

AllianceScan: identificação de disfluências em português brasileiro em áudios transcritos de psicoterapia para prever aliança terapêutica

Alana S. Luna¹, Ariane Machado-Lima¹, Fátima L. S. Nunes¹

¹Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Caixa Postal 03828-000 – São Paulo – SP – Brazil

alana.luna@usp.br, ariane.machado@usp.br, fatima.nunes@usp.br

Abstract. *Speech disfluencies, such as pauses and repetitions, can indicate both cognitive processes and underlying emotional states. This study proposes AllianceScan, an approach to identify disfluencies in transcriptions of Brazilian Portuguese audio, applied in psychotherapy sessions. Using speech recognition, the approach classifies four types of disfluencies and employs regression models to predict therapeutic alliance. The results, evaluated by metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Median Absolute Error (MDAE), indicate that the approach can enhance the understanding of therapeutic interactions and identify signs of treatment progress.*

Resumo. *As disfluências na fala, como pausas e repetições, podem indicar tanto processos cognitivos quanto estados emocionais subjacentes. Este estudo propõe a AllianceScan, uma abordagem para identificar disfluências em transcrições de áudios em Português Brasileiro, aplicada em sessões de psicoterapia. Utilizando reconhecimento de voz, a abordagem classifica quatro tipos de disfluências e utiliza modelos de regressão para prever a aliança terapêutica. Os resultados, avaliados por métricas como Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Median Absolute Error (MDAE), indicam que a abordagem pode melhorar a compreensão das interações terapêuticas e identificar sinais de progresso no tratamento.*

1. Introdução

Com o avanço das tecnologias multimídia, interfaces conversacionais entre humanos e computadores têm ganhado destaque, impulsionando estudos sobre representações vocais, faciais e corporais [Ming et al. 2023, Swamy et al. 2022, Sakurai and Kosaka 2021]. No caso da fala, aspectos como entonação, pronúncia e disfluências são fundamentais para a análise de aspectos comunicativos e cognitivos da interação.

As disfluências, que são interrupções espontâneas no fluxo discursivo caracterizadas por pausas, repetições, prolongamentos sonoros e outras ocorrências [Rocholl et al. 2021], podem ser identificadas por tecnologias de Reconhecimento Automático de Fala (*Automatic Speech Recognition - ASR*). Além de melhorar a qualidade das transcrições [Rocholl et al. 2021, Lee et al. 2021, Chen et al. 2020], essa detecção possibilita investigar relações entre estados emocionais do falante e seu desempenho comunicativo [Clavel 2016, Almeida et al. 2011].

O Inglês predomina nas pesquisas, conforme demonstrado na revisão sistemática desenvolvida em [Luna et al. 2025], e são poucas as abordagens para identificação e classificação de disfluências em Português Brasileiro (pt-br). Neste contexto, este estudo propõe e valida uma abordagem para identificar disfluências em transcrições textuais de áudios em pt-br, utilizando reconhecimento de voz. Especificamente, foi realizado um estudo de caso em sessões gravadas de psicoterapia do COMVC-19, um programa de apoio à saúde mental de profissionais da saúde durante a pandemia de COVID-19 [Fukuti et al. 2021], que utilizou o *Working Alliance Inventory* (WAI) para avaliar a relação terapeuta-paciente. O objetivo do estudo de caso foi verificar o potencial de contribuição das disfluências na prevenção de desengajamento dos pacientes. De modo geral, os resultados incluem uma abordagem para identificar disfluências em transcrições reais, prever o nível de aliança terapêutica por meio de regressão e demonstrar o potencial das disfluências como indicadores de estados emocionais e cognitivos.

2. Fundamentação Teórica

Esta seção visa a introduzir os principais conceitos que envolvem o escopo da abordagem *AllianceScan*: disfluências, transcrição de áudios e aliança terapêutica.

A fluência da fala é caracterizada pela continuidade semântica e coerência lógica, enquanto a fala disfluente apresenta interrupções perceptíveis [Khara et al. 2018]. Embora todos possam experimentar momentos de disfluência na fala, nem toda disfluência indica gagueira, pois esta é um transtorno de comunicação persistente [Maguire et al. 2020, Frigerio-Domingues and Drayna 2017]. As disfluências são categorizadas em padrões como o modelo “*Reparandum* - Ponto de interrupção - *Interregnum* - Reparo” [Shriberg 1994] (Figura 1). Nesse modelo, o “*Reparandum*” é a parte do enunciado descartada ou corrigida, enquanto o “Ponto de Interrupção” marca sua interrupção, podendo incluir revisões ou repetições. O “*Interregnum*” é opcional e reflete o desenvolvimento do raciocínio ou confusão, e o “Reparo” refaz ou corrige o “*Reparandum*”. Além desse modelo, a revisão sistemática de [Luna et al. 2025] consolidou dez categorias distintas de disfluências, ampliando a compreensão sobre o tema.

