresentando os melhores desempenhos especialmente em cenários com múltiplos erros (duas ou três falhas por frase), o que valida a viabilidade da técnica para tornar interfaces de voz mais precisas, eficientes e satisfatórias para o usuário.

Palavras-chave: Reconhecimento de fala, Correção de erros, Comunicação homem-máquina, Interfaces controladas pela fala, Correção por similaridade, Correção por proximidade espacial.

Abstract

The objective of this work is to propose an error correction system for speech-controlled human-machine communication interfaces, capable of improving recognition accuracy and reducing interaction failures. The methodology employs algorithms that process words recognized by the system, perform comparisons against a command database, and apply approximation and substitution techniques to correct potential discrepancies. This approach aims to create a more precise and efficient interface, increasing the correct recognition rate and providing a more satisfactory user interaction. Experimental results demonstrated that it is possible to achieve high success rates, even with sentences containing artificially inserted errors. Among the methods evaluated, the multi-dimensional space method stood out as the most robust, showing the best performance especially in scenarios with multiple errors (two or three failures per sentence), which validates the viability of the technique for making voice interfaces more precise, efficient, and satisfactory for the user.

Keywords: Speech recognition, Error correction, Human-machine communication, Speech-controlled interfaces, Similarity-based correction, Spatial proximity correction.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Interfaces homem-máquina (HMI, do inglês *Human-Machine Interface*) controladas pela fala estão cada vez mais presentes no cotidiano, oferecendo facilidade, praticidade e uma comunicação intuitiva para a maioria dos usuários. Segundo (MALLIOS; BOURBAKIS, 2016), avanços em áreas da ciência da computação, como a Inteligência Artificial, possibilitaram a condução de diálogos entre pessoas e computadores, consolidando os sistemas de diálogo como uma nova categoria de interação humano-computador. Essa tendência é impulsionada não apenas pela busca por acessibilidade, mas também pela melhoria na qualidade de vida e pela eficiência operacional dos sistemas.

Conforme apresentado por (RUSAN; MOCANU, 2022), o uso da fala proporciona uma interação mais natural, permitindo uma troca ágil de informações. O futuro das HMI deve priorizar cenários centrados no usuário, promovendo uma comunicação recíproca e eficaz. Nesse sentido, a escolha das tecnologias de interação deve minimizar distrações e otimizar a transmissão de dados, garantindo que as tarefas sejam realizadas de forma eficiente (mensurada pelo tempo de resposta e pela Taxa de Erro de Palavras – *Word Error Rate*), prática (avaliada pela Taxa de Sucesso na Execução das tarefas) e satisfatória (mensurada pela percepção do usuário via *System Usability Scale* – SUS).

Diante da complexidade da comunicação verbal, este trabalho estabelece a meta de 90% de acerto no reconhecimento dos comandos. O desafio central reside nas limitações dos algoritmos atuais, que, ao cometerem erros de reconhecimento, comprometem a resposta e a experiência do usuário.

Um dos principais problemas ao desenvolver sistemas de diálogo é garantir a correção das interações baseadas em comandos de fala entre humanos e máquinas. Com as limitações da tecnologia de fala atual e da tecnologia de *Spoken Language Understanding* (SLU), erros na identificação da entrada do usuário e na compreensão das transcrições de palavras a partir da saída do reconhecedor podem levar a dificuldades tanto para o sistema quanto para seus usuários no diálogo subsequente. De acordo com Zhou et al. (2006), a necessidade de elaborar estratégias melhores para detectar problemas em diálogos entre humanos e máquinas e, em seguida, lidar com eles de maneira elegante, tornou-se primordial para os sistemas de diálogo falado. O desenvolvimento de um sistema baseado na fala enfrenta o desafio de lidar ou identificar variações nos padrões dos comandos, reconhecendo possíveis erros e ajustando-os para garantir uma interpretação precisa e coesa pelo sistema.

Nessa perspectiva, o sistema proposto neste trabalho visa não apenas reconhecer a diversidade nos padrões de comandos impostos pelo usuário, mas também corrigir potenciais equívocos, elevando assim a taxa de acerto do sistema. Essa abordagem busca atingir um equilíbrio entre o objetivo do usuário e a clareza dos comandos impostos por ele, ao

mesmo tempo em que amplia a capacidade do sistema em interpretar a complexidade intrínseca da fala do usuário.

O objetivo é alcançar uma taxa de acerto superior a 90% na identificação correta das intenções do usuário. Este aprimoramento visa atender às demandas do mercado por sistemas mais avançados e também marca um avanço na busca por interfaces mais intuitivas e adaptáveis.

Portanto, este trabalho tem por finalidade o desenvolvimento de métodos para aprimoramento de um sistema de diálogo, em um contexto no qual ocorram erros no reconhecimento de fala.

1.1 Problema

Na prática, a utilização de assistentes virtuais, busca-se proporcionar uma interação intuitiva e dinâmica, em um ambiente mais confortável e prático. Entretanto, em alguns momentos, essas ferramentas cometem erros ou enfrentam dificuldades na interpretação dos comandos a elas impostos.

A comunicação pela fala enfrenta um desafio significativo: a perda ou transmissão incorreta de informações, seja por falhas no método utilizado (sistema de reconhecimento de fala) ou por erros do usuário. Essa limitação é destacada por Errattahi, Hannani e Ouahmane (2018),a persistência de altas taxas de erro compromete diretamente a usabilidade dos sistemas, uma vez que a descontextualização de palavras e frases resulta em respostas inadequadas. Palavras e frases podem perder seu sentido quando inseridas em contextos diferentes, resultando em respostas inadequadas e levando o usuário ao erro, à frustração e à impaciência. Visando aprimorar a experiência do usuário nessa forma de comunicação, os sistemas incorporam métodos de análise e quantificação de erros, permitindo ajustes por meio de algoritmos de previsão de dados, análise do banco de dados, repetição ou até mesmo solicitação de um novo comando.

Um dos desafios neste trabalho reside na criação de uma abordagem em que o sistema possa identificar o erro e compreender qual comando o usuário pretendia transmitir por meio de sua fala. Isso inclui situações em que o usuário utiliza palavras diferentes das cadastradas, mas com o mesmo significado, como sinônimos, que podem ser interpretados como palavras distintas e fora do contexto. A pronúncia incorreta de uma palavra também introduz a possibilidade de gerar um novo comando que pode ou não ser reconhecido pela máquina, resultando em respostas insatisfatórias ou até mesmo na ausência de resposta. Lidar com as lacunas na comunicação impostas por essas variações representa um dos aspectos desafiadores a serem superados no desenvolvimento desses sistemas.

1.2 Motivação e justificativa

O crescente protagonismo dos sistemas de Reconhecimento Automático de Fala (ASR, do inglês *Automatic Speech Recognition*) na interação Humano–Computador tem transformado significativamente a forma como os indivíduos acessam informações e interagem com dispositivos tecnológicos. Desde assistentes virtuais de uso doméstico até interfaces de voz empregadas em setores críticos, como saúde, centros de contato e sistemas embarcados, observa-se uma crescente dependência da precisão e confiabilidade desses sistemas.

Apesar dos avanços expressivos proporcionados por técnicas modernas de aprendizado, os sistemas ASR ainda enfrentam desafios relevantes, dentre os quais se destaca a ocorrência de erros de transcrição. Tais erros podem ser decorrentes de diversos fatores, como a presença de ruído ambiente, a variabilidade acústica entre falantes e fenômenos linguísticos, como a homofonia. Como resultado, o sistema pode gerar transcrições foneticamente semelhantes à fala original, porém semanticamente incorretas e desalinhadas da real intenção do usuário.

Diante desse cenário, este estudo propõe ir além da simples identificação de erros, concentrando-se na correção ativa e inteligente; ou seja, na capacidade de identificar falhas e tratá-las corretamente nas transcrições produzidas pelo ASR. Para isso, são desenvolvidos e comparados três modelos distintos de correção de erros, com o objetivo de identificar um algoritmo capaz de refinar as saídas do sistema, ajustando frases incorretas para aquelas que sejam mais prováveis e semanticamente coerentes dentro de uma base de dados previamente estabelecida. Dessa forma, busca-se contribuir para o aumento da robustez e da confiabilidade dos sistemas de interação por voz, especialmente em aplicações nas quais a precisão é um requisito essencial.

1.3 Objetivos

O objetivo geral deste trabalho é o desenvolvimento de métodos para a melhoria na interpretação de sequências de palavras ou comandos em um sistema de diálogo baseado na fala. Os objetivos específicos são:

- Estabelecer uma representação numérica para as palavras do dicionário do sistema, a partir da qual seja possível determinar pontos correspondentes as ideias ou frases produzidas pelo usuário, as quais demandam uma resposta do sistema (ação ou informação);
- Determinar a taxa de acerto do sistema de diálogo convencional e daquele que emprega métodos de melhoria na interpretação visando uma taxa superior a 90%.

- Comparar os modelos de busca implementados via código neste trabalho, identificando aquele que apresenta maior robustez e desempenho frente ao aumento do número de erros nas transcrições.

