INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A Voz da Eficiência: Avaliação do impacto do speech-to-text na celeridade dos relatórios clínicos e estudo de caso da implementação do Invox Medical Dictation

Diana Fernandes dos Anjos

Mestrado em Gestão Aplicada à Saúde

Orientador
Professor Nuno Miguel Pascoal Simões Crespo,
Professor Associado com Agregação,
ISCTE – Instituto Universitário de Lisboa

Outubro, 2025

Departamento de Marketing, Operações e Gestão Geral

A Voz da Eficiência: Avaliação do impacto do speech-to-text na celeridade dos relatórios clínicos e estudo de caso da implementação do Invox Medical Dictation

Diana Fernandes dos Anjos

Mestrado em Gestão Aplicada à Saúde

Orientador
Professor Nuno Miguel Pascoal Simões Crespo,
Professor Associado com Agregação,
ISCTE – Instituto Universitário de Lisboa

Outubro, 2025

Agradecimentos

A realização desta dissertação representa o culminar de uma etapa importante da minha vida académica e pessoal, que não teria sido possível sem o apoio, a presença e a generosidade de várias pessoas, a quem gostaria de expressar o meu mais profundo agradecimento.

Agradeço, em primeiro lugar, aos meus pais, cuja dedicação, valores e incentivo constante foram sempre fundamentais ao longo do meu percurso. Um agradecimento muito especial ao meu pai, pelo apoio incondicional, nos momentos em que mais foi necessário.

Agradeço à empresa Quattro, pelo voto de confiança e pela abertura demonstrada desde o início. Um agradecimento particular à Filipa Correia e Diana Costa, cuja disponibilidade, simpatia e colaboração foram essenciais para o desenvolvimento desta investigação.

Ao Professor Nuno Crespo, meu orientador, agradeço a partilha de conhecimento, o rigor, a dedicação e o acompanhamento atento e disponível ao longo de todo este processo.

À Professora Generosa do Nascimento, deixo um especial agradecimento pela sua amabilidade, dedicação e constante disponibilidade, qualidades que marcaram positivamente o meu percurso académico e pessoal.

Gostaria de expressar o meu reconhecimento à Técnica Clara Rodrigues, o meu contacto direto no Serviço de Anatomia Patológica do Hospital Curry Cabral, cuja colaboração constante e dedicação foram determinantes para o desenvolvimento desta tese. Estendo igualmente os meus agradecimentos à Administrativa Cristina Soares, à Administrativa Margarida Sousa, à Técnica Mariana Neves, à Doutora Ana Carvalho, ao Doutor Rodrigo Cavalcanti, ao Doutor Miguel Cristóvão, à Técnica Helena Romão e à Técnica Elisabete Balau, pelo tempo disponibilizado, pela partilha de conhecimento e pela generosidade em integrar este processo.

A todos, o meu sincero obrigado.

Resumo

A presente dissertação tem como objetivo analisar o impacto da tecnologia *speech-to-text* (STT) na celeridade da emissão de relatórios clínicos, com foco no *software* Invox Medical Dictation. A elaboração manual de relatórios por profissionais de saúde, nomeadamente Médicos e Técnicos de Diagnóstico e Terapêutica (TSDT), constitui um processo moroso e suscetível a erros de transcrição, contribuindo para atrasos significativos na entrega de resultados e afetando negativamente o fluxo de trabalho e a qualidade dos cuidados prestados. Este estudo propõe avaliar se a implementação de uma solução de reconhecimento de voz, adaptada ao vocabulário médico, permite reduzir o tempo de produção documental e aumentar a eficiência organizacional.

O projeto será desenvolvido em colaboração com a empresa Quattro, representante da tecnologia em Portugal, e inclui a realização de revisão de literatura, entrevistas, inquéritos e a discussão de potenciais de melhoria. Será conduzido um estudo de caso numa unidade de saúde onde o *software* já se encontra em utilização, com análise qualitativa da experiência dos utilizadores, complementada por uma análise quantitativa das perceções qualitativas dos colaboradores, de forma a sistematizar e mensurar a sua experiência.

A metodologia segue uma abordagem mista, combinando dados qualitativos e quantitativos para aferir o impacto da solução ao nível da precisão, fiabilidade e satisfação dos utilizadores. Espera-se que os resultados contribuam para uma melhor compreensão do papel das tecnologias STT na modernização da documentação clínica, evidenciando o seu potencial enquanto ferramenta de apoio à decisão, à gestão eficiente de recursos e à melhoria contínua dos serviços de saúde.

Palavras-chave: Reconhecimento de voz; Documentação clínica; Eficiência hospitalar; Tecnologias da saúde; Transcrição médica; Invox Medical Dictation.

Código de Classificação JEL: I10; O33.

Abstract

This dissertation aims to analyse the impact of speech-to-text technology on the timeliness of clinical report production, with a focus on the Invox Medical Dictation software. The manual preparation of reports by healthcare professionals, particularly Physicians and Diagnostic and Therapeutic Technicians, is a time-consuming process prone to transcription errors, which contributes to significant delays in the delivery of results and negatively affects workflow and the quality of care provided. This study seeks to evaluate whether the implementation of a voice recognition solution, adapted to medical vocabulary, can reduce documentation time and enhance organisational efficiency.

The project will be developed in collaboration with Quattro, the company representing the technology in Portugal, and will include a literature review, interviews, surveys, and the discussion of potential improvements. A case study will be conducted in a healthcare unit where the software is already in use, comprising a qualitative analysis of users' experiences, complemented by a quantitative analysis of the qualitative perceptions of the collaborators, in order to systematize and measure their experience.

The methodology follows a mixed-methods approach, combining qualitative and quantitative data to assess the solution's impact in terms of accuracy, reliability, and user satisfaction. It is expected that the findings will contribute to a better understanding of the role of speech-to-text technologies in the modernisation of clinical documentation, highlighting their potential as tools for decision support, efficient resource management, and the continuous improvement of healthcare services.

Keywords: Speech recognition; Clinical documentation; Hospital efficiency; Health Technologies; Medical transcription; Invox Medical Dictation.

JEL Classification Codes: I10; O33.

Índice

Índice de Quadros e Figuras

Glossário de Siglas

AED – Attention-based Encoder-Decoders

ASR – Automatic Speech Recognition

CER – Character Error Rate

CNN – Redes Neuronais Convolucionais

DL – Deep Learning

DNN-HMM – Hybrid Deep Neural Network-Hidden Markov Model

EHR – Sistemas de Registo Eletrónico de Saúde

GMM – Gaussian Mixture Models

HCI – Interação Humano-Computador

HMM – Hidden Markov Models

HMM-GMM – Hidden Markov Models and Gaussian Mixture Models

HMM/NN – Hidden Markov Models (HMM) and Neural Networks

LLMs – Large Language Models

LSTM – Redes Neuronais de Memória de Longo e Curto Prazo

MC-WER – Medical Concept WER

MFCC – Coeficientes Cepstrais em Frequência de Mel

ML – Machine Learning

MM-LLMs – Modelos Multimodais de Grande Escala

NLP – Processamento de Linguagem Natural

PLP – Predição Linear Perceptual

PoC – Proof of Concept

RNN – Redes Neuronais Recorrentes

RTF – Real-Time Factor

STT – Speech-to-text

TAT – Turnaround Time

TSDT – Técnico Superior de Diagnóstico e Terapêutica

UBM - Universal Background Model

WER – Word Error Rate

WFST – Weighted Finite-State Transducers

Introdução

A presente dissertação de Mestrado, intitulada "A Voz da Eficiência: Avaliação do impacto do *speech-to-text* na celeridade dos relatórios clínicos e estudo de caso da implementação do Invox Medical Dictation", surge no âmbito da necessidade crescente de otimização dos processos de documentação médica nas unidades de saúde. Com o avanço das tecnologias de reconhecimento de voz e a sua integração progressiva em ambientes clínicos, torna-se crucial compreender o seu impacto.

A elaboração de relatórios por parte dos profissionais de saúde, nomeadamente Médicos e TSDTs, constitui uma etapa demorada, exigente e propensa a erros de transcrição. O tempo despendido na digitação de relatórios compete com outras tarefas clínicas, contribuindo para a acumulação de exames pendentes e atrasos na entrega dos resultados aos utentes.

A morosidade da documentação clínica não só sobrecarrega os profissionais, como também tem impacto nos indicadores de desempenho institucional, como os tempos médios de resposta, a produtividade dos serviços e a satisfação dos utentes. Há um impacto negativo direto sobre a cadeia de prestação de cuidados: as consultas de seguimento são frequentemente adiadas devido à indisponibilidade dos relatórios necessários e, em casos mais graves, verifica-se o adiamento de tratamentos que dependem desses resultados.

Adicionalmente, o crescente volume de informação clínica, aliado à escassez de recursos humanos em saúde, impõe uma reflexão sobre a dependência de processos administrativos tradicionais baseados em digitação manual. Apesar de algumas instituições terem informatizado parcialmente os seus sistemas, a documentação continua, em muitos contextos, a ser um processo moroso e suscetível a falhas.

Neste contexto, as tecnologias de reconhecimento automático de voz têm ganho destaque como soluções promissoras para a automatização da documentação clínica. Estas ferramentas, com base em modelos avançados de Inteligência Artificial (IA) e processamento de linguagem natural (NLP), convertem a voz do profissional em texto estruturado ou livre, reduzindo o tempo de transcrição e permitindo maior fluidez na prática clínica. Torna-se imperativo encontrar soluções na adoção de novas tecnologias, como o *software* Invox Medical Dictation, que corresponde a uma ferramenta *speech-to-text*, que mitiga os problemas mencionados.

Assim, o foco do projeto incide sobre o *software* Invox Medical Dictation, desenvolvido pela empresa espanhola VÓCALI SISTEMAS INTELIGENTES S.L. e comercializado em Portugal pela empresa Quattro. A VÓCALI foi fundada em 2007 como *spin-off* de um grupo de investigação da Universidade de Múrcia, com competências centrais em NLP e IA. A empresa dedica-se à criação de motores próprios de reconhecimento de voz e modelos linguísticos adaptados ao vocabulário médico, utilizando técnicas de *Machine Learning* (ML) e *Deep Learning* (DL).

A solução preconizada distingue-se pelo compromisso da VÓCALI com a privacidade e a segurança da informação, sendo detentora das certificações ISO/IEC 27001, que atesta a conformidade com normas internacionais de gestão de segurança da informação, e ENS (*Esquema Nacional de Seguridad*) de nível alto, concedida pelo governo de Espanha. Estas certificações garantem que a solução cumpre rigorosos requisitos técnicos e organizacionais no tratamento de dados sensíveis.

Importa também sublinhar que a VÓCALI coloca a relação humana no centro da sua abordagem de desenvolvimento e suporte tecnológico. De acordo com Santiago Bermúdez, responsável pela área de *Customer Success*, "a relação humana é essencial para poder oferecer um serviço excelente e orientado a conhecer e satisfazer as necessidades do cliente". Esta filosofia revela-se particularmente relevante no contexto da saúde, onde a empatia e a capacidade de adaptação às necessidades dos profissionais e das instituições são tão importantes quanto a componente técnica.

Com uma presença internacional consolidada em mais de 600 instituições de saúde em 21 países, a VÓCALI afirma-se como um dos principais *players* na digitalização do setor da saúde. O presente projeto assume, assim, particular relevância ao propor uma avaliação do impacto desta tecnologia numa realidade concreta do Sistema de Saúde Português, promovendo uma análise crítica dos seus benefícios, limitações e condições ótimas de implementação.

A presente tese de Mestrado conta com a colaboração da empresa Quattro IP Quatro Soluções, Unipessoal, Lda. A Quattro foi fundada em 2020 e é uma consultora portuguesa, multissetorial, com especialização no setor da Saúde. A empresa tem como foco o fornecimento de produtos e serviços de tecnologias de informação que permitem às organizações obter uma gestão mais eficaz e eficiente dos seus processos.

Ainda que esta solução seja um ganho na área da saúde, a adoção destas tecnologias ainda encontra resistências, quer pela necessidade de integração com os sistemas de

informação hospitalares existentes, quer pela perceção de perda de controlo sobre o conteúdo e forma do relatório clínico.

Deste modo, torna-se pertinente investigar o real impacto da implementação de uma ferramenta de reconhecimento de voz em contexto clínico, nomeadamente numa unidade hospitalar de grande dimensão. O presente estudo propõe avaliar a eficiência, fiabilidade e satisfação associadas à utilização do Invox Medical Dictation no Hospital Curry Cabral, pertencente à Unidade Local de Saúde de São José, na especialidade de Anatomia Patológica, explorando o seu potencial para mitigar os desafios associados à documentação manual.

A evidência apresentada visa responder às seguintes questões de investigação:

1. Qual o impacto da utilização do Invox Medical Dictation no tempo de elaboração dos relatórios clínicos?
2. Qual o nível de satisfação dos utilizadores relativamente à experiência com o *software*?
3. Que tipo de erros são mais comuns na transcrição *speech-to-text* e como se comparam com os erros do processo manual?

O objetivo central deste estudo é avaliar o desempenho do *software* Invox Medical Dictation em contexto real de utilização hospitalar. Procura-se compreender de que forma esta ferramenta influencia a eficiência dos processos de documentação clínica e a produtividade dos serviços hospitalares onde é implementada. Para além disso, pretende-se analisar a perceção dos profissionais de saúde, enquanto interface de apoio à prática clínica. A investigação inclui também a identificação de potenciais limitações operacionais e técnicas, com vista à sua superação em futuras implementações, e pretende contribuir para uma reflexão alargada sobre a viabilidade da adoção generalizada deste tipo de tecnologia no Sistema de Saúde.

O projeto está estruturado em seis partes principais, de forma a refletir uma abordagem progressiva, que parte da caracterização do problema até à avaliação empírica da solução e análise do seu potencial de expansão futura.

O Capítulo I expõe a revisão da literatura, baseada em 40 artigos científicos, permitindo uma visão transversal sobre a adoção de sistemas de reconhecimento de voz em diversos contextos clínicos internacionais. Este capítulo inclui a análise do estado de arte, os benefícios e limitações observados, as métricas mais relevantes de desempenho, e os avanços mais recentes em tecnologias de *speech-to-text*.

O Capítulo II caracteriza a solução tecnológica proposta, o Invox Medical Dictation, descrevendo a sua arquitetura, o enquadramento técnico e o percurso de implementação nos hospitais selecionados.

De seguida, no capítulo III é apresentada a metodologia do estudo em causa, onde se inclui a seleção da amostra, a recolha de dados e a forma como os mesmos foram tratados e analisados.

O Capítulo IV apresenta o estudo de caso realizado no serviço de Anatomia Patológica do Hospital Curry Cabral, com o objetivo de aferir o impacto da tecnologia Invox Medical Dictation no tempo de elaboração de relatórios, perceção de eficiência por parte dos profissionais, e fiabilidade dos resultados. Esta parte inclui os resultados obtidos, bem como o tratamento e análise de dados.

O Capítulo V é dedicado à empresa Quattro e integra a entrevista realizada a colaboradores, constituindo um contributo complementar para a análise desenvolvida no presente estudo de caso.

O Capítulo VI é direcionado para a discussão dos resultados e discute o potencial de aprofundamento da solução, abordando melhorias futuras na arquitetura baseada em redes neuronais, e estratégias de adaptação à variabilidade do discurso médico, bem como desafios como a escassez de dados anotados ou diferenças no estilo de ditado entre médicos.

CAPÍTULO I - Revisão de Literatura

1.1 – Estado de Arte e arquitetura das ferramentas de reconhecimento de voz

O reconhecimento automático de voz, conhecido internacionalmente como *Automatic Speech Recognition* (ASR), tem vindo a consolidar-se como uma das áreas mais promissoras da IA aplicada à saúde. Esta tecnologia visa a conversão da linguagem falada em texto escrito de forma automática, através de algoritmos que processam o sinal de áudio e criam transcrições textuais com um elevado grau de precisão.

De acordo com Kumar (2024), os avanços no poder computacional, aliados à disponibilidade de grandes volumes de dados clínicos e à evolução de modelos estatísticos baseados em redes neuronais profundas, têm impulsionado significativamente o desenvolvimento de sistemas ASR especificamente concebidos para o setor da saúde.

A arquitetura tradicional dos sistemas de ASR é composta por três fases fundamentais: pré-processamento, extração de características e reconhecimento propriamente dito. Na fase de pré-processamento, o sinal de voz é digitalizado, normalizado e segmentado, eliminando-se ruídos e silêncios desnecessários. Segue-se a extração de características acústicas, como os Coeficientes Cepstrais em Frequência de Mel (MFCC) ou a Predição Linear Perceptual (PLP), que representam matematicamente o conteúdo fonético do sinal. Estas características alimentam, depois, um modelo de reconhecimento, onde algoritmos de ML ou DL são responsáveis por mapear os sinais acústicos nas suas correspondentes unidades linguísticas, como fonemas, palavras ou frases (Latif et al., 2021; Kumar et al., 2022).

Tradicionalmente, os sistemas ASR seguiam uma arquitetura modular, composta por modelos acústicos, modelos linguísticos e mecanismos de decodificação baseados, geralmente, em Modelos Ocultos de Markov (HMM) combinados com Modelos de Mistura Gaussiana (GMM) para a modelação acústica, e modelos *n-gram* para a componente linguística (Ogundokun et al., 2020). Contudo, os avanços recentes conduziram ao desenvolvimento de arquiteturas mais integradas, como os modelos baseados em Redes Neuronais Recorrentes (RNNs), Convolucionais (CNNs) e, mais recentemente, *Transformers*. Estas abordagens *end-to-end* visam otimizar o processo de conversão de voz em texto de forma mais direta, minimizando a necessidade de alinhamentos forçados e de múltiplos componentes independentes (Bahdanau et al., 2016). Esta evolução permitiu o desenvolvimento de sistemas de transcrição de voz mais precisos, com capacidade de aprendizagem contextual e adaptabilidade a diferentes locutores e ambientes acústicos.

Particularmente relevantes são os modelos baseados em CNNs, que conseguem operar diretamente sobre o sinal acústico bruto, sem necessidade de etapas intermediárias de extração manual de características. Palaz et al. (2015) demonstram que as CNNs conseguem aprender representações discriminativas diretamente a partir da forma de onda do áudio, proporcionando resultados competitivos com modelos tradicionais, e com a vantagem adicional de uma maior robustez a variações do sinal, como ruído de fundo, sotaques e alterações no ritmo da fala.

No domínio da saúde, o reconhecimento de voz tem aplicações estratégicas, nomeadamente na documentação clínica, permitindo a criação automática de relatórios clínicos, notas de evolução e prescrições, entre outros documentos. Este contexto exige uma elevada especialização dos modelos, nomeadamente em relação ao vocabulário técnico e à sensibilidade semântica dos conteúdos. Segundo Latif et al. (2021), os modelos STT aplicados à saúde beneficiam de corpora específicos, como o MIMIC-III, que incluem registos sonoros com terminologia médica especializada, permitindo treinar sistemas que reconhecem termos complexos, expressões clínicas e estruturas frásicas típicas da linguagem dos profissionais de saúde.

Bahdanau et al. (2016) apresentam um modelo baseado em atenção (*Attention-based Recurrent Sequence Generators*), no qual uma RNN é capaz de identificar autonomamente os segmentos relevantes do sinal de entrada para cada unidade textual a produzir, eliminando a dependência dos HMM. Esta abordagem elimina a necessidade de alinhamentos prévios entre áudio e texto e oferece maior flexibilidade no tratamento de línguas com estruturas fonéticas complexas. Adicionalmente, a integração de modelos de linguagem baseados em *n-gram* ao processo de decodificação por meio *de Weighted Finite-State Transducers* (WFSTs) contribui para melhorar a precisão final do sistema.

Além disso, os sistemas modernos tendem a integrar mecanismos de adaptação por locutor e contextos multimodais, combinando áudio com texto clínico ou imagens médicas. De acordo com Kumar et al. (2022), este tipo de integração aumenta significativamente a fiabilidade da transcrição e a sua utilidade para fins clínicos.

Do ponto de vista da arquitetura funcional, os sistemas de reconhecimento de voz modernos podem ser classificados em unimodais e multimodais. Os sistemas unimodais, particularmente os baseados em áudio, incluem tecnologias como o reconhecimento de voz e de locutor, sendo projetados para interagir com o utilizador através de um único canal sensorial. Por outro lado, os sistemas multimodais integram diferentes modalidades (áudio,

visual e/ou tátil), possibilitando interações mais robustas e adaptáveis, sobretudo em contextos médicos e assistivos (Ogundokun et al., 2020).

Os avanços na interação humano-computador (HCI) também influenciaram o desenvolvimento de sistemas de reconhecimento de voz, com enfoque na melhoria da usabilidade, eficiência e adequação ao contexto do utilizador (Ogundokun et al., 2020). Estes sistemas são agora concebidos para operar com dados de entrada variáveis e para lidar com desafios como ruído ambiental, sotaques regionais e fluência do discurso, fatores críticos para a aplicação em ambientes clínicos.

O reconhecimento automático de voz não é apenas uma ferramenta técnica, é uma alavanca para a eficiência organizacional e para a transformação digital da saúde. O seu impacto na produtividade clínica, na redução da carga administrativa e na qualidade da informação registada torna-o um dos pilares emergentes da modernização dos serviços de saúde, como argumentado por Latif et al. (2021) e reforçado por Kumar (2024) no seu estudo de referência sobre ASR no setor clínico.

Em síntese, o panorama atual das ferramentas STT revela uma transição progressiva de sistemas híbridos baseados em modelos estatísticos para arquiteturas *end-to-end* profundamente integradas, suportadas por avanços no campo da IA. Esta evolução tem potenciado o desempenho, a escalabilidade e a aplicabilidade dos sistemas ASR em contextos cada vez mais exigentes, como a documentação clínica automatizada.

1.2 – Impacto do uso de *softwares* de conversão de voz em texto na documentação clínica

A adoção de ferramentas de conversão de voz em texto na documentação clínica tem vindo a transformar de forma significativa os processos de registo de informação nos cuidados de saúde. Tradicionalmente centrada na digitação manual ou ditado transcrito por terceiros, a documentação clínica tem gradualmente incorporado tecnologias de reconhecimento de voz, com o objetivo de otimizar a eficiência, qualidade e segurança da informação médica.

Estudos recentes têm demonstrado que a utilização de *softwares* ASR pode reduzir o tempo necessário para concluir registos clínicos, enquanto aumenta a quantidade e a riqueza de informação documentada. Vogel et al. (2015) realizaram um ensaio clínico que demonstrou um aumento de 26% na velocidade de documentação quando os clínicos utilizaram um sistema *web-based* de ASR, em comparação com a digitação tradicional. Além disso, os documentos criados com reconhecimento de voz foram significativamente mais extensos, indicando uma maior expressividade e detalhe da informação clínica.

Do ponto de vista qualitativo, Blackley et al. (2020) observaram que as notas clínicas ditadas continham mais palavras, maior variedade vocabular e obtiveram pontuações superiores em critérios como clareza, completude e suficiência de informação, em comparação com notas digitadas. Ainda que o tempo de documentação não tenha diferido de forma estatisticamente significativa entre os métodos, os profissionais relataram uma perceção de maior eficiência e preferiram o uso de ASR para relatar episódios clínicos mais detalhados.

Ajami (2016) argumenta que os sistemas de reconhecimento de voz oferecem oportunidades claras para melhorar a qualidade da documentação, reduzir os custos e o tempo de registo, e apoiar os profissionais de saúde na produção de registos clínicos mais legíveis, precisos e acessíveis. Este estudo salienta que os benefícios destas tecnologias não se limitam à eficiência, estendendo-se à melhoria da qualidade dos serviços prestados, contribuindo para a satisfação dos profissionais. O estudo ressalva, contudo, que a adoção eficaz exige atenção a fatores como o treino dos utilizadores, a qualidade do dicionário técnico e a gestão de ruído nos ambientes clínicos.

Relativamente à dimensão económica, Ajami (2016) refere que a adoção da tecnologia resultou em reduções de custo significativas em instituições que implementaram ASR em larga escala, como relatado por Antiles et al. no *Massachusetts General Hospital*, onde se registou uma poupança de 530 mil dólares em dois anos. Ainda assim, outros estudos, como o de Pezzullo et al. (2008), referem um aumento do tempo de ditado e de frustração dos utilizadores quando a tecnologia não está suficientemente madura ou não se adapta ao contexto clínico específico.

Importa ainda considerar a variabilidade de impacto conforme o local de aplicação. Na Anatomia Patológica, a tecnologia demonstrou reduzir o tempo de produção e melhorar o fluxo de trabalho (Ajami, 2016). Já nos cuidados ambulatórios, Molnar et al. (2000) identificaram um aumento inicial no tempo de edição, mas uma redução significativa no tempo total para finalização dos relatórios. Em ambientes de urgência, a tecnologia revelou-se capaz de reduzir o tempo médio de registo de mais de 12 horas para cerca de 2 horas (Ajami, 2016).