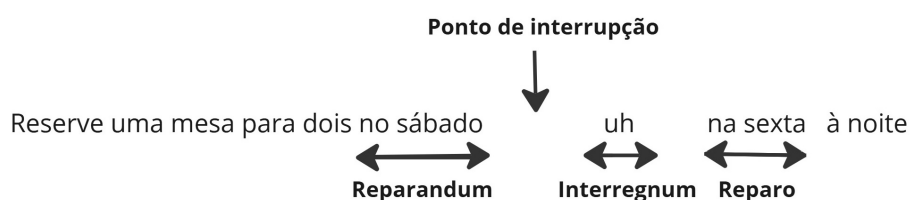


Figura 1. Exemplo da Estrutura Disfluente de [Shriberg 1994]: Reparandum, Ponto de Interrupção, Interregnum e Reparo.

A identificação de disfluências em gravações de áudio pode ser feita por ferramentas ASR, que convertem fala em texto com base em modelos acústicos e de linguagem [Kumar and Singh 2019]. Algumas ferramentas populares incluem *HDecode* [Young et al. 2002]; *Sphinx* [Lee et al. 1990]; e *Kaldi* [Povey et al. 2011], além de soluções comerciais como *Amazon Transcribe*¹, *Google Speech*² e *IBM Watson*³. A es-

¹<https://docs.aws.amazon.com/transcribe/index.html>

²<https://cloud.google.com/speech-to-text>

³<https://www.ibm.com/cloud/watson-speech-to-text>

colha da ferramenta ASR depende do contexto específico e dos requisitos de precisão necessários para a análise.

No contexto psicoterapêutico, a análise da fala é fundamental para avaliar a aliança terapêutica, um preditor da eficácia do tratamento [Maia et al. 2017, Zilcha-Mano 2017]. Instrumentos como o WAI mensuram a relação entre terapeuta e paciente por meio de questionários com escala Likert [Serralta et al. 2020, Maia et al. 2017, Prado and Meyer 2006]. Nesse contexto, estudos sugerem que uma forte aliança está associada a maior envolvimento emocional, e as disfluências podem atuar como indicadores de estados emocionais subjacentes [Negri et al. 2019, Clavel 2016, Almeida et al. 2011]. Dessa forma, o estudo das disfluências tem potencial para enriquecer a análise da interação terapêutica e revelar padrões de comunicação relacionados ao sucesso do tratamento.

3. Trabalhos Relacionados

Diferentes abordagens foram propostas para detecção de disfluências na fala. Modelos baseados em redes do tipo *Long Short-Term Memory* (LSTMs), como o de [Wang et al. 2017], capturaram disfluências do tipo “Reparandum” em Inglês e Chinês. Para aprimorar a detecção dessas disfluências, estratégias auxiliares, como reconhecimento de entidades nomeadas (*Named Entity Recognition* - *NER*) e a marcação de classes gramaticais (*Part-Of-Speech* - *POS*), foram exploradas, com foco em Inglês [Lee et al. 2021].

Em Japonês, pesquisas como a de [Horii et al. 2022] desenvolveram métodos que substituem disfluências por etiquetas reconhecíveis pelo ASR, focando em “Preenchimentos”. Outros estudos, como o de [Yoshikawa et al. 2016], alinham anotações entre transcrições e saídas em sistemas ASR, abordando disfluências de “Reparo”, “Marcador de Discurso” e “Preenchimento” em Inglês. Em dispositivos móveis, versões reduzidas de arquiteturas *Bidirectional Encoder Representations from Transformers* (BERT) foram testadas para otimizar a detecção de disfluências do tipo “Reparandum” e “Preenchimento” em Inglês [Rocholl et al. 2021].

Além de aplicações em processamento de fala, as pesquisas avançam em aplicações clínicas. Um exemplo disso é o trabalho de [Treviso and Aluísio 2018], que desenvolveu um modelo baseado em Redes Neurais Convolucionais Recorrentes (*Recurrent Convolutional Neural Networks* - *RCNNs*) para segmentação de sentenças e detecção de disfluências em testes neuropsicológicos, abordando disfluências de “Preenchimento” e “Edição” em pt-br para diagnóstico de demências.