1.4 Organização do trabalho

Este Trabalho de Conclusão de Curso encontra-se estruturado em capítulos que descrevem desde a fundamentação teórica até a validação experimental do sistema proposto, conforme detalhado a seguir:

- **Capítulo 1: Introdução** - Apresenta o contexto do reconhecimento de voz, a motivação para o desenvolvimento de sistemas de correção de erros em interfaces homem-máquina, bem como os objetivos e a estrutura do trabalho.
- **Capítulo 2: Revisão teórica** - Aborda os conceitos fundamentais de processamento de fala, métricas de similaridade textual e os métodos de correção de erros, incluindo a base matemática para o método de espaço multi-dimensional.
- **Capítulo 3: Desenvolvimento do trabalho** - Detalha a arquitetura do sistema desenvolvido. Aqui, descreve-se o fluxo de processamento, desde o banco de dados de comandos até os algoritmos de aproximação e substituição de termos.
- **Capítulo 4: Resultados** - Apresenta a análise experimental. Inclui a metodologia de testes com erros inseridos artificialmente e a comparação de desempenho entre os métodos avaliados, evidenciando a robustez da técnica proposta.
- **Capítulo 5: Conclusão e trabalhos futuros** - Sintetiza as principais contribuições do trabalho, discute as limitações encontradas e sugere caminhos para pesquisas e desenvolvimentos futuros.

2 Revisão teórica

2.1 Introdução

A integração de interfaces homem-máquina baseadas em comandos de voz tem se consolidado no cotidiano, abrangendo uma ampla gama de aplicações voltadas ao aumento da comodidade e à satisfação das necessidades dos usuários. De acordo com El-Azazy et al. (2025), os sistemas de controle baseados na fala revolucionaram a forma como as pessoas interagem com os computadores, permitindo o uso de comandos baseados na fala e linguagem natural. O emprego de uma interface intuitiva e de fácil utilização surge como alternativa às interfaces textuais ou dependentes da digitação. A adoção da fala como mecanismo de interação mostra-se natural para grande parte dos indivíduos, permitindo uma comunicação direta e simplificada entre o usuário e o sistema (JURAFSKY; MARTIN, 2013).

2.2 Interfaces homem-máquina baseadas em fala

A comunicação homem-máquina por meio de interfaces baseadas na fala tem se consolidado como uma área de grande relevância no campo da computação e da engenharia. Tais interfaces possibilitam que os usuários interajam com dispositivos e aplicações utilizando a linguagem natural, tornando a operação mais intuitiva e acessível. Entretanto, apesar dos avanços tecnológicos observados nas últimas décadas, os sistemas de reconhecimento de fala ainda enfrentam desafios consideráveis relacionados à interpretação semântica dos comandos e à capacidade de adaptação a diferentes contextos de uso. Segundo Wilpon e Roe (1994), embora tais interfaces promovam uma interação mais natural entre humanos e máquinas, elas continuam limitadas por fatores como ambiguidades contextuais e variações no padrão de fala dos usuários.

2.2.1 Reconhecimento de fala

O reconhecimento automático de fala ASR é a tecnologia responsável por converter sinais acústicos em texto ou comandos compreensíveis pelo computador. Esse processo envolve diversas etapas, como o pré-processamento do sinal de áudio, a extração de características relevantes e a classificação do sinal utilizando métodos probabilísticos ou baseados em aprendizado de máquina, como os Modelos Ocultos de Markov (HMM, do inglês *Hidden Markov Models*) e as Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*).

Enquanto os HMMs são amplamente utilizados para modelar sequências temporais de sinais de fala, permitindo o reconhecimento de padrões mesmo na presença de ruído, as SVMs atuam na classificação de padrões complexos, sendo eficazes na distinção de fonemas ao estabelecer hiperplanos de separação em espaços de alta dimensionalidade. Já os coeficientes extraídos na escala Mel são eficazes na representação das propriedades espectrais do sinal de áudio de forma compacta e discriminativa, servindo como entrada para os classificadores utilizados no processo de reconhecimento.

De forma simplificada, um sistema de reconhecimento de fala recebe como entrada um sinal sonoro captado por um microfone e produz como saída uma representação simbólica, como uma palavra ou comando. Para isso, o sinal acústico passa por uma série de transformações que visam extrair informações relevantes da fala humana e compará-las com modelos previamente treinados, permitindo ao sistema decidir qual comando possui maior probabilidade de ter sido pronunciado.

2.2.2 Escala mel e extração de coeficientes cepstrais

A escala Mel é uma escala perceptual de frequências que busca aproximar matematicamente a forma como o ouvido humano percebe os sons, representando de maneira mais fiel a sensibilidade auditiva em diferentes faixas de frequência. Em sistemas de reconhecimento de fala, o uso dessa escala permite que a representação do sinal acústico esteja mais alinhada à percepção humana (DAVIS; MERMELSTEIN, 1980).

A partir da escala Mel, são extraídos os Coeficientes Cepstrais na Escala Mel (MFCC, do inglês *Mel-Frequency Cepstral Coefficients*), amplamente utilizados em aplicações de processamento de sinais de voz.

Além dos coeficientes MFCC, também podem ser calculadas suas derivadas de primeira e segunda ordem, conhecidas como delta e delta-delta, que capturam a dinâmica temporal da fala e contribuem para uma melhor caracterização do sinal ao longo do tempo.

Em termos conceituais, estes podem ser entendidos como uma forma de resumir o som da fala em um conjunto reduzido de valores numéricos que representam como o ouvido humano percebe variações de frequência e intensidade. Essa representação facilita a distinção entre diferentes sons da fala, ao mesmo tempo em que reduz a quantidade de dados a serem processados pelos classificadores.

2.2.3 Modelos ocultos de markov

Os HMM são modelos estatísticos amplamente utilizados para representar sequências temporais de eventos cujos estados reais não são diretamente observáveis. De acordo com Rabiner (2002), o conceito de estados ocultos e observações probabilísticas permite representar fenômenos complexos de forma estatística.

No reconhecimento de fala, um HMM pode representar uma unidade fonética ou subfonética, enquanto as observações correspondem às características acústicas extraídas do sinal, como os vetores MFCC. Segundo Yared (2006), os HMMs modelam de forma eficiente a variabilidade temporal e espectral dos padrões acústicos, isto é, as diferentes realizações de uma mesma unidade acústica e as transições entre diferentes unidades representadas em um espaço de características.

Os estados do HMM são denominados ocultos porque não correspondem diretamente a unidades observáveis no sinal de áudio, como palavras ou fonemas isolados. Em vez disso, cada estado representa uma etapa abstrata do processo de produção da fala, enquanto as observações correspondem aos vetores de características acústicas extraídos do sinal.

Para determinar qual sequência de estados de um HMM é mais provável dado um sinal de entrada, utiliza-se o algoritmo de Viterbi (VITERBI, 1967). Esse algoritmo calcula o caminho de estados que maximiza a probabilidade de gerar a sequência de observações acústicas extraídas da fala. De maneira intuitiva, o algoritmo avalia as possíveis sequências de estados e seleciona aquela que apresenta a maior probabilidade de ocorrência.

No contexto do reconhecimento de comandos de voz, o Algoritmo de Viterbi permite comparar o sinal de entrada com diferentes modelos previamente treinados e identificar aquele que apresenta a maior verossimilhança. Assim, o comando reconhecido corresponde ao modelo cujo caminho de estados mais provável explica melhor as características acústicas observadas, tornando o processo de decisão robusto mesmo na presença de ruídos ou variações na pronúncia.

Dessa forma, os HMMs mostram-se particularmente adequados ao reconhecimento de fala por lidarem naturalmente com a natureza sequencial do sinal acústico, permitindo modelar a evolução temporal da fala por meio de estados e transições probabilísticas.

2.2.4 Correção de erros em interfaces baseadas na fala

ASR podem apresentar falhas na identificação de comandos, seja por ambiguidades na pronúncia, sotaques regionais, ruídos ambientais ou limitações modelo acústico. Conforme destaca Yared (2006), o reconhecimento de fala continua enfrentando desafios significativos relacionados à precisão, à interpretação semântica e à adaptação a diferentes contextos de uso, especialmente em ambientes não controlados. Tais falhas podem gerar insatisfação do usuário e comprometer a eficiência da interação homem-máquina.

A literatura aponta que métodos de correção de erros, baseados em aproximação de palavras e substituição inteligente, podem aumentar a taxa de acerto sem exigir reprocessamento completo do sinal, como demonstrado por Leng et al. (2022) e Leng et al. (2021). Esses métodos analisam os comandos reconhecidos, comparam-nos a um banco de dados de frases esperadas ou a modelos de linguagem treinados e ajustam palavras incorretas ou parcialmente reconhecidas, estratégia também observada em abordagens de correção

contextual e ortográfica discutidas por Wang et al. (2021) e Bassil e Alwani (2012). De acordo com a revisão apresentada por Rahhal, Hannani e Ouahmane (2016), essas técnicas contribuem significativamente para o aprimoramento da precisão e da robustez dos sistemas de reconhecimento de voz.

