Sumarização Multimodal de Diálogos Clínicos na Atenção Primária Digital: Integrando Mensagens Textuais e Áudios

Davi Reis[1], Anderson A. Ferreira[2], Washington Cunha[3], Victor Macul[4], Olivio Neto[4], Jussara Almeida[1], Leonardo Rocha[5], Marcos André Gonçalves[1]

[1] Universidade Federal de Minas Gerais, [2]Universidade Federal de Ouro Preto, [3]Universidade Estadual de Campinas, [4]Ana Health, [5]Universidade Federal de São João del-Rei

{davireis,jussara,mgoncalv}@dcc.ufmg.br

anderson.ferreira@ufop.edu.br, wcunha@unicamp.br

{victor,osouzaneto}@anahealth.app, lcrocha@ufsj.edu.br

Resumo. *Plataformas de mensagens na saúde digital ampliaram o volume de interações, tornando a gestão e a recuperação de informações clínicas um desafio central na Atenção Primária Digital. Embora a sumarização automática de diálogos textuais com Grandes Modelos de Linguagem (LLMs) tenha sido explorada, parte relevante do diálogo ocorre por áudio. Assim, este trabalho propõe um pipeline multimodal para integrar fala e texto na sumarização com LLMs. Foi investigado (i) como extrair automaticamente informações clínicas de áudios com qualidade variável e (ii) o impacto dessa integração na qualidade do resumo. A metodologia foi desenvolvida a partir de 706 áudios reais, com base anotada manualmente e classificadores para filtrar transcrições inadequadas. Os resultados mostram que incorporar áudios enriquece os resumos, aumentando contextualização e detalhamento das informações clínicas.*

Abstract. *Instant messaging platforms in digital health have increased the volume of interactions, making the management and retrieval of clinical information a central challenge in digital primary care. Although automatic summarization of text-based dialogues with Large Language Models (LLMs) has been explored, a substantial portion of these exchanges occurs through audio messages. In this work, we propose a multimodal pipeline that integrates speech and text for LLM-based dialogue summarization. It was investigated (i) how to automatically extract clinically relevant information from audio messages with varying quality and (ii) the impact of this integration on summary quality. The methodology was developed using 706 real-world audio messages, a manually annotated dataset, and classifiers to filter out inadequate transcriptions. Results show that incorporating audio messages enriches the summaries by increasing contextualization and the level of clinical detail.*

1. Introdução

O crescente uso de plataformas de mensagens instantâneas (e.g., WhatsApp) tem impulsionado a consolidação de portais de e-saúde voltados à comunicação entre pacientes e equipes de saúde [Liu et al. 2024]. Esse cenário vem aumentando o volume de interações

digitais, tornando sua gestão um desafio. Estima-se que médicos da atenção primária dediquem cerca de 2 horas por dia à análise de aproximadamente 150 mensagens enviadas por pacientes [Liu et al. 2024]. Para oferecer respostas adequadas, os profissionais precisam interpretar o conteúdo atual à luz do histórico de conversas e de outras informações clínicas, garantindo compreensão contextual e continuidade do cuidado.

No contexto brasileiro, limitações estruturais ampliam essa complexidade: cerca de 34% da população não possui acesso à atenção primária e projeções indicam que a universalização desse atendimento demandaria aproximadamente 236.900 profissionais de saúde, com custo estimado em $R\$$ 22,9 bilhões por ano [Hone et al. 2017]. Diante desse cenário, modelos de Atenção Primária Digital surgem como alternativa para ampliar cobertura. Um exemplo é a empresa Ana Health[1], que concentra grande parte da comunicação com pacientes no WhatsApp. Ainda assim, quando uma nova mensagem chega, é necessário revisar o histórico para compreender o contexto e integrar informações antes de responder, o que impacta tanto a precisão clínica quanto o engajamento do paciente [Keszthelyi et al. 2023]. Além disso, essa comunicação não é apenas textual: mensagens de áudio são frequentes e podem conter detalhes clínicos e nuances relevantes, além de favorecer acessibilidade para pacientes com baixo letramento e dificuldades de escrita [Esquivel et al. 2024]. Ignorar esse conteúdo pode levar à perda de informações importantes para a tomada de decisão e para a continuidade do cuidado.