Diferente das pesquisas anteriores, este estudo explora como as disfluências na fala se relacionam com a aliança terapêutica, oferecendo novas perspectivas sobre dinâmicas interpessoais em psicoterapia. A abordagem proposta pode auxiliar os terapeutas na avaliação do progresso da terapia, identificando precocemente possíveis rupturas na relação com o paciente e contribuindo para a personalização do tratamento, o que pode otimizar os resultados terapêuticos.

4. Materiais e métodos

Este estudo utiliza uma abordagem quantitativa para investigar a relação entre disfluências na fala e a aliança terapêutica em sessões de psicoterapia, conforme detalhado a seguir.

4.1. Materiais

O programa COMVC-19, criado em 2020 e registrado na Comissão Nacional de Ética em Pesquisa (CAAE 30710820.8.0000.0068), ofereceu apoio em saúde mental aos profissionais do Hospital das Clínicas da Faculdade de Medicina da USP (HCF-MUSP) durante a pandemia de COVID-19. As sessões, conduzidas por residentes de Psiquiatria com supervisão e gravadas em vídeo, duravam entre 40 e 60 minutos [Fukuti et al. 2021] e foram avaliadas por uma versão simplificada do instrumento WAI [Hatcher and and 2006, Horvath and Greenberg 1989] aplicada a terapeutas (10 questões) e pacientes (12 questões), usando escala Likert de cinco pontos (1 = raramente a 5 = sempre). A Tabela 1 fornece informações adicionais sobre o número de participantes, duração média das sessões e a quantidade de díades (pares de terapeutas e pacientes). No contexto da presente pesquisa, o material do programa COMVC-19 foi transcrito pela ferramenta *Amazon Transcribe* (conforme apresentado no módulo 4.2.1).

Tabela 1. Projeto COMVC-19

Categoria	Informações
Participantes	14 terapeutas (7 homens e 7 mulheres) 23 pacientes (7 homens e 16 mulheres)
Sessões	70 registros tempo médio de 50 minutos
Díades	23 díades Homem x Homem (2) Mulher x Mulher (6) Mulher x Homem (15)

4.2. Métodos

A visão geral da *AllianceScan* é apresentada na Figura 2, na qual se verifica sua composição em quatro módulos. Na seção 4.2.1, descreve-se o módulo responsável pela captura, processamento e manipulação dos áudios. A seção 4.2.2 aborda o módulo dedicado à identificação de disfluências, que funciona em paralelo com o módulo de extração de características, descrito na seção 4.2.3. O resultado desse processamento é um conjunto de características, que são utilizadas na previsão do nível de aliança terapêutica, conforme ilustrado na seção 4.2.4, com a escala WAI.

4.2.1. Módulo de transcrição

Este módulo processa o áudio e o converte em texto. Um estudo anterior [Luna et al. 2022] analisou o desempenho de ferramentas de transcrição para pt-br, destacando a ferramenta *Google Speech* por apresentar os melhores resultados tanto em sentenças fluentes quanto disfluências, além de sua capacidade de diarização. A diarização é fundamental no contexto atual para separar as falas de pacientes e terapeutas no material do programa COMVC-19 (seção 4.1). No entanto, o *Google Speech* apresentou uma limitação ao suprimir silêncios, prejudicando a detecção de algumas disfluências. Assim, a ferramenta *Amazon Transcribe* foi escolhida para transcrever o material do COMVC-19 por oferecer diarização de locutores, suporte a pt-br e não suprimir os intervalos de silêncio, superando as limitações das outras opções.