2.2.5 Sistemas de interação homem-máquina baseados na fala

As interfaces controladas por fala, inseridas no contexto de Interação Homem-Computador (IHC, ou HCI, do inglês *Human-Computer Interaction*), promovem interações mais naturais e eficientes, colocando o usuário no centro do processo. De acordo com Oviatt (2000), ao reduzir os erros de reconhecimento e incorporar mecanismos automáticos de correção, essas interfaces diminuem a frustração e o esforço necessário para resolver problemas, permitindo que o usuário se concentre em suas tarefas com maior confiança. A utilização de múltiplas modalidades de interação contribui para evitar e recuperar erros de forma mais eficaz, ampliando a acessibilidade da tecnologia a pessoas de diferentes idades, habilidades e condições sensoriais. Dessa forma, a robustez e a adaptabilidade desses sistemas tornam a experiência mais confortável, inclusiva e confiável, melhorando a qualidade de vida digital do usuário.

2.2.6 Assistência e acessibilidade

Os sistemas de diálogo baseados na fala são projetados não apenas com o objetivo de promover a inclusão de pessoas com deficiência visual, motora ou com limitações que dificultem o uso de interfaces convencionais, mas também para oferecer maior praticidade e facilidade de uso ao público em geral. Dessa forma, tal sistema de diálogo busca tornar a interação mais intuitiva, eficiente e acessível a todos os usuários, independentemente de suas habilidades ou familiaridade com tecnologias digitais.

Essa abordagem oferece maior autonomia, conforto e segurança, ao mesmo tempo em que promove a inclusão de públicos que, de outra forma, poderiam enfrentar barreiras no acesso às informações e aos sistemas informatizados.

Atualmente, uma variedade de sistemas de reconhecimento pela fala integram-se de maneira significativa ao cotidiano da população. Dentre esses, destacam-se a Siri, desenvolvida pela Apple, o Google Assistant, desenvolvido pelo Google, e a Alexa, desenvolvida pela Amazon, consolidando-se como tecnologias que simplificam algumas tarefas da vida diária. Essas assistentes virtuais não apenas se popularizaram, mas também se tornaram instrumentos úteis ao oferecerem funcionalidades como realizar pesquisas rápidas e gerenciar dispositivos conectados à rede.

Essas plataformas, que inicialmente surgiram como tecnologias inovadoras, evoluíram para se tornarem partes integrantes das rotinas das pessoas. Sua presença contribui diretamente para o aumento da produtividade, oferecendo soluções ágeis e práticas para

diversas demandas do dia a dia. A capacidade de realizar tarefas simples por meio de comandos pela fala, sendo este o meio mais natural de comunicação dos seres humanos (SCHAFER, 1995), simplifica significativamente as interações tecnológicas, promovendo uma experiência mais eficiente e conveniente para o usuário. Assim, os assistentes desempenham um papel importante na criação de um ambiente mais integrado e acessível para os usuários modernos.

2.2.7 Correção e adaptação de comandos de voz

Um dos principais desafios em sistemas de reconhecimento de fala é lidar com comandos mal compreendidos ou expressões ambíguas. Nesse contexto, a capacidade do sistema de corrigir e ajustar comandos de forma inteligente torna-se essencial para garantir uma interação contínua com o usuário.

Com base nos conceitos apresentados, observa-se que a combinação entre técnicas clássicas de reconhecimento de fala e métodos de correção de erros constitui uma abordagem viável para sistemas de interação por voz com vocabulário restrito. Esses fundamentos teóricos servem como base para a metodologia adotada neste trabalho, detalhada no capítulo seguinte.

2.3 Abordagem do sistema: Dependência de locutor

A implementação desenvolvida neste trabalho restringe-se a um sistema dependente de locutor. Esta escolha metodológica não reflete uma limitação técnica, mas sim uma decisão deliberada pautada na praticidade e na simplicidade, essenciais para a validação dos objetivos propostos.

O foco central desta monografia é a eficácia de algoritmos de correção de erros em comandos de voz, e não o desenvolvimento de um motor de reconhecimento universal. Para isolar e validar a performance da técnica de correção e a robustez da substituição de termos, um sistema controlado é suficiente para demonstrar o comportamento do algoritmo sem o ruído das variáveis presentes em sistemas independentes de falante.

A implementação de sistemas independentes de locutor exigiria um volume de dados de treinamento significativamente maior e uma infraestrutura complexa de modelagem acústica, aspectos que excedem o escopo de validação do método de correção proposto. Ao restringir o sistema a um locutor específico, é possível concentrar a análise no processamento semântico e na resiliência do sistema frente às discrepâncias de reconhecimento, garantindo uma avaliação clara e mensurável da viabilidade da técnica.

2.4 Trabalhos relacionados

A correção automática de erros em sistemas ASR tem sido amplamente investigada na literatura como uma etapa complementar essencial para aumentar a robustez e a confiabilidade das interfaces baseadas em voz. Diante das limitações inerentes aos modelos acústicos e linguísticos, especialmente em ambientes ruidosos ou com variações linguísticas, diversos trabalhos propõem estratégias de pós-processamento capazes de mitigar erros de transcrição sem a necessidade de reprocessar o sinal acústico original.

Nesse contexto, abordagens recentes exploram diferentes paradigmas para a correção de erros, incluindo modelos não-autorregressivos baseados em alinhamento de edição, o aproveitamento de múltiplas hipóteses geradas pelo sistema ASR e *frameworks* não supervisionados que reduzem a dependência de grandes volumes de dados anotados. Tais estratégias buscam equilibrar precisão, eficiência computacional e aplicabilidade em tempo real, aspectos fundamentais para sistemas de interação homem-máquina.

Os trabalhos apresentados a seguir ilustram essas tendências e evidenciam avanços significativos na área de correção de erros em ASR, servindo como base conceitual e comparativa para a abordagem proposta neste trabalho.

Dentro desse contexto, destaca-se a proposta do *FastCorrect* (LENG et al., 2021), que é um método eficiente para correção de erros em sistemas ASR. Diferentemente das abordagens tradicionais baseadas em modelos autoregressivos, que apresentam alta latência, o método *FastCorrect* adota um modelo não-autorregressivo guiado por alinhamento de edição, explorando diretamente as operações de inserção, remoção e substituição identificadas por meio da distância de edição entre a transcrição incorreta e o texto de referência. Essa estratégia permite a geração paralela das correções, resultando em ganhos significativos de velocidade sem comprometer a precisão. Os resultados experimentais demonstram reduções relevantes na Taxa de Erro de Palavras (WER, do inglês *Word Error Rate*), com ganhos de precisão situados entre 8% e 14%. Além da melhoria na qualidade das transcrições, a abordagem evidencia um alto potencial para aplicações em tempo real, uma vez que o processamento se mostrou até 9 vezes mais veloz que os modelos de correção convencionais.

Enquanto o *FastCorrect* prioriza a eficiência computacional por meio do alinhamento direto das operações de edição, outras abordagens exploram o uso de informações adicionais fornecidas pelo próprio sistema ASR. Nesse sentido, Ma et al. (2023) propõem o método N-best T5, um modelo para correção de erros em reconhecimento automático de fala que utiliza listas de hipóteses de saída (N-best) em vez de apenas a melhor transcrição. Ao permitir a comparação entre múltiplas transcrições candidatas, o modelo melhora a detecção e correção de erros. A adaptação do modelo pré-treinado T5 e a implementação de um processo de decodificação restrita resultam em reduções superiores a 20% na WER. Os resultados evidenciam ganhos significativos de robustez e precisão ao utilizar múltiplas hipóteses (N-best) em comparação às abordagens tradicionais baseadas apenas

na hipótese mais provável (*1-best*), consolidando o potencial de modelos de linguagem para a correção de erros de fala.

Apesar dos avanços obtidos com modelos supervisionados e pré-treinados, a dependência de grandes conjuntos de dados anotados ainda representa um desafio relevante. Com o objetivo de mitigar essa limitação, Guo et al. (2024) desenvolveram o método UCorrect, um *framework* não supervisionado para correção de erros em sistemas ASR. A abordagem é composta por três módulos — detector, gerador e seletor — responsáveis por identificar erros potenciais, gerar candidatos de correção e selecionar a alternativa mais confiável. Essa estrutura reduz a necessidade de dados pareados, aumenta a explicabilidade do processo e evita a modificação indevida de *tokens* corretos — unidades fundamentais de texto, como palavras ou subpalavras, resultantes do processo de segmentação (*tokenização*). Os resultados obtidos em bases públicas demonstram reduções de até 10,2% na WER, mantendo baixa latência e desempenho consistente em diferentes configurações de decodificação.

Em suma, a literatura recente tem explorado diversas vertentes para a correção de erros em sistemas ASR, destacando-se o uso de modelos não-autorregressivos (LENG et al., 2021), estratégias baseadas em múltiplas hipóteses (*N-best*) (MA et al., 2023) e *frameworks* não supervisionados (GUO et al., 2024). Embora tais abordagens tenham alcançado avanços expressivos na redução da WER, ainda persistem desafios relacionados ao equilíbrio entre a capacidade de generalização para domínios específicos e a viabilidade de implementação em dispositivos com recursos computacionais limitados.

2.5 Considerações parciais

A revisão teórica realizada neste capítulo permitiu estabelecer uma base sólida para a compreensão dos sistemas de interação homem-máquina baseados em voz. Observou-se que, embora o ASR tenha atingido um estágio de maturidade que permite sua ampla adoção em assistentes virtuais cotidianos, a precisão da transcrição ainda é sensível a fatores como ruído ambiente, variabilidade linguística e limitações dos modelos acústicos e linguísticos subjacentes.