A perceção de ganho de produtividade e de bem-estar dos profissionais de saúde também tem sido destacada. Galloway et al. (2024), num estudo piloto com tecnologia baseada em IA generativa, observaram uma melhoria significativa na perceção dos profissionais relativamente à usabilidade do sistema de documentação e ao impacto do mesmo no seu bem-estar. Os participantes relataram que, com a utilização da ferramenta

de reconhecimento de voz, sentiram uma redução do impacto negativo que os processos de registo tinham na sua saúde mental, referindo também uma melhoria significativa na experiência global proporcionada aos utentes, evidenciando o potencial destas ferramentas para mitigar fatores associados ao *burnout* clínico.

A eficiência percebida pelo utilizador foi também salientada no estudo de Goss et al. (2019), onde mais de 75% dos clínicos relataram que o ASR aumentava a sua produtividade e lhes permitia cumprir as exigências de documentação com maior agilidade. Apesar da presença de erros em parte das transcrições, a maioria foi considerada pouco significativa do ponto de vista clínico, e os utilizadores demonstraram elevada satisfação com a tecnologia. A preferência pelo ASR esteve positivamente associada à perceção de eficiência e negativamente correlacionada com o tempo de edição e a frequência de erros.

Joseph et al. (2020), numa revisão sistemática centrada na documentação de enfermagem, identificaram que o reconhecimento de voz pode aumentar significativamente a produtividade e a precisão do registo clínico dos utentes. A análise de dez estudos revelou que, apesar da heterogeneidade metodológica, a maioria dos resultados apontava para melhorias consistentes na exatidão e na eficiência das notas clínicas. Contudo, os autores sublinham que a eficácia destas ferramentas depende da compatibilidade com os sistemas eletrónicos de saúde já existentes, bem como da formação adequada dos profissionais e da disponibilidade de suporte técnico no local.

A revisão sistemática conduzida por Hammana et al. (2015), focada em departamentos de Radiologia, identificou melhorias consideráveis na produtividade após a adoção de ASR, com reduções nos tempos de *turnaround* (TAT) que variaram entre 35% e 99%. No entanto, a produtividade individual dos Radiologistas tendeu a diminuir, com um aumento do tempo médio necessário para produzir cada relatório. Os autores referem ainda uma elevada heterogeneidade nas taxas de erro: entre 4,8% e 89% dos relatórios com ASR apresentavam pelo menos um erro, enquanto nos métodos de transcrição tradicional os erros variavam entre 2,1% e 22%. Estes dados sugerem que, apesar dos ganhos sistémicos, os sistemas STT exigem maior esforço de edição por parte dos clínicos, o que pode comprometer a produtividade individual e implicar estratégias de adaptação operacional.

Deepa e Khilar (2022) destacam que os sistemas ASR podem ser configurados como "companheiros médicos digitais", capazes de interagir com os profissionais de forma contextualizada, apoiando a decisão clínica e enriquecendo a experiência do registo.

Uma revisão sistemática conduzida por Hodgson e Coiera (2016) identificou uma heterogeneidade substancial nos resultados dos estudos avaliados, revelando que o tempo de ditado e edição pode aumentar em certos contextos. Apesar de uma melhoria consistente nos tempos de TAT dos documentos, os erros por relatório tendem a ser superiores nas transcrições por ASR em comparação com os métodos tradicionais. Este aumento nos erros, por vezes de natureza clínica relevante, levanta preocupações sobre a segurança da documentação automatizada e aponta para a necessidade de monitorização e validação contínua dos conteúdos criados por estas tecnologias.

Por outro lado, estudos mais recentes têm revelado melhorias significativas nas taxas de erro e precisão das transcrições, à medida que os sistemas evoluem. Busch et al. (2025), num estudo de *proof of concept* sobre a utilização do modelo GPT-4o para transcrição de relatórios radiológicos em inglês e alemão, concluíram que a tecnologia é capaz de produzir transcrições com elevada precisão sem diferenças significativas entre idiomas, sendo os erros na maioria dos casos menores e técnicos. Os resultados sugerem que os modelos de linguagem multimodais de última geração são particularmente promissores para uso clínico, sobretudo em ambientes multilíngues.

Em síntese, o impacto dos *softwares* de conversão de voz em texto na documentação clínica é, de forma geral, positivo, proporcionando ganhos em eficiência, conteúdo e experiência do utilizador. No entanto, os benefícios estão condicionados à maturidade tecnológica das ferramentas, à sua integração nos sistemas clínicos e à capacidade dos profissionais para as utilizar de forma eficaz e crítica. A investigação futura deve continuar a explorar os efeitos destas tecnologias em contextos clínicos diversos, considerando não apenas os indicadores de desempenho técnico, mas também as implicações para a qualidade assistencial e segurança do utente.

1.3 – Métricas de avaliação de desempenho de ferramentas *speech-to-text*

A avaliação do desempenho de ferramentas de reconhecimento de voz em contexto clínico exige métricas objetivas que quantifiquem tanto a fiabilidade da transcrição como a sua utilidade prática, para garantir a qualidade das transcrições, sobretudo em contextos clínicos. Diversos estudos na área da saúde têm utilizado um conjunto de métricas que incluem: *Word Error Rate* (WER), *Character Error Rate* (CER), *Medical Concept* WER (MC-WER), *Turnaround Time* (TAT), *Real-Time Factor* (RTF), Taxa de erro global, Precisão, Sensibilidade (*Recall*), *F1-score*, *Accuracy, BERTScore* e o *ROUGE*.

A métrica amplamente utilizada na avaliação da *performance* de ferramentas de reconhecimento de voz é o WER, que calcula a proporção de palavras transcritas incorretamente em relação ao total de palavras no texto de referência. Quanto menor o WER, melhor o desempenho do sistema. Em aplicações clínicas, valores elevados de WER podem comprometer significativamente a qualidade e a segurança da informação (Haboussi et al., 2025). A métrica é determinada pela fórmula:

$$WER = \frac{Substituições + Inserções + Omissões}{Número\ total\ de\ palavras\ de\ referência} \quad (1.1)$$

A CER é uma métrica semelhante à WER, mas calcula os erros ao nível de caracteres em vez de palavras. A CER é particularmente útil para línguas com elevada complexidade ou com termos técnicos longos, como o alemão ou o português. Assim como a WER, a CER considera inserções, omissões e substituições de caracteres (Haboussi et al., 2025). No estudo de Busch et al. (2025), o modelo GPT-4o atingiu uma CER de apenas 1,98% em transcrições, demonstrando elevada precisão fonética em contextos linguísticos complexos.

O MC-WER é uma métrica apresentada no artigo Adedeji et al., 2024, que adapta a tradicional WER ao contexto clínico, focando-se na precisão dos conceitos médicos presentes na transcrição. Em vez de contabilizar todos os erros de palavras, o MC-WER avalia especificamente os erros que afetam termos médicos relevantes, como diagnósticos, procedimentos ou localizações anatómicas. Esta métrica é particularmente útil porque pequenas falhas na transcrição de conceitos clínicos podem ter impacto direto na interpretação médica.

O TAT é outra métrica crítica, que mede o tempo total gasto entre ditar um relatório e a sua disponibilização final no sistema. Esta métrica é central na avaliação de eficiência operacional, especialmente em contextos como a Radiologia ou Anatomia Patológica, em que há elevada rotatividade de exames. Estudos como os de Ringler et al. (2017) e Zuchowski et al. (2022) destacam ganhos significativos na redução do TAT com o uso de ASR, reforçando o valor desta métrica para decisões clínicas mais ágeis. Inclusive no estudo de Zuchowski e Göller (2022), o TAT médio foi reduzido de 8,9 minutos (digitação manual) para 5,11 minutos com uso de reconhecimento de voz, representando uma poupança de 43%.

O RTF mede a relação entre o tempo de processamento da transcrição e a duração do áudio original. Um RTF < 1 indica que o sistema consegue operar em tempo real, o que

é crucial em ambientes clínicos dinâmicos (Haboussi et al., 2025). No estudo de Busch et al. (2025), os autores salientam que o GPT-4o atingiu um RTF inferior a 1 na maioria dos casos, mesmo em transcrições multilíngues, o que comprova a sua viabilidade para aplicações médicas em tempo real.

A percentagem de erros totais por relatório ou por linha também é frequentemente reportada como métrica. Segundo Ringler et al., 2017, a taxa de erro global representa a proporção de relatórios com pelo menos um erro, e pode ser segmentada em erros materiais, potencialmente críticos para a segurança do utente, e imateriais de menor impacto. Esta métrica é particularmente útil em auditorias internas e programas de controlo de qualidade. No estudo de Ringler et al. (2017), cerca de 9,7% dos relatórios apresentavam erros, sendo 1,9% classificados como materiais. No estudo de Busch et al. (2025), que avaliou o GPT-4o para transcrição de relatórios de TC e RM, verificou-se uma taxa de erro média de 38% a 44% por utilizador, em 100 relatórios, sendo a maioria dos erros classificados à posteriori como menores e técnicos ou tipográficos.

Segundo Salifu et al., 2025, no domínio da IA aplicada ao reconhecimento de voz, são frequentemente utilizadas as métricas clássicas de classificação, tendo em conta os seguintes elementos: *True Positive* (TP) que representa os casos em que o sistema previu corretamente um resultado positivo, ou seja, quando um evento clínico relevante foi corretamente reconhecido pelo *software*; *True Negative* (TN), que corresponde às situações em que o sistema previu corretamente um resultado negativo, ou seja, reconheceu corretamente a ausência de um evento; *False Positive* (FP), que ocorre quando o sistema identifica um evento como presente, quando na realidade este não ocorreu; *False Negative* (FN), que representa os casos em que o sistema não reconhece um evento que de facto ocorreu, podendo comprometer a segurança do utente por omissão de informação relevante. Com base nesses conceitos, calculam-se as seguintes métricas: a Precisão, que se refere à proporção de palavras corretamente transcritas entre todas as palavras que o sistema reconheceu, ou seja, mede a proporção de verdadeiros positivos entre todas as previsões positivas feitas pelo sistema, indicando o quão confiável é o sistema quando identifica um evento; a Sensibilidade ou *Recall*, que mede a proporção de verdadeiros positivos entre todos os casos em que o evento realmente ocorreu, avaliando a capacidade do sistema em captar todos os casos relevantes; o *F1-score*, que resulta de uma média entre a Precisão e a Sensibilidade, especialmente útil quando há desequilíbrio entre tipos de erros; a *Accuracy*, uma métrica global que mede a proporção de predições corretas, tanto positivas quanto

negativas, em relação ao total de palavras avaliadas, ou seja, indica quantas vezes o sistema ASR acertou nas suas previsões, independentemente da classe.

Seguem-se as fórmulas:

$$Precisão = \frac{TP}{TP + FP} \tag{1.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{1.3}$$

$$F1 - score = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{1.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.5}$$

Além das métricas tradicionais com base em contagem de palavras ou caracteres, métricas mais recentes e sofisticadas têm sido utilizadas para avaliar a fiabilidade semântica das transcrições criadas por sistemas STT. Entre estas, destacam-se o *BERTScore* e o *ROUGE*. O *BERTScore* utiliza modelos de linguagem como o *BERT* para comparar o significado das palavras entre a transcrição criada e o texto de referência, em vez de apenas verificar correspondência exata de palavras, o que permite avaliar se o conteúdo principal foi mantido mesmo quando há variações na forma de expressão, como o uso de sinónimos ou reordenação de frases. Já o *ROUGE* mede a sobreposição entre partes do texto, como palavras ou sequências de palavras, entre a transcrição e o original. No contexto clínico é essencial manter o significado exato de termos técnicos e achados médicos, logo o uso combinado de *BERTScore* e *ROUGE* permite uma avaliação mais completa da fiabilidade semântica da transcrição (T. Zhang et al., 2020). Estudos como o de Busch et al. (2025) demonstram que o *BERTScore* apresenta resultados mais alinhados com julgamentos humanos sobre a qualidade do conteúdo transcrito.

A inclusão destas métricas na avaliação de soluções como o Invox Medical Dictation, objeto de estudo deste projeto, seria importante para a análise do seu impacto na celeridade, exatidão e qualidade dos relatórios clínicos.

1.4 – Limitações estruturais de ferramentas *speech-to-text*

Embora os sistemas de reconhecimento automático de voz tenham evoluído substancialmente nas últimas décadas, persistem limitações estruturais que comprometem a sua eficácia e precisão em ambientes clínicos. Tais limitações podem ser agrupadas em três categorias principais: fatores ambientais e contextuais, falhas tecnológicas intrínsecas aos modelos de reconhecimento, e desafios de integração nos fluxos de trabalho clínicos.

Um dos principais entraves à *accuracy* das ferramentas ASR é a influência de variáveis ambientais, como ruído de fundo, sobreposição de vozes e a velocidade do discurso. Estudos demonstram que o desempenho de *softwares* de reconhecimento é significativamente reduzido em ambientes ruidosos ou com fala acelerada. Por exemplo, Karbasi et al. (2019) reportaram que a *accuracy* do *software* NEVISA caiu de 82,08% em ambientes silenciosos com discurso pausado para 59,81% em ambientes movimentados com discurso rápido, evidenciando a sensibilidade dos sistemas a condições acústicas adversas.

Os sistemas com base em modelos híbridos HMM/NN e redes neuronais *end-to-end*, como RNN ou *attention-based encoder-decoders* (AED), sofrem deteriorização de *performance* quando confrontados com condições acústicas divergentes dos dados de treino, o que evidencia a necessidade de algoritmos robustos de adaptação, especialmente em contextos clínicos onde o ambiente sonoro é imprevisível (Bell et al., 2021).

Além disso, os sistemas multilíngues enfrentam interferência entre línguas e desequilíbrios de dados, o que afeta a generalização do modelo. Em particular, a escassez de dados anotados em português (de Portugal) prejudica o desempenho em comandos específicos como pausas e início de parágrafos, revelando fragilidade das ferramentas atuais para usos especializados. A inadequação das ferramentas também se manifesta em ambientes com múltiplos utilizadores ou interrupções. Sistemas que operam em modo contínuo e holístico, como os que têm na sua base de construção *Transformers*, por vezes falham na segmentação e na interpretação contextual, especialmente quando há comandos implícitos ou interrupções na fala (Li et al., 2025).

O estudo de Paul et al. (2025) evidencia as limitações estruturais enfrentadas por sistemas ASR em línguas com poucos recursos, como o Sadri, uma língua falada no leste da Índia. Os autores demonstraram que, mesmo com a aplicação de técnicas avançadas de *Transfer Learning* e Redes Neuronais como CNN, o desempenho do sistema é fortemente condicionado pelo volume e qualidade dos dados disponíveis. A escassez de corpora treináveis impede a construção de modelos robustos e bem generalizados. Embora o uso de

dados auxiliares em *hindi* e *bengali* tenha melhorado substancialmente os resultados, reduzindo a taxa de erro de palavras de 22,08% para 15,25%, o estudo salienta que, sem estratégias específicas de adaptação linguística e contextual, as ferramentas de reconhecimento de voz permanecem estruturalmente limitadas na sua capacidade de suportar múltiplos idiomas regionais com precisão clínica.

Uma das limitações estruturais mais críticas destacadas por Latif et al. (2021) diz respeito à ausência de soluções verdadeiramente integradas e orientadas à saúde nos sistemas atuais de reconhecimento de voz. Apesar dos avanços proporcionados pelas Redes Neuronais Profundas e pelo uso de modelos como CNNs e RNNs, os autores referem que a maioria das ferramentas ASR ainda são desenvolvidas com foco genérico e carecem de componentes adaptativos específicos ao domínio clínico, o que inclui limitações em léxicos médicos, falta de contextualização semântica e dificuldades em interpretar sinais paralinguísticos, que são cruciais para a compreensão adequada do discurso clínico. Além disso, sublinha-se que muitos sistemas existentes não foram concebidos para processar linguagem natural em contextos com múltiplos oradores ou interrupções frequentes, dificultando a sua aplicabilidade em cenários hospitalares de alta complexidade.

Em alguns casos, a qualidade e diversidade dos *datasets* utilizados para treino das ferramentas de ASR são insuficientes, particularmente em línguas com menos recursos, como o português. A utilização de dados de baixa qualidade ou com variações acentuadas de sotaque compromete a eficácia dos modelos (Kumar et al., 2023).

A acrescentar, a natureza probabilística dos modelos de linguagem utilizados em ASR, como os baseados em *Hidden Markov Models* ou Redes Neuronais Profundas, não é imune a falhas. A revisão sistemática de Hodgson e Coiera (2016) identificou taxas de erro consideravelmente mais elevadas em documentos criados por ASR em comparação com a transcrição tradicional, com uma média de 0,05 a 6,66 erros por documento, contra 0,02 a 0,40 na transcrição manual. Em ambientes de elevada exigência clínica esses erros podem comprometer a precisão diagnóstica e a segurança do utente. Erros classificados como clínicos significativos, como troca de lado anatómico, medicação ou valor laboratorial, foram particularmente preocupantes (Blackley et al., 2019).

Outro obstáculo estrutural relevante diz respeito à integração ineficiente das ferramentas ASR aos Sistemas de Registo Eletrónico de Saúde (EHR). No estudo experimental de Hodgson, Magrabi e Coiera (2017) foi demonstrado que o uso de reconhecimento de voz aumentou o tempo total de documentação em 18,11% e triplicou o número de erros documentais em comparação com teclado e rato, mesmo em tarefas

simples. Esta ineficiência decorre, em parte, da fraca integração entre o *software* de reconhecimento e o EHR, exigindo melhorias tanto na interoperabilidade quanto na ergonomia da interface (Hodgson et al., 2018).

Estudos sobre a adoção de ASR por Médicos revelam perceções mistas. Apesar da expectativa inicial de poupança de tempo, apenas 51% dos Clínicos confirmaram essa vantagem após seis meses de uso, sinalizando uma discrepância entre as capacidades percebidas e reais do sistema (Lyons et al., 2016). A aceitação crescente está mais associada à melhoria das interfaces e à redução de erros do que a ganhos de produtividade imediatos.

Em suma, embora promissoras, as ferramentas de conversão de voz em texto ainda enfrentam limitações estruturais que comprometem a sua fiabilidade, especialmente em ambientes clínicos complexos. Superar tais limitações requer avanços técnicos contínuos, melhoria de adaptação contextual e uma integração mais fluida com os sistemas clínicos existentes.

1.5 – Avanços técnicos de ferramentas *speech-to-text*

O reconhecimento automático de voz tem passado por transformações significativas nas últimas décadas, impulsionado por inovações tecnológicas com modelos de redes neuronais profundas, capacidades computacionais e técnicas de pré-treino auto-supervisionado. Estas inovações têm contribuído para melhorias substanciais na precisão, robustez e capacidade de generalização dos sistemas de ASR.

Historicamente, os sistemas ASR tinham como base modelos HMM-GMM, os quais foram progressivamente substituídos por arquiteturas híbridas DNN-HMM (Modelos Ocultos de Markov combinados com Redes Neuronais Profundas), permitindo uma melhor modelação das probabilidades acústicas sem pressupostos rígidos sobre a distribuição dos dados (Li et al., 2025). Por sua vez, estas estruturas deram lugar a abordagens *end-to-end*, como os modelos *Connectionist Temporal Classification* (CTC), *RNN-Transducer* e *attention-based encoder-decoders* (AED), os quais eliminam a necessidade de componentes separados como dicionários de pronúncia e modelos de linguagem (Bell et al., 2021).

Segundo o estudo de Bell et al., 2021, o CTC permite mapear as sequências do áudio de entrada para sequências de saída de menor dimensão (como caracteres ou palavras), introduzindo um símbolo de "em branco" e assumindo independência temporal entre os *frames* do *input*, o que facilita o alinhamento entre o *input* e o *output* através de uma função de perda que considera todas as possíveis segmentações. Já o *RNN-Transducer* expande

esta abordagem ao incluir uma rede preditora que modela dependências linguísticas, combinando-a com o codificador acústico por meio de uma *joint network*, permitindo prever a próxima unidade de *output* com base no áudio e no histórico de símbolos criados previamente. Por sua vez, os modelos AED utilizam um mecanismo *attention-based* para aceder a toda a sequência de representações acústicas codificadas, capaz de oferecer uma maior flexibilidade ao decodificador na criação da sequência textual, embora com maior custo computacional e requisitos de memória. Estes três modelos têm sido amplamente adotados por permitirem uma aprendizagem direta entre o sinal acústico e a transcrição, e por demonstrarem melhor desempenho em múltiplos contextos de aplicação de ASR.

Complementarmente, Bahdanau et al. (2016) introduzem um modelo *end-to-end* com base em *Attention-Based Recurrent Sequence Generator*, que elimina a necessidade de alinhamentos explícitos, automatizando o mapeamento entre *frames* de entrada e caracteres de saída.

Um dos avanços mais marcantes foi a introdução de técnicas de pré-treino auto-supervisionado, nos quais os rótulos não são fornecidos por humanos, mas sim criados automaticamente a partir do próprio dado bruto. Estes modelos alcançaram melhorias significativas no desempenho, especialmente em contextos com poucos recursos linguísticos. Esta capacidade tem sido explorada com sucesso no reconhecimento de voz em português, permitindo melhores resultados mesmo em conjuntos de dados anotados escassos (Li et al., 2025).

Mais recentemente, Modelos Generativos de Linguagem, como o GPT-4, têm sido explorados como ferramentas auxiliares para deteção de erros em transcrições criadas por sistemas ASR. Estes modelos demonstraram elevada precisão na identificação de erros clinicamente significativos em relatórios radiológicos, com *F1-score* de 86,9% para erros críticos e 94,3% para erros não críticos (Schmidt et al., 2024). Estes resultados ilustram o potencial dos *Large Language Models* (LLMs) em contextos sensíveis como a saúde, funcionando como uma camada adicional de validação semântica e contextual.

Zhao et al. (2025) sistematizam os principais avanços nos LLMs, incluindo o seu papel crescente na resolução de tarefas de transcrição e interpretação de linguagem falada, dando destaque à integração dos LLMs como componentes centrais em *pipelines* modernos de ASR, especialmente na correção e validação semântica pós-transcrição.

O artigo de Hu et al. (2024) propõe um novo paradigma de correção de erros em ASR com LLMs multimodais, com a integração direta do sinal de áudio com hipóteses alternativas de transcrição para melhorar a fiabilidade do texto criado. Esta abordagem é

reforçada pelo trabalho de Zhang et al. (2024), que fornece uma visão abrangente sobre modelos multimodais de grande escala (MM-LLMs) e o seu potencial para interpretar, corrigir e criar linguagem com base em entradas complexas como fala, texto e imagem. Também no contexto clínico, Adedeji et al. (2024) demonstram como LLMs podem melhorar significativamente a precisão da transcrição médica, não só na redução do WER, mas também na preservação de conceitos clínicos e na distinção de utilizadores.

Paralelamente, Latif et al. (2021) destacam o papel promissor da tecnologia de voz no setor da saúde, não só para transcrição, mas também para diagnóstico e monitorização, sublinhando que os avanços recentes em DL, combinados com infraestruturas de dados adequadas, podem transformar a aplicabilidade da conversão de voz em texto em contextos clínicos.

A adaptação a variações de fala, incluindo sotaques e características individuais dos utilizadores, é outro eixo de desenvolvimento relevante. Estudos recentes propuseram arquiteturas que combinam características acústicas intra-nativas com informações translinguísticas para melhorar a classificação de sotaques. A integração dessas estratégias em sistemas de ASR resultou numa melhoria significativa na métrica *accuracy*, atingindo até 99% em tarefas específicas (Salifu et al., 2025).

No contexto da diversidade linguística, Paul et al. (2025) demonstram a eficácia da aprendizagem por *Transfer Learning* na construção de um sistema de ASR para Sadri, uma língua pouco documentada, com a utilização de dados de línguas relacionadas para colmatar a escassez de corpora específicos. Também no âmbito da inclusão e acessibilidade, o projecto *CloudCAST*, descrito por Cunningham et al. (2017), apresenta uma plataforma que permite a personalização de modelos de ASR para utilizadores com perturbações da fala, promovendo a adaptabilidade clínica e o treino colaborativo entre Especialistas.

A investigação de Agrawal e Ganapathy (2019) contribui com um método inovador de filtragem espectro-temporal por *Convolutional Variational Autoencoders* (CVAE), que preserva modulações relevantes da voz e melhora a robustez em ambientes ruidosos.