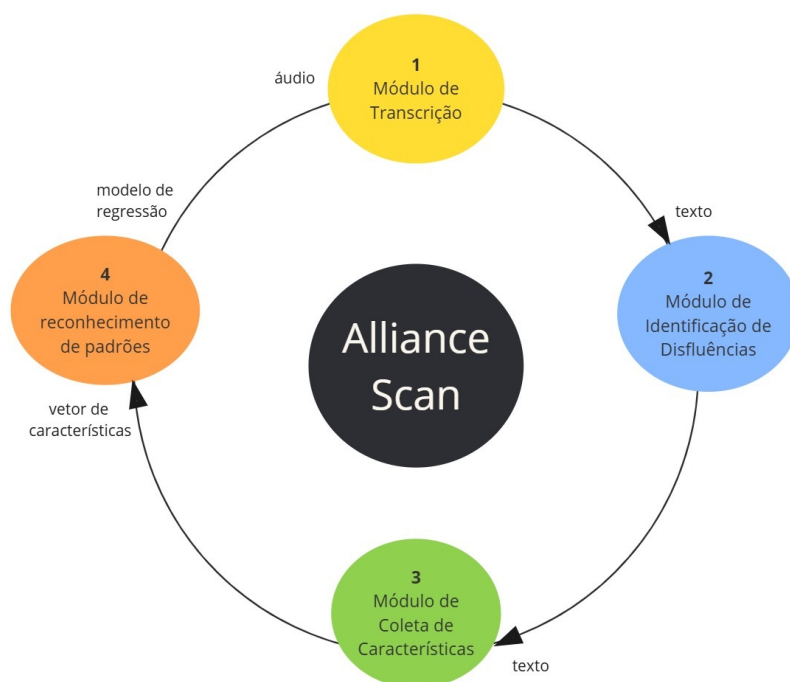


Figura 2. Módulos da *AllianceScan*.

4.2.2. Módulo de identificação de disfluências

Este módulo tem como objetivo identificar as disfluências presentes nos dados de entrada adquiridos no módulo anterior. No contexto da *AllianceScan*, o escopo foi delimitado para identificar quatro categorias de disfluências: “*Interregnum*”, “Ponto de Interrupção”, “Reparo” e “Repetição”.

Disfluências da categoria *Interregnum*: o “*Interregnum*” é uma parte opcional da fala que indica desenvolvimento de raciocínio ou confusão. A estratégia adotada para a detecção desse tipo de disfluência combina um dicionário de expressões — que inclui palavras como “então,” “tal”, “bom,” “tá?”, “viu?”, “enfim,” “assim”, “sabe?”, “né”, “ahn”, “ah”, “uh”, “hun”, “oh”, “hunhun”, “hum”, “heim”, “uhn”, “hrum”, “hinhun”, “uhm”, “hanhan”, “han”, “ohr”, “ha”, “eh” — com o algoritmo BERT [Devlin et al. 2019, Vaswani et al. 2017], que calcula a similaridade semântica entre as palavras transcritas e as do dicionário. O dicionário de expressões foi construído com base em uma revisão da literatura [Luna et al. 2025] e um estudo anterior [Treviso and Aluísio 2018], considerando também a influência das pontuações gramaticais, que refletem o processo cognitivo na fala. Após testes com diferentes limiares de similaridade semântica (60%, 70%, 85% e 90%), o valor de 85% foi escolhido por minimizar falsos positivos e falsos negativos.

Disfluências da categoria Ponto de Interrupção: ocorre quando o falante interrompe o enunciado para revisar, repetir ou abandonar parte do conteúdo. Por exemplo, em “Eu estava pensando em estudar... ou talvez descansar um pouco”, as reticências indicam dúvida. O modelo utiliza as marcações temporais (*timestamps*) da ferramenta *Amazon Transcribe* para registrar um “Ponto de Interrupção” quando detecta silêncios.

Disfluências da categoria Reparo: o “Reparo” corrige ou modifica um enunciado, como em “Eu estava empolgado... quer dizer, eu estava mais nervoso do que empolgado”, em que “nervoso” substitui “empolgado”. Inicialmente, sua detecção baseava-se em um “Ponto de Interrupção” seguido por um “*Interregnum*”, conforme a literatura [Wu et al. 2015, Wang et al. 2016, Zayats and Ostendorf 2019], mas essa sequência nem sempre ocorria. Por isso, incorporou-se o modelo *Deepbond* [Treviso and Aluísio 2018], cuja complementaridade justificou a adoção de ambos os métodos na marcação dos “Reparos”.

Disfluências da categoria Repetição: a “Repetição” ocorre quando um enunciado é repetido totalmente ou parcialmente, como em “Eu acho que... que podemos tentar”. Para detectá-la, o módulo compara palavras sucessivas e registra as correspondências. Termos compostos com repetições naturais, como “reco-reco”, “lero-lero”, “lenga-lenga”, “quero-quero”, foram propositalmente excluídos para evitar marcações indevidas.

4.2.3. Módulo de coleta de características

Esta fase extrai e quantifica características do áudio transcrito e das disfluências identificadas. Com base nos dados dos módulos de transcrição (seção 4.2.1) e disfluências (seção 4.2.2), foram extraídas características relacionadas à estrutura frasal, distribuição percentual, duração, desvio padrão, densidade relativa e absoluta, conforme a Tabela 2.