Conforme discutido, a extração de características acústicas via coeficientes MFCC, aliada ao poder de modelagem sequencial dos Modelos Ocultos de Markov (HMM) e do algoritmo de Viterbi, constitui o paradigma clássico que viabiliza a conversão de sinais analógicos em representações simbólicas compreensíveis por máquinas. Contudo, a persistência de erros de reconhecimento impõe a necessidade de mecanismos de pós-processamento.

A análise dos trabalhos relacionados evidenciou que a fronteira tecnológica atual busca superar a dependência de grandes conjuntos de dados rotulados e reduzir a latência de sistemas de correção. Abordagens como o *FastCorrect* (LENG et al., 2021), o uso de listas N-best (MA et al., 2023) e *frameworks* não supervisionados como o UCorrect

(GUO et al., 2024) demonstram que é possível otimizar a robustez das interfaces sem a necessidade de reprocessar o sinal de áudio original.

Dessa forma, a convergência entre técnicas clássicas de processamento de sinais e métodos modernos de correção de erros define o escopo metodológico para o presente trabalho, servindo como alicerce para o desenvolvimento de uma solução que busque equilibrar eficiência computacional, precisão na interpretação semântica e adaptabilidade a diferentes contextos de interação.

3 Desenvolvimento do trabalho

3.1 Introdução

Primeiramente, torna-se imprescindível criar um banco de dados para o desenvolvimento do sistema de diálogo. Visando um modelo introdutório mais simples, optou-se pela implementação de um sistema dependente de locutor, ou seja, criado a partir dos dados obtidos de um único locutor, o qual será o usuário do sistema. Essa condição de contorno visa reduzir a variabilidade acústica, permitindo que a análise se concentre exclusivamente na etapa de correção textual dos comandos reconhecidos.

Nesta fase, desenvolveu-se um algoritmo para a aquisição de sinais de fala fornecidos pelo usuário, formando, assim, um banco de dados a ser utilizado no sistema de reconhecimento de comandos e correção de erros.

Ao implementar o sistema de correção de erros, alguns parâmetros devem ser considerados. O primeiro deles é a quantidade de erros permitidos em uma frase. Inicialmente, estabeleceu-se como premissa a ocorrência de apenas um erro por frase, sendo esta formada por quatro palavras, com o objetivo de simplificar a implementação. Contudo, durante a fase experimental, os métodos desenvolvidos passaram a ser avaliados em cenários com múltiplos erros, a fim de analisar sua robustez frente ao aumento da complexidade do problema. Um segundo fator relevante é o tamanho de cada frase; por uma análise de frases comuns do dia a dia foi adotado um limite de quatro palavras por frase nesta etapa visando praticidade de implementação. Esses parâmetros fundamentais são essenciais para a realização dos experimentos, a análise dos resultados e a simplicidade da operação do sistema.

3.1.1 Aquisição de dados

Para que o algoritmo possa operar, é essencial a criação de um banco de dados. Nesta etapa, definiu-se que o banco de dados utilizado neste trabalho seria composto por 19 palavras, formando 6 frases distintas. As gravações foram realizadas em um único dia, em um ambiente doméstico com pouco ruído, utilizando um microfone de headset Redragon Minos H210, caracterizado por uma resposta de frequência de 100 Hz a 10 kHz e padrão de captação omnidirecional. As especificações do sistema de aquisição para gravar os dados são apresentadas na Tabela 1, contando com 20 gravações de cada palavra, a fim de proporcionar modelos robustos para o treinamento do algoritmo de reconhecimento de comandos. A padronização das condições de aquisição minimiza interferências externas, garantindo a integridade dos sinais para as etapas subsequentes de extração de características.

Tabela 1 – Especificações do sistema de aquisição utilizado durante a etapa de gravação da base de dados.

Parâmetro	**Valor**
Taxa de amostragem (Fs)	16 kHz
Resolução (bits)	16 bits
Canais	1 (mono)
Duração da gravação	3 segundos
Formato de armazenamento	`.mat`

Fonte: Autoria própria.

Para a etapa de treinamento, foram utilizadas 15 amostras das palavras gravadas para o aprendizado do modelo e 5 amostras para a validação. Foram realizadas 5 partições entre a base de dados de treinamento e a de verificação, variando-se aleatoriamente as amostras utilizadas em cada conjunto. A codificação numérica atribuída a cada palavra, bem como a estrutura das frases que compõem o banco de dados, são apresentadas na Tabela 2, que mapeia cada índice à sua respectiva palavra no vocabulário.

Tabela 2 – Estrutura do banco de frases: mapeamento entre índices e palavras.

1ª Palavra		**2ª Palavra**		**3ª Palavra**		**4ª Palavra**	
Índ.	Palavra	Índ.	Palavra	Índ.	Palavra	Índ.	Palavra
01	Bom	02	Dia	06	Tudo	07	Bem
03	Boa	04	Tarde	06	Tudo	07	Bem
08	Onde	09	Fica	10	A	11	Biblioteca
12	Qual	13	Horário	14	De	15	Almoço
16	Me	17	Informe	18	O	19	Clima

Fonte: Autoria própria.

3.1.2 Tratamento de dados

Após a etapa de aquisição do sinal de voz, foi realizado o pré-processamento dos dados com o objetivo de preparar o sinal para a extração de características e o posterior reconhecimento de comandos. Esse processo foi implementado em ambiente MATLAB por meio de um algoritmo que executa diversas operações de pré-processamento e parametrização do sinal gravado.

Inicialmente, o sistema realiza a aquisição do sinal de áudio em tempo real utilizando o microfone. O algoritmo verifica a potência do sinal adquirido, garantindo que o áudio possua um nível de energia mínimo para ser considerado como fala. A energia local $E(n)$ de um sinal x em uma janela de tamanho L é calculada por:

$$E(n) = \sum_{k=n}^{n+L-1} x(k)^2 \tag{3.1}$$

Caso a energia seja inferior a um limiar pré-definido, o sistema interpreta o sinal como ruído e interrompe o processamento.

Na sequência, o sinal é reamostrado de 16kHz para 8kHz, reduzindo o volume de dados a serem processados e submetido a uma função de detecção de energia para isolar o intervalo útil da locução. Aplica-se então um filtro de pré-ênfase de primeira ordem, definido por $y(n) = x(n) - 0{,}95x(n-1)$, visando atenuar componentes de baixa frequência e destacar as regiões espectrais mais relevantes. O sinal é segmentado em janelas de 20 ms com sobreposição de 50%, sobre as quais é aplicada a janela de Hamming $w(n)$:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{3.2}$$

A partir de cada janela, extraem-se os coeficientes MFCCs. O cálculo inicia-se com a Transformada Rápida de Fourier (FFT, do inglês *Fast Fourier Transform*) para obter o espectro de potência $|X(k)|^2$. Em seguida, aplica-se um banco de filtros triangulares na escala Mel. O logaritmo da energia de cada filtro m, denotado por $S(m)$, é dado por:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right) \tag{3.3}$$

onde $H_m(k)$ é a resposta em frequência do m-ésimo filtro. Os coeficientes MFCC (C_n) são obtidos pela Transformada Discreta de Cosseno (DCT, do inglês *Discrete Cosine Transform*):

$$C_n = \sum_{m=1}^{M} S(m)\cos\left[n\,(m-0{,}5)\,\frac{\pi}{M}\right] \tag{3.4}$$

Desses coeficientes calcula-se também a energia logarítmica e as derivadas de primeira (Δ) e segunda ordem ($\Delta\Delta$). A derivada de primeira ordem é aproximada por:

$$\Delta_n = \frac{\sum_{k=-K}^{K} k \cdot c_{n+k}}{\sum_{k=-K}^{K} k^2} \tag{3.5}$$

onde c_n representa o coeficiente no quadro n. Os vetores resultantes, concatenando MFCC, energia, Δ e $\Delta\Delta$, formam o padrão acústico que é comparado aos modelos estatísticos baseados em HMM.

3.1.3 Propostas

Com base no problema apresentado, é crucial definir qual tipo de erro o algoritmo criado visa solucionar. Nesse contexto, optou-se por focar na detecção e correção de erros causados por palavras diferentes em uma sentença esperada, a qual pode estar associada a uma ação demandada pelo usuário, ou a uma resposta que possa ser compreendida pelo sistema. A lógica de operação do sistema, ilustrada na Figura 1, demonstra o fluxo de decisão entre a aceitação de um comando válido e o acionamento dos mecanismos de correção para sentenças desconhecidas.

É importante ressaltar que eventuais erros provenientes do sistema de reconhecimento de fala não comprometem a proposta deste trabalho. Pelo contrário, falhas na captação que resultem em transcrições incorretas dos comandos constituem o objeto de estudo central da pesquisa. Portanto, tais erros são considerados válidos e esperados, uma vez que o objetivo é justamente a análise da eficácia dos algoritmos em corrigi-los.

A estratégia consiste em utilizar o banco de dados como referência para criar um mapa de comandos e palavras, facilitando a identificação do erro e as possíveis correções que o sistema pode efetuar.