Adicionalmente, técnicas de adaptação com origem em *embeddings* dos oradores, *fine-tuning* e *data augmentation* são fundamentais para ajustar modelos a novos domínios, sotaques e condições acústicas. Com base no artigo de Bell et al., 2021, os *embeddings* dos oradores consistem na utilização de representações vetoriais das características vocais individuais para personalizar o modelo de reconhecimento de voz; o *fine-tuning* envolve o reajuste dos pesos do modelo com dados específicos do domínio ou do utilizador,

permitindo melhorar o desempenho em contextos particulares; o *data augmentation* refere-se à criação de variações artificiais dos dados de treino, como alterações no ruído ou da velocidade da fala, para aumentar a robustez e a generalização do modelo.

O uso de Redes Neuronais Profundas continua a ser o pilar central da inovação técnica em ASR. Trabalhos que exploram melhorias nos Coeficientes Cepstrais (MFCC), integração de vetores de identidade do utilizador (*i-vectors*) e adaptação no espaço de características mostraram reduções substanciais na WER, reforçando a importância de escolhas precisas na engenharia de atributos (Song, 2023).

Jorg et al. (2023) desenvolveram uma ferramenta que integra um sistema de diálogo inteligente com ASR e NLP, que permite converter automaticamente ditados em relatórios estruturados. Este sistema não só preenche automaticamente *templates* clínicos como envia mensagens de *feedback* ao utilizador sempre que dados relevantes estejam em falta. A abordagem inovadora de Jorg et al. (2023) permite combinar os benefícios do ditado livre, como maior fluidez e naturalidade, com as vantagens do relatório estruturado, como a uniformidade e completude.

No domínio da eficiência computacional, Zhang et al. (2021) propõem o *Tiny Transducer*, um modelo extremamente leve, otimizado para dispositivos periféricos, que utiliza técnicas como *blank deweighting* e compressão via SVD, para proporcionar um desempenho competitivo com apenas 0,9 milhões de parâmetros.

Em suma, os avanços técnicos no reconhecimento automático da voz refletem uma convergência entre a inovação algorítmica, aumento da capacidade computacional e integração de modelos de linguagem. A transição de arquiteturas clássicas para modelos *end-to-end*, combinada com técnicas de pré-treino auto-supervisionado e adaptação contextual, permitiu não só uma melhoria na precisão e robustez das transcrições, mas também uma maior aplicabilidade em ambientes clínicos, multilíngues e personalizados. A incorporação de LLMs, quer na validação semântica, quer na correção de erros e integração multimodal, reforça o papel destas ferramentas como elementos centrais em *pipelines* modernos de ASR. À medida que se desenvolvem soluções cada vez mais eficientes, inclusivas e adaptáveis, torna-se evidente que o reconhecimento de voz caminha para uma nova era, mais inteligente, contextual e centrada no utilizador.

CAPÍTULO II – Arquitetura da Solução

O reconhecimento de voz refere-se à tecnologia que permite converter automaticamente comunicação verbal para texto digital, viabilizando a interação com sistemas informáticos através da linguagem falada. Os sistemas de ASR modernos combinam dois elementos principais, o reconhecimento automático de voz e modelos de linguagem. O primeiro componente capta o sinal de áudio através de um microfone e converte-o em texto básico, utilizando modelos acústicos e léxicos que identificam fonemas, palavras e frases. O uso clínico do ASR tem evoluído de forma significativa nas últimas duas décadas.

Compreender a arquitetura tradicional e moderna dos sistemas de reconhecimento de voz, é importante para compreender como funcionam sistemas, como o Invox Medical Dictation. A arquitetura típica de um sistema de ASR, segundo Haboussi et al., 2025, é um processo complexo de reconhecimento de fala descrito como uma *pipeline* modular, isto é, uma forma de organizar um sistema em módulos sequenciais, em que cada módulo executa uma tarefa específica e envia o resultado para o módulo seguinte. No contexto de reconhecimento de voz, o processo de transcrição de fala integra vários componentes autónomos, que cooperam entre si.

Em primeiro lugar é necessária a entrada de áudio (*speech input*), que corresponde ao sinal de voz gravado através de um microfone. O *speech input* pode conter ruído de fundo, sotaques ou hesitações, o que afeta a qualidade do reconhecimento.

Em segundo lugar, é realizada a extração de características (*feature extraction*), que converte o sinal de áudio num conjunto de vetores numéricos representativos, geralmente Coeficientes Cepstrais em Frequência de Mel, que captam as propriedades fonéticas da fala. Os MFCC permitem comprimir a informação acústica de um sinal sonoro num conjunto de vetores numéricos mais fáceis de processar por modelos computacionais. O processo de cálculo dos MFCC envolve várias etapas: inicialmente, o sinal de voz é dividido em pequenos segmentos temporais, aos quais é aplicada uma Transformada de Fourier, que converte o som do domínio temporal para o domínio da frequência. Em seguida, as amplitudes dessas frequências são analisadas por um conjunto de filtros dispostos na escala Mel, uma escala que aproxima a forma como o ouvido humano compreende as frequências sonoras, dando mais peso às frequências médias, onde se concentra a maior parte da informação linguística, como as vogais e consoantes. Os MFCC permitem ignorar ruídos irrelevantes, como tons muito agudos ou graves que não afetam a fala, e, concentram-se nas frequências que são mais utilizadas para falar. Posteriormente, é aplicado o logaritmo da energia de cada banda de frequência, que simula a perceção logarítmica da intensidade

sonora pelo ouvido, e é aplicada uma Transformada Discreta do Cosseno (DCT), que comprime a informação e elimina redundâncias. O resultado é um vetor de coeficientes que descreve a forma do espectro sonoro em termos de frequência percebida, como se tratasse de uma assinatura acústica da fala. Estes coeficientes são altamente eficazes para distinguir diferentes sons da fala, fonemas, mesmo em ambientes com ruído moderado, sendo por isso amplamente utilizados como entrada para modelos acústicos em sistemas ASR, incluindo aplicações médicas como ditados clínicos.

Em terceiro lugar, são necessários modelos acústicos (*acoustic models*), responsáveis por associar os vetores de características a unidades linguísticas, como fonemas ou subpalavras. Tradicionalmente baseiam-se em HMM-GMM, mas atualmente as *Deep Neural Networks* (DNNs), RNNs, CNNs ou *Transformers* têm ganho destaque.

Em quarto lugar, o componente léxico de pronúncia é responsável por estabelecer a correspondência entre a forma ortográfica das palavras e a respetiva representação fonética, por exemplo "colangiopancreatografia" é representado por /ko.lan.dʒi.o.../. Este elemento é particularmente importante para garantir o correto reconhecimento de terminologia médica especializada.

Em quinto lugar, assume particular relevância o Processamento de Linguagem Natural, nomeadamente o Modelo de Linguagem (*language model* – LM), cuja função consiste em interpretar corretamente palavras proferidas de forma rápida ou em corrigir automaticamente eventuais erros de pronúncia. O LM corresponde a um conjunto de algoritmos treinados, que podem basear-se em modelos estatísticos tradicionais (*n-gram*) ou redes neuronais, que analisam a frequência de ocorrência de palavras e auxiliam o sistema na seleção da hipótese mais plausível entre múltiplas possibilidades de transcrição. Os *n-gram* funcionam com base em frequência de ocorrência de palavras em sequência, limitando-se ao contexto de 2, 3 ou 4 palavras anteriores (*2-gram; 3-gram; 4-gram*). Estes modelos, apesar de mais céleres e simples, não têm a perceção do contexto, apenas têm em conta padrões. Atualmente, os modelos mais avançados recorrem a redes neuronais profundas, como RNNs ou *Transformers*, que permitem compreender o contexto semântico alargado de frases clínicas complexas. Neste caso, os modelos conseguem analisar contextos longos e compreender relações semânticas entre palavras. Os modelos neuronais apresentam a vantagem de serem altamente precisos, no entanto são mais pesados computacionalmente.

Em sexto lugar, o decodificador (*decoder*) integra os resultados provenientes dos modelos acústico, linguístico e léxico, de modo a criar a transcrição final. Este componente

é responsável por selecionar a sequência de palavras mais provável entre as várias hipóteses consideradas, constituindo assim a etapa final do processo de reconhecimento, da qual resulta a saída textual (*text output*) apresentada ao utilizador.

O diagrama seguinte representa o fluxo de processamento típico num sistema modular de reconhecimento de fala.

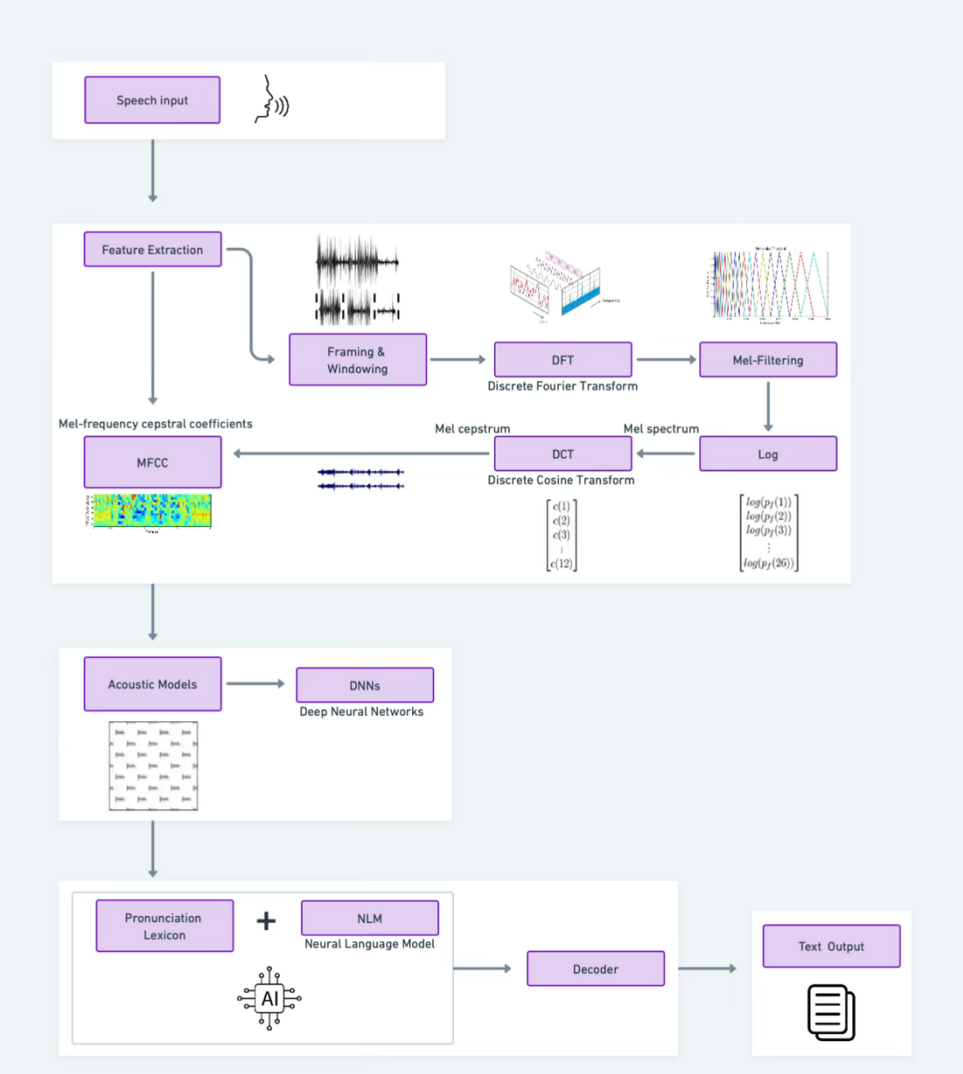

Figura 2.1 – *Pipeline* de um sistema ASR

A arquitetura modular dos sistemas de reconhecimento de voz permite compreender de forma estruturada os fatores que influenciam o seu desempenho em contexto clínico. A qualidade do microfone utilizado, a clareza da articulação do utilizador e a presença de ruído no ambiente têm impacto direto na fiabilidade da transcrição, uma vez que interferem na fase de captação e extração de características acústicas. No caso do Invox Medical Dictation, o reconhecimento de terminologia médica complexa depende, adicionalmente, da capacidade do sistema em utilizar um vocabulário técnico adaptado e de interpretar

corretamente o contexto clínico da frase. É nesta fase que entram os Modelos de Linguagem e os mecanismos de IA, os quais ajudam a prever a sequência mais provável de palavras com base em dados anteriores, aumentando a coerência e a correção dos relatórios criados.

Importa salientar que, nos sistemas de reconhecimento de voz, podemos encontrar *pipelines* modulares, mas também existem arquiteturas *end-to-end*. Nos sistemas modulares tradicionais, o processo é composto pelas várias etapas sequenciais descritas anteriormente, designadamente: a extração de características acústicas, o modelo acústico, o léxico de pronúncia, o modelo de linguagem e o decodificador. Cada um destes componentes é desenvolvido, treinado e ajustado de forma independente. Por sua vez, nas arquiteturas *end-to-end*, todo o processo, desde o sinal de áudio até à transcrição final, é aprendido de forma integrada por uma rede neuronal profunda, dispensando a necessidade de módulos separados e promovendo uma maior otimização global do sistema.

O Invox Medical Dictation é capaz de compreender a linguagem natural e transcrevê-la em tempo real, e disponibiliza dicionários próprios para cada especialidade clínica, o que eleva a sua precisão de reconhecimento e a distingue das demais.

Das principais funcionalidades, destacam-se as substituições, que permitem criar comandos de voz associados a textos pré-definidos, modelos, campos personalizáveis e/ou atalhos do teclado; as transformações, que permitem ditar termos por extenso e transcrever o mesmo como sigla ou vice-versa, por exemplo HISTORIAL MÉDICO → HM; e o dicionário de palavras, que permite acrescentar novos termos que ainda não existam no dicionário do *software*. Estas funcionalidades, aliadas a um reconhecimento de voz com uma elevada taxa de conformidade, tornam o Invox Medical Dictation numa ferramenta que aporta múltiplas vantagens para as instituições, profissionais de saúde e utentes.

Salienta-se, entre muitas vantagens, a otimização de tempo, o aumento da produtividade, e consequente redução de tarefas administrativas, disponibilização de relatórios no momento e redução de custos. À aplicação do *software* acresce uma melhoria da qualidade do relatório, uma vez que a ferramenta compreende a linguagem natural e processa a informação a nível semântico, garantindo elevados padrões de qualidade do relatório clínico.

Como qualquer *software*, o Invox Medical Dictation estabelece os seus requisitos para implementação. A utilização da ferramenta requer que os postos de trabalho da unidade de saúde cumpram os seguintes requisitos técnicos mínimos: sistema operativo Windows 7 ou superior (64-bit), um processador Intel Core i3-4130 (ou equivalente AMD) com dois núcleos e quatro threads, 6 GB de memória RAM, 100 GB de espaço disponível

em disco, .NET Framework 4.6.2 ou superior, e compatibilidade com navegadores modernos como Edge, Chrome ou Firefox. A instalação da aplicação em cada posto de trabalho é um processo expedito com duração estimada de cerca de dois minutos por terminal. A tecnologia subjacente ao Invox Medical Dictation baseia-se em reconhecimento de voz com recurso a NLP e a modelos de IA com *deep learning*, que permitem ao sistema adaptar-se progressivamente ao estilo de ditado de cada utilizador.

O *software* já se encontra implementado em diversas instituições do setor público (ULS São José, ULS São João, ULS Braga, ULS Algarve, ULS Castelo Branco, ULS Médio Tejo) e privado (Joaquim Chaves Saúde, Germano de Sousa e LifePlus) nas especialidades de Radiologia, Anatomia Patológica, Gastrenterologia e Medicina Nuclear. Adicionalmente, a solução encontra-se em avaliação em outras instituições de saúde do setor público.

CAPÍTULO III – Metodologia

A metodologia adotada nesta investigação assumiu uma natureza qualitativa, orientada para a compreensão aprofundada da experiência dos utilizadores com o *software* em estudo. Para a recolha de dados foram implementados inquéritos por questionário (Anexo A) e entrevistas semiestruturadas (Anexo B), instrumentos que permitiram recolher testemunhos diretos dos profissionais e analisar as suas perceções relativamente ao impacto da tecnologia ao nível da eficiência, da satisfação e da fiabilidade. Esta abordagem possibilitou a obtenção de informação contextualizada, permitindo identificar vantagens, limitações e oportunidades de melhoria associadas à implementação do sistema.

Os inquéritos por questionário serviram como fonte de dados quantitativos, e foram desenhados para quantificar perceções qualitativas dos colaboradores. Com o intuito de reduzir vieses de respostas fornecidas pelos utilizadores abordados, foi explicado cautelosamente o objetivo do questionário, reforçando o anonimato e a rigorosidade nas respostas às questões colocadas. Os dados obtidos foram analisados através de estatística descritiva, com base em escalas ordinalizadas (1 a 5, 1 a 4 ou 1 a 3, dependendo da questão), que seguem o princípio de gradação, permitindo atribuir valores numéricos e calcular médias. Aliado a isso, foram estabelecidas diversas relações com as variáveis em estudo.

As entrevistas foram conduzidas com representantes da empresa *Quattro*, bem como com profissionais de saúde utilizadores do *software*, com o objetivo de compreender as perceções, experiências e desafios associados à implementação da tecnologia.

Para apreciação do potencial da ferramenta foi desenvolvido um estudo de caso na especialidade de Anatomia Patológica do Hospital Curry Cabral (ULS São José), local em que o *software* já se encontra implementado desde 2023, permitindo uma análise contextualizada dos efeitos da solução tecnológica no processo de emissão de relatórios clínicos.

CAPÍTULO IV – Estudo de Caso ULS São José

O Serviço de Anatomia Patológica do Hospital Curry Cabral, enquanto centro de referência para análises hepatobiliopancreáticas e colorretais, representa um contexto particularmente exigente no que respeita à precisão e celeridade dos relatórios clínicos. A implementação do *software* Invox Medical Dictation neste serviço trouxe vantagens substanciais que se refletem tanto na qualidade técnica da documentação como na eficiência organizacional. O *software* Invox constitui uma ferramenta inovadora para a otimização do fluxo de trabalho no laboratório de Anatomia Patológica, ao permitir a transcrição automática, em tempo real, dos relatórios clínicos ditados.

Tradicionalmente, no serviço de Anatomia Patológica, cada análise implica a elaboração de dois relatórios distintos: um macroscópico e outro microscópico. Este processo era historicamente moroso, exigindo múltiplos passos e a intervenção de diferentes categorias profissionais, nomeadamente Técnicos Superiores de Diagnóstico e Terapêutica, Médicos Patologistas e uma Administrativa.

O processo era iniciado com a receção da peça cirúrgica pela Administrativa, que procedia ao registo no sistema. Seguidamente, o TSDT iniciava o relatório macroscópico, ditando-o durante a dissecação da peça. O áudio era então transcrito manualmente pelo Administrativa, que desempenhava funções de datilógrafa, convertendo a gravação em texto no sistema informático e realizando a sua impressão. Para garantir a fiabilidade do registo, o relatório era devolvido ao TSDT, que verificava a transcrição e, em caso de incorreções, procedia à respetiva correção e nova impressão.

Concluída esta fase, o Médico responsável iniciava a elaboração do relatório microscópico, redigindo-o manualmente durante a observação das lâminas. A natureza manuscrita deste documento constituía um fator de risco para a segurança do utente, uma vez que a caligrafia nem sempre era facilmente legível. Este relatório manuscrito era, posteriormente, entregue à Administrativa, que o transcrevia para o sistema informático. Tal como sucedia no relatório macroscópico, o documento era devolvido ao Médico para revisão e correção, antes de seguir para o Secretariado, que procedia à introdução dos códigos de faturação. Só após este ciclo de múltiplas etapas e sucessivos retornos entre profissionais, o relatório ficava disponível no processo clínico eletrónico do utente (SClínico).

Assim, o método convencional envolvia ditado, transcrição manual por parte da Administrativa, revisões sucessivas, impressões e correções, tanto no relatório macroscópico ditado pelo TSDT como no relatório microscópico manuscrito pelo Médico.

Acresce ainda o facto de a Administrativa acumular várias funções simultâneas (receção de peças cirúrgicas, atendimento telefónico, transcrição de relatórios e gestão de e-mails), cenário que favorecia a sobrecarga laboral e aumentava a probabilidade de erros de transcrição.

Com a introdução do sistema STT, o processo tornou-se substancialmente mais ágil e eficiente, passando de nove para apenas cinco etapas principais. Atualmente, após a receção das peças cirúrgicas pela Administrativa, o relatório macroscópico é elaborado pelos TSDT através da análise da peça em simultâneo com o ditado diretamente no Invox Medical Dictation, sendo transcrito automaticamente pelo *software* e, corrigido no momento, pelo Técnico, caso necessário. De igual modo, o relatório microscópico é ditado diretamente pelo Médico no Invox durante a observação das lâminas, ficando também registado em tempo real no sistema. Seguidamente, a Administrativa procede apenas à formatação de ambos os relatórios, à correção pontual de aspetos formais de acordo com as normas instituídas pela ULS de São José, e à inserção dos códigos de faturação. Por fim, o Médico valida e assina a análise, que fica imediatamente disponível no processo clínico eletrónico do utente (SClínico).

Neste contexto, o ditado com transcrição em tempo real revelou-se mais célere e eficiente do que o processo manual de escrita, apresentando-se como uma ferramenta essencial para reduzir tempos de resposta e eliminar atrasos nas várias etapas de elaboração dos relatórios. Paralelamente, a utilização do *software* contribui para a produção de registos clínicos mais precisos e completos, diminuindo de forma significativa a probabilidade de erros de transcrição e reforçando, consequentemente, a qualidade e segurança dos cuidados de saúde prestados. Adicionalmente, a sua utilização caracteriza-se pela simplicidade e intuitividade.

A implementação do *software* trouxe melhorias significativas na qualidade e especificidade dos relatórios, uma vez que estes podem ser elaborados em simultâneo com a análise laboratorial, reduzindo o risco de omissões ou imprecisões decorrentes de registos diferidos.

De momento existem dificuldades associadas à compatibilidade do Invox com o *software* ANAPAT utilizado pela Anatomia Patológica, conforme descreverei de forma mais detalhada na análise qualitativa com base nas entrevistas realizadas. Caso a interoperabilidade entre os sistemas não fosse um *handicap*, bastava ao utilizador posicionar o cursor no campo da ficha do utente onde pretende desenvolver o relatório, garantindo assim integração fluida no sistema informático existente.

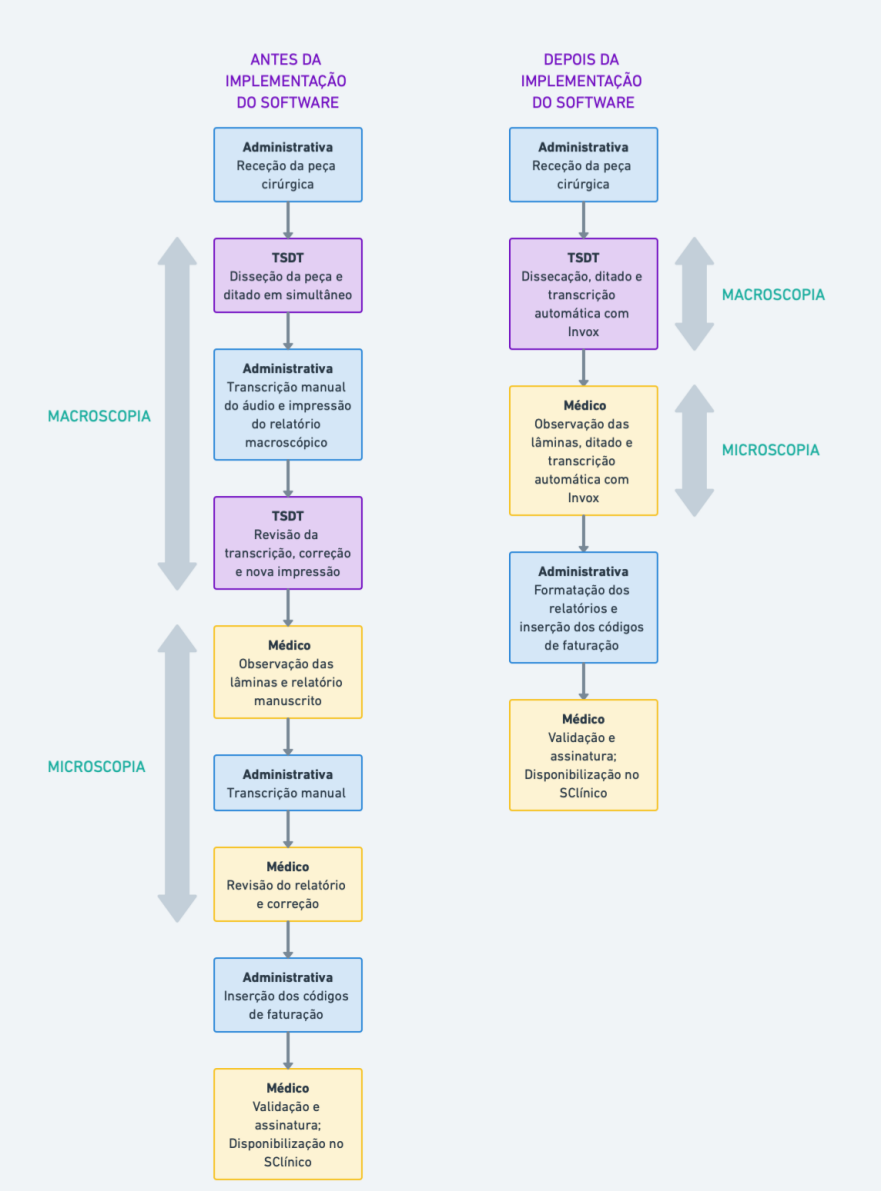

Figura 4.1 – Fluxograma que apresenta, de forma comparativa, as etapas do processo antes e depois da implementação do *software* de reconhecimento de voz.