As características extraídas foram organizadas em uma planilha com 70 linhas (sessões gravadas do COMVC-19) e 35 colunas, sendo uma delas o nome do arquivo da sessão, 17 colunas para as características do paciente e outras 17 colunas para as características do terapeuta. A normalização *Min-Max* do pacote *sklearn preprocessing* foi aplicada para garantir comparabilidade dos dados. Essas características foram usadas como variáveis independentes no reconhecimento de padrões, enquanto a variável dependente foi derivada do questionário WAI (seção 4.1), aplicado tanto aos pacientes quanto aos terapeutas.

4.2.4. Módulo de reconhecimento de padrões

Este módulo prevê o nível de aliança terapêutica com base nas características previamente obtidas. As informações foram organizadas em três conjuntos: uma base completa, com características de pacientes e terapeutas; uma apenas com características dos pacientes; e outra apenas com as características dos terapeutas. A modelagem, feita em *Python* com *Scikit-Learn*⁴, usou cinco algoritmos (Random Forest, SVR, Linear, KNN e Árvore de Decisão) para testar diferentes cenários, aplicando PCA e Correlação de Pearson. O PCA manteve 99% da variabilidade, enquanto a Correlação de Pearson excluiu variáveis com coeficiente maior ou igual a 0,98. A Tabela 3 resume os modelos implementados.

Para validar os modelos, foi usada uma validação cruzada tradicional com $k = 10$. No entanto, devido ao risco de vazamento de informações entre os conjuntos de treino e teste causado pela repetição de diádes, adotou-se uma variação do método *leave-one-out*,

⁴<https://scikit-learn.org/stable/>

Tabela 2. Resumo das características extraídas a partir dos áudios transcritos e das disfluências identificadas

#	Nome do descritor	Descrição
1	Total de palavras na frase transcrita	O número total de palavras presentes na frase transcrita, que representa o tamanho da frase.
2	Quantidade de “ <i>Interregnum</i> ”	O número de “ <i>Interregnum</i> ” durante a fala.
3	Percentual de “ <i>Interregnum</i> ”	Porcentagem da fala transcrita que é ocupada por “ <i>Interregnum</i> ”.
4	Quantidade de “Repetições”	O número de vezes que palavras ou frases são repetidas durante a fala.
5	Percentual de “Repetições”	Porcentagem da fala transcrita que é ocupada por repetições.
6	Quantidade de “Pontos de Interrupção”	O número de momentos em que a fala é interrompida ou pausada abruptamente.
7	Percentual de “Pontos de Interrupção”	Porcentagem da fala transcrita que é ocupada por interrupções bruscas.
8	Tempo total de “Pontos de Interrupção”	Quantidade total de tempo gasto em todas as interrupções durante a fala.
9	Percentual de tempo de “Pontos de Interrupção”	Porcentagem do tempo total de fala ocupado pelas interrupções em relação ao tempo total da fala
10	Média de tempo de “Pontos de Interrupção”	Média do tempo gasto em cada ponto de interrupção durante a fala.
11	Desvio padrão dos “Pontos de Interrupção”	Medida de dispersão dos pontos de interrupção, representando a variabilidade dessas pausas.
12	Quantidade de “Reparos” na frase	Número de correções ou revisões feitas na fala.
13	Percentual de “Reparos”	Porcentagem da fala transcrita que é ocupada por correções ou revisões.
14	Quantidade de “Reparos” (método <i>Deepbond</i>)	Número de correções ou revisões feitas na fala contabilizadas pelo método <i>Deepbond</i> .
15	Percentual de “Reparos” (método <i>Deepbond</i>)	Porcentagem da fala transcrita que é ocupada por correções ou revisões contabilizadas pelo método <i>Deepbond</i> .
16	Densidade relativa	Relação entre palavras e tempo de fala específico do orador.
17	Densidade absoluta	Relação entre palavras e a duração total do áudio, sem levar em consideração quem está falando.

Tabela 3. Modelos implementados na *AllianceScan* utilizando diferentes bases de dados

Variável resposta	Cenário
WAI Paciente	Base Completa Base Completa + Pearson Base Completa + Pearson + PCA Base do Paciente + Pearson Base do Terapeuta + Pearson
WAI Terapeuta	Base Completa Base Completa + Pearson Base Completa + Pearson + PCA Base do Paciente + Pearson Base do Terapeuta + Pearson

denominada *leave-one-dyad-out*, que isolou as ocorrências de cada díade no conjunto de teste. O desempenho dos modelos foi avaliado com base nas métricas de erro *MAE*, *MSE*, *RMSE*, *MDAE* e erro máximo, conforme detalhado a seguir.