Um trabalho recente investigou o uso de Grandes Modelos de Linguagem (LLMs) para sumarizar diálogos clínicos textuais em cenários reais e ruidosos, incluindo português brasileiro [Ferreira et al. 2025]. Este artigo avança nesse problema ao propor uma metodologia para incorporar mensagens de áudio ao processo de sumarização do histórico clínico, buscando responder às seguintes questões de pesquisa: **RQ1**. Como extrair automaticamente informações relevantes a partir de mensagens de áudio trocadas entre pacientes e equipe de saúde, considerando variações na qualidade do áudio? **RQ2**. Qual o impacto da incorporação de mensagens de áudio na qualidade da sumarização do histórico clínico, em comparação a abordagens baseadas apenas em texto?

2. Trabalhos Relacionados

A utilização de dados oriundos de interações digitais em saúde tem recebido crescente atenção, especialmente no contexto de plataformas de e-saúde e comunicação assíncrona entre pacientes e equipes clínicas. [Liu et al. 2024] mostram que o volume de mensagens em portais de saúde impõe desafios relevantes à prática clínica, motivando estratégias automáticas para organização e sumarização dessas interações, com potencial para apoiar a tomada de decisão e reduzir a sobrecarga dos profissionais. Nesse contexto, [Anibal et al. 2025] exploram a aplicação de LLMs para apoiar a análise de mensagens em atenção primária, destacando que a interpretação adequada requer compreensão contextual, integração com históricos prévios e robustez a ruídos linguísticos. Essas evidências reforçam o potencial de abordagens baseadas em LLMs para melhorar eficiência e qualidade do cuidado em ambientes de saúde digital. Por fim, [Ferreira et al. 2025] avaliam a sumarização de diálogos clínicos textuais reais com LLMs, discutindo desafios de concisão, completude e veracidade em dados do mundo real. Este trabalho se diferencia ao tratar explicitamente a modalidade de áudio, propondo mecanismos de filtragem de transcrições e integração multimodal para incorporar

[1] `https://www.anahealth.com.br/`

informações clínicas presentes em mensagens de voz ao processo de sumarização.

3. Metodologia

Esta seção descreve o delineamento experimental adotado para inclusão dos dados/mensagens de áudio como fonte de informação, para auxiliar em uma comunicação mais efetiva entre profissionais de saúde e pacientes[2].

3.1. Caracterização do conjunto de dados de áudio

Para a etapa de caracterização, foi utilizado um conjunto composto por 706 arquivos de áudio proprietários fornecidos pela Ana Health, que constitui o conjunto de dados analisado neste estudo. Desses áudios, foram extraídas características que pudessem ser úteis à identificação de qualidade dos áudios. As características extraídas foram: duração, correspondente ao tempo total do áudio em segundos, útil para detectar mensagens excessivamente curtas (silenciosas ou acidentais) ou longas; frequência de amostragem, taxa, em Hz, utilizada na digitalização do sinal, relacionada ao nível de detalhamento e inteligibilidade; amplitude média, média dos valores do sinal, que auxilia na identificação de áudios muito baixos ou predominantemente silenciosos; amplitude máxima, maior valor observado, útil para detectar picos de volume e possíveis distorções; razão de amplitude (média/máximo), relação entre a amplitude média e a máxima, indicando estabilidade do volume; SNR estimado (*Signal-to-Noise Ratio*), estimativa da razão sinal-ruído em decibéis, que mensura a presença de ruídos; *Zero Crossing Rate* (ZCR), taxa de cruzamento por zero, auxiliando na distinção entre fala, silêncio e ruído; e regiões de silêncio, intervalos de baixa intensidade sonora, relevantes para avaliar a fluidez da fala e possíveis falhas técnicas. Para cada uma dessas métricas, foram calculadas estatísticas descritivas (média, desvio-padrão, valores mínimo e máximo e quartis) com o objetivo de fornecer uma caracterização quantitativa detalhada do conjunto de dados. Esse procedimento permite avaliar a variabilidade, a qualidade técnica e os padrões estruturais dos áudios.