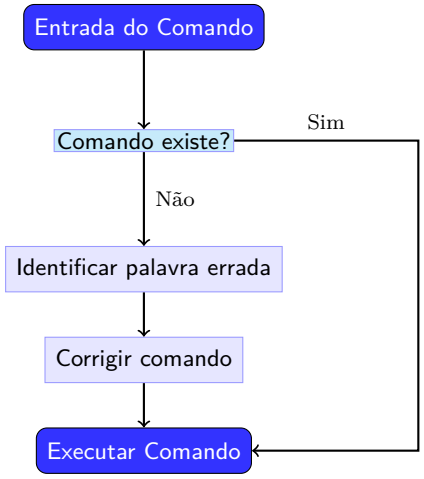


Figura 1 – Fluxo de decisão e correção de comandos.

3.2 Método 1 - busca palavras

No sistema desenvolvido, todas as palavras são armazenadas em um único vetor, em que cada posição representa uma palavra específica conforme apresentado na Tabela 2. As frases são formadas pela combinação ordenada dessas posições, o que permite identificá-las numericamente. Assim, o algoritmo não busca diretamente as palavras, mas sim suas posições no vetor, verificando se a sequência numérica gerada corresponde a uma frase previamente cadastrada no banco de dados. Caso a sequência seja reconhecida, o sistema executa o comando associado. Se não houver correspondência, inicia-se o processo de variação da frase, no qual o algoritmo busca a maior similaridade no banco de dados, de forma a corrigir possíveis erros de reconhecimento de fala.

Embora o método proposto tenha sido projetado para corrigir erros de reconhecimento de fala, o método implementado baseia-se em variações exaustivas das combinações possíveis, utilizando substituições sucessivas das posições no vetor sem empregar técnicas de otimização ou estratégias avançadas de busca, o que caracteriza uma abordagem de busca exaustiva.

Assim, o sistema recebe as palavras ditas pelo usuário, transformando-as em uma frase seguindo a estrutura da Tabela 3. Em seguida, realiza-se a conversão dessas palavras em valores numéricos com base na posição em que foram alocadas no vetor principal.

Tabela 3 – Estrutura da frase numérica.

1ª Palavra	**2ª Palavra**	**3ª Palavra**	**4ª Palavra**
Palavra 1	Palavra 2	Palavra 3	Palavra 4

Fonte: Autoria própria.

A função realiza a busca sequencial das palavras no banco de dados, mapeando suas posições para compor um vetor de identificação. Este vetor representa uma instrução cadastrada ou uma variação válida da mesma. Conforme ilustrado na Figura 2, o sistema primeiramente verifica a existência do comando original no banco de dados. Caso não seja localizado, inicia-se um processo iterativo de ajuste, onde o algoritmo modifica sistematicamente valores do vetor até encontrar uma correspondência válida ou atingir o limite de tentativas. Após a validação, o sistema prossegue para a execução da tarefa correspondente. Em cenários onde a validação direta falha, o sistema prioriza a busca pela sequência com maior similaridade semântica em relação à entrada do reconhecimento de fala, garantindo que uma resposta adequada seja entregue ao usuário.

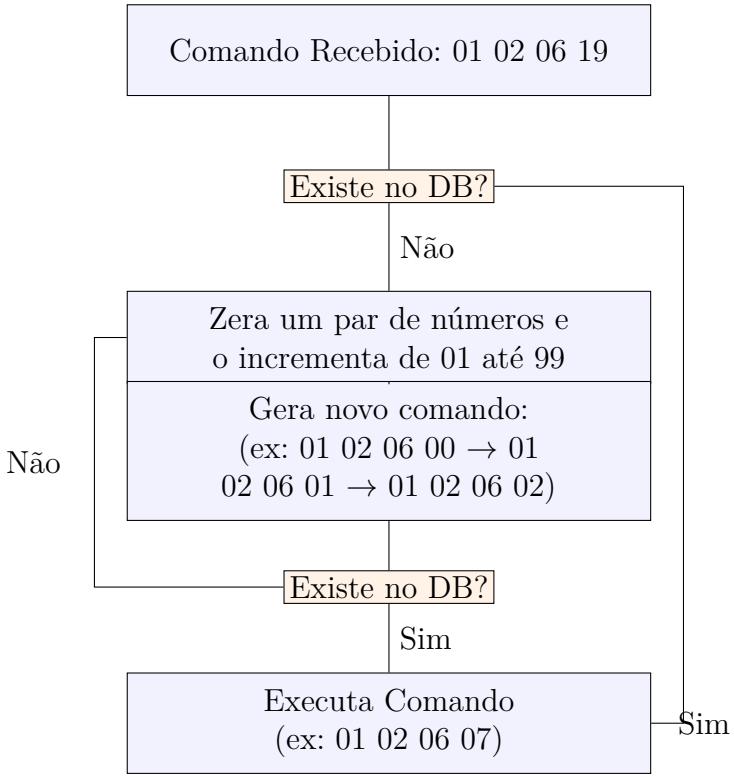


Figura 2 – Fluxo de processamento e correção do comando numérico.

Quando o sistema recebe uma entrada que diverge do padrão esperado, conforme ilustrado na Tabela 4, o algoritmo inicia um processo de ajuste iterativo. Este procedi-

mento modifica os índices da frase reconhecida incorretamente até localizar uma combinação equivalente armazenada no banco de dados, permitindo que a intenção do usuário seja processada corretamente mesmo diante de variações ou erros de reconhecimento.

Tabela 4 – Comparação entre a entrada recebida e a referência do banco de dados.

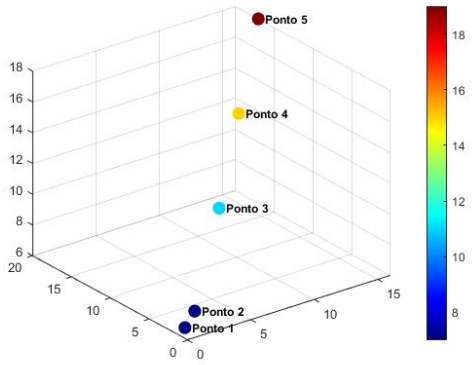
Origem	**1ª Palavra**	**2ª Palavra**	**3ª Palavra**	**4ª Palavra**
Referência (Certa)	Bom	Dia	Tudo	Bem
Entrada (Erro)	Bom	Dia	Tudo	Clima

Fonte: Autoria própria.

3.3 Método 2 - espaço multidimensional

Este algoritmo realiza a conversão da solicitação do usuário em um vetor de posições, representando a ordem das palavras fornecidas. A partir de uma matriz de frases, em que cada linha corresponde a um comando previamente reconhecido pelo sistema, o algoritmo define um espaço multidimensional, conforme ilustrado na Figura 3, no qual cada palavra ocupa uma coordenada em um dos eixos usando da cor para representar a quarta coordenada. A combinação dessas coordenadas gera um ponto único no espaço, representando um comando específico.

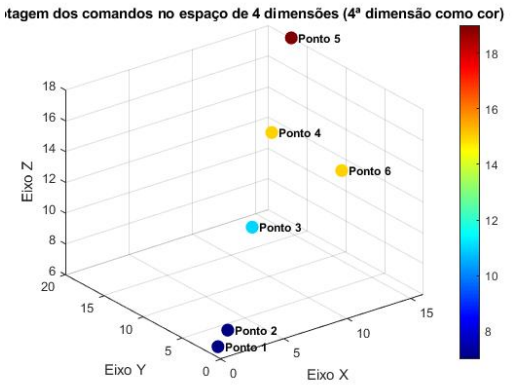
Figura 3 – Comandos no espaço multidimensional.



Fonte: Autoria própria.

Ao receber uma sequência de quatro palavras, o sistema gera um novo ponto no espaço correspondente à frase fornecida pelo usuário conforme ilustra a Figura 4 ao inserir o **Ponto 6** no espaço e, em seguida, identifica o ponto mais próximo mediante o cálculo da distância euclidiana em relação aos pontos existentes. Com base nesse cálculo, o sistema determina o ponto mais próximo como sendo o comando mais provável, o qual é então interpretado como a instrução fornecida pelo usuário.

Figura 4 – Erro no espaço multidimensional.



Fonte: Autoria própria.

3.3.1 Verificação de erro

Para evitar interpretações incorretas causadas pela proximidade euclidiana de comandos distintos, o ponto candidato é submetido a um processo de validação de similaridade. Este procedimento compara o vetor de entrada com o vetor de referência, exigindo uma coincidência mínima de coordenadas (25%, representando pelo menos uma coordenada correta) para validar a correção. Caso a similaridade seja inferior a este limiar, o sistema prossegue a busca pelo próximo vizinho mais próximo até encontrar uma correspondência que atenda ao critério estabelecido.

3.4 Método 3 - busca por similaridade

Diferentemente do método busca palavra, que realizava uma variação iterativa da sequência numérica com o objetivo de corrigir a frase reconhecida, o método proposto baseia-se na comparação direta entre a frase fornecida pelo usuário e um conjunto de frases previamente cadastradas em um banco de dados. O algoritmo retorna como resultado a frase que apresenta o maior grau de similaridade em relação à entrada.