5. Discussões

Os resultados da validação cruzada e da aplicação da avaliação *leave-one-dyad-out*, são apresentados nas seções a seguir.

5.1. Resultado da validação cruzada

A Tabela 4 resume os resultados da validação cruzada. A primeira coluna apresenta a variável resposta, a segunda descreve o cenário do experimento, e a terceira destaca o algoritmo com o melhor desempenho. As demais colunas mostram as métricas calculadas, com os melhores resultados em negrito.

Tabela 4. Resultados da validação cruzada tradicional com $k=10$ (experimento 1, considerando cada variável resposta, ordenados por MAE)

Variável resposta	Cenário	Melhor algoritmo	MAE	MSE	RMSE	MDAE	Erro Máximo	Número de características
WAI Paciente	Base Completa, Pearson, PCA	Árvore de Decisão	6.914	64.343	8.021	7.500	17.000	13.000
	Base Paciente, Pearson	Árvore de Decisão	6.986	83.414	9.133	5.000	23.000	15.000
	Base Terapeuta, Pearson	Árvore de Decisão	8.514	105.514	10.272	8.000	24.000	15.000
	Base Completa	Regressão Linear	8.793	132.485	11.510	6.592	30.576	33.000
	Base Completa, Pearson	Regressão Linear	9.472	152.177	12.336	7.657	34.172	31.000
WAI Terapeuta	Base Completa, Pearson, PCA	Árvore de Decisão	4.734	34.205	5.848	5.047	15.584	13.000
	Base Paciente, Pearson	Árvore de Decisão	5.371	42.686	6.533	5.000	16.000	15.000
	Base Completa	Regressão Linear	5.928	50.481	7.105	5.495	15.731	33.000
	Base Completa, Pearson	Regressão Linear	6.068	55.526	7.452	5.847	16.135	31.000
	Base Terapeuta, Pearson	Árvore de Decisão	6.114	57.143	7.559	6.000	17.000	15.000

Para o WAI do paciente, o modelo de Árvore de Decisão com Correlação de Pearson e PCA na base completa ofereceu os melhores resultados, exceto para o *MDAE*, que foi melhor na base do paciente com Correlação de Pearson. Para o WAI do terapeuta, a base completa com essas técnicas apresentou os melhores resultados em *MAE*, *MSE* e *RMSE*, enquanto o *MDAE* e *Erro Máximo* foram melhores na base do paciente com Correlação de Pearson.

A comparação entre as variáveis resposta mostrou que o WAI do terapeuta apresentou melhorias nas métricas de erro, sugerindo que terapeutas, devido ao seu treinamento e experiência, são mais sensíveis a nuances na comunicação, como disfluências, e podem perceber, mesmo de forma inconsciente, essas informações como indicativos da aliança terapêutica.

5.2. Resultado da abordagem *leave-one-dyad-out*

Um resumo dos resultados da abordagem *leave-one-dyad-out* é apresentado na Tabela 5. Para o WAI do paciente, o modelo de Regressão Linear apresentou o melhor desempenho em todas as métricas de erro, usando a correlação de Pearson na base do paciente. Para o WAI do terapeuta, o melhor resultado foi obtido também usando a base do paciente com a correlação de Pearson porém aplicado ao algoritmo KNN.

A melhoria na precisão dos modelos para o WAI do paciente no segundo experimento sugere que, ao evitar vazamentos de informações, os modelos identificaram padrões mais gerais. No entanto, a leve queda no desempenho para o WAI do terapeuta pode ter sido causada pela distribuição desigual das sessões na base de dados. Assim como no experimento anterior, o uso do WAI do terapeuta como variável resposta melhorou as métricas de erro. Além disso, os resultados apontam uma relação entre disfluências e a percepção da aliança terapêutica medida pelo WAI, com potencial para prever a aliança durante o tratamento, apesar da influência de fatores como comunicação não verbal e características individuais.