3.2. Classificador para identificação prévia de transcrições inadequadas

Primeiramente, foi realizada a anotação de instâncias a serem usadas para treinar o classificador. Em seguida, inferiu-se diversos classificadores afim de selecionar um que fosse eficiente para indicar a adequação do áudio a uma boa transcrição. Para a construção do classificador, foram realizados experimentos com diferentes algoritmos: DecisionTree, KNN, SVM, Logistic Regression e Gradient Boosting, visando avaliar qual estratégia melhor se adapta aos dados da tarefa. Para avaliar os resultados utilizou-se a técnica *Leave-One-Out Cross-Validation*, devido à quantidade pequena de exemplos rotulados, e a métrica *Macro F1-Score* para quantificar a qualidade dos modelos, por avaliar o desempenho médio entre todas as classes de forma equilibrada, sendo especialmente adequada para cenários com desbalanceamento de dados.

3.2.1. Construção e anotação de um conjunto de exemplos rotulados

Iniciou-se com a recuperação dos áudios enviados pelos pacientes. Em seguida, foram extraídas características acústicas, descritas anteriormente, para avaliar a qualidade das gravações. Representado por essas características, os áudios devem ser classificados

[2]Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Universidade Federal de Minas Gerais, parecer de número CAAE 80632524.4.0000.5149.

automaticamente em *BOM* ou *RUIM*, indicando se cada gravação pode ser utilizada como fonte de informação. Para viabilizar o treinamento de um classificador, foi construído um conjunto de dados, anotado manualmente, composto por 100 arquivos de áudio e suas respectivas transcrições, geradas utilizando o modelo Whisper (*openai/whisper-large-v3*). Esse modelo foi escolhido após uma avaliação qualitativa entre diversos modelos, incluindo Whisper, Vosk e Ultravox. Em seguida, cada um dos 100 áudios, juntamente com sua respectiva transcrição, foi avaliado por 3 avaliadores independentes em áudio e transcrição "boa" ou "ruim". A avaliação majoritária foi a classe atribuída ao áudio.

Embora a anotação manual tenha produzido a base inicial, o problema de desbalanceamento entre as classes pode enviesar o treinamento de um classificador em favor da classe majoritária, prejudicando o desempenho e qualidade do resultado final almejado. Para mitigar esse problema, realizou-se uma busca na base por exemplos adicionais da classe *RUIM*. Todos os áudios disponíveis foram representados em um espaço vetorial definido pelas características descritas anteriormente e, a partir das instâncias *RUIM* já anotadas, selecionaram-se os 25 áudios mais próximos (valor escolhido por representar um acréscimo significativo em relação às 100 amostras iniciais). Em seguida, esses 25 áudios foram avaliados manualmente e os classificados como *RUIM* foram incorporados à base anotada manualmente, reduzindo o desbalanceamento para o processo de treinamento. Além disso, foi aplicada a técnica SMOTE [Chawla et al. 2002], visando também aumentar, de forma sintética, o número de instâncias da classe RUIM na base.

3.3. Incorporação de áudios ao processo de sumarização

Para incorporar os arquivos de áudio ao processo de sumarização das mensagens, para cada paciente, cada um de seus arquivos é submetido ao classificador, para filtrar áudios com ruídos excessivos, baixa inteligibilidade ou outras limitações que possam comprometer a interpretação do conteúdo, e, sendo considerado adequado, é fornecido como entrada ao Whisper para produzir a sua transcrição. Essa transcrição é incorporada ao conjunto de mensagens textuais. Esse conjunto resultante de mensagens textuais é então enviado como entrado a um LLM em um *prompt* com instruções para realizar a sumarizarização.

4. Resultados Experimentais

Esta seção contém os resultados da avaliação experimental deste trabalho.