Uma das características distintivas da sala de Macroscopia é o ruído ambiente constante, resultante do funcionamento da câmara de extração de formol. Apesar deste desafio, o *software* demonstrou elevada robustez, conseguindo filtrar o ruído de fundo e captar exclusivamente a voz do utilizador, assegurando a qualidade da transcrição.

A este desempenho acresce o facto de a ferramenta estar equipada com um dicionário específico da especialidade, o que permite a correta transcrição de terminologia altamente especializada, incluindo estrangeirismos e epónimos frequentemente utilizados

em Anatomia Patológica. Tal funcionalidade representa uma melhoria significativa face ao cenário anterior à implementação, em que erros ortográficos em estruturas anatómicas e doenças específicas, como "*Wirsung*" ou "*Síndrome de Sjögren*", eram recorrentes. Para além disso, possibilita a introdução de termos e expressões adicionais, adaptados ao léxico próprio do Serviço de Anatomia, o que assegura maior flexibilidade e adequação ao contexto profissional.

Outro contributo fulcral reside na transcrição em tempo real: o texto ditado surge transcrito no relatório em apenas dois a três segundos, o que possibilita que a análise seja finalizada e disponibilizada quase de imediato aos clínicos. Esta celeridade elimina diversos passos burocráticos, reduz o número de intervenientes no processo e assegura uma maior fluidez no fluxo de trabalho. Como consequência, os relatórios são entregues de forma mais célere, potenciando consultas médicas mais eficazes, orientações terapêuticas mais atempadas e contribuindo, de forma indireta, para a redução das listas de espera.

Do ponto de vista da proteção de dados, o *software* acrescenta ainda uma camada de segurança. Quando o problema de interoperabilidade entre o *software* Invox e o *software* ANAPAT for solucionado, o TSDT conseguirá aceder à ficha do utente através da leitura do código de barras da requisição e ditar diretamente no campo correspondente ao relatório. Assim, mesmo que ocorra acesso indevido a registos de áudio, estes não podem ser associados a um utente específico, reforçando a confidencialidade da informação clínica. O mesmo se sucede na análise microscópica.

Por fim, destaca-se também a contribuição do Invox para a sustentabilidade. Ao reduzir drasticamente a necessidade de múltiplas impressões ao longo do processo, sendo atualmente necessária apenas uma impressão final. O *software* promove uma prática mais ecológica e alinhada com as diretrizes de digitalização e desmaterialização documental nos serviços de saúde.

Em suma, a adoção do Invox Medical Dictation no Hospital Curry Cabral traduziu-se numa transformação significativa do processo anátomo-patológico, tornando-o mais digital, fluido e eficiente, garantindo diagnósticos de elevada qualidade técnica, disponibilizados atempadamente aos clínicos, com impacto direto na segurança e no bem-estar dos utentes. Comparando o processo pré e pós-implementação, torna-se evidente a magnitude do impacto do Invox. Anteriormente, a elaboração dos relatórios implicava múltiplas etapas, sucessivas transcrições manuais, revisões e impressões, consumindo tempo significativo de profissionais de diferentes categorias e aumentando a probabilidade de erros. Com a adoção do *software*, este fluxo foi simplificado, reduzindo etapas

redundantes, eliminando falhas decorrentes da escrita manual ou da transcrição indireta e assegurando a disponibilização quase imediata dos resultados. Assim, o Serviço passou de um modelo moroso e burocrático para um processo digital, ágil e fiável, reforçando simultaneamente a precisão diagnóstica, a segurança do doente e a eficiência organizacional.

Para avaliar qualitativamente a ferramenta *speech-to-text*, foram realizados 7 inquéritos e 8 entrevistas a colaboradores (Anexo C), seguem-se os resultados obtidos.

A análise das entrevistas realizadas com Técnicos, Médicos e Administrativa do Serviço de Anatomia Patológica do Hospital Curry Cabral permite evidenciar uma experiência globalmente positiva com a utilização do Invox, ainda que persistam constrangimentos técnicos relevantes. Os relatos apontam para ganhos expressivos em termos de celeridade na elaboração dos relatórios, redução da sobrecarga administrativa, aumento da completude e legibilidade dos registos clínicos e melhoria na fluidez do trabalho em equipa. Contudo, a incompatibilidade do *software* com o sistema informático ANAPAT e a variabilidade da experiência entre diferentes perfis de utilizadores constituem limitações que condicionam o aproveitamento pleno do potencial da ferramenta.

De forma transversal, os Técnicos destacaram a transformação no modo como os relatórios são produzidos. Antes da implementação do *software*, o processo era fragmentado, exigia pausas constantes, gravações manuais suscetíveis de falhas técnicas e dependia de várias interações com a Administrativa para correções. Com o Invox, este fluxo tornou-se mais direto e contínuo, permitindo ditar diretamente durante a dissecação das peças, evitando pausas forçadas no raciocínio, erros associados à gravação manual ou interrupções para esclarecimento junto da Administrativa. Um TSDT referiu que "atualmente enquanto disseco a peça, estou a ditar o que estou a ver e demoro entre 40 a 45 segundos a elaborar o relatório", o que traduz uma aceleração evidente do processo. Outro testemunho reforça que "antes tínhamos de estar sempre a pausar e ativar a gravação, e agora é um processo muito mais contínuo". Esta fluidez reduziu significativamente o tempo de resposta entre a análise e a disponibilização do relatório, permitindo aos clínicos acederem mais rapidamente à informação necessária para a tomada de decisão.

Os Médicos confirmaram igualmente esta perceção. Um deles assinalou que, ao finalizar o ditado, "num espaço de segundos fica fechado, ou seja, fica elaborado e disponível em tempo real". Outro destacou "um grande impacto na celeridade dos relatórios", permitindo analisar um maior número de casos no mesmo intervalo de tempo.

A celeridade foi assim identificada como um dos ganhos mais significativos da introdução do *software*, com efeitos diretos na eficiência global do Serviço.

A possibilidade de visualizar a transcrição em tempo real teve impacto positivo na qualidade e completude dos relatórios. Para os Técnicos, o facto de o texto estar imediatamente disponível permite rever se todos os elementos necessários foram incluídos, reduzindo omissões involuntárias. Um deles reconheceu que, ao visualizar a transcrição, "vejo se abordei todos os pontos essenciais para o diagnóstico".

Também os Médicos sentiram que a ferramenta promove relatórios mais completos. Casos específicos, como o da Patologista que relatou dificuldades crescentes em escrever manualmente devido a artroses, demonstram o contributo direto da ferramenta para a continuidade e qualidade do trabalho clínico. O impacto revelou-se também geracional: médicos mais experientes valorizaram a eliminação das dificuldades da caligrafia e da dependência de terceiros na transcrição, enquanto alguns médicos mais jovens, com maior familiaridade com o uso direto do computador, percecionaram o ganho como menos transformador, embora igualmente vantajoso.

Por outro lado, nem todos os entrevistados percecionaram melhorias significativas na qualidade intrínseca da informação, defendendo que a diferença se situa sobretudo na fluidez e rapidez do processo. Ainda assim, é consensual que a redução de erros de transcrição e de falhas decorrentes de caligrafia ilegível ou ditados mal compreendidos resultou numa documentação mais clara e fidedigna.

A implementação do Invox teve também um efeito claro na redistribuição de tarefas e na diminuição do stress laboral associado ao trabalho administrativo. A Administrativa entrevistada foi explícita ao afirmar que "a minha carga de trabalho é bastante mais reduzida. Tenho mais tempo para utilizar noutras funções que não sejam a transcrição dos relatórios". O desaparecimento da necessidade de descodificar letras ilegíveis ou ditados pouco percetíveis, muitas vezes interrompidos por chamadas telefónicas ou outras tarefas paralelas, contribuiu para uma gestão mais equilibrada das funções administrativas. Esta mudança foi também valorizada pelos Técnicos e Médicos, que reconheceram que a ferramenta libertou a Administrativa de uma das tarefas mais demoradas e suscetíveis a erros.

Um dos constrangimentos mais salientados pelos utilizadores prende-se com a incompatibilidade entre o Invox Medical Dictation e o sistema informático ANAPAT, utilizado no Serviço de Anatomia Patológica para registo dos relatórios. Embora o *software* de transcrição apresente bom desempenho em termos de reconhecimento de voz e criação

do texto, a ausência de interoperabilidade entre as duas plataformas compromete a sua eficácia plena.

Na prática, o processo deveria decorrer da seguinte forma: após o ditado e eventual correção da transcrição no Invox, o utilizador deveria confirmar "Aceitar texto", permitindo que a versão final fosse automaticamente transferida para o ANAPAT. Este passo teria uma dupla função: integrar diretamente o relatório no sistema hospitalar e fornecer ao Invox um mecanismo de aprendizagem contínua, assimilando as correções efetuadas pelo utilizador, adaptando-se progressivamente à sua dicção, estilo de linguagem e terminologia médica.

Contudo, esta integração não ocorre de forma funcional. Os profissionais relatam que, sempre que tentam aceitar o texto, o ANAPAT entra em sobrecarga, permanecendo em processamento durante longos períodos. Frequentemente, esse processo culmina na perda total da transcrição corrigida, obrigando o utilizador a repetir o trabalho de correção desde o início. Para evitar esta duplicação de tarefas e a perda de informação, a solução encontrada foi realizar manualmente um *copy-paste* do texto criado pelo Invox para o ANAPAT. Esta alternativa, ainda que funcional, gera um efeito colateral relevante: o *software* de transcrição não aprende com as correções introduzidas, mantendo os erros recorrentes, como interpretações incorretas de medidas, de expressões negativas ou de abreviaturas técnicas.

Deste modo, a ausência de comunicação eficaz entre o Invox e o ANAPAT cria um ciclo paradoxal: embora o *software* reduza etapas burocráticas e aumente a celeridade do processo, a falta de integração digital perpetua ineficiências e compromete a evolução do sistema, impedindo que este atinja o seu potencial máximo enquanto tecnologia adaptativa. A interoperabilidade entre plataformas hospitalares e soluções de *speech-to-text* emerge, assim, como condição essencial para a consolidação da transformação digital na Anatomia Patológica.

Entre os erros mais relatados encontram-se alterações indevidas em medidas ("23x14x5" interpretado como texto corrido, transcrevendo o que deveria ser "x" como "por"), adição incorreta de algarismos ("50 gramas" transcrito como "150 gramas") ou interpretações erradas de expressões negativas ("não se identificam lesões" transformado em "identifica-se"). Apesar destas experiências, os utilizadores consideram que estes erros são facilmente e rapidamente corrigidos.

A variabilidade da experiência entre utilizadores também foi notória. Um Médico Interno salientou que o *software* "não me percebe bem, penso que seja pelo meu sotaque brasileiro, tenho de fazer bastantes correções". A presença de sotaques distintos, bem como

o ruído ambiental em gabinetes partilhados, constituíram fatores adicionais de variabilidade na fiabilidade da transcrição. Ainda assim, em contextos controlados e com perfis vocais mais bem reconhecidos, os utilizadores relataram níveis de precisão próximos dos 90%.

Em termos de adaptação, a maioria dos entrevistados considerou o *software* intuitivo e de fácil utilização, mesmo para profissionais com menor literacia digital. As formações iniciais foram descritas como breves e suficientes para garantir a adoção. Ainda assim, alguns reconheceram resistência entre colaboradores menos habituados à digitalização, apontando que a aceitação foi mais imediata em gerações mais jovens e mais familiarizadas com tecnologia.

As entrevistas também forneceram contributos relevantes quanto a potenciais evoluções do *software*. Entre as sugestões mais recorrentes destacaram-se a necessidade de uma melhor interoperabilidade com o sistema ANAPAT, a possibilidade de distinguir vozes em ditados simultâneos e a criação de textos pré-definidos para relatórios repetitivos. Um Médico sugeriu precisamente que seria útil ter um comando no *software* para "texto pré-definido sobre gastrite crónica", com o qual automaticamente o Invox ia procurar essa redação e o Médico apenas alterava a localização da lesão e intensidade, o que facilitaria a fluidez do processo.

Em síntese, a experiência qualitativa dos utilizadores demonstra que o Invox Medical Dictation produziu uma transformação significativa no fluxo de trabalho em Anatomia Patológica. A ferramenta permitiu reduzir burocracias, acelerar processos e reforçar a clareza dos relatórios, enquanto contribuiu para diminuir a sobrecarga administrativa e melhorar a distribuição de tarefas. Persistem, contudo, barreiras tecnológicas relacionadas com a interoperabilidade com sistemas hospitalares, a adaptação a sotaques e ambientes ruidosos, e a ausência de certas funcionalidades de apoio, aspetos cuja otimização poderá maximizar ainda mais o potencial da solução. Globalmente, os testemunhos revelam entusiasmo e reconhecimento pelo contributo do *software* para a modernização e eficiência do serviço, embora também ressalvem a necessidade de contínua evolução e adaptação tecnológica.

A análise dos inquéritos realizados a Técnicos, Médicos e à Administrativa do Serviço de Anatomia Patológica do Hospital Curry Cabral permite constatar que a experiência de utilização do Invox Medical Dictation é globalmente positiva, embora não isenta de limitações relevantes (Anexo E). Os dados quantitativos revelam uma perceção clara de ganhos de eficiência: 57% dos inquiridos afirmaram "concordar totalmente" que o tempo necessário para a elaboração dos relatórios diminuiu, 14% "concordaram" e 29%

mantiveram-se neutros, o que corresponde a uma média de 4,3 numa escala de 1 a 5. De forma consistente, 71% reconheceram que o *software* contribui para a celeridade da emissão dos relatórios, com a média de 4,1/5, confirmando o impacto da ferramenta na aceleração dos processos internos. Este resultado está em consonância com os testemunhos qualitativos, nos quais os Técnicos descreveram a possibilidade de ditar e dissecar em simultâneo, reduzindo o tempo médio de elaboração de relatórios para menos de um minuto, e os Médicos salientaram que os relatórios ficam disponíveis "num espaço de segundos".

No que diz respeito à facilidade de utilização, 43% responderam "concordo", 43% "concordo totalmente" e 14% neutro, resultando numa média de 4,3/5. A avaliação da interface revelou também resultados positivos: 86% dos colaboradores consideraram-na adequada ("sim") e apenas 14% responderam "parcialmente", o que corresponde a uma média de 2,9 numa escala de 1 a 3. Esta perceção reforça a ideia, presente nas entrevistas, de que o *software* é intuitivo mesmo para profissionais com menor literacia digital, embora alguns colaboradores tenham reconhecido dificuldades iniciais de adaptação.

Relativamente à precisão da transcrição, todos os inquiridos concordaram que o *software* apresenta elevada fiabilidade na reprodução de terminologia médica, correspondendo a uma média de 4,3/5. Contudo, a confiança global na transcrição revelou-se moderada, com 43% a afirmar confiar parcialmente, 43% a confiar plenamente e 14% a não confiar, resultando numa média de 2,3/3. Este aparente paradoxo entre precisão técnica e confiança subjetiva encontra explicação nas entrevistas, onde se destacaram erros pontuais, mas significativos, como a adição de algarismos nas medidas, a interpretação incorreta de expressões negativas ou a dificuldade em compreender determinados sotaques. Ainda assim, a maioria dos utilizadores referiu que tais erros não comprometem a utilização diária do *software*, sobretudo porque podem ser corrigidos em tempo real.

A satisfação geral com o desempenho do Invox é elevada: 43% dos respondentes declararam-se "muito satisfeitos", 43% "satisfeitos" e apenas 14% "indiferentes", com uma média de 3,3 numa escala de 1 a 4. De forma unânime, todos os inquiridos afirmaram que recomendariam a utilização da ferramenta a outros profissionais, o que reforça a perceção de valor acrescentado do *software*. No entanto, a formação inicial, embora considerada suficiente pela maioria, não foi universal, com 57% a indicar ter recebido formação formal e os restantes a referirem orientações parciais ou ausência de treino estruturado (média de 2,3/3).

A principal fragilidade identificada tanto nos inquéritos como nas entrevistas prende-se com a integração do Invox com o sistema ANAPAT. Todos os participantes discordaram da afirmação de que o *software* se encontra bem integrado com o sistema informático, com uma média de apenas 1,7/5. Os relatos qualitativos confirmam esta limitação, explicando que, ao invés de permitir a validação e a aprendizagem contínua do *software* através da aceitação das correções, os utilizadores têm de recorrer a procedimentos manuais de *copy-paste*. Tal situação impede que o Invox assimile erros corrigidos, perpetuando inconsistências na transcrição e aumentando a carga de trabalho em casos específicos.

Em síntese, os dados recolhidos demonstram que a ferramenta *speech-to-text* teve um impacto significativo na otimização dos fluxos de trabalho do Serviço, reduzindo tempos de resposta e aumentando a eficiência global. O *software* foi bem aceite pelos profissionais, independentemente da sua área de atuação. Todavia, a confiança plena na ferramenta continua a ser condicionada por erros residuais e, sobretudo, pela ausência de interoperabilidade com o sistema ANAPAT, apontada unanimemente como a principal barreira à sua utilização plena.

CAPÍTULO V – A Perspetiva da Empresa Fornecedora: O Caso da Quattro

A Quattro é uma empresa portuguesa resultante da fusão de duas organizações com mais de 30 anos de experiência no fornecimento de produtos e serviços de tecnologias de informação, Bravantic e PaIdata, tendo sido criada com o propósito de promover a transformação digital do setor da saúde em Portugal. A sua missão centra-se em oferecer soluções de valor acrescentado capazes de responder aos desafios emergentes no contexto hospitalar, com foco em inovação, segurança, infraestrutura robusta e serviço de excelência.

A empresa destaca-se pela sua atuação na interconectividade dos sistemas de informação de saúde, sendo reconhecida como uma solução de interoperabilidade adotada por diversos grupos hospitalares em Portugal, com milhões de registos clínicos integrados. Além da interoperabilidade, a Quattro oferece serviços de consultoria tecnológica e de negócio, incluindo automação de processos, sistemas de gestão hospitalar, sistemas laborais, suporte à transformação digital, e soluções como telemedicina, IoT e robótica, reforçando a sua posição como um parceiro estratégico integral na modernização dos serviços de saúde.

Neste contexto, a colaboração na implementação do Invox Medical Dictation reflete o alinhamento da Quattro com a sua missão de promover soluções inovadoras que potenciem ganhos de eficiência, reduzam burocracias e contribuam para a segurança e fiabilidade da documentação clínica, aspetos centrais do estudo de caso desenvolvido nesta investigação.

No âmbito desta investigação, foi realizada uma entrevista a duas colaboradoras da empresa Quattro (anexo D), com o objetivo de compreender a perspetiva da entidade fornecedora relativamente à solução em estudo e à sua implementação no contexto clínico. Esta entrevista constituiu um contributo complementar para a análise do caso, permitindo integrar a visão da empresa responsável pela comercialização e suporte da tecnologia.

As entrevistadas explicaram que a parceria da Quattro com o Invox Medical Dictation surgiu naturalmente, uma vez que, “quando integrámos a Quattro esta solução já estava identificada, contudo dado o interesse e a diferenciação da solução perante outras de mercado e a recetividade de potenciais utilizadores, fez todo o sentido a relação de parceria”. Entre os principais diferenciais competitivos da ferramenta, foi destacada a existência de “dicionários especificamente afetos a cada especialidade médica com léxico complexo e nomenclatura própria”, o que assegura maior precisão no reconhecimento de termos técnicos. Foi ainda sublinhada a flexibilidade do modelo de licenciamento, que pode

ser nominal (licenças para até 20 utilizadores diferentes) ou concorrencial (licenças ilimitadas, mas apenas 20 utilizadores podem aceder ao *software* em simultâneo), permitindo que a solução seja utilizada tanto em clínicas de pequena dimensão como em grandes instituições hospitalares.

No que respeita ao *feedback* dos clientes, a empresa relatou uma receção globalmente positiva em várias especialidades médicas, com especial destaque para a Radiologia, onde os profissionais reconhecem uma precisão superior da transcrição face a outras soluções do mercado. Quanto aos processos de implementação, as colaboradoras reforçaram que estes decorrem de forma geralmente simples, embora a infraestrutura tecnológica de cada instituição possa constituir um desafio. No caso concreto da ULS de São José, "a implementação correu bastante bem, sempre em coordenação com o departamento de informática da instituição".

Outro ponto abordado foi a formação e acompanhamento disponibilizados pela Quattro. Segundo as entrevistadas, a empresa assegura sessões de formação a todos os utilizadores aquando da implementação, incluindo demonstrações práticas e esclarecimento de dúvidas, mantendo posteriormente canais de contacto permanentes para apoio técnico. Esta proximidade é considerada essencial para lidar com a resistência natural que alguns profissionais de saúde apresentam perante novas tecnologias. Como afirmaram, "procuramos sempre ter em portfólio soluções intuitivas, fáceis de utilizar e que a curto prazo se demonstrem capazes de otimizar processos", sendo fundamental "promover relações de proximidade com os profissionais de saúde para que possam confiar na solução e na nossa empresa".

Relativamente à interoperabilidade, questão levantada pelos utilizadores do caso em estudo, a Quattro esclareceu que o Invox é compatível com qualquer *software* hospitalar de inserção de texto, ainda que, na ausência de integração nativa, recorra a uma janela auxiliar de transcrição. Contudo, salientaram que existe a possibilidade de integrar o Invox Medical Dictation com sistemas como RIS ou LIS, o que abre perspetivas de maior integração futura.

Apesar do enquadramento positivo apresentado pela Quattro, a análise cruzada com os testemunhos dos utilizadores revela algumas divergências entre a perspetiva institucional e a experiência prática no terreno. Enquanto a empresa enfatiza a simplicidade da implementação e a compatibilidade do *software* com os sistemas hospitalares, os profissionais destacaram, de forma unânime, a ausência de interoperabilidade plena com o ANAPAT como a principal limitação do Invox Medical Dictation. Esta discrepância sugere

que, embora a solução seja tecnicamente versátil e disponha de mecanismos alternativos, como a utilização de janelas auxiliares, a falta de integração nativa compromete a perceção de eficiência por parte dos utilizadores e perpetua erros recorrentes. Assim, torna-se evidente que a evolução futura da ferramenta deverá não apenas reforçar as características técnicas já reconhecidas como diferenciadoras pela Quattro, como a existência de dicionários especializados e a flexibilidade de licenciamento, mas também responder de forma concreta às necessidades práticas expressas pelos profissionais de saúde, sobretudo no domínio da interoperabilidade e da aprendizagem contínua do sistema.

Por fim, relativamente à evolução da solução, as entrevistadas destacaram que esta faz parte de um portefólio mais vasto de tecnologias da VÓCALI, que incluem já sistemas mais avançados como o Invox Medical Genesis, um progresso natural do Dictation, concebido para otimizar especificamente o registo de informação em contexto de consulta médica. Esta solução incorpora módulos de Inteligência Artificial capazes de analisar em tempo real a interação verbal entre médico e utente, identificar automaticamente os elementos relevantes da anamnese e do exame objetivo e, com base numa estrutura de relatório previamente configurada, extrair e organizar a informação clínica no processo eletrónico do doente. Trata-se, assim, de um sistema que vai além da simples transcrição do ditado, assumindo uma dimensão interpretativa e de estruturação semântica, aproximando-se de um modelo de apoio à decisão clínica e reforçando a tendência de evolução das tecnologias *speech-to-text* para soluções integradas de processamento inteligente da linguagem natural em saúde.