Tabela 5. Resultados do *leave-one-dyad-out* (experimento 2, considerando cada variável resposta, ordenados por MAE)

Variável resposta	Cenário	Melhor algoritmo	MAE	MSE	RMSE	MDAE	Erro Máximo	Número de características
WAI Paciente	Base Paciente, Pearson	Regressão Linear	6.026	55.091	7.422	5.987	8.227	15.000
	Base Completa, Pearson, PCA	Regressão Linear	6.864	74.891	8.653	6.988	9.365	13.000
	Base Terapeuta, Pearson	Árvore de Decisão	6.907	74.057	8.605	6.580	10.120	15.000
	Base Completa, Pearson	Regressão Linear	8.234	98.831	9.941	8.008	11.639	31.000
	Base Completa	Regressão Linear	8.234	98.831	9.941	8.008	11.639	33.000
WAI Terapeuta	Base Paciente, Pearson	KNN	5.169	39.437	6.279	5.176	6.912	15.000
	Base Terapeuta, Pearson	Árvore de Decisão	6.753	63.997	7.999	6.620	8.960	15.000
	Base Completa, Pearson	Árvore de Decisão	6.763	69.211	8.319	6.960	9.040	31.000
	Base Completa, Pearson, PCA	Árvore de Decisão	6.839	67.177	8.196	6.400	9.120	13.000
	Base Completa	Árvore de Decisão	6.941	72.976	8.542	7.140	9.040	33.000

O estudo apresenta algumas limitações, como o tamanho reduzido da base de dados, possíveis erros de transcrição da ferramenta *Amazon Transcribe* e imprecisões na detecção de disfluências. No entanto, a abordagem *AllianceScan* demonstra potencial para analisar a relação entre disfluências e a aliança terapêutica, com possíveis aplicações clínicas para ajustes no tratamento.

Por fim, a comparação dos resultados obtidos com estudos prévios é limitada pela natureza exploratória da metodologia, mas algumas pesquisas com abordagens similares contextualizam a proposta da *AllianceScan*. Por exemplo, a investigação de [Ryu et al. 2021] sugere o uso de aprendizado de máquina e neurociência social para avaliar com maior precisão a aliança terapêutica e os resultados do tratamento. Além disso, o trabalho de [Flemotomos et al. 2022] desenvolveu uma plataforma para avaliar competências terapêuticas usando reconhecimento de fala e Processamento de Linguagem Natural, visando *feedbacks* imediatos e maior qualidade nos serviços.

6. Considerações finais

Neste trabalho um método foi desenvolvido e validado para identificar disfluências em textos transcritos de áudios de psicoterapia em pt-br, com o objetivo de prever o nível da aliança terapêutica. Os resultados mostram potencial para melhorar a compreensão das interações terapêuticas e detectar sinais de progresso no tratamento. Além da psicoterapia, o modelo pode beneficiar áreas que envolvem interação entre pessoas tais como saúde, educação e serviços em geral. Por exemplo, na área de educação pode monitorar o engajamento de alunos; no atendimento ao cliente pode detectar sinais de insatisfação.

Esta pesquisa faz parte de um projeto maior com o objetivo de analisar diferentes tipos de dados e integrar os artefatos desenvolvidos a uma plataforma para avaliar suas contribuições ao processo terapêutico. Trabalhos futuros podem combinar características lexicais e prosódicas, testar novos algoritmos, adaptar o método a outros idiomas e contextos, além de investigar quais disfluências mais influenciam a previsão da escala WAI e possíveis vieses ligados às díades, como idade, gênero ou tipo de terapia.

Referências

Almeida, A. A. F. d., Behlau, M., and Leite, J. R. (2011). Correlação entre ansiedade e performance comunicativa. *Revista da Sociedade Brasileira de Fonoaudiologia*, 16:384–389.

- Chen, Q., Chen, M., Li, B., and Wang, W. (2020). Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8069–8073.
- Clavel, C. (2016). Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Flemotomos, N., Martinez, V. R., Chen, Z., Singla, K., Ardulov, V., Peri, R., Caperton, D. D., Gibson, J., Tanana, M. J., Georgiou, P., et al. (2022). Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, 54(2):690–711.
- Frigerio-Domingues, C. and Drayna, D. (2017). Genetic contributions to stuttering: the current evidence. *Molecular genetics & genomic medicine*, 5(2):95–102.
- Fukuti, P., Uchôa, C. L. M., Mazzoco, M. F., Cruz, I. D., Echegaray, M. V., Humes, E. d. C., Silveira, J. B., Santi, T. D., Miguel, E. C., and Corchs, F. (2021). Comvc-19: A program to protect healthcare workers’ mental health during the covid-19 pandemic. what we have learned. *Clinics*, 76.
- Hatcher, R. L. and and, J. A. G. (2006). Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research*, 16(1):12–25.
- Horii, K., Fukuda, M., Ohta, K., Nishimura, R., Ogawa, A., and Kitaoka, N. (2022). End-to-end spontaneous speech recognition using disfluency labeling. In *Proc. Interspeech 2022*, pages 4108–4112.
- Horvath, A. O. and Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223.
- Khara, S., Singh, S., and Vir, D. (2018). A comparative study of the techniques for feature extraction and classification in stuttering. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 887–893.
- Kumar, Y. and Singh, N. (2019). A comprehensive view of automatic speech recognition system - a systematic literature review. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pages 168–173.
- Lee, D., Ko, B., Shin, M. C., Whang, T., Lee, D., Kim, E., Kim, E., and Jo, J. (2021). Auxiliary sequence labeling tasks for disfluency detection. In *Proc. Interspeech 2021*, pages 4229–4233.
- Lee, K.-F., Hon, H.-W., and Reddy, R. (1990). An overview of the sphinx speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45.