Nesse método, as frases são representadas por vetores de índices inteiros, obtidos a partir de um vocabulário previamente definido. A similaridade entre a frase de entrada e cada frase do banco de dados é determinada pelo número de palavras coincidentes na mesma posição da sequência, caracterizando uma comparação posicional direta entre os vetores numéricos. Conforme ilustrado na Figura 5, o processo compreende desde a conversão da entrada para vetores de índices até a resolução de empates e a reconversão final para a execução do comando correspondente.

Dessa forma, o sistema mantém a representação das frases baseada em índices numéricos, porém substitui a abordagem de correção iterativa por um mecanismo de seleção direta da frase mais semelhante, conforme ilustrado na Tabela 5. Essa estratégia reduz

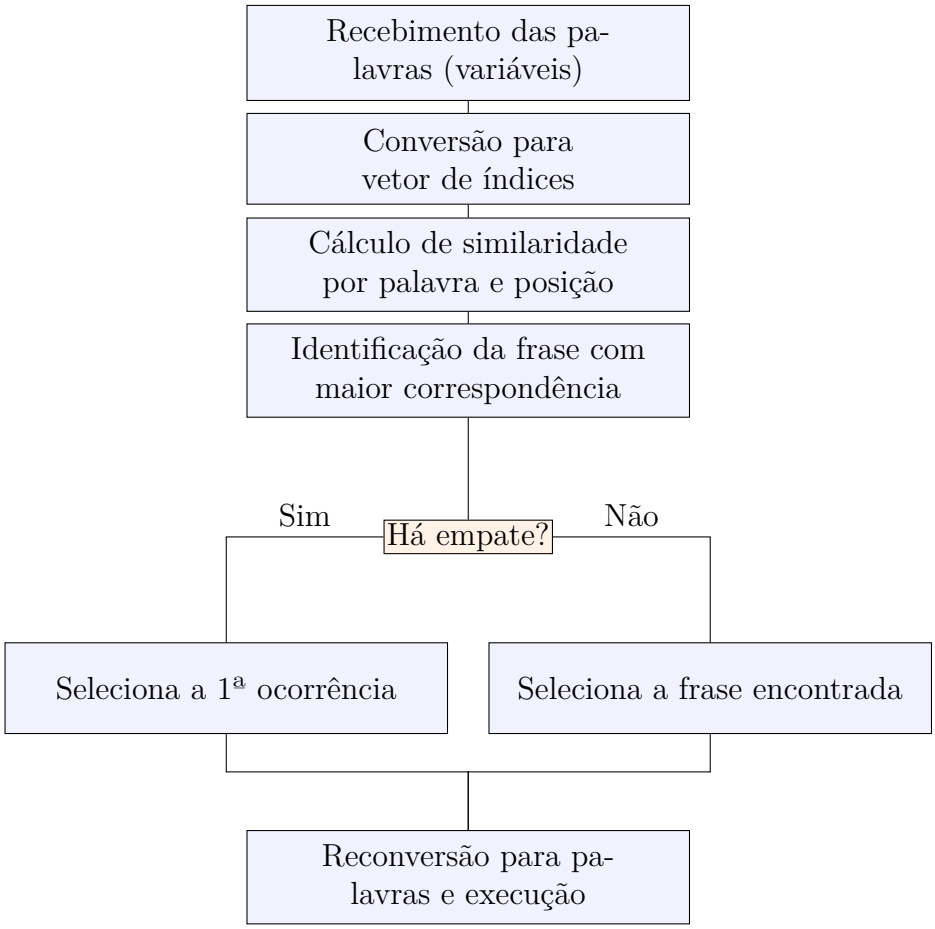


Figura 5 – Fluxograma do algoritmo de busca por similaridade posicional.

a complexidade do processo de correção, uma vez que elimina a necessidade de variações combinatórias, limitando o número de comparações ao total de frases cadastradas no banco de dados, e apresenta-se como uma solução simples e eficiente para o reconhecimento e validação de comandos de voz em sistemas com vocabulário e estrutura de frases previamente definidos.

Tabela 5 – Exemplo de correspondência de frase pelo método de similaridade posicional.

1ª Palavra	**2ª Palavra**	**3ª Palavra**	**4ª Palavra**
Bom	Dia	Tudo	Biblioteca
Frase mais próxima encontrada: Bom Dia Tudo Bem			

Fonte: Autoria própria.

4 Resultados

4.1 Introdução

Nesta seção são apresentados os resultados obtidos a partir dos testes realizados com o sistema de reconhecimento de fala e correção automática de erros desenvolvido neste trabalho. As análises contemplam tanto a validação do modelo de reconhecimento quanto a avaliação do desempenho dos algoritmos de correção, considerando métricas como taxa de acerto e tempo médio de processamento. Os resultados obtidos permitem discutir a eficiência, a robustez e a viabilidade do sistema em cenários com diferentes níveis de complexidade, evidenciando seu potencial de aplicação em interfaces homem-máquina baseadas em comandos de voz.

4.2 Modelo de reconhecimento

O conjunto de dados foi composto por locuções de áudio processadas digitalmente, totalizando 20 amostras para cada um dos 19 comandos de voz cadastrados no dicionário do sistema. Para a construção e validação dos modelos, aplicou-se a técnica de validação por reamostragem aleatória (*repeated random sub-sampling validation*), na qual as amostras foram particionadas aleatoriamente: 15 locuções (75%) foram destinadas à etapa de treinamento dos HMM e as 5 restantes (25%) foram utilizadas para a fase de testes e avaliação de desempenho.

Este procedimento foi repetido em cinco iterações independentes (variações de treinamento), permitindo que o algoritmo explorasse diferentes combinações de dados e evitasse vieses de seleção. Ao final das iterações, o sistema avaliou a acurácia de cada modelo treinado, selecionando aquele que apresentou o desempenho superior. Os modelos gerados exibiram taxas de reconhecimento situadas entre 94% e 100%. Como critério de otimização, foi selecionado o modelo que apresentou o melhor desempenho médio ao longo das iterações, atingindo taxa máxima de acerto de 100% nos testes realizados.

A adoção de múltiplas iterações de treinamento contribuiu para reduzir a dependência de uma única partição dos dados, tornando a avaliação mais robusta. Dessa forma, o modelo selecionado apresenta maior capacidade de generalização, fator essencial para aplicações práticas de reconhecimento de fala em ambientes reais.

4.3 Reconhecimento de comandos

Antes da aplicação dos algoritmos de correção, o sistema de captação e reconhecimento de fala passou por uma fase de validação de desempenho. Os parâmetros de pré-processamento, como filtros de ruído e normalização de volume, foram definidos através de um processo iterativo de ajuste empírico, visando maximizar a clareza do sinal de entrada. É importante destacar que este refinamento configurou apenas o estágio de recepção de dados, não exercendo influência direta na lógica interna ou na estrutura dos algoritmos de correção de erros subsequentes.

Os experimentos foram conduzidos em ambiente acústico controlado, idêntico ao da captação das amostras originais, assegurando a consistência experimental. A robustez do modelo foi avaliada frente a variações controladas de pronúncia, intensidade sonora e ruído de fundo. Visto que a integridade da extração de comandos é determinante para a confiabilidade global, esta etapa de verificação estabelece a base necessária para que a correção de erros atue sobre dados de entrada.

4.4 Inserção artificial e tratamento dos erros

Com o sistema de reconhecimento validado, os testes de correção consistiram na inserção controlada de erros nas frases, simulando situações em que o usuário pronuncia palavras incorretas ou com erros produzidos durante o reconhecimento de fala. Essa abordagem permite avaliar a capacidade do sistema em identificar divergências e corrigir automaticamente as frases. Para isso, os erros foram inseridos de forma controlada durante a pronúncia das frases, buscando reproduzir situações em que determinadas palavras fossem pronunciadas de maneira divergente do banco de dados, conforme ilustrado na Tabela 6.

Tabela 6 – Exemplo de frase numérica com erro inserido.

1ª Palavra	**2ª Palavra**	**3ª Palavra**	**4ª Palavra**
01	02	06	17
Bom	Dia	Tudo	Informe

Fonte: Autoria própria.

Esse procedimento permitiu avaliar a capacidade do sistema em reconhecer padrões incorretos, localizar a discrepância e substituí-la pela palavra correta registrada no banco de dados. Dessa forma, o método de correção de erros pôde ser analisado em sua função principal, a detecção e correção automática de erros de fala, o que representa um passo essencial para o aprimoramento do desempenho do modelo e para a correta compreensão, pelo sistema de diálogo, da frase produzida pelo usuário.

Foram conduzidos testes experimentais com o objetivo de avaliar o desempenho dos algoritmos de correção de frases implementados. Em cada teste, foi inserido um ou mais erros propositais na sequência correspondente à frase original, de modo a simular situações reais de falha no reconhecimento de fala.

Embora a inserção dos erros tenha sido realizada de forma controlada, essa abordagem permite isolar o comportamento dos algoritmos de correção, garantindo reprodutibilidade dos testes e possibilitando a análise sistemática do impacto do aumento do número de erros no desempenho dos métodos avaliados.

Para avaliar a robustez dos algoritmos, foram realizados ensaios exaustivos baseados no número total de combinações possíveis para cada nível de erro. O experimento compreendeu 76 testes para frases com uma alteração, 2.166 para duas, e 27.436 para três alterações, totalizando um espaço amostral que abrange as substituições sistemáticas de cada sentença original.