A comparação entre ambos os *softwares* evidencia um claro salto qualitativo em termos de inovação: o Dictation contribui para a eficiência da documentação através da transcrição imediata, mas permanece dependente da revisão e estruturação pelo utilizador; já o Genesis introduz um patamar de automação e análise semântica, em que a IA não apenas transcreve, mas também interpreta e organiza o conteúdo clínico. Esta evolução posiciona o Genesis como uma solução mais avançada, com potencial para reduzir ainda mais a carga administrativa, aumentar a padronização dos relatórios e reforçar a qualidade dos registos clínicos.

Em suma, a entrevista às colaboradoras da Quattro permitiu compreender que a empresa reconhece no Invox Medical Dictation uma solução diferenciadora e adaptável, destacando a precisão técnica, a flexibilidade de licenciamento, a simplicidade de implementação e o apoio próximo aos profissionais como fatores críticos para o sucesso da sua adoção no contexto da saúde.

CAPÍTULO VI – Discussão dos Resultados e Potencial de Aprofundamento

No contexto clínico, onde a precisão da documentação é crítica, persistem desafios significativos relacionados com a variabilidade do discurso médico, a escassez de dados anotados e as diferenças nos estilos de ditado entre profissionais. Para mitigar esses desafios e potenciar a eficácia de soluções como o Invox Medical Dictation, diversas inovações tecnológicas têm vindo a ser exploradas.

No Capítulo II foi explorada a arquitetura do *software* Invox, especificando cada componente e o seu estado de arte. Após a análise dos resultados obtidos, é relevante falar em evoluções que garantam uma maior robustez a hesitações, repetições e variações no ritmo de fala, características típicas do discurso clínico.

A empresa VÓCALI avaliou a solução com recurso à métrica WER, obtendo um valor aproximado de 5% nos dicionários de todas as especialidades disponíveis. Importa, contudo, salientar que estes testes foram realizados em ambiente controlado, com utilizador de dicção clara, microfone compatível e infraestrutura otimizada, assegurando todas as condições técnicas necessárias para maximizar o desempenho do sistema. Tal resultado traduz-se num indicador de elevada precisão, mas não é totalmente representativo da realidade hospitalar. De facto, no estudo de caso desenvolvido no Serviço de Anatomia Patológica do Hospital Curry Cabral, os profissionais relataram uma experiência globalmente positiva quanto à precisão, mas identificaram erros recorrentes. Estes testemunhos confirmam que, fora de condições laboratoriais, a taxa de erro tende a aumentar, sobretudo em ambientes com ruído de fundo ou variabilidade de estilos de ditado. Assim, pode concluir-se que o desempenho observado em contexto real é coerente com o potencial demonstrado pela métrica WER, mas condicionado por fatores ambientais e humanos, reforçando a necessidade de evolução tecnológica em áreas como modelos *attention-based encoder-decoder* (AED), *feedback* com base em NLP**,** *fine-tuning* para o português europeu, técnicas de *data augmentation*, pré-treino auto-supervisionado e personalização por sotaque.

6.1 – Modelos Attention-Based Encoder Decoder

Os modelos *attention-based encoder-decoder* constituem uma das abordagens mais avançadas e eficazes na arquitetura de sistemas *end-to-end* de reconhecimento automático de voz. Este paradigma propõe-se a modelar diretamente a relação entre a sequência de entrada (sinal acústico) e a sequência de saída (texto transcrito), sem a necessidade de componentes intermediários como modelos acústicos ou léxicos explicados anteriormente.

No domínio da documentação clínica, os modelos AED mostram-se especialmente úteis pela sua capacidade de lidar com frases longas, complexas e com estrutura irregular, e pela sua robustez perante variações individuais no estilo de ditado, pausas, reformulações e uso de siglas médicas.

Neste sentido, segundo Watanabe et al., 2017 a arquitetura AED é composta por diversos passos. Em primeiro lugar é recebida uma sequência de entrada de comprimento *T*, com vetores de características acústicas, como representado pela seguinte equação:

$$O = \{o_t\} \tag{6.1}$$

Neste caso cada vetor representa, por exemplo, filtros Mel extraídos do som a cada instante de tempo. Em segundo lugar, o codificador, também designado de *encoder*, recebe como entrada uma sequência de vetores de características acústicas extraídas do sinal de voz, e transforma essa sequência numa nova sequência de representações:

$$H = \{h_l\} \tag{6.2}$$

Cada h_l representa um estado codificado, sendo que para reduzir a carga de processamento, é realizada uma *subsampling technique* na qual o modelo ignora parte dos dados menos relevantes.

No passo seguinte, o mecanismo *attention-based* cria um vetor de contexto c_n, com base em pesos de atenção a_n, que informam o modelo em que parte da entrada se deve focar para prever o próximo símbolo (letra ou palavra) e utiliza um tipo de atenção designado de *location-based*, que considera não só o conteúdo, mas também a posição dos elementos. Seguidamente, decorre a decodificação, no qual o *decoder* vai criando cada símbolo de saída um a um utilizando o que já foi escrito (*tokens* anteriores), o vetor de contexto c_n, que corresponde à atenção sobre o áudio original, e o seu próprio estado interno, ou também designado de estado oculto, uma vez que o *decoder* tem a capacidade de recorrer à sua memória interna.

Todos os passos incluídos no modelo são treinados para maximizar a probabilidade da sequência de texto correta (*Y*), utilizando uma função de perda designada de entropia cruzada. A probabilidade *Y* é calculada dada a entrada *O* e é representada pela seguinte equação:

$$P\,(Y|O) = \prod_n P\,(y_n|O, y_1, \dots, y_{n-1}) \tag{6.3}$$

Todos os componentes, *encoder*, mecanismo *attention-based* e *decoder*, são treinados em conjunto, o que permite ao modelo aprender a realizar todo o procedimento, desde ouvir, focar e transcrever corretamente, sem depender de módulos separados.

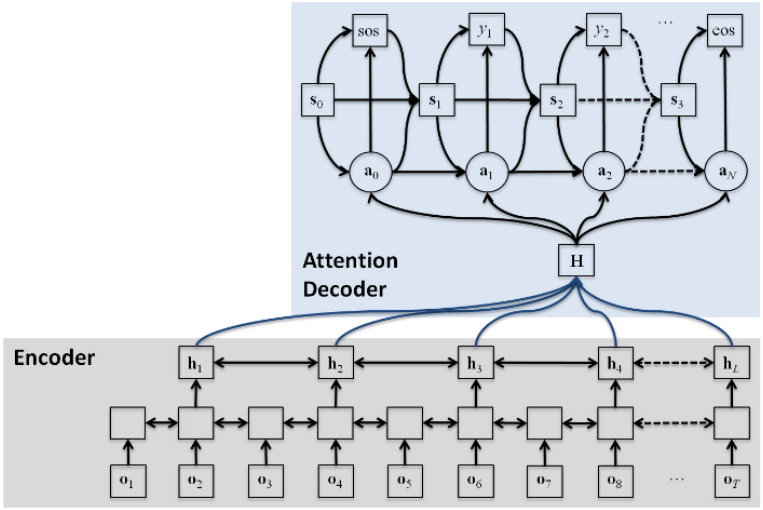

Figura 6.1 – Framework de um modelo encoder-decoder com mecanismo attention-based. O *encoder* transforma a sequência de entrada acústica *O* numa sequência de representações *H*. Em seguida, o *decoder* cria a sequência de saída *Y*, por exemplo, caracteres ou palavras, utilizando o mecanismo *attention-based*, que permite focar dinamicamente nos segmentos relevantes da entrada a cada passo da criação da transcrição. (Watanabe, S., et al. (2017). Multichannel End-to-End Speech Recognition.)

6.2 – Feedback com base em NLP

Uma inovação complementar de elevado potencial para reforçar a fiabilidade da documentação clínica automatizada é a introdução de mecanismos de *feedback* inteligente baseados em NLP. Embora o Invox Medical Dictation já demonstre uma notável robustez na transcrição, pois quando o utilizador soletra incorretamente um termo o *software* reconhece e corrige, a incorporação de um módulo de análise semântica permitiria avaliar a coerência, clareza e completude da transcrição criada em tempo real.

Com base em modelos de linguagem treinados em contexto clínico, o sistema poderia detetar, por exemplo, a ausência de elementos essenciais numa descrição ecográfica ou outra modalidade de imagem, a utilização de termos pouco claros ou inconsistentes com a lógica médica, ou mesmo possíveis lapsos linguísticos associados à fluência do ditado. Estas análises dariam origem a mensagens de *feedback* automatizadas, apresentadas ao utilizador sob a forma de sugestões, alertas ou anotações contextuais, promovendo uma revisão assistida mais eficaz.

Este tipo de funcionalidade encontra fundamento em estudos como o de Schmidt et al. (2024), que demonstraram a capacidade de *Large Language Models* como o GPT-4 para identificar erros clinicamente significativos em relatórios radiológicos, com desempenhos comparáveis aos de profissionais de saúde especializados na revisão de documentação clínica. Assim, o *feedback* com NLP reforçaria não só a precisão linguística, mas também a segurança e qualidade do registo clínico, contribuindo para um ciclo de aprendizagem contínuo e personalizado para cada utilizador.

6.3 – Fine-tuning

Uma das limitações mais significativas na aplicação de sistemas de reconhecimento de voz em ambientes clínicos é a disponibilidade limitada de dados anotados, ou seja, gravações de voz com as respetivas transcrições corretas, especialmente para as diversas especialidades médicas. Este desafio acentua-se de forma particularmente evidente em línguas com menor representação digital. A língua portuguesa, nomeadamente na variante europeia (português de Portugal), constitui um exemplo paradigmático desta limitação. Em comparação com línguas amplamente faladas e digitalmente dominantes, como o inglês, o mandarim ou o espanhol, o português europeu apresenta um volume significativamente reduzido de corpora anotados de voz, tanto em quantidade como em diversidade de sotaques e contextos clínicos. Esta disparidade tem impacto direto na eficácia dos sistemas ASR, dificultando a generalização e reduzindo a robustez dos modelos quando aplicados a contextos clínicos portugueses.

Neste cenário, estratégias como o *transfer learning* com *fine-tuning* específico para português europeu revelam-se particularmente úteis, permitindo reaproveitar modelos previamente treinados num domínio de larga escala e adaptá-los a um novo domínio com menos dados, neste caso, o clínico. Como descrito por Bell et al. (2021), o *fine-tuning* "permite adaptar modelos complexos a novos domínios com quantidades limitadas de dados de adaptação, através da atualização parcial ou total dos parâmetros do modelo pré-treinado".

O artigo de Howard & Ruder, 2018, apresenta uma proposta detalhada para a aplicação de *fine-tuning* em modelos de linguagem, através da metodologia denominada ULMFiT (*Universal Language Model Fine-tuning*). Esta abordagem parte da premissa de que um modelo de linguagem pode ser treinado inicialmente com um grande corpus genérico, como a Wikipedia, para aprender as estruturas fundamentais da língua, incluindo sintaxe, dependências gramaticais e padrões de uso lexical. Este modelo, já pré-treinado,

pode depois ser adaptado a um domínio específico, como a linguagem médica, mesmo quando há escassez de dados anotados nesse domínio.

O processo divide-se em três etapas principais. A primeira consiste no pré-treino do modelo de linguagem em dados genéricos, permitindo ao modelo adquirir uma compreensão ampla e versátil da linguagem. A segunda fase, correspondente ao *fine-tuning* do modelo de linguagem no domínio-alvo, envolve o ajuste dos parâmetros do modelo com dados pertencentes ao contexto específico, como ditados clínicos, de modo que o modelo se familiarize com o vocabulário técnico, as estruturas frásicas típicas e os estilos próprios do domínio. A terceira e última etapa diz respeito ao ajuste final do modelo para a tarefa concreta, como por exemplo a classificação de textos clínicos, a criação automática de relatórios ou a transcrição do ditado médico.

Para assegurar a eficácia deste processo de adaptação, os autores apresentam três técnicas inovadoras. A primeira é o ajuste discriminativo por camada (*discriminative fine-tuning*), que consiste em aplicar diferentes taxas de aprendizagem a diferentes camadas do modelo, reconhecendo que cada camada representa níveis distintos de abstração linguística, das mais gerais às mais específicas. A taxa de aprendizagem é um parâmetro de controlo que determina a velocidade com que um modelo ajusta os seus pesos durante o treino, com base no erro que comete ao fazer previsões.

A segunda técnica é a de agendamento triangular inclinado da taxa de aprendizagem (*slanted triangular learning rates*), que permite acelerar o processo de adaptação nas fases iniciais do treino e estabilizar as alterações nas fases posteriores.

Por fim, a técnica de descongelamento gradual das camadas (*gradual unfreezing*) orienta o processo de libertação progressiva das camadas do modelo para treino, evitando que as camadas inferiores, onde reside o conhecimento linguístico mais geral, sejam desajustadas prematuramente, o que poderia causar o chamado *catastrophic forgetting*. Quando se procede ao *fine-tuning* de um modelo treinado em linguagem geral para o domínio médico, sem técnicas de controlo, o modelo pode perder a sua capacidade de compreender estruturas linguísticas gerais ou realizar bem outras tarefas anteriores.

Os resultados experimentais demonstram que esta abordagem é altamente eficaz, permitindo obter desempenhos comparáveis ou superiores aos de modelos treinados de raiz, mesmo com conjuntos de treino significativamente mais pequenos. Através de 100 exemplos anotados, o modelo afinado com ULMFiT alcançou resultados que, tradicionalmente, requereriam dez a cem vezes mais dados. Estes resultados reforçam o

potencial do *fine-tuning* como solução prática e eficiente para domínios com escassez de dados, como é o caso da documentação clínica em português.

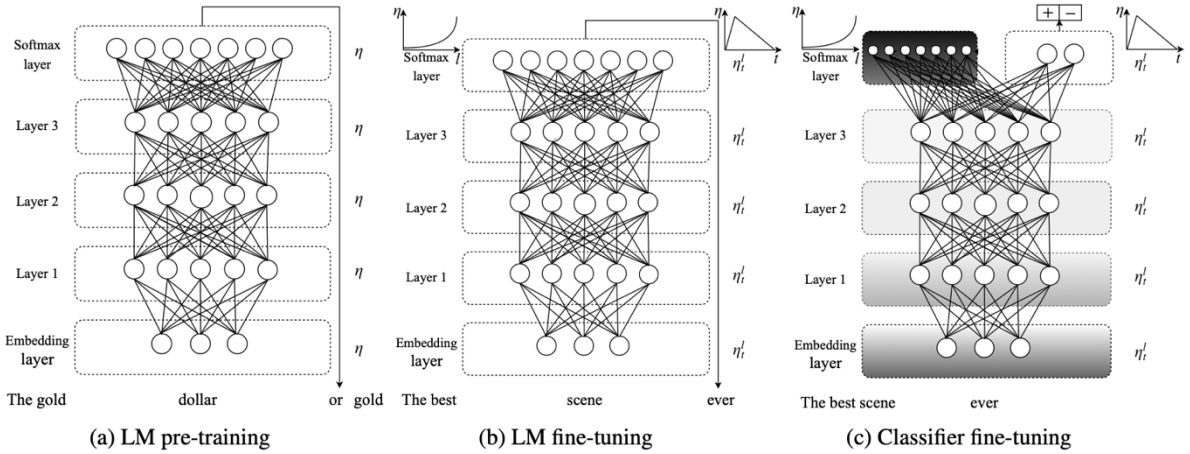

Figura 6.2 – *Framework* do ULMFiT.

Composto por três etapas: (a) o modelo de linguagem (LM) é treinado num corpus de domínio geral e capta características linguísticas gerais em diferentes camadas da rede; (b) o LM é afinado com dados da tarefa-alvo, utilizando *discriminative fine-tuning* e *slanted triangular learning rates* para aprender representações específicas da tarefa; (c) o classificador é posteriormente afinado com recurso ao *gradual unfreezing*, permitindo preservar representações de baixo nível e adaptar representações de alto nível. As áreas sombreadas indicam as fases de descongelamento progressivo e as áreas a preto correspondem a camadas ainda congeladas. (Howard & Ruder, 2018)

6.4 – Data Augmentation

A técnica *data augmentation* tem assumido um papel central na evolução dos sistemas de reconhecimento automático de voz, especialmente em contextos onde os dados anotados são escassos ou pouco variados. Esta abordagem prossupõe a criação de novas circunstâncias de dados de treino a partir de exemplos reais existentes, aplicando transformações acústicas controladas que simulam variações típicas do mundo real, sem comprometer o conteúdo linguístico.

Entre as transformações mais comuns encontram-se a adição de ruído de fundo, a alteração da velocidade ou do tom da fala, a simulação de reverberação, e a variação de características do canal de gravação. Ao introduzir estas variações nos dados de treino,

pretende-se expor o modelo a uma maior diversidade de situações, promovendo uma aprendizagem mais robusta e generalizável.

No caso específico do reconhecimento de ditado clínico, como é o caso do Invox Medical Dictation, a *data augmentation* pode mitigar significativamente os efeitos de ruído hospitalar, vozes sobrepostas, sotaques variados ou alterações na velocidade e clareza da fala. Treinar o modelo com dados artificialmente ruidosos pode prepará-lo para transcrever corretamente ditados realizados no contexto da Anatomia Patológica, local onde o ambiente acústico é instável.

A *data augmentation* pode incluir transformações como a adição de ruído de fundo, como o som do extrator do formol, conversas paralelas, equipamentos médicos; alteração da velocidade da fala; variação do tom da voz, designado de *pitch shifting*; variação da reverberação, com simulação de diferentes tipos de ambiente com propriedades acústicas diferentes como salas pequenas, salas grandes ou vazias, ambientes com superfícies absorventes ou superfícies reflexivas; conversão de canal, através da simulação de diferentes microfones ou dispositivos de gravação; e remoção ou inserção de pausas, para simular interrupções típicas do discurso espontâneo. Estas transformações podem ser aplicadas isoladamente ou combinadas, e são geralmente implementadas aquando da realização do treino.

Uma das técnicas mais eficazes e amplamente utilizadas é o *SpecAugment* (Park et al., 2019), que mascara diretamente partes do espectrograma do sinal de áudio, através da perturbação das representações sem afetar o conteúdo semântico. Esta técnica, juntamente com outras formas de *augmentation* baseadas em manipulação do sinal, tem demonstrado eficácia na melhoria do desempenho de modelos em tarefas de transcrição em ambientes adversos, bem como em contextos com dados limitados (Ko et al., 2015; Ragni et al., 2014).

6.5 – Técnicas de pré-treino auto-supervisionado

O pré-treino auto-supervisionado tem emergido como uma abordagem determinante para o avanço dos sistemas de reconhecimento automático de voz, uma vez que permite a aprendizagem de representações acústicas de elevada qualidade a partir de grandes volumes de áudio não anotado. Ao contrário do treino supervisionado convencional, que requer transcrições detalhadas para cada amostra de áudio, o paradigma auto-supervisionado é uma forma de ensinar um modelo ASR sem precisar de transcrições no início, uma vez que cria tarefas internas, também designadas de *proxy tasks*, que o modelo resolve utilizando apenas o próprio áudio como fonte de informação.

Um exemplo reconhecido neste campo é o *wav2vec 2.0* (Baevski et al., 2020), que opera em duas fases. Em primeiro lugar decorre o pré-treino, uma fase não supervisionada. Neste passo, o áudio é convertido em representações contínuas através de uma rede convolucional inicial e, alguns segmentos temporais dessas representações são mascarados, isto é, escondidos do modelo. A função do *wav2vec 2.0 é* adivinhar quais seriam as representações corretas dos segmentos mascarados, utilizando apenas o contexto dos segmentos visíveis.

De forma mais detalhada, posteriormente ao áudio ser processado pela primeira parte da rede (o extrator de características convolucional), é obtida uma sequência de vetores contínuos que representam o som ao longo do tempo. Estes vetores apresentam valores reais e muita informação, no entanto para criar uma tarefa clara para o modelo durante o pré-treino, é útil convertê-los nas chamadas unidades discretas, como símbolos que representem segmentos do som. Este processo é designado de quantização.

O resultado é uma sequência de unidades discretas, semelhante a um alfabeto fonético aprendido automaticamente. A tarefa de decifrar os segmentos não visíveis força o modelo a compreender o contexto temporal da fala. Desta forma, o modelo tem a capacidade de aprender padrões fonéticos, estrutura da pronúncia, e até elementos prosódicos, mesmo sem nunca ter acesso ao texto escrito.

Em segundo lugar, existe uma fase supervisionada de *fine-tuning*, já explicada no capítulo 6.3., na qual o modelo já está sensibilizado para a estrutura acústica da fala e aprende com maior celeridade a associar os padrões ao texto. Neste passo é fornecido ao *wav2vec 2.0* um conjunto mais pequeno de dados com transcrições corretas, e são ajustados os pesos do modelo para associar os padrões de áudio que aprendeu previamente, às palavras ou caracteres corretos. Uma vez que o modelo já compreende a estrutura acústica da fala, requer menos dados anotados para atingir um alto desempenho.

No contexto clínico, obter dados de áudio anotados, isto é, áudio e respetiva transcrição, é dispendioso e sensível do ponto de vista da privacidade, contrariamente a áudios não anotados, como é o caso de gravações de ditados médicos anonimizadas, obtidas com consentimento. Através da utilização do pré-treino auto-supervisionado, podem ser utilizados grandes volumes de áudio não anotado para ensinar o modelo a reconhecer padrões de fala em português europeu e, posteriormente, recorrer a um conjunto relativamente reduzido de dados anotados do domínio-alvo, realizando o *fine-tuning*, e adaptando o modelo à terminologia médica e ao estilo de ditado característico de cada especialidade.

No contexto do Invox Medical Dictation, a adoção desta técnica permitiria desenvolver um sistema menos dependente de grandes volumes de transcrições clínicas, promovendo simultaneamente maior adaptabilidade a diferentes utilizadores e ambientes de gravação.

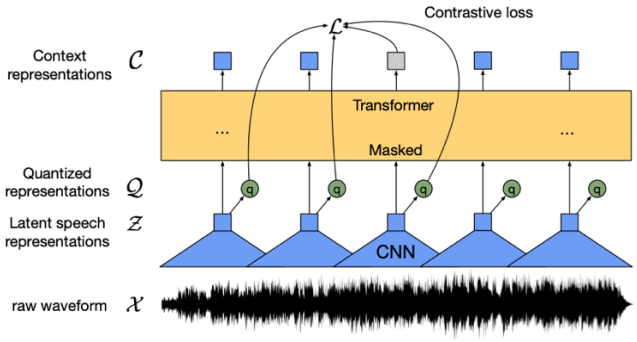

Figura 6.3 – *Framework* do modelo wav2vec 2.0. (Baevski et al., 2020)

6.6 – Personalização de sistemas ASR

A adaptação de sistemas de reconhecimento de voz ao perfil vocal individual de cada utilizador poderá ser um elemento importante na melhoria da precisão em contextos clínicos. Entre as técnicas mais relevantes para esse fim destacam-se os *speaker embeddings* e os *i-vectors*, ambos concebidos para representar de forma compacta as características acústicas e prosódicas de um orador.

Os *speaker embeddings* consistem em vetores de dimensão fixa criados por redes neuronais profundas, treinadas para extrair representações discriminativas da voz de cada indivíduo (Snyder et al., 2018). Estas representações captam padrões característicos como timbre, entoação, ritmo de fala e pronúncia, permitindo que o sistema se adapte a variações linguísticas e a particularidades da fala. No contexto do Invox Medical Dictation, esta tecnologia possibilita, por exemplo, que o *software* reconheça e interprete de forma consistente abreviaturas ou termos técnicos específicos utilizados por um determinado médico, mesmo que estes não sigam a norma fonética ou ortográfica convencional.

Os *speaker embeddings*, também designados *x-vectors* na implementação proposta por Snyder et al. (2018), recorrem a redes neuronais profundas para aprender representações discriminativas da voz. Neste caso, o áudio é primeiro convertido em características acústicas, como MFCCs, que são processados por camadas convolucionais temporais capazes de modelar dependências de curto e longo prazo. Segue-se uma camada de *pooling* estatístico, que agrega a informação ao longo de toda a duração do segmento,

capaz de criar um vetor de dimensão fixa que é posteriormente refinado por camadas densas. O treino é supervisionado, tratando cada orador do conjunto de treino como uma classe distinta, de modo a otimizar diretamente a distinção entre vozes diferentes, ou seja, trata-se de um modelo *one-to-one*. O *embedding* é posteriormente extraído de uma das últimas camadas da rede, antes da camada de classificação, e pode ser utilizado para adaptação de modelos de reconhecimento de voz.