- Luna, A. S., Machado-Lima, A., and Nunes, F. L. S. (2022). Analysis of transcription tools for brazilian portuguese with focus on disfluency detection. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10.
- Luna, A. S., Machado-Lima, A., and Nunes, F. L. S. (2025). Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges. *Journal of the Brazilian Computer Society*, 31(1):154–173.
- Maguire, G. A., Nguyen, D. L., Simonson, K. C., and Kurz, T. L. (2020). The pharmacologic treatment of stuttering and its neuropharmacologic basis. *Frontiers in neuroscience*, 14:158.
- Maia, R. d. S., Araújo, T. C. S. d., Silva, N. G. d., and Maia, E. M. C. (2017). Instrumentos para avaliação da aliança terapêutica. *Revista Brasileira de Terapias Cognitivas*, 13:55 – 63.
- Ming, F. J., Shabana Anhum, S., Islam, S., and Keoy, K. H. (2023). Facial emotion recognition system for mental stress detection among university students. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6.
- Negri, A., Christian, C., Mariani, R., Belotti, L., Andreoli, G., and Danskin, K. (2019). Linguistic features of the therapeutic alliance in the first session: a psychotherapy process study. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, 22(1).
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- Prado, O. Z. and Meyer, S. B. (2006). Avaliação da relação terapêutica na terapia assíncrona via internet. *Psicologia em estudo*, 11:247–257.
- Rocholl, J. C., Zayats, V., Walker, D. D., Murad, N. B., Schneider, A., and Liebling, D. J. (2021). Disfluency detection with unlabeled data and small bert models. In *Proc. Interspeech 2021*, pages 766–770.
- Ryu, J., Banthin, D. C., and Gu, X. (2021). Modeling therapeutic alliance in the age of telepsychiatry. *Trends in Cognitive Sciences*, 25(1):5–8.
- Sakurai, M. and Kosaka, T. (2021). Emotion recognition combining acoustic and linguistic features based on speech recognition results. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pages 824–827.
- Serralta, F. B., Benetti, S., Laskoski, P. B., and Abs, D. (2020). The brazilian-adapted working alliance inventory: preliminary report on the psychometric properties of the original and short revised versions. *Trends in Psychiatry and Psychotherapy*, 42:256 – 261.
- Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley.

- Swamy, T. J., Nandini, M., B, N., Karthika K, V., Anvitha, V. L., and Sunitha, C. (2022). Voice and gesture based virtual desktop assistant for physically challenged people. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 222–226.
- Treviso, M. V. and Aluísio, S. M. (2018). Sentence segmentation and disfluency detection in narrative transcripts from neuropsychological tests. In *Computational Processing of the Portuguese Language (PROPOR)*, pages 409–418. Springer International Publishing.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Wang, S., Che, W., and Liu, T. (2016). A neural attention model for disfluency detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 278–287, Osaka, Japan. The COLING 2016 Organizing Committee.
- Wang, S., Che, W., Zhang, Y., Zhang, M., and Liu, T. (2017). Transition-based disfluency detection using lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2794, Copenhagen, Denmark. Association for Computational Linguistics.
- Wu, S., Zhang, D., Zhou, M., and Zhao, T. (2015). Efficient disfluency detection with transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 495–503, Beijing, China. Association for Computational Linguistics.
- Yoshikawa, M., Shindo, H., and Matsumoto, Y. (2016). Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1041, Austin, Texas. Association for Computational Linguistics.
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The htk book. *Cambridge university engineering department*, 3(175):12.
- Zayats, V. and Ostendorf, M. (2019). Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 86–95, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zilcha-Mano, S. (2017). Is the alliance really therapeutic? revisiting this question in light of recent methodological advances. *American Psychologist*, 72:311–325.