A Tabela 7 sistematiza a correlação entre a frase de referência (armazenada no sistema) e a taxa de recuperação obtida por cada método, evidenciando o impacto do incremento de erros no desempenho algorítmico. A coluna 'Erro' define a severidade da distorção, contemplando tanto o deslocamento estrutural quanto a inserção de *tokens* estranhos a estrutura da sentença original. A geração automatizada dessas instâncias foi executada por um algoritmo de repetição estruturado em ambiente MATLAB, garantindo a cobertura total do espaço de estados das combinações possíveis dentro dos limites definidos para cada configuração de teste.

Tabela 7 – Comparação do desempenho dos métodos de correção em função do número de erros.

Frase Original	**Erros**	**Mét. 1 (%)**	**Mét. 2 (%)**	**Mét. 3 (%)**
Bom Dia Tudo Bem	1	98,68	100,00	100,00
Bom Dia Tudo Bem	2	10,11	98,38	96,31
Bom Dia Tudo Bem	3	0,79	92,89	63,44
Boa Tarde Tudo Bem	1	98,68	97,37	94,74
Boa Tarde Tudo Bem	2	10,11	81,44	74,28
Boa Tarde Tudo Bem	3	0,79	51,11	39,15
Onde Fica A Biblioteca	1	100,00	100,00	100,00
Onde Fica A Biblioteca	2	10,25	99,49	92,24
Onde Fica A Biblioteca	3	0,80	78,96	47,67
Qual Horario De Almoço	1	100,00	100,00	100,00
Qual Horario De Almoço	2	10,25	99,22	87,82
Qual Horario De Almoço	3	0,80	68,44	43,30
Me Informe O Clima	1	100,00	100,00	100,00
Me Informe O Clima	2	10,25	98,94	83,38
Me Informe O Clima	3	0,80	59,22	40,32

Fonte: Autoria própria.

4.4.1 Tratamento do erro

O tratamento de erros no sistema inicia-se com a conversão da frase captada em sua representação numérica. Nesse processo, cada palavra é substituída por um valor inteiro que a identifica unicamente com base na posição que ocupa no vetor do banco de dados. Como ilustrado na Tabela 6, a sequência resultante é expressa como uma série numérica (ex: 01020617), cujos pares correspondem aos índices das palavras na tabela de referência. A partir dessa representação, o sistema pode aplicar um dos três métodos independentes de correção:

Busca palavras: Nesta abordagem, o sistema realiza uma variação controlada, alterando iterativamente os valores numéricos por outros presentes na tabela de referência. A cada modificação, a sequência é comparada com as frases cadastradas até encontrar uma correspondência válida. Este método foca na correção palavra por palavra através de tentativas sucessivas, sendo monitorado pela eficiência temporal de cada busca.

Busca por similaridade: Este método baseia-se na análise direta das diferenças entre os códigos numéricos. Ao receber uma frase com erros, o algoritmo compara seu índice com as frases armazenadas, contabilizando o número de disparidades posicionais. A frase do banco de dados que apresenta a menor quantidade de discrepâncias em relação à entrada é selecionada como a correção mais provável.

Espaço multi-dimensional: O sistema também pode representar frases como pontos em um espaço multi-dimensional. Utilizando a distância euclidiana, o algoritmo identifica qual ponto no espaço vetorial está mais próximo da entrada do usuário.

Neste modelo, a correção é considerada bem-sucedida quando a similaridade entre o vetor ajustado e o de referência supera um limiar de 25%. Este valor foi definido experimentalmente para equilibrar a tolerância a erros e a precisão, evitando tanto o excesso de falsos positivos quanto a restrição severa na capacidade de correção do sistema.

4.4.2 Dados e Comparações

A Tabela 8 apresenta um resumo comparativo do desempenho dos três métodos de correção de frases numéricas em função da quantidade de erros inseridos. Foram avaliados cenários contendo um, dois e três erros, considerando diferentes volumes de testes, o que permite analisar não apenas a taxa de acerto de cada abordagem, mas também sua robustez à medida que a complexidade do problema aumenta. O aumento no número de erros inseridos amplia o espaço de busca e a possibilidade de ambiguidades na correspondência das frases, tornando o processo de correção progressivamente mais desafiador.

Dessa forma, os resultados apresentados na tabela possibilitam uma avaliação mais abrangente da eficiência e da estabilidade dos métodos propostos em diferentes condições de operação.

Tabela 8 – Desempenho dos métodos em função da quantidade de erros inseridos.

Método	**Erros Inseridos**	**Quantidade de Testes**	**% de Acertos**
Método 1	1	380	99,47 %
Método 1	2	10830	10,19 %
Método 1	3	137180	0,80 %
Método 2	1	380	99,47 %
Método 2	2	10830	95,49 %
Método 2	3	137180	70,10 %
Método 3	1	380	98,95 %
Método 3	2	10830	86,82 %
Método 3	3	137180	46,77 %

Fonte: Autoria própria.

No cenário com a inserção de um único erro, cujos resultados detalhados são apresentados na Tabela 9, observa-se que todos os métodos obtiveram desempenho elevado e bastante semelhante, indicando que, em situações simples, os três métodos são eficazes na correção de erros isolados.

Tabela 9 – Desempenho dos métodos considerando uma alteração por frase.

Método	**Total de Testes**	**% de Acertos**
Método 1	380	99,47 %
Método 2	380	99,47 %
Método 3	380	98,95 %

Fonte: Autoria própria.

Quando analisado o cenário com dois erros inseridos, conforme apresentado na Tabela 10, nota-se uma diferença significativa no desempenho entre os métodos. O método de busca palavras apresentou queda acentuada na taxa de acerto, alcançando apenas 10,19%, evidenciando sua limitação para lidar com múltiplos erros simultâneos. Em contrapartida, o método de busca por similaridade manteve uma taxa de acerto de 86,82%, enquanto o método de espaço multi-dimensional obteve o melhor desempenho nesse cenário, com 95,49% de acertos.

Tabela 10 – Desempenho dos métodos considerando duas alterações por frase.

Método	**Total de Testes**	**% de Acertos**
Busca palavras	10830	10,19 %
Espaço multi-dimensional	10830	95,49 %
Busca por similaridade	10830	86,81 %

Fonte: Autoria própria.

No cenário mais complexo, com três erros inseridos, os resultados apresentados na Tabela 11 indicam um aumento ainda mais expressivo das diferenças entre os métodos.

O método de busca palavras apresentou desempenho praticamente inviável, com apenas 0,80% de acertos. O método de busca por similaridade demonstrou maior robustez, alcançando 46,77% de acertos, enquanto o método de espaço multi-dimensional obteve o melhor desempenho entre os métodos avaliados, com uma taxa de acerto de 70,10%.

Tabela 11 – Desempenho dos métodos considerando três alterações por frase.

Método	**Total de Testes**	**% de Acertos**
Busca palavras	137180	0,80 %
Espaço multi-dimensional	137180	70,12 %
Busca por similaridade	137180	46,78 %

Fonte: Autoria própria.

Observa-se que, à medida que o número de erros aumenta, o desempenho do método de busca palavras decai de forma acentuada, enquanto os método de busca por similaridade e espaço multi-dimensional mantêm taxas de acerto mais elevadas, com destaque para o método de espaço multi-dimensional nos cenários com dois e três erros.

De forma geral, os resultados evidenciam que o método de busca palavras apresenta desempenho satisfatório apenas em cenários com baixo nível de erro, tornando-se inadequado à medida que a complexidade aumenta. O método de busca por similaridade de monstra maior robustez intermediária, mantendo taxas de correção razoáveis mesmo com múltiplos erros, porém com queda significativa em cenários mais severos. O método de espaço multi-dimensional destaca-se como o mais robusto entre os avaliados, apresentando as maiores taxas de acerto nos cenários com dois e três erros, o que indica maior capacidade de generalização e tolerância a falhas no reconhecimento de fala. Esses resultados confirmam a adequação do método baseado em similaridade para aplicações práticas que exigem maior confiabilidade na interpretação de comandos de voz.

4.5 Discussão dos resultados

A análise comparativa entre as abordagens implementadas revela que a estratégia de correção deve ser adaptada à severidade dos erros esperados no ambiente de operação.

- **Comportamento em cenários de baixa complexidade:** A eficácia equivalente dos três métodos para erros unitários (superior a 98%) sugere que, quando a ambiguidade é mínima, a complexidade computacional do algoritmo pode ser um diferencial. O método busca palavras, embora simples, é suficiente para aplicações onde a taxa de erro do reconhecedor de fala é controlada.

- **A escalabilidade da falha:** O declínio drástico no desempenho do método busca palavras à medida que o número de erros aumenta (de 99,47% para 0,80%) demons-

tra que a busca exaustiva por palavra é vulnerável à explosão combinatória e a falsos positivos.

- **Robustez e Vetorização:** O superior desempenho das abordagens baseadas em similaridade vetorial (métodos de espaço multi-dimensional e busca por similaridade) comprova que a representação geométrica ou posicional das sentenças permite que o sistema capture o contexto da frase, em vez de depender apenas da exatidão de *tokens* isolados. Esta característica é fundamental para a resiliência do sistema em ambientes acústicos ruidosos, onde múltiplos fonemas podem ser confundidos simultaneamente.