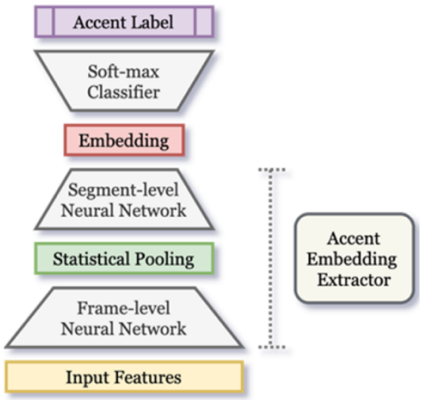

Figura 6.4 – Utilização de *speaker embeddings* para um *software speech-to-text* mais personalizado. (https://aihub.org/ by Tugtekin Turan)

Por sua vez, os *i-vectors* representam uma abordagem estatística consolidada, que permite extrair vetores de pequena dimensão capazes de representar simultaneamente as características do orador e as condições acústicas da gravação (Dehak et al., 2011). Embora tenham sido amplamente desenvolvidos antes da popularização das redes neuronais profundas, os *i-vectors* continuam a ser relevantes na adaptação de modelos acústicos, podendo ser utilizados isoladamente ou como complemento a métodos mais recentes.

O processo inicia-se com a construção de um *Universal Background Model* (UBM), geralmente implementado como um *Gaussian Mixture Model* (GMM), treinado com um conjunto de dados de voz provenientes de múltiplos utilizadores. Este modelo serve como uma referência estatística que descreve a distribuição geral das características acústicas da fala. Quando é fornecida uma nova gravação, são extraídas as suas características acústicas (MFCCs), e são calculadas estatísticas de adaptação em relação ao UBM. As estatísticas calculadas permitem criar um *supervetor*, que consiste na ligação das médias adaptadas de todas as gaussianas do GMM.

Como este supervetor é de dimensão muito elevada e contém variações tanto do utilizador como do canal de gravação, utiliza-se um modelo linear para projetá-lo num

espaço de variabilidade total, conhecido como *Total Variability Space*. Este espaço foi aprendido previamente e procura captar todas as fontes de variação de um vetor compacto. O resultado desta projeção é o *i-vector*, uma representação densa e de pequena dimensão que condensa as características essenciais da voz e do ambiente de gravação. Para finalizar, o *i-vector* é normalizado para ter um comprimento unitário, antes de ser utilizado em tarefas como reconhecimento do locutor, ou como informação adicional para adaptação de modelos de reconhecimento de voz.

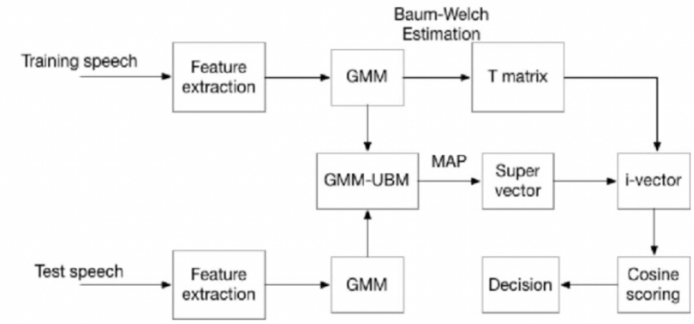

Figura 6.5 – Diagrama de um sistema de reconhecimento de orador com base em *i-vector*. (Yuan et al., 2021)

Assim, embora partilhem o mesmo objetivo, discriminar a identidade vocal de um orador, estas abordagens diferem significativamente na sua origem tecnológica e na forma como representam a informação. Enquanto os *i-vectors* derivam de técnicas de modelação estatística clássica, os *speaker embeddings* exploram a capacidade das redes neuronais para aprender representações não lineares mais ricas e adaptáveis. A integração de ambas as técnicas pode, assim, potenciar a robustez e a personalização do sistema, garantindo um desempenho consistente mesmo perante variações significativas de voz, sotaque ou ambiente de gravação (Anexo F).

6.7 – Considerações finais sobre as inovações tecnológicas propostas

As inovações tecnológicas mencionadas configuram um conjunto complementar de abordagens capazes de elevar significativamente a precisão, a adaptabilidade e a resiliência de sistemas de conversão de voz em texto no domínio clínico.

Desde a aplicação de arquiteturas avançadas como os modelos *attention-based encoder-decoder*, capazes de lidar com a complexidade e variabilidade do discurso médico, até ao uso de *feedback* inteligente baseado em NLP para assegurar coerência e completude

semântica, as soluções propostas respondem a limitações estruturais e operacionais frequentemente observadas no contexto hospitalar.

A integração de estratégias de *fine-tuning*, *data augmentation* e pré-treino auto-supervisionado mitiga de forma eficaz a escassez de dados anotados em português europeu, promovendo não apenas um melhor desempenho técnico, mas também a equidade linguística.

Por fim, os mecanismos de personalização por *speaker embeddings* e *i-vectors* consolidam a capacidade do sistema em adaptar-se a perfis vocais individuais, assegurando consistência e fiabilidade em cenários clínicos de elevada exigência.

Concluindo, o conjunto destas inovações não apenas reforça a maturidade tecnológica de soluções de *speech-to-text* na saúde, como também evidencia um caminho estratégico para o seu desenvolvimento futuro, ancorado na personalização, na robustez e na eficiência. A conjugação destas técnicas, suportada por evidência empírica e alinhada com tendências emergentes de IA aplicada à saúde, sustenta um modelo de evolução contínua, no qual a otimização da documentação clínica se torna indissociável da qualidade assistencial e da segurança do doente.

6.8 – Discussão dos resultados face às inovações tecnológicas propostas

Os resultados obtidos no estudo de caso no Serviço de Anatomia Patológica do Hospital Curry Cabral evidenciam que o Invox Medical Dictation trouxe ganhos expressivos em termos de celeridade, simplicidade de utilização e qualidade dos relatórios clínicos, mas também revelou limitações, sobretudo no que respeita à interoperabilidade com o sistema ANAPAT e à variabilidade de desempenho consoante sotaques e perfis de utilizadores. Quando confrontados com os avanços tecnológicos descritos no presente capítulo, estes achados permitem delinear caminhos claros para a evolução futura da solução.

Em primeiro lugar, a elevada perceção de precisão global (média de 4,3 numa escala de 1 a 5) contrasta com erros pontuais e recorrentes em medidas, frases construídas na negativa ou epónimos. Os modelos *attention-based encoder-decoder* (AED) representam uma resposta tecnológica robusta, dado que permitem maior resiliência a frases longas, pausas e reformulações típicas do discurso médico, diminuindo a propensão para erros estruturais na transcrição.

Por outro lado, a confiança parcial dos utilizadores, que em alguns casos não excedeu 2,3 numa escala de 1 a 3, aponta para a necessidade de mecanismos que reforcem a validação semântica. Os módulos de *feedback* baseados em NLP representam uma

oportunidade de evolução, permitindo que o sistema não apenas transcreva, mas também interprete e valide o conteúdo em tempo real, sinalizando incoerências, omissões ou inconsistências clínicas. Este tipo de apoio seria particularmente útil em contextos ruidosos ou em ditados extensos, mitigando erros clínicos potencialmente críticos e aumentando a confiança dos utilizadores na ferramenta.

Outro aspeto transversal às entrevistas foi a perceção de que a fiabilidade do *software* varia consoante o utilizador, especialmente em função do sotaque e da clareza de articulação. Este desafio encontra resposta nas estratégias de *fine-tuning*, que permitem adaptar modelos genéricos ao português europeu, e de data augmentation, que simulam diferentes condições acústicas e variações de fala, ampliando a robustez do modelo perante sotaques regionais e ruído hospitalar. Estas técnicas não só responderiam às dificuldades relatadas por alguns médicos internos com sotaques distintos, como também aumentariam a consistência da ferramenta em gabinetes partilhados, onde a sobreposição de vozes foi identificada como problemática.

De igual modo, a escassez de corpora clínicos em português de Portugal constitui uma barreira estrutural, amplamente discutida na literatura. O recurso a pré-treino auto-supervisionado permitiria aproveitar grandes volumes de áudio não anotado, facilitando a aprendizagem de padrões fonéticos e prosódicos da língua sem necessidade de extensas bases de dados anotadas. Posteriormente, a aplicação de *fine-tuning* com dados médicos específicos garantiria que o sistema aprendesse terminologia técnica e estruturas linguísticas típicas da Anatomia Patológica, área em que se exige elevada precisão. Desta forma, problemas com frases descritivas padronizadas como “mucosa gástrica com aspetos de gastrite crónica moderada”, enumerações frequentes para o número de fragmentos biopsados como “EDA1, EDA2 (...)”, descrições de dimensões como “23x14x5mm” e localizações anatómicas, combinações de adjetivos técnicos como “limites bem definidos”, abreviaturas e siglas que seguem padrões específicos como “HPB”, poderiam ser solucionados.

A personalização dos sistemas de reconhecimento, por via de *speaker embeddings* e *i-vectors*, também surge como via promissora para ultrapassar as dificuldades relatadas por utilizadores com sotaques ou dicções específicas. Ao permitir a adaptação do modelo ao perfil vocal de cada profissional, estas técnicas assegurariam que tanto Médicos Seniores, com estilos de ditado mais pausados, como Patologistas Internos, com sotaques distintos, obtivessem um desempenho homogéneo e fiável.

Por fim, os resultados de satisfação elevados (média de 3,3/4) e a unanimidade na recomendação do *software* refletem a aceitação da solução e criam um terreno fértil para a introdução destas inovações tecnológicas. Conforme sintetizado no Capítulo 5.7, a conjugação de abordagens como AED, feedback semântico, *fine-tuning*, *data augmentation*, pré-treino auto-supervisionado e personalização vocal delineia um caminho de desenvolvimento contínuo. A sua adoção permitiria não apenas mitigar as fragilidades atuais, mas também consolidar o papel do *speech-to-text* como tecnologia central na modernização da documentação clínica, ancorada na precisão, adaptabilidade e eficiência.

Para além da métrica WER, tradicionalmente utilizada para aferir a precisão de sistemas de *speech-to-text*, seria relevante a inclusão de outras métricas referidas na revisão de literatura, capazes de fornecer uma avaliação mais abrangente do impacto da solução em contexto clínico. Entre estas destacam-se o CER, o MC-WER, o TAT, o RTF, bem como indicadores clássicos de desempenho como a taxa de erro global, precisão, sensibilidade (recall), F1-score e accuracy. Adicionalmente, métricas mais recentes como o BERTScore e o ROUGE permitiriam avaliar a qualidade semântica e a completude dos relatórios criados. A utilização combinada destes indicadores possibilitaria uma análise mais rigorosa da solução não apenas ao nível da celeridade, mas também da sua exatidão terminológica e da qualidade global da documentação clínica.

Relativamente à interoperabilidade com o sistema ANAPAT, importa enquadrar esta limitação no contexto mais amplo dos processos de digitalização no Serviço Nacional de Saúde (SNS). A implementação de novas soluções tecnológicas em ambiente hospitalar é frequentemente condicionada por restrições de recursos humanos e materiais nos departamentos de informática, que se encontram sobrecarregados com múltiplas solicitações e prioridades concorrentes. Esta realidade torna os processos de integração mais lentos e complexos, mesmo quando a solução em si apresenta um elevado potencial de inovação. Assim, a dificuldade não decorre apenas das características técnicas do *software*, mas também da necessidade de garantir que a infraestrutura e o suporte interno das instituições de saúde acompanham o ritmo das inovações, assegurando uma transição digital progressiva e sustentável.

Referências Bibliográficas

Adedeji, A., Joshi, S., & Doohan, B. (2024). The Sound of Healthcare: Improving Medical Transcription ASR Accuracy with Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2402.07658

Agrawal, P., & Ganapathy, S. (2019). Deep Variational filter learning models for speech recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5731–5735. https://doi.org/10.1109/icassp.2019.8682520

Ajami, S. (2016). Use of speech-to-text technology for documentation by healthcare providers. Natl Med J India, Vol. 29(3). https://pubmed.ncbi.nlm.nih.gov/27808064/

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Neural Information Processing Systems, 33,12449–12460. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2015, August 18). End-to-End attention-based large vocabulary speech recognition. arXiv.org. http://arxiv.org/abs/1508.04395

Bell, P., Fainberg, J., Klejch, O., Li, J., Renals, S., & Swietojanski, P. (2020). Adaptation Algorithms for Neural Network-Based Speech Recognition: An Overview. IEEE Open Journal of Signal Processing, 2, 33–66. https://doi.org/10.1109/ojsp.2020.3045349

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., & Zhou, L. (2018). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. Journal of the American Medical Informatics Association, 26(4), 324–338. https://doi.org/10.1093/jamia/ocy179

Blackley, S. V., Schubert, V. D., Goss, F. R., Assad, W. A., Garabedian, P. M., & Zhou, L. (2020). Physician use of speech recognition versus typing in clinical documentation: A controlled observational study. International Journal of Medical Informatics, 141, 104178. https://doi.org/10.1016/j.ijmedinf.2020.104178

Busch, F., Prucker, P., Komenda, A., Ziegelmayer, S., Makowski, M. R., Bressem, K. K., & Adams, L. C. (2024). Multilingual feasibility of GPT-4o for automated Voice-to-Text CT and MRI report transcription. European Journal of Radiology, 182, 111827. https://doi.org/10.1016/j.ejrad.2024.111827

Cunningham, S., Green, P., Christensen, H., Atria, J. J., Coy, A., Malavasi, M., Desideri, L., & Rudzicz, F. (2017). Cloud-Based Speech Technology for Assistive Technology Applications (CloudCAST). Studies in Health Technology and Informatics, 242, 322–329. https://doi.org/10.3233/978-1-61499-798-6-322

Deepa, P., & Khilar, R. (2022). Speech technology in healthcare. Measurement Sensors, 24, 100565. https://doi.org/10.1016/j.measen.2022.100565

Dehak, N., J. Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End factor analysis for speaker verification. IEEE Journals & Magazine | IEEE Xplore, 19(4). https://ieeexplore.ieee.org/document/5545402

Galloway, J. L., Munroe, D., Vohra-Khullar, P. D., Holland, C., Solis, M. A., Moore, M. A., & Dbouk, R. H. (2024). Impact of an Artificial Intelligence-Based Solution on Clinicians' clinical Documentation experience: Initial findings using ambient listening technology. Journal of General Internal Medicine, 39(13), 2625–2627. https://doi.org/10.1007/s11606-024-08924-2

Goss, F. R., Blackley, S. V., Ortega, C. A., Kowalski, L. T., Landman, A. B., Lin, C., Meteer, M., Bakes, S., Gradwohl, S. C., Bates, D. W., & Zhou, L. (2019). A clinician survey of using speech recognition for clinical documentation in the electronic health record.

International Journal of Medical Informatics, 130, 103938. https://doi.org/10.1016/j.ijmedinf.2019.07.017

Haboussi, S., Oukas, N., Zerrouki, T., & Djettou, H. (2025). Arabic speech recognition using neural networks: concepts, literature review and challenges. Journal of Umm Al-Qura University for Applied Sciences. https://doi.org/10.1007/s43994-025-00213-w

Hammana, I., Lepanto, L., Poder, T., Bellemare, C., & Ly, M. (2015). Speech Recognition in the Radiology Department: A Systematic review. Health Information Management Journal, 44(2), 4–10. https://doi.org/10.1177/183335831504400201

Hodgson, T., & Coiera, E. (2015). Risks and benefits of speech recognition for clinical documentation: a systematic review. Journal of the American Medical Informatics Association, 23(e1), e169–e179. https://doi.org/10.1093/jamia/ocv152

Hodgson, T., Magrabi, F., & Coiera, E. (2017). Efficiency and safety of speech recognition for documentation in the electronic health record. Journal of the American Medical Informatics Association, 24(6), 1127–1133. https://doi.org/10.1093/jamia/ocx073

Hodgson, T., Magrabi, F., & Coiera, E. (2018). Evaluating the Efficiency and Safety of Speech Recognition within a Commercial Electronic Health Record System: A Replication Study. Applied Clinical Informatics, 09(02), 326–335. https://doi.org/10.1055/s-0038-1649509

Howard, J., & Ruder, S. (2018, January 18). Universal Language model fine-tuning for text classification. arXiv.org. https://arxiv.org/abs/1801.06146

Hu, Y., Chen, C., Qin, C., Zhu, Q., Chng, E. S., & Li, R. (2024, May 16). Listen Again and Choose the Right Answer: A New Paradigm for Automatic Speech Recognition with Large Language Models. arXiv.org. http://arxiv.org/abs/2405.10025

Jorg, T., Kämpgen, B., Feiler, D., Müller, L., Düber, C., Mildenberger, P., & Jungmann, F. (2023). Efficient structured reporting in radiology using an intelligent dialogue system based on speech recognition and natural language processing. Insights Into Imaging, 14(1), 47. https://doi.org/10.1186/s13244-023-01392-y

Joseph, J., Moore, Z. E. H., Patton, D., O'Connor, T., & Nugent, L. E. (2020). The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review. Journal of Clinical Nursing, 29(13–14), 2125–2137. https://doi.org/10.1111/jocn.15261

Karbasi, Z., Bahaadinbeigy, K., Ahmadian, L., Khajouei, R., & Mirzaee, M. (2019). Accuracy of speech recognition system's medical report and physicians' experience in hospitals. Frontiers in Health Informatics, 8(1), 19. https://doi.org/10.30699/fhi.v8i1.199

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. https://www.semanticscholar.org/paper/Audio-augmentation-for-speech-recognition-Ko-Peddinti/66661a68dbf1d98d794fd025113b103683510303

Kumar, Y. (2024). A Comprehensive analysis of Speech recognition Systems in healthcare: Current research challenges and future Prospects. SN Computer Science, 5(1). https://doi.org/10.1007/s42979-023-02466-w

Kumar, Y., & Singh, N. (2019). A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature review. International Conference on Automation, Computational and Technology Management (ICACTM 2019), Session 3, 168–173. https://www.proceedings.com/content/049/049759webtoc.pdf

Kumar, Y., Koul, A., & Singh, C. (2022). A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimedia Tools and Applications, 82(10), 15171–15197. https://doi.org/10.1007/s11042-022-13943-4

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech Technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering, 14, 342–356. https://doi.org/10.1109/rbme.2020.3006860

Li, Y., Wang, Y., Hoi, L. M., Yang, D., & Im, S. (2025). A review on speech recognition approaches and challenges for Portuguese: exploring the feasibility of fine-tuning large-scale end-to-end models. EURASIP Journal on Audio Speech and Music Processing, 2025(1). https://doi.org/10.1186/s13636-024-00388-w

Lyons, J. P., Sanders, S. A., Cesene, D. F., Palmer, C., Mihalik, V. L., & Weigel, T. (2015). Speech recognition acceptance by physicians: A temporal replication of a survey of expectations and experiences. Health Informatics Journal, 22(3), 768–778. https://doi.org/10.1177/1460458215589600

Ochiai, T., Watanabe, S., Hori, T., & Hershey, J. R. (2017). Multichannel end-to-end speech recognition. arXiv (Cornell University), 2632–2641. https://doi.org/10.48550/arxiv.1703.04783

Ogundokun, R. O., Abikoye, O. C., Adegun, A. A., & Awotunde, J. B. (2020). Speech Recognition System: Overview of the State-Of-The-Arts. International Journal of Engineering Research and Technology, 13(3), 384. https://doi.org/10.37624/ijert/13.3.2020.384-392

Palaz, D., Magimai-Doss, M. M., & Collobert, R. (2015, April 1). Convolutional Neural Networks-based continuous speech recognition using raw speech signal. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7178781

Park, D. S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). SpecAugment: a simple data augmentation method for automatic speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, 2613–2617. https://doi.org/10.21437/interspeech.2019-2680

Paul, S., Bhattacharjee, V., & Saha, S. K. (2025). A transfer learning approach for continuous speech recognition system in Indian language sadri. International Journal of Information Technology, 17(7), 3835–3844. https://doi.org/10.1007/s41870-025-02516-x

Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.F. (2014) Data augmentation for low resource languages. Proc. Interspeech 2014, 810-814, doi: 10.21437/Interspeech.2014-207

Ringler, M. D., Goss, B. C., & Bartholmai, B. J. (2015). Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics Journal, 23(1), 3–13. https://doi.org/10.1177/1460458215613614

Salifu, A., Mensah, H. N., Tchao, E. T., Acheampong, F. A., Agbemenu, A. S., & Kponyo, J. J. (2025). Enhancing speech recognition through diverse shared features accent classification. International Journal of Speech Technology, 28(2), 461–481. https://doi.org/10.1007/s10772-025-10198-w

Schmidt, R. A., Seah, J. C. Y., Cao, K., Lim, L., Lim, W., & Yeung, J. (2024). Generative large language models for detection of speech recognition errors in radiology reports. Radiology Artificial Intelligence, 6(2), e230205. https://doi.org/10.1148/ryai.230205

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. https://doi.org/10.1109/icassp.2018.8461375

Song, Y. (2023). Chinese speech recognition system based on neural network acoustic network model. Procedia Computer Science, 228, 144–154. https://doi.org/10.1016/j.procs.2023.11.018

Vogel, M., Kaisers, W., Wassmuth, R., & Mayatepek, E. (2015). Analysis of documentation speed using Web-Based Medical Speech Recognition Technology: Randomized Controlled trial. Journal of Medical Internet Research, 17(11), e247. https://doi.org/10.2196/jmir.5072

Wu, Z., Valentini-Botinhao, C., Watts, O., King, S., & University of Edinburgh, United Kingdom. (2015, April 1). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. 2015 IEEE International Conference on

Acoustics, Speech and Signal Processing, Queensland, South Brisbane, Australia. pp. 4460-4464. https://www.proceedings.com/content/027/027049webtoc.pdf.

Yuan, X., Li, G., Han, J., Wang, D., & Tiankai, Z. (2021). Speaker identification based on iVector and Xvector. Journal of Physics Conference Series, 1827(1), 012133. https://doi.org/10.1088/1742-6596/1827/1/012133

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024, January 24). MM-LLMS: Recent Advances in MultiModal Large Language Models. arXiv.org. http://arxiv.org/abs/2401.13601

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019, April 21). BERTScore: Evaluating Text Generation with BERT. arXiv.org. http://arxiv.org/abs/1904.09675

Zhang, Y., Sun, S., & Ma, L. (2021). Tiny Transducer: a highly-efficient speech recognition model on edge devices. IEEE International Conference on Acoustics, Speech and Signal Processing, 6024–6028. https://doi.org/10.48550/arxiv.2101.06856

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Wen, J. (2023, March 31). A survey of large language models. arXiv.org. http://arxiv.org/abs/2303.18223

Zuchowski, M., & Göller, A. (2022). Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. British Journal of Healthcare Management, 28(1), 30–36. https://doi.org/10.12968/bjhc.2021.0074

Anexos

Anexo A – Inquérito de Satisfação para utilizadores do Invox Medical Dictation

Este inquérito tem como finalidade recolher a opinião dos profissionais de saúde que utilizam o software Invox Medical Dictation, no sentido de avaliar o seu impacto na prática clínica. A sua colaboração é essencial para o sucesso desta investigação. Todas as respostas são confidenciais e utilizadas exclusivamente para fins académicos.

1. Dados Gerais

Profissão: ☐ Médico ☐ Técnico ☐ Outro: _____

Especialidade: _____

Tempo de utilização do Invox: ☐ < 1 mês ☐ 1–6 meses ☐ 6–12 meses ☐ > 1 ano

2. Eficiência Operacional

2.1. O tempo necessário para elaborar relatórios clínicos diminuiu após a adoção do Invox?

☐ Concordo totalmente ☐ Concordo ☐ Neutro ☐ Discordo ☐ Discordo totalmente

2.2. Considera que o Invox contribui para uma maior celeridade na emissão de resultados para os utentes?

☐ Concordo totalmente ☐ Concordo ☐ Neutro ☐ Discordo ☐ Discordo totalmente

3. Usabilidade e Experiência de Utilização

3.1. Considera o software fácil de utilizar?

☐ Concordo totalmente ☐ Concordo ☐ Neutro ☐ Discordo ☐ Discordo totalmente

3.2. Encontra frequentemente erros de reconhecimento de voz?

☐ Raramente ☐ Ocasionalmente ☐ Frequentemente ☐ Sempre

3.3. Considera a interface intuitiva e adequada à sua prática?

☐ Sim ☐ Parcialmente ☐ Não

4. Fiabilidade e Qualidade dos Relatórios

4.1. A transcrição automática apresenta uma elevada precisão dos termos médicos utilizados?

☐ Concordo totalmente ☐ Concordo ☐ Neutro ☐ Discordo ☐ Discordo totalmente

4.2. Necessita de realizar muitas correções após a transcrição automática?

☐ Nunca ☐ Raramente ☐ Frequentemente ☐ Sempre

4.3. Confia na fiabilidade da informação gerada pelo sistema?

☐ Sim ☐ Parcialmente ☐ Não

5. Satisfação Geral

5.1. Está satisfeito com o desempenho geral do Invox?

☐ Muito satisfeito ☐ Satisfeito ☐ Indiferente ☐ Insatisfeito

5.2. Recomendaria o uso do Invox a outros profissionais?

☐ Sim ☐ Não

6. Adoção e Integração na Rotina

6.1. O tempo de adaptação ao software foi:

☐ Muito curto ☐ Curto ☐ Adequado ☐ Demorado ☐ Ainda não me adaptei

6.2. Recebeu formação ou apoio inicial suficiente para usar o Invox com confiança?

☐ Sim ☐ Parcialmente ☐ Não

6.3. Considera que o software está bem integrado com os sistemas informáticos que utiliza no seu dia a dia (ex.: ANAPAT, SISPAT)?