4.1. Conjunto de dados de áudio

A Tabela 1 mostra dados estatísticos dos 706 arquivos de áudio considerados neste trabalho. Observa-se, de modo geral, que as estatísticas descritivas indicam que a base é composta majoritariamente por áudios curtos, com boa qualidade técnica (frequência padronizada e SNR adequado), baixa incidência de ruído estrutural (ZCR controlado) e dinâmica compatível com fala espontânea, reforçando sua adequação e potencial para tarefas de transcrição e análise automática.

4.2. Conjunto de dados de áudios rotulados

Dos 100 arquivos de áudio inicial utilizados para treinar os classificadores, 92 foram considerados adequados (BOM) e 8 inadequados (RUIM), pelos avaliadores. Sendo assim, a base obtida evidencia que, de acordo com os avaliadores, os áudios possuem boa qualidade e podem ser utilizados como fonte de informação, entretanto, esse desbalanceamento pode prejudicar o processo de treinamento de um classificador.

Tabela 1. Dados estatísticos sobre as métricas aplicadas aos arquivos de áudio.

	Média	**Desvio padrão**	**Mínimo**	**Máximo**	**25%**	**50%**	**75%**
Duração (s)	34,22	34,99	0,89	354,25	11,96	23,81	43,85
Frequência de amostragem (Hz)	48000	48000	48000	48000	48000	48000	48000
Amplitude média	0,04	0,03	0,000003	0,23	0,02	0,03	0,06
Amplitude máxima	0,76	0,27	0,00003	1,00	0,54	0,87	1,00
Razão de amplitude	0,06	0,03	0,004	0,23	0,04	0,06	0,08
SNR estimado (dB)	33,61	21,62	17,97	185,26	25,73	29,37	33,99
Zero Crossing Rate	0,04	0,01	0,013	0,13	0,03	0,03	0,04
Regiões de silêncio	7,29	12,05	0,00	149,00	0,00	3,00	9,00

4.3. Avaliação dos classificadores

Considerando o conjunto de dados inicial, foram treinados e avaliados, usando a técnica leave-one-out e a métrica Macro-F1, os classificadores mostrados na Tabela 2. Ressalta-se que os hiperparâmetros de cada técnica de aprendizado foram ajustados usando *grid search*. Sem a adição de novos exemplos, o classificador baseado em árvore de decisão obteve a maior Macro-F1 e também Macro-recall, mostrando um acerto mais balanceado entre as duas classes. Esse aspecto é fundamental, uma vez que o objetivo principal do classificador é detectar áudios que serão utilizados para agregar informações nas sumarizações de histórico dos pacientes.

Visando melhorar os resultados, foram adicionados mais 7 exemplos considerados ruins e aplicada a técnica SMOTE de adição de novos exemplos também ruins ao conjunto de treinamento, adicionando 40% de novos exemplos de treinamento. Neste caso o melhor resultado foi obtido pelo classificador inferido pela técnica Gradient Boost, com os seguintes resultados: macro-precision=0,8971, macro-recall=0,7279 e macro-F1=0,7808. Assim, optou-se por utilizar o classificador inferido por meio do Gradient Boost na próxima etapa, por ter maior eficácia e mitigar o problema do desbalanceamento.

4.4. Avaliação qualitativa preliminar da incorporação dos áudios à sumarização

Como uma avaliação preliminar da incorporação dos áudios à sumarização das mensagens, é mostrado na Figura 1 um exemplo contendo uma sumarização das mensagens originalmente textuais e outro envolvendo a adição dos áudios ao processo de sumarização.

Observa-se que, a sumarização baseada exclusivamente em mensagens originalmente textuais apresenta menor extensão quando comparada àquela que incorpora também os dados provenientes de áudios. Contudo, as diferenças identificadas não se restringem ao comprimento do texto gerado. Verifica-se um aumento na riqueza e na granularidade das informações clínicas quando os áudios são incluídos, resultando em descrições mais detalhadas do contexto do paciente. Especificamente, a versão enriquecida com dados de áudio contempla aspectos adicionais relatados pela paciente, como a sobrecarga mencionada, bem como informações relativas a indivíduos de seu convívio que podem influenciar seu estado de saúde. Esses achados sugerem que a

Tabela 2. Resultados dos classificadores com e sem oversampling.