5 Conclusão e Trabalhos Futuros

A análise dos algoritmos propostos demonstrou que os métodos de correção desenvolvidos são capazes de atuar de forma eficaz na correção automática de erros em comandos reconhecidos por fala, especialmente quando integrados a uma etapa de verificação e validação das frases reconhecidas. Dessa forma, os objetivos propostos neste trabalho foram plenamente alcançados nos casos de até dois erros (representando 50% da frase), evidenciando a viabilidade do uso de estratégias de correção como complemento aos sistemas de reconhecimento automático de fala.

Os resultados experimentais indicaram que, em cenários com baixo nível de complexidade, caracterizados pela inserção de um único erro por frase, todos os métodos apresentaram desempenho elevado e estatisticamente semelhante, com taxas de acerto próximas a 99%. Esses resultados demonstram que abordagens simples de correção são suficientes para aplicações com baixo índice de erro no reconhecimento das palavras.

Entretanto, à medida que a complexidade das frases aumenta, com a inserção de dois e três erros simultâneos, observam-se diferenças significativas entre os métodos avaliados. O método de busca palavras mostrou-se limitado a cenários simples, apresentando queda acentuada na taxa de acerto quando submetido a múltiplos erros, o que restringe sua aplicabilidade em situações mais complexas. O método de busca por similaridade apresentou robustez intermediária, mantendo taxas de correção razoáveis mesmo com o aumento do número de erros, porém com redução significativa de desempenho nos cenários mais severos. Por sua vez, o método de espaço multidimensional destacou-se como o mais robusto entre os avaliados, apresentando as maiores taxas de acerto nos cenários com dois e três erros, evidenciando sua maior tolerância a falhas no reconhecimento das frases.

De forma geral, os resultados obtidos indicam que a escolha do método de correção deve considerar o contexto de aplicação e o nível de complexidade esperado nas frases reconhecidas. Enquanto métodos mais simples podem ser adequados para aplicações com baixo índice de erro, cenários mais exigentes demandam abordagens mais robustas, capazes de lidar eficientemente com múltiplos erros simultâneos. Assim, o sistema proposto demonstra potencial significativo para aprimorar interfaces de comunicação homem-máquina baseadas em fala, contribuindo para interações mais naturais, confiáveis e eficientes.

5.1 Etapas Futuras

Para dar continuidade a esta pesquisa e ampliar os resultados obtidos, algumas direções para trabalhos futuros podem ser consideradas:

- **Expansão do banco de dados:** Ampliar o conjunto de palavras e frases reconhecidas pelo sistema, incluindo variações linguísticas, sotaques regionais e diferentes estruturas frasais, de modo a avaliar a escalabilidade e a robustez, especialmente dos **métodos de espaço multidimensional** e **busca por similaridade**, em contextos mais diversos.

- **Aprimoramento da correção de comandos:** Implementar técnicas adicionais de verificação e correção de erros, como algoritmos probabilísticos ou redes neurais, para reduzir ainda mais a taxa de erros e aumentar a confiabilidade do sistema em tempo real.

- **Avaliação em cenários reais:** Submeter o sistema a testes com usuários reais, verificando o desempenho da interface homem-máquina em situações práticas e avaliando métricas de usabilidade, tempo de resposta e satisfação do usuário.

- **Integração com interfaces multimodais:** Explorar a integração do reconhecimento de comandos de voz com outros canais de entrada, como gestos ou texto, permitindo uma interação mais natural e eficiente com o sistema.

- **Avaliação de desempenho em sistemas mais complexos:** Investigar o comportamento dos algoritmos de correção frente a um aumento significativo no dicionário de palavras e à transição para um sistema independente de locutor. Esta etapa é crucial para medir a acurácia quando a variabilidade acústica aumenta — devido a diferentes timbres e entonações — e quando a densidade do espaço de busca cresce com o maior número de comandos. Espera-se avaliar se o **método de espaço multidimensional** mantém sua taxa de acerto conforme os pontos (frases) se tornam mais próximos e densos no espaço vetorial, exigindo maior precisão nos cálculos de distância euclidiana.

- **Otimização computacional:** Investigar técnicas de otimização para reduzir o tempo de processamento e a complexidade dos algoritmos, especialmente ao lidar com bancos de dados maiores e maior número de comandos simultâneos.

Essas etapas futuras visam consolidar o sistema desenvolvido, aprimorar a interação entre usuário e máquina, e fornecer bases para pesquisas mais amplas no campo de interfaces de comunicação homem-máquina.

Referências

BASSIL, Y.; ALWANI, M. Post-editing error correction algorithm for speech recognition using bing spelling suggestion. *arXiv*, 2012.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980. doi: 10.1109/TASSP.1980.1163420.

EL-AZAZY, A. A. M. E.-H. et al. A survey on advancements in voice control systems enhancing human-computer interaction through speech recognition and ai. *Engineering Research Journal (Shoubra)*, Benha University, Faculty of Engineering (Shoubra), v. 54, n. 1, p. 95–102, 2025.

ERRATTAHI, R.; HANNANI, A. E.; OUAHMANE, H. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, Elsevier, v. 128, p. 32–37, 2018.

GUO, J. et al. Ucorrect: An unsupervised framework for automatic speech recognition error correction. *arXiv preprint arXiv:2401.05689*, 2024.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: Pearson New International Edition PDF eBook*. [S.l.]: Pearson Higher Ed, 2013.

LENG, Y. et al. Softcorrect: Error correction with soft detection for automatic speech recognition. 2022.

LENG, Y. et al. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *Advances in Neural Information Processing Systems*, v. 34, p. 21708–21719, 2021.

MA, R. et al. N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space. *arXiv preprint arXiv:2303.00456*, 2023.

MALLIOS, S.; BOURBAKIS, N. A survey on human machine dialogue systems. In: *2016 7th International Conference on Information, Intelligence, Systems Applications (IISA)*. [S.l.: s.n.], 2016. p. 1–7. doi: 10.1109/IISA.2016.7785371.

OVIATT, S. Taming recognition errors with a multimodal interface. *Communications of the ACM*, ACM New York, NY, USA, v. 43, n. 9, p. 45–51, 2000.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Ieee, v. 77, n. 2, p. 257–286, 2002.

RAHHAL, E.; HANNANI, A. E.; OUAHMANE, H. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, v. 128, p. 32–37, 2016.

RUSAN, H.-A.; MOCANU, B. Human-computer interaction through voice commands recognition. In: *2022 International Symposium on Electronics and Telecommunications (ISETC)*. [S.l.: s.n.], 2022. p. 1–4. doi: 10.1109/ISETC56213.2022.10010253.

SCHAFER, R. W. Scientific bases of human-machine communication by voice. *Proceedings of the National Academy of Sciences*, v. 92, n. 22, p. 9914–9920, 1995.

VITERBI, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, v. 13, n. 2, p. 260–269, 1967. doi: 10.1109/TIT.1967.1054010.

WANG, X. et al. A light-weight contextual spelling correction model for customizing transducer-based speech recognition systems. *arXiv*, 2021.

WILPON, J. G.; ROE, D. B. *Voice communication between humans and machines*. [S.l.]: National Academies Press, 1994.

YARED, G. F. G. *Método para a Determinação do Número de Gaussianas em Modelos Ocultos de Markov para Sistemas de Reconhecimento de Fala Contínua*. Tese (Doutorado) — Universidade Estadual de Campinas, 2006.

ZHOU, W. et al. Error correction via phonetic similarity-based processing for chinese spoken dialogue system. In: *2006 8th international Conference on Signal Processing*. [S.l.: s.n.], 2006. v. 3. doi: 10.1109/ICOSP.2006.345742.

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado do Curso de Engenharia Elétrica

TERMO DE RESPONSABILIDADE

O texto do trabalho de conclusão de curso intitulado Interface homem-máquina controlada pela fala com correção de erro é de minha inteira responsabilidade. Declaro que não há utilização indevida de texto, material fotográfico ou qualquer outro material pertencente a terceiros sem a devida citação ou consentimento dos referidos autores.

João Monlevade, 10 de março de 2026.

Documento assinado digitalmente
gov.br
NICOLAS VIEIRA MAGALHAES
Data: 15/04/2026 09:28:48-0300
Verifique em https://validar.iti.gov.br

Nicolas Vieira Magalhães

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Colegiado do Curso de Engenharia Elétrica

DECLARAÇÃO DE CONFERÊNCIA DA VERSÃO FINAL

Declaro que conferi a versão final a ser entregue pelo aluno Nicolas Vieira Magalhães, autor do trabalho de conclusão de curso intitulado Interface homem-máquina controlada pela fala com correção de erro quanto à conformidade nos seguintes itens:

1. A monografia corresponde a versão final, estando de acordo com as sugestões e correções sugeridas pela banca e seguindo as normas ABNT;

2. A versão final da monografia inclui a ata de defesa (Anexo IV), a ficha catalográfica e o termo de responsabilidade (ANEXO X) devidamente assinado.

João Monlevade, 10 de março de 2026.

Prof. Dr. Glauco Ferreira Gazel Yared