☐ Concordo totalmente ☐ Concordo ☐ Neutro ☐ Discordo ☐ Discordo totalmente

7. Barreiras e Dificuldades

7.1. Quais são os principais obstáculos que sente na utilização do Invox? (pode assinalar mais do que uma opção)

☐ Reconhecimento de voz impreciso

☐ Problemas técnicos (microfone, sistema)

☐ Integração com outros sistemas

☐ Falta de tempo para aprender

☐ Resistência à mudança

☐ Outros: _____

8. Sugestões e Melhoria Contínua

8.1. Que funcionalidades ou melhorias gostaria de ver no Invox Medical Dictation?

_____

_____

_____

9. Comentários Finais (facultativo)

_____

_____

_____

Anexo B – Guião para entrevista semi-estruturada para avaliar qualitativamente a performance, integração, usabilidade e impacto do Invox Medical Dictation na rotina dos profissionais de saúde, com foco na documentação clínica.

1. Experiência de uso e adoção

1.1. Há quanto tempo utiliza o Invox Medical Dictation?

1.2. Como descreveria a sua experiência global com a ferramenta?

1.3. Que tipo de formação recebeu para utilizar o sistema?

2. Eficiência e impacto no tempo de trabalho

2.1. Nota alguma diferença no tempo que dedica à elaboração dos relatórios clínicos desde que começou a usar o Invox?

2.2. O tempo entre o exame e a emissão do relatório sofreu alterações?

2.3. Sente que a ferramenta lhe permite dedicar mais tempo a outras atividades clínicas?

3. Usabilidade e integração com outros sistemas

3.1. Como avalia a facilidade de utilização do software no seu dia a dia?

3.2. O Invox está bem integrado com os sistemas informáticos da sua instituição (ex: PACS, RIS, SClínico, Byme)?

3.3. Já teve dificuldades técnicas ou problemas de compatibilidade?

4. Qualidade e precisão da transcrição

4.1. Considera que a transcrição por voz é geralmente precisa e fiável?

4.2. Há erros frequentes? De que tipo (terminologia médica, pontuação, ortografia, estrutura da frase)?

4.3. Costuma ter de rever ou corrigir muito os relatórios após ditá-los?

5. Impacto na qualidade da documentação

5.1. Acha que os relatórios gerados com Invox têm a mesma qualidade (ou superior/inferior) do que os anteriores?

5.2. Já recebeu feedback de colegas, técnicos ou doentes sobre a qualidade dos relatórios?

6. Satisfação e impacto pessoal/profissional

6.1. De que forma o uso do Invox influenciou a sua carga administrativa ou nível de stress?

6.2. Sente-se mais satisfeito com o processo de documentação clínica?

6.3. Recomendaria esta ferramenta a outros colegas?

7. Sugestões e perspetivas de melhoria

7.1. Há algo que gostaria de ver melhorado ou alterado?

7.2. Que funcionalidades adicionais considera que poderiam melhorar a ferramenta?

Anexo C – Respostas às entrevistas realizadas aos TSDTs, Patologistas e Administrativas do Serviço de Anatomia Patológica do Hospital Curry Cabral

---

Entrevista 1(TSDT)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 1: Há cerca de 2 anos. Tivemos um período experimental de 1 ano e meio e foi adquirido oficialmente há 6 meses pelo Serviço.

Entrevistadora: Como descreve o impacto que o software teve?

Utilizador 1: O impacto foi totalmente positivo ainda que existam ajustes necessários. Desde a primeira semana de demonstração que ficou muito claro que era uma ferramenta essencial. Foram notadas algumas dificuldades de adaptação de alguns colegas que têm algumas limitações a nível de conhecimento informático. O nosso sistema informático também apresentou algumas barreiras, no entanto à exceção destes dois pormenores o impacto foi surreal. Existem erros pontuais, no entanto o software vai aprendendo conforme a utilização.

Entrevistadora: Recebeu alguma formação para trabalhar no software?

Utilizador 1: No primeiro dia que instalaram no nosso Serviço a responsável da empresa conseguiu explicar em meia hora o funcionamento simples do software. Quando foi oficialmente implementado após o período experimental tivemos uma sessão de formação todos juntos.

Entrevistadora: A aplicação é intuitiva?

Utilizador 1: É muito simples e intuitiva. Posso destacar que o mais complexo foi perceber onde se encontrava a funcionalidade de diagnóstico de erros, uma definição mais oculta, mas que após percecionar uma vez ficou clarificado. Na minha opinião é uma mais valia ser uma funcionalidade oculta para que os utilizadores não estejam constantemente a aceder e alterar parâmetros.

Entrevistadora: Nota alguma diferença no tempo que dedica à elaboração dos relatórios?

Utilizador 1: Sim muita. Quando tinha uma peça cirúrgica grande, eu ia ditando à medida que ia dissecando, e tinha de dar tempo que o outro colega conseguisse acompanhar e escrever tudo o que eu estava a dizer ou então ditar devagar para que quando a Administrativa estivesse a ouvir conseguisse percecionar o que eu estava a dizer. Portanto era um ditado mais pausado. O gravador antigo dava para fazer pausa e continuar a gravar, mas muitas vezes de forma não propositada eu tocava no botão stop e finalizava aquele

ficheiro de forma incompleta e tinha de criar um novo de seguida e pedir desculpa à Administrativa fazendo referência que o novo áudio era a parte restante do áudio anterior da análise x e, com a utilização do Invox esse dilema acabou. Por vezes com esta dificuldade existiam gravações que se perdiam e, consequentemente, acabávamos por perder tempo a relatar novamente. Atualmente enquanto disseco a peça, estou a ditar o que estou a ver e demoro entre 40 a 45 segundos a elaborar o relatório. O colega que está ao meu lado por vezes corrige algumas palavras, mas de forma mais fluída. Algumas palavras não são bem percecionadas porque eu sibilo bastante e o software ainda não se adaptou.

Entrevistadora: Então quando, antes do software, elaborava os relatórios, sentia que o seu raciocínio estava sempre a ser interrompido?

Utilizador 1: Sim, para além de que me questionava várias vezes se já tinha feito referência a determinados pormenores quando efetuava o relato para o gravador para depois ser transcrito pela Administrativa e então constantemente afirmava na gravação "Peço desculpa se já referi isto, mas não me recordo". Perdia-se muito tempo e envolvia mais disponibilidade por parte da Administrativa, tornando os áudios produzidos mais extensos.

Entrevistadora: Depois de realizar a disseção e o respetivo ditado o relatório é elaborado em tempo real. Sente que consegue dedicar mais tempo à parte técnica/clínica?

Utilizador 1: Sim porque como enquanto eu estou a ditar já estou a dissecar a peça, não dito por partes, é muito mais fluido. O processo é muito mais contínuo, sem interrupções e quebras de raciocínio. Para além disso, imagine que já ditei tudo, mas depois corto uma fatia mais fina do fígado e descubro um nódulo, em vez de ter de referir novamente no áudio, tendo de voltar atrás na análise, que a Administrativa deve acrescentar mais uma determinada informação, com o Invox é mais fácil. Eu coloco o cursor na parte do ditado que pretendo detalhar mais e dito essa informação adicional. Ou seja, mesmo que depois de analisar a peça haja um novo achado é mais fácil acrescentar essa informação do que referir novamente no áudio para a Administrativa que numa determinada parte anterior do discurso deve acrescentar uma dada informação.

Entrevistadora: Como funcionou a integração com o sistema informático ANAPAT?

Utilizador 1: Não correu bem, mas adaptámos-mos. Quando obtemos a transcrição no Invox e efetuamos algumas correções ele aprende mais rápido se aceitarmos o ditado. Como existe esta incompatibilidade, em vez de aceitarmos o ditado na própria ferramenta temos de fazer copy past da transcrição para o ANAPAT. Se aceitarmos logo no software speech-to-text corremos o risco de ele não passar para os registos clínicos do doente e inclusive

quando tentamos fazer essa passagem podemos perder a transcrição. Infelizmente não estamos a utilizar o potencial máximo da ferramenta devido a este problema, caso o fizéssemos este aprenderia muito mais rápido.

Entrevistadora: Então podemos afirmar que o software não está a obter feedback da vossa parte para aprender a corrigir os erros de transcrição pois como não existe essa ligação correta com o ANAPAT vocês não aceitam a transcrição corrigida, mas sim fazem copy-past do texto que foi elaborado pelo Invox e corrigido por vós?

Utilizador 1: Exato. Como não queremos perder o texto e, inclusive, já aconteceram situações em que tentámos aceitar a transcrição do Invox, mas depois ficámos cerca de 30 segundos à espera que essa passagem de informação ocorresse e a nossa aplicação ANAPAT foi abaixo, não suportando esta partilha de informação, acabámos por contornar o problema desta forma.

Entrevistadora: O que acha da transcrição? É precisa e fiável? Que tipos de erros surgem? São mais relacionados com terminologia médica, pontuação e ortografia ou estruturação de frases?

Utilizador 1: O erro mais comum, que já foi corrigido, é que o software acrescentava o número 1 ao peso que mencionávamos, por exemplo se disséssemos que um órgão pesava 50 gramas ele transcrevia para 150 gramas, um erro grave. Outro erro é quando começamos uma frase por "Não se identificam lesões" o software assume "Identifica-se". Isto apenas acontece quando começamos a frase pela negativa, não sei se é porque há pessoas que sibilam mais que outras e não pronunciam tão bem o "Não". Outro erro que me recordo é quando dizemos "limites bem definidos" e o software assume apenas como "limites definidos", tornando a informação mais dúbia, porque é necessário perceber se estão ou não bem definidos para um correto diagnóstico.

Entrevistadora: Sente que perde muito tempo a corrigir estes erros?

Utilizador 1: Como tenho um Técnico de apoio, enquanto eu estou a dissecar e a ditar e, o Invox gera a transcrição, o Técnico vai corrigindo caso seja necessário. Tentamos otimizar ao máximo o tempo, sendo que a função desse Técnico é no fundo um controlo de qualidade. Mas no geral, não sinto que haja uma grande perda de tempo porque os erros não são muito frequentes.

Entrevistadora: Acha que os relatórios criados pelo Invox apresentam mais qualidade, por exemplo em termos de completude?

Utilizador 1: Sim porque quando eu dito digo muitas coisas que, se efetivamente soubesse que seria alguém a escrever manualmente, acharia que para não sobrecarregar esse colega podia encurtar a análise. Sinto que sou muito mais detalhada com o software.

Entrevistadora: A experiência dos seus colegas também tem sido positiva?

Utilizador 1: Sim. Inclusive sinto que para a Administrativa reduziu muito o seu stress laboral porque antes acumulava muitas funções.

Entrevistadora: Recomendaria esta ferramenta a outros colegas?

Utilizador 1: Sem dúvida.

Entrevistadora: Sente que há algum aspeto que deveria ser melhorado ou alguma funcionalidade que deveria ser acrescentada ao Invox?

Utilizador 1: Por exemplo pré-textos. Nas biopsias gástricas ajudaria muito. Nestes casos vêm sempre 4 tubos e nós temos de ser muito repetitivos em relação aos títulos – por exemplo temos sempre de iniciar dizendo "abrir aspas, ativar maiúsculas, nome do tubo, dois pontos". Se disséssemos gástricas 4 tubos e ele automaticamente reconhecesse que precisa de gerar os títulos sempre da mesma forma iria tornar a abordagem mais fluida.

Entrevista 2 (Patologista)

Entrevistadora: Há quanto tempo é que utiliza o Invox?

Utilizador 2: Há cerca de 3 meses.

Entrevistadora: Como descreve a sua experiência de forma geral até agora?

Utilizador 2: Para mim foi uma grande mais-valia. A minha letra é difícil de entender e escrevia muito. A acrescentar que tenho artroses devido à minha atividade profissional, com tendência a piorar, logo quando escrevia à mão tinha muita dificuldade, a letra estava a ficar cada vez mais ilegível e, portanto, detetava muitos erros nos relatórios após a transcrição da Administrativa para formato digital, por não perceber exatamente o que eu tinha escrito manualmente. Desde que utilizo o Invox, eu dito, vejo logo se existe algum erro ou não e, por isso, o relatório é muito mais rápido, eu tenho muito menos trabalho e liberto muito mais a Administrativa. Para além disso os erros são escassos, pontualmente numa ou outra palavra. Foram só vantagens.

Entrevistadora: Recebeu alguma formação para funcionar com o software?

Utilizador 2: Sim, veio ao Serviço uma representante da Quattro que nos deu as regras básicas de funcionamento e nos forneceu uma folha com algumas orientações para podermos trabalhar com a ferramenta. Eu confesso que não sou muito expert em

informática, uso o básico do software, sei que tem outras funcionalidades que ainda não tive tempo para aprender, mas só a parte básica já me ajuda imenso.

Entrevistadora: Acha que a ferramenta é simples de funcionar e intuitiva?

Utilizador 2: Sim, nesta parte mais básica da ferramenta que mencionei é bastante simples de funcionar, inclusive para alguém que não tem grandes bases em termos de conhecimento digital.

Entrevistadora: O tempo que dedica agora à elaboração dos relatórios é então muito menor?

Utilizador 2: Sim, sem dúvida. A partir do momento que eu acabo o ditado, num espaço de segundos fica fechado, ou seja, fica elaborado e disponível logo em tempo real.

Entrevistadora: O tempo que agora é poupado na elaboração dos relatórios, consegue dedicá-lo mais à parte clínica?

Utilizador 2: Consigo ver muito mais análises em menos tempo e, portanto, ter as coisas mais em ordem. O workflow melhorou muito.

Entrevistadora: Como foi a integração com o sistema informático ANAPAT?

Utilizador 2: Isso foi péssimo. Eu desisti confesso. Inicialmente eu importava o relatório, ou seja, aceitava o ditado para ver se ele passava para o nosso sistema, mas depois desisti porque nem sempre gravava e, inclusive aconteceu por várias vezes o Invox ficar a pensar durante muito tempo e, por incompatibilidade com o ANAPAT, apagava todo o relatório e era perdida a transcrição. Atualmente após a transcrição do meu ditado ser elaborada, eu faço copy past para o campo do ANAPAT. Infelizmente essa funcionalidade não está a trabalhar. Se nós aceitarmos o texto às vezes passa para o ANAPAT, outras vezes apaga simplesmente e, como não queremos correr esse risco, achámos mais seguro assim.

Entrevistadora: Sente que a transcrição com o software é mais precisa e fiável? Os erros que surgem estão mais relacionados com terminologia médica, erros ortográficos ou pontuação, estruturação de frases?

Utilizador 2: São erros muito pontuais que, pela minha perceção, surgem quando falo mais depressa. Sinto que os erros surgem mais quando digo números. Mas não considero erros graves.

Entrevistadora: Sente que perde muito tempo a corrigir estes erros?

Utilizador 2: Não de todo. E sinto que perco muito menos tempo do que antes, quando corrigia um relatório transcrito pela Administrativa. Ela (a Administrativa) sentia muita dificuldade em perceber a minha letra e interrompia-me algumas vezes quando estava a analisar outras amostras para me questionar o que estava escrito em determinados relatórios

ou, no final das minhas análises, enumerava todas as dúvidas de todos os relatórios que estavam pendentes. Era uma perda de tempo para mim e para ela.

Entrevistadora: Sente que agora os seus relatórios têm mais qualidade, no sentido de completude?

Utilizador 2: Tenho a noção que sim, atendendo a que quando a pessoa escreve, sobretudo quando tem dores na mão como é o meu caso, inconscientemente tem tendência a reduzir aquilo que escreve. Neste momento consigo ditar tudo. Não é que antes os relatórios não tivessem a informação mais importante, mas agora são muito mais detalhados. Eu, ao final de um dia de trabalho, tinha mesmo muitas dores na mão, a letra ficava cada vez pior e a tendência era escrever cada vez menos. Este problema foi agora contornado.

Entrevistadora: O feedback que tem dos seus colegas tem sido positivo?

Utilizador 2: Tem sido positivo, sobretudo entre as pessoas mais velhas aqui no Serviço. Penso que as pessoas mais velhas tiram mais partido disto porque nós estamos habituados desde sempre (estou aqui há quase 40 anos), a ditar e alguém escrever inclusive à máquina! Tem sido uma transformação brutal. Os meus colegas mais novos, de gerações mais recentes, funcionam muito melhor com computadores e, portanto, eles já digitavam muito rápido no computador e por vezes preferiam ser eles a escrever o próprio relatório no processo do doente em vez de ditar e, depois, a Administrativa transcrever ou de escrever à mão para depois a Administrativa transcrever no computador.

Entrevistadora: Recomendaria esta ferramenta a outros colegas?

Utilizador 2: Eu recomendo, sobretudo aos colegas da minha geração que escrevem ainda manualmente as análises, são os que vão tirar mais partido disto. Aos mais novos, reconheço que possa não ser uma transformação assim tão grande, mas considero que ainda assim com o Invox o processo deve ser mais rápido.

Entrevistadora: Sente que algo deve ser melhorado no software?

Utilizador 2: Sobretudo a incompatibilidade com o ANAPAT. Como ainda não explorei todas as funcionalidades do programa, penso que as minhas dificuldades podem estar relacionadas com isso. Quero salientar que nunca pensei ter um programa que transcreveria aquilo que eu dito, estou muito contente com esta transformação digital. Sou uma entusiasta deste software!

Entrevista 3 (Patologista)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 3: Há quase 1 mês.

Entrevistadora: E como considera a sua experiência?

Utilizador 3: Pessoalmente não é positiva. Tenho tido algumas dificuldades. O software não me percebe bem, penso que seja pelo meu sotaque visto que sou brasileiro. Tenho de fazer bastantes correções.

Entrevistadora: Recebeu alguma formação para a utilização do software?

Utilizador 3: Não recebi nenhuma formação da empresa que comercializa, mas recebi alguma orientação das colegas Técnicas.

Entrevistadora: Sente que o tempo que dedica, atualmente, à elaboração dos relatórios foi reduzido ou no seu caso não, uma vez que pode perder mais tempo com a correção dos erros?

Utilizador 3: Ainda que a minha perceção do trabalho dos outros colegas seja que o software aumenta a fluidez do processo, eu em particular não tenho essa experiência. Acho que a ferramenta tem muito potencial, porque com o ditado direto para o software penso que seja mais rápido do que ditar pausadamente para uma colega escrever. No meu caso em particular, como tenho de verificar constantemente se o Invox está a percecionar as palavras que lhe estou a dizer, sinto que o tempo de elaboração acaba por ser mais ou menos o mesmo. O tempo entre realizar o exame e ter o relatório é mais rápido, mas como perco mais tempo a corrigir, no geral não sinto nem que lentifique o processo nem que melhore.

Entrevistadora: Acha que o software é intuitivo e prático?

Utilizador 3: Parece-me que sim, porque digamos que o que eu vejo do software quando estou a utilizar é apenas uma caixa de texto que ele abre, enquanto estamos a fazer o ditado. Nós dizemos "Iniciar ditado" e o software abre de imediato uma caixa de texto por cima do programa ANAPAT, nós narramos e ele transcreve. Quando digo "Terminar ditado" ele produz a transcrição e nós fazemos copy past para o campo do ANAPAT pretendido.

Entrevistadora: Sente que o Invox está bem integrado com o vosso sistema informático?

Utilizador 3: Não, porque infelizmente o texto produzido pelo Invox não passa automaticamente para o programa que utilizamos. Inclusive, como não estamos a aceitar a transcrição que o software produz ele não está a aprender com os erros e isso no meu caso específico seria útil uma vez que ele não me percebe como aos outros colegas. Outra coisa que eu noto é que, para o nosso Diretor de Serviço, que utiliza bastante o software, a transcrição acaba por ser mais assertiva porque ele tem um gabinete próprio. No meu caso eu tenho um gabinete partilhado, chegando a ter 4 internos na mesma sala, o que influencia logo a qualidade da transcrição.

Entrevistadora: Considera que transcrição criada é precisa e fiável?

Utilizador 3: No meu caso não. Eu tenho entendido ao longo deste mês os termos que ele costuma transcrever errado e tento adaptar-me.

Entrevistadora: Sente que os erros ocorrem mais em terminologia médica, pontuação e ortografia ou estrutura de frases?

Utilizador 3: Um pouco de tudo. Ainda que, o programa tem tendência para utilizar mais terminologia médica, por exemplo se eu utilizar palavras mais comuns e o Invox não percebe, o software interpreta como um termo médico.

Entrevistadora: Podemos afirmar que perde muito tempo na revisão dos ditados?

Utilizador 3: Sim, por isso é que considero que o tempo que demoro a rever ou o tempo que anteriormente demorava a ditar, de forma pausada, para uma pessoa transcrever, acaba por ser equivalente.

Entrevistadora: Em termos de qualidade, sente que consegue produzir relatórios mais completos com o Invox?

Utilizador 3: Penso que a qualidade se manteve no meu caso, porque no final de contas, tem de existir sempre uma pessoa a reler e corrigir a transcrição.

Entrevistadora: Acha que a implementação teve algum impacto no stress laboral?

Utilizador 3: Penso que em termos de stress é equiparável. Antes quando eu ditava e a pessoa não percebia bem o que estava a dizer, então já perdia algum tempo a corrigir. Nesse aspeto o software manteve a dinâmica. Ou seja, para mim o nível de stress manteve-se.

Entrevistadora: Recomendava esta ferramenta a outros colegas?

Utilizador 3: Sim, principalmente porque eu vejo que há potencial, nas condições ideais, por exemplo quando o profissional tem um gabinete exclusivo. Apesar de a minha análise ser mais negativa, eu acho que devido ao sotaque não estou a utilizar a ferramenta no seu potencial máximo. Para além disso, se conseguíssemos que os dois sistemas se coordenassem entre si, o Invox e o ANAPAT, com os mecanismos de aprendizagem que o software tem, conseguiríamos tirar o seu partido máximo. Em Portugal sei que também existem muitos sotaques diferentes, pelo que se o nosso sistema informático fosse mais colaborativo com o ANAPAT e pudéssemos aplicar os mecanismos de aprendizagem da ferramenta ajudaria muito. Como disse, eu tenho entendido ao longo deste mês os termos que ele costuma transcrever errado. Se houvesse um mecanismo de feedback mais prático que pudéssemos aplicar para que ele aprendesse, já que não conseguimos propriamente aceitar o ditado devido à dificuldade de interoperabilidade, seria útil.

Entrevistadora: Considera que o software deva ter alguma funcionalidade adicional?

Utilizador 3: O feedback mais simplificado como mencionei, acho que seria um ganho.

Entrevista 4 (TSDT)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 4: Cerca de 2 anos.

Entrevistadora: Como descreve a sua experiência global com a utilização da ferramenta?

Utilizador 4: Foi positiva. Existem pequenas coisas que ainda têm de ser aferidas, mas penso que faz parte. O fluxo de trabalho melhorou muito, quer entre nós, quer para a Administrativa, porque muitas vezes elas tinham dúvidas em relação ao nosso ditado e agora esse problema foi resolvido.

Entrevistadora: Recebeu alguma formação para utilizar o software?

Utilizador 4: Sim, pela empresa Quattro.

Entrevistadora: Acha o software prático e intuitivo?

Utilizador 4: Sim, é muito simples de utilizar. Houve algumas dúvidas durante a utilização, mas foram sempre esclarecidas pela empresa.

Entrevistadora: Nota alguma diferença no tempo que dedica à elaboração dos relatórios?

Utilizador 4: Sim, porque antes tínhamos de estar sempre a pausar e ativar a gravação, e agora é um processo muito mais contínuo.

Entrevistadora: O tempo entre realizar o exame e ter o relatório é agora mais rápido?

Utilizador 4: Sim, porque muitas vezes existiam problemas em conseguir passar a informação à Administrativa com um ditado percetível para a mesma, temos todos dicções diferentes o que dificultava o seu trabalho. O Invox é agora mais prático, tem uma melhor perceção do nosso ditado, salvo algumas exceções, mas são mesmo pontuais.