	Sem *Oversampling*			**Com *Oversampling***		
Modelo	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Decision Tree	0,7188	0,6141	0,6454	0,5736	0,5736	0,5736
KNN	0,4600	0,5000	0,4792	0,6833	0,5279	0,5208
SVM	0,4600	0,5000	0,4792	0,5603	0,6029	0,5569
Logistic Regression	0,7143	0,5571	0,5789	0,6906	0,5837	0,6040
Gradient Boosting	0,9646	0,5625	0,5928	0,8971	0,7279	0,7808

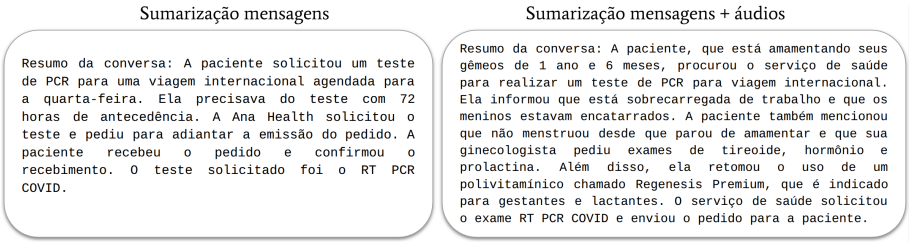

Figura 1. Comparação entre sumarização apenas textual e com áudios

incorporação de mensagens de áudio contribui para uma representação mais abrangente e contextualizada do histórico clínico. Esse resultado é confirmado também pela equipe de saúde da empresa AnaHealth, a qual confirma a utilidade prática da abordagem proposta.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma metodologia para incorporar mensagens de áudio ao processo de sumarização de diálogos clínicos na Atenção Primária Digital, estendendo abordagens previamente focadas apenas em texto. A caracterização de 706 áudios, a criação de uma base manualmente anotada e o desenvolvimento de classificadores permitiram estruturar um fluxo multimodal robusto. Após tratar o desbalanceamento de classes, demonstrou-se a viabilidade de identificar automaticamente áudios com qualidade clínica. Além disso, a análise qualitativa mostrou que a incorporação das informações da fala enriquece os resumos, ampliando a contextualização e a granularidade dos dados clínicos disponíveis. Como trabalhos futuros, pode-se extender o processo de inclusão à diversos casos e realizar avaliações quantitativas adicionais da qualidade dos resumos, incluindo métricas automatizadas e validação por profissionais de saúde, visando fornecer evidências mais robustas sobre o impacto clínico da solução proposta.

Agradecimentos INCT-TILDIAR(# 408490/2024-1), CIIA-Saúde, Ana Health, SEBRAE, EMBRAPII, FINEP, CAPES, CNPq, FAPEMIG e FAPESP.

Referências

Anibal, J., Huth, Wood, B., et al. (2025). Voice EHR: introducing multimodal audio data for health. *Frontiers in Digital Health*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*

Esquivel, P., Gill, K., Goldberg, M., Sundaram, S. A., Morris, L., and Ding, D. (2024). Voice assistant utilization among the disability community for independent living: A rapid review of recent evidence. *Human Behavior and Emerging Technologies*.

Ferreira, A. A., Rocha, L., et al. (2025). A comprehensive qualitative analysis of patient dialogue summarization using large language models applied to noisy, informal, non-english real-world data. *Scientific Reports*.

Hone, T., Rasella, D., Barreto, M. L., Majeed, A., and Millett, C. (2017). Association between expansion of primary healthcare and racial inequalities in mortality amenable to primary care in brazil: a national longitudinal analysis. *PLoS medicine*.

Keszthelyi, D., Gaudet-Blavignac, C., Bjelogrlic, M., and Lovis, C. (2023). Patient information summarization in clinical settings: Scoping review. *JMIR Medical Informatics*.

Liu, S., McCoy, A. B., Wright, A., et al. (2024). Leveraging large language models for generating responses to patient messages-a subjective analysis. *JAMIA*.