Entrevistadora: Sente que o tempo que dedicava em interromper o trabalho para esclarecer as dúvidas da Administrativa é neste momento otimizado para se dedicar mais à parte mais técnica do seu trabalho?

Utilizador 4: Sim, estamos muito mais focados agora. Por vezes tínhamos de fazer a alteração de uma descrição de uma peça, então no áudio referíamos que tínhamos de voltar atrás ao período x, e voltávamos a descrever a peça, o que se tornava mais confuso para a Administrativa. Atualmente paramos o ditado, colocamos o cursor no sítio que queremos alterar e descrevemos novamente a peça, o que torna o processo muito mais simples. O

facto de fazermos o ditado, obtermos de imediato a transcrição e podermos colocar logo no campo correto do nosso sistema informático facilita muito mais o processo.

Entrevistadora: Como foi a integração com o vosso sistema informático?

Utilizador 4: Foi e ainda é má. Torna o processo menos produtivo. Há pequenos erros que o software comete que seriam muito fáceis de ser aprendidos se pudéssemos aceitar a transcrição, mas como não existe compatibilidade com o nosso sistema, o Invox não aprende. Por exemplo, constantemente dizemos "fragmentos em cunha", e o software reconhece que cunha é um apelido e escreve com letra maiúscula. Este erro não tem influência no diagnóstico e, se pudéssemos corrigi-lo aceitando a transcrição, o Invox aprenderia facilmente a não voltar a cometer a mesma gralha.

Entrevistadora: Sente que a transcrição é fiável? Quais são os tipos de erros que o software comete mais: é a terminologia médica, ortografia, pontuação ou estrutura de frases?

Utilizador 4: Um dos erros mais frequentes é nas medidas. Por exemplo nós referimos muitas vezes que um determinado órgão mede "23x14x5" e ele nem sempre assume o termo "por" como um "x". Aqui vamos bater no mesmo assunto da aprendizagem do software. Se fosse compatível com o ANAPAT isto era facilmente colmatado.

Entrevistadora: Sente que perde muito tempo a corrigir os erros?

Utilizador 4: Penso que não tem um impacto significativo. Como os erros são repetitivos por o software não estar a aprender connosco, corrigir estas pequenas coisas otimizaria o processo, ainda assim não sinto que haja uma grande perda de tempo.

Entrevistadora: Sente que a qualidade dos relatórios é melhor com a implementação do software, por exemplo em termos de completude?

Utilizador 4: Eu acho que não teve muito impacto nesse sentido, porque apesar de o tempo do processo ser diferente e mais facilitado, nós já ditávamos a mesma informação que depois era transcrita pela Administrativa. Penso que o real impacto é em termos de otimização de tempo e não da qualidade dos relatórios produzidos.

Entrevistadora: Sente que feedback dos seus colegas é positivo?

Utilizador 4: Sim, não sinto que os colegas tenham tido dificuldade em adaptar-se à ferramenta.

Entrevistadora: Recomendaria esta ferramenta a outros colegas?

Utilizador 4: Sim, sem dúvida.

Entrevistadora: Gostaria que algo melhorasse no software ou que existisse alguma funcionalidade adicional?

Utilizador 4: A comunicação com o ANAPAT é sobretudo o que eu mudaria, para que o software estivesse mais avançado e adaptado à nossa realidade. De resto estou muito satisfeita.

Entrevista 5 (Administrativa)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 5: Há cerca de 3 semanas porque apenas trabalho neste Serviço para fazer faltas e férias da Administrativa que habitualmente trabalha aqui. Habitualmente trabalho na Anatomia Patológica no Polo do Hospital São José que, infelizmente, não trabalha com esta ferramenta.

Entrevistadora: Durante a sua curta experiência, acha que a ferramenta tem um impacto positivo ou negativo?

Utilizador 5: Bastante positivo, principalmente comparando com o Serviço de Anatomia Patológica do Hospital de São José.

Entrevistadora: Sente que o stress laboral, decorrente da acumulação de funções, é reduzido com a utilização do software?

Utilizador 5: A minha carga de trabalho é bastante mais reduzida. Consigo mais tempo para utilizar noutras funções que não sejam a transcrição dos relatórios, o que é muito positivo. Para além disso, como cada Técnico e Médico escreve e dita à sua maneira, por exemplo nas abreviaturas, uns profissionais fazem de uma forma e outros doutra, o que nos confunde e induz em erro.

Entrevistadora: Costumava ter muitas dúvidas na realização da transcrição dos relatórios?

Utilizador 5: Sim, às vezes para não interromper constantemente o trabalho dos Técnicos e Médicos sublinhávamos as palavras em dúvida a vermelho e, depois quando devolvíamos a transcrição feita por nós eles tinham de ver se de facto a palavra em dúvida estava correta. Perdia muito tempo a desvendar as palavras por causa das letras menos legíveis ou então por não perceber o que ditavam. O software tem muitas vantagens neste sentido. Poupa imenso tempo e evita o erro. Muitas vezes, quando estava a tentar decifrar as palavras nos ditados, o telefone tocava e perdia completamente a linha de raciocínio, inclusive na construção das frases ou nas numerações. Acontecia-me também, quando estava a transcrever o relatório feito manualmente pelo Médico, o telefone tocava ou chegava uma nova inscrição e eu saltava uma linha de texto e perdia-me completamente.

Entrevistadora: Recomendaria isto a outros colegas?

Utilizador 5: Sem sombra de dúvida! Eu ando a fazer pressão para que no Hospital de São José comecem a utilizar esta ferramenta porque é maravilhosa.

Entrevista 6 (TSDT)

Entrevistadora: Há quanto tempo é que utiliza o Invox?

Utilizador 6: Desde Fevereiro/Março (há 5/6 meses).

Entrevistadora: Como descreve a experiência com a ferramenta?

Utilizador 6: Tem sido uma experiência bastante positiva. Nós antes tínhamos um gravador e depois a Administrativa transcrevia tudo, o que lhe ocupava uma grande parte do dia só para fazer essa função. Agora, com o software, é mais simples porque transcreve rapidamente e nós conseguimos copiar e colar o texto para o processo do doente, poupando imenso tempo. É menos um passo e menos uma pessoa que tem de mexer no texto, o que ajuda bastante. Por vezes o software troca algumas palavras porque cada pessoa tem a sua forma de ditar, considero que não são erros graves. A nível de peças é espetacular, a nível de biópsias temos alguma dificuldade porque como vêm vários frascos e cada um está identificado com letras maiúsculas, o software nem sempre consegue interpretar bem o nosso ditado. Imagine, um frasco é referenciado como "EDA1", nós tentamos dizer de várias formas, mas o software não interpreta corretamente o que atrapalha o processo pois temos de corrigir. Apesar disto, este erro tem um impacto mínimo.

Entrevistadora: Recebeu alguma formação para trabalhar com o Invox?

Utilizador 6: A formação decorreu durante 1 dia em que uma representante da empresa veio explicar o produto e explicar alguns comandos. Não é uma aplicação difícil, é bastante intuitiva. Nunca tivemos muitos problemas com o seu funcionamento, o maior problema está relacionado com a compatibilidade do Invox com o ANAPAT. Os dois programas não comunicam corretamente, por isso tivemos de dar a volta à questão. Quando nós ditamos aparece uma caixa de texto do Invox, que transcreve o que foi dito e nós depois corrigimos. Para o Invox aprender deveríamos "Aceitar texto", para este assimilar tudo o que corrigirmos, inclusive a adaptar o texto produzido à nossa dicção e linguagem. Como nós temos de fazer copy-past e não aceitamos o texto criado, o software não aprende. Se aceitarmos o texto o que acontece é que o ANAPAT fica a processar durante muito tempo e, não só o texto não é partilhado de um para outro sistema como muitas vezes se perde o trabalho de correção que efetuámos e temos de voltar à transcrição feita pelo Invox e corrigir novamente. Nesse caso demoramos o dobro do tempo a efetuar as correções. Não

considero que esta incompatibilidade seja um problema do Invox, mas sim da nossa informática.

Entrevistadora: Nota alguma diferença no tempo que dedica à elaboração do relatório?

Utilizador 6: Como não éramos nós que transcrevíamos o relatório, mas sim a Administrativa, onde noto mais diferença é na correção do relatório que ela elaborava. Como só temos uma Administrativa, ela tinha de acumular muitas funções e a transcrição acabava por demorar muito mais tempo e, consequentemente a elaboração do relatório. Para além disso, por vezes no gravador carregávamos em botões de forma não propositada e perdíamos o texto e, tínhamos de voltar a refazer a gravação toda. Com o Invox, os áudios ficam todos registados. Mesmo que tenha erros, a transcrição fica registada e não se perde como no gravador.

Entrevistadora: O tempo entre fazer o exame e emitir o relatório foi alterado?

Utilizador 6: O tempo útil de saída do relatório é muito menor. Antes a Administrativa tinha de transcrever, imprimir, entregar-nos para fazermos as correções e depois anexar ao processo. Agora a Administrativa só tem de imprimir a transcrição que já foi corrigida por nós e anexar ao processo do doente.

Entrevistadora: A utilização da ferramenta e poupança de tempo permitiu a que se dedicassem mais a outras competências técnicas?

Utilizador 6: Sim, porque a Administrativa após reunir algumas dúvidas que tinha nos nossos ditados, tinha de nos abordar para nos questionar e tínhamos de interromper o nosso trabalho e perder mais tempo. Neste momento, logo após o ditado nós corrigimos o que temos a corrigir, não precisamos de interromper várias vezes o nosso trabalho, logo flui muito melhor e dedicamos-mos mais à parte técnica do que burocrática.

Entrevistadora: Sente que a transcrição é fiável e precisa? Os erros que mais se sucedem estão relacionados com terminologia médica, ortografia, pontuação ou estrutura de frases?

Utilizador 6: Os erros que geralmente surgem penso que estão relacionados com a dicção. Os problemas mais frequentes são o software acrescentar o algarismo 1 ao peso ou medidas que nós ditamos, e o programa assumir que quando digo, por exemplo, “10x20x30” é para escrever por extenso “por” em vez de “x”. Creio que apesar destes erros a transcrição é fiável, tendo em conta o ruído sentido nas nossas salas seja de campainhas a tocar ou o barulho de fundo do extrator de formol. São erros que facilmente são corrigidos. Outro erro que por vezes se sucede é quando lhe damos um comando, por exemplo para o software

assumir aspas ou numeração romana, ele fica a pensar e não assume de imediato o que dizemos a seguir, mas penso que é fácil nos adaptarmos a este erro, basicamente é dar tempo à ferramenta que processe o comando e depois é que continuamos o ditado.

Entrevistadora: Sente que os relatórios com o software têm mais qualidade, por exemplo em termos de completude?

Utilizador 6: Na minha perspetiva têm muito mais qualidade. Sinto que não me perco tanto no meu discurso porque como visualizo logo a transcrição vejo se abordei todos os pontos essenciais para o diagnóstico. Quando ditávamos para um gravador por vezes não me recordava se já tinha falado num ou noutro ponto então ou repetia novamente ou por lapso não o abordava. O facto de termos acesso ao texto em tempo real é muito vantajoso.

Entrevistadora: O feedback dos seus colegas também tem sido positivo?

Utilizador 6: Sim, inclusive dos Médicos que têm menos capacidades informáticas. Acelerou o processo de todos, aumentou a qualidade dos relatórios e retirou uma grande carga de trabalho à Administrativa.

Entrevistadora: A implementação e adaptação envolveu mais stress laboral?

Utilizador 6: Para mim foi bastante tranquilo, talvez porque na minha geração é mais fácil dominar a parte digital. Para os colegas mais velhos que não têm tanta aptidão informática talvez tenha sido mais stressante, ainda assim acho a ferramenta tão intuitiva que não acho que tenha sido um grande problema.

Entrevistadora: Recomendaria o software a outros colegas?

Utilizador 6: Sim. Eu tive oportunidade de trabalhar também na Anatomia Patológica do Hospital de São José e lá eles escrevem à mão, o que torna tudo menos percetível, com rasuras e chavetas. A informação é muito mais confusa. A única questão que se coloca é que no Hospital de São José existem duas mesas de Macroscopia, ou seja, dois Técnicos a ditar em simultâneo e não sei se essa logística não influencia a qualidade da transcrição. O software não sabe distinguir vozes, ele simplesmente ouve e transcreve, logo seria um problema.

Entrevistadora: O que gostava que fosse melhorado ou alterado? Alguma funcionalidade adicional que se possa acrescentar?

Utilizador 6: Em termos de funcionalidades eu acho que o software já é bastante completo, nós é que, decorrente da incompatibilidade com o ANAPAT, acabamos por não retirar o máximo partido da ferramenta. O que melhorava era talvez a distinção de vozes de utilizadores, permitiria que não existissem interferências e os relatórios pudessem ser realizados em simultâneo por vários profissionais.

Entrevista 7 (TSDT)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 7: Desde há 2 anos.

Entrevistadora: Como descreve a sua experiência com esta ferramenta?

Utilizador 7: No geral a ferramenta é muito positiva. Ao início ele não reconhecia a minha voz, mas agora já está mais afinado.

Entrevistadora: Recebeu alguma formação para a utilização do software?

Utilizador 7: Eu não tive formação, mas como outros colegas tiveram passaram-me os conhecimentos aprendidos.

Entrevistadora: Nota alguma diferença no tempo que dedica à elaboração dos relatórios?

Utilizador 7: Acho que só não é mais rápido porque como o software não está a aprender pois não aceitamos a transcrição do Invox, temos de corrigir erros repetitivos que caso esta incompatibilidade fosse resolvida já não era necessário. O que para mim é maravilhoso é não ter de escrever, isso torna o processo bem mais rápido.

Entrevistadora: Sente que consegue dedicar mais tempo à parte técnica, agora que a parte burocrática foi reduzida?

Utilizador 7: Sim. O Invox não veio retirar trabalho, veio otimizar os processos, sem tanta entropia. Havia mais desordem, pormenores que faltavam, a minha letra era péssima, para além de que sentia muita dor na mão e ao final do dia a minha letra já era praticamente ilegível. Foi uma mais-valia.

Entrevistadora: Considera que o software é fácil e intuitivo?

Utilizador 7: Sim muito fácil de utilizar. Mesmo não tendo feito nenhuma formação consegui adaptar-me muito bem à ferramenta.

Entrevistadora: O software está bem integrado com o vosso sistema informático?

Utilizador 7: Não, infelizmente não está. Era uma mais-valia se estivesse, para que ele aprendesse com as nossas correções.

Entrevistadora: Considera que a transcrição é precisa e fiável? Que tipo de erros costumam surgir?

Utilizador 7: Na minha experiência quanto mais difícil é a palavra mais o software percebe. Na terminologia médica é maravilhoso. Em termos de pontuação e ortografia nem sempre é bom. O erro mais frequente é acrescentar medidas, por exemplo o algarismo 1. Também

em relação às medidas, o software nem sempre compreende. Às vezes quando dizemos "por" na verdade a ferramenta deveria assumir como "x", mas nem sempre isso acontece. Por isso é que reforço que é um handicap o Invox não estar a aprender, porque estes erros são facilmente solucionáveis.

Entrevistadora: Acha que os relatórios têm agora mais qualidade? São mais completos?

Utilizador 7: Para mim, nesse sentido, contínua igual. Acho que não está relacionado com o software, mas sim com o operador, ou seja, a pessoa que está a ditar.

Entrevistadora: O feedback que recebe dos seus colegas também é positivo?

Utilizador 7: Depende se estamos a falar de alguém que tem um à-vontade com a tecnologia ou não. Há pessoas que têm mais facilidade em trabalhar com a parte digital. Há pessoas mais adversas à digitalização e à mudança.

Entrevistadora: Sente que a ferramenta reduziu o stress laboral?

Utilizador 7: Sim, sobretudo para a Administrativa. A única coisa que me stressa e faz com que o processo não seja tão fluído é o facto de ele não estar a aprender. O tempo que perdemos a corrigir podíamos estar a fazer outras coisas. Isto precisa urgentemente de ser otimizado.

Entrevistadora: Recomendaria esta ferramenta a outros colegas?

Utilizador 7: Sim, ainda que alguns colaboradores possam oferecer mais resistência exatamente por não serem digitalmente tão aptos.

Entrevistadora: O que deveria ser melhorado e que funcionalidades adicionais gostava que o Invox tivesse?

Utilizador 7: Melhorado seria a compatibilidade com o nosso sistema informático para que o software aprendesse, para eu não corrigir regularmente o básico. Pelo que percebi a fusão implica que o Invox "invada" o ANAPAT que tem antivírus e estratégias de segurança que não permitem ao software partilhar informação. Em relação a funcionalidades adicionais, acho que o programa já é bastante completo, por isso estou satisfeita. Na minha opinião já temos uma ferramenta muito boa, mas sabemos que ainda pode ser melhor, nós queremos sempre mais, logo se conseguíssemos que existisse interoperabilidade ficávamos muito satisfeitos.

Entrevista 8 (Patologista)

Entrevistadora: Há quanto tempo utiliza o Invox?

Utilizador 8: Uso há 2 meses.

Entrevistadora: Como descreve a sua experiência?

Utilizador 8: É muito boa. Notei um grande impacto na celeridade dos relatórios.

Entrevistadora: Recebeu alguma formação para a utilizar o software?

Utilizador 8: Não formalmente. Os colegas explicaram-me como funcionava e como é muito intuitiva foi fácil aprender.

Entrevistadora: Nota alguma diferença no tempo que dedica à elaboração dos relatórios?

Utilizador 8: Sim, eu só utilizo o Invox para os relatórios de Macroscopia. Eu faço o processo em 2 tempos, analiso a peça e depois descrevo e tenho logo a transcrição. Antigamente eu analisava a peça, descrevia, e o técnico escrevia manualmente o que eu ditava. Depois esse relatório manual era encaminhado para a Administrativa e retornava para correções. Agora é mais fácil, reduzindo os passos necessários. Há quem inclusive enquanto vê a peça a esteja a descrever e ditar.

Entrevistadora: Podemos então afirmar que o tempo entre realizar o exame e a emissão do relatório sofreu alterações significativas?

Utilizador 8: É mais rápido e mais cómodo. O médico acaba por conseguir ver mais exames devido à otimização dos tempos.

Entrevistadora: Sente que há uma boa integração com os vossos sistemas informáticos?

Utilizador 8: Não, idealmente a transcrição deveria ser diretamente partilhada do Invox para o ANAPAT, mas não é isso que está a acontecer. Nós temos de fazer copy past da transcrição criada para o campo correto do nosso sistema informático.

Entrevistadora: Sente que a transcrição é precisa e fiável? Costuma corrigir muitos erros?

Utilizador 8: Eu diria que é 90% fiável. No meu caso os erros estão mais relacionados com terminologia médica. Há palavras semelhantes que o software interpreta de uma maneira diferente. No entanto é pontual, não costumo ter muitas correções a fazer.

Entrevistadora: Acha que os relatórios criados pelo Invox têm uma qualidade superior?

Utilizador 8: Sim, existe mais fluidez da descrição. Antes eu tinha de ditar frase a frase e tinha de parar para dar tempo à pessoa que estava ao meu lado escrever manualmente. Sentia que quando este processo acontecia existiam mais erros em pesos. Agora consigo ditar do início ao fim, sem fazer pausas.

Entrevistadora: O feedback dos seus colegas é positivo?

Utilizador 8: É muito positivo.

Entrevistadora: Recomendaria o software aos seus colegas?

Utilizador 8: Sim, inclusive para o ano vou continuar o internato no Hospital de São José e vou sentir que vou voltar uns passos atrás.

Entrevistadora: O que acha que deve ser melhorado ou acrescentado à ferramenta?

Utilizador 8: Acho que seria interessante que o software tivesse textos pré-definidos. Existem certas coisas que dizemos que são muito repetitivas, por exemplo, para descrever uma gastrite crónica, nós temos de dizer em que parte do estômago é e qual a intensidade, e é sempre o mesmo esquema. Eu podia ter um comando no software para "texto pré-definido gastrite crónica" e automaticamente o Invox ia buscar esse texto e eu só tinha de alterar as medidas. Facilitava ainda mais a fluidez do processo.

Anexo D – Respostas às entrevistas realizadas aos colaboradores representantes da Empresa Quattro

Como surgiu a parceria da Quattro com o Invox Medical Dictation em Portugal?

Quando integrámos a Quattro esta solução já estava identificada, contudo dado o interesse e a diferenciação da solução perante outras de mercado e a recetividade de potenciais utilizadores, fez todo o sentido a relação de parceria.

Quais consideram ser os principais diferenciais competitivos do software face a outras soluções speech-to-text no mercado da saúde?

O Invox Medical Dictation destaca-se das diferentes soluções do mercado, principalmente pelo facto de ter dicionários especificamente afetos a cada especialidade médica com léxico complexo e nomenclatura própria de cada valência da medicina. Esse detalhe torna o reconhecimento dos termos ditados, mais preciso e adequado. Para além disso, disponibiliza a possibilidade de optar por licenciamento nominal (licenças para até 20 utilizadores diferentes) ou concorrencial (licenças ilimitadas, mas apenas 20 utilizadores podem aceder ao *software* em simultâneo), o que possibilita a comercialização a clínicas de pequena dimensão, mas também a grandes instituições hospitalares.

Que feedback têm recebido dos clientes em diferentes especialidades médicas para além da Anatomia Patológica?

O feedback recebido até à data é muito positivo. De todas as especialidades implementadas até à data, a Radiologia destaca-se pelo volume de licenças que já disponibilizámos e porque é uma especialidade médica que já utiliza o reconhecimento de voz há muito tempo, o que faz com que tenham muitos modelos para comparação. Nesses casos, os profissionais de saúde têm destacado a precisão de reconhecimento do Invox Medical Dictation em comparação com os restantes softwares já testados.

Quais foram os maiores desafios encontrados na implementação do Invox no Hospital Curry Cabral?

Podemos dizer que o processo de implementação da solução Invox Medical Dictation é relativamente simples, contudo cada projeto é único e, por vezes deparamo-nos com alguns desafios, muitas vezes a infraestrutura menos moderna é um entrave à performance do sistema. Contudo durante o projeto de implementação na ULS São José (Hospital São José e Hospital Curry Cabral) não nos deparámos com nenhum entrave técnico, a implementação correu bastante bem, sempre em coordenação com o departamento de

informática da instituição. Com o envolvimento das equipas conseguimos ultrapassar os desafios com os quais por vezes nos deparamos.

Que tipo de formação e acompanhamento a Quattro disponibiliza aos profissionais de saúde durante a fase de adoção da ferramenta?

A Quattro, no final de cada implementação, disponibiliza uma formação a todos os profissionais de saúde utilizadores da solução. Essa formação inclui uma breve apresentação, uma demonstração e esclarecimento de questões. Posteriormente, a Quattro disponibiliza um contacto de e-mail para o qual todos os profissionais podem expor as suas dúvidas ou problemas com a solução e encontra-se disponível para se deslocar ao local de implementação para resolução mais precisa e ajuda mais personalizada, se necessário, como aconteceu, por exemplo, no Hospital de S. José com uma médica Patologista que não estava familiarizada com novas tecnologias, pelo que foi necessária uma ajuda mais próxima.

Como a empresa lida com a resistência inicial à digitalização que alguns profissionais de saúde podem manifestar?

O processo de digitalização da saúde já decorre há algum tempo. Ainda assim, existem muitos profissionais de saúde que manifestam uma clara resistência a novas implementações. O volume de trabalho dos profissionais de saúde é muito considerável, pelo que a Quattro tem plena noção de que todas as soluções implementadas na área da saúde têm de ser intuitivas, fáceis de utilizar, que a curto prazo se demonstrem capazes de otimizar processos e que não criem constrangimentos iniciais. Assim sendo, em primeiro lugar, procuramos sempre ter em portfólio, soluções com essas características. Posteriormente, utilizamos essas características a nosso favor e promovemos relações de proximidade com os profissionais de saúde para que possam confiar na solução e na nossa empresa. A resistência é natural, lidamos com isso como um risco na fase de projeto, mas sabemos que estabelecendo uma relação de proximidade com os profissionais, esse risco diminui substancialmente e facilita muito a rápida adaptação.

A incompatibilidade com sistemas como o ANAPAT foi destacada pelos utilizadores. Que soluções a Quattro prevê para melhorar a interoperabilidade com sistemas hospitalares?

O sistema Invox Medical Dictation é compatível com todos os softwares hospitalares para inserção de texto. Seja em contexto de integração ou apenas como aplicação isolada. Quando o Invox MD encontra uma solução com a qual não é diretamente nativo, o mesmo abre uma janela auxiliar onde o ditado é transcrito e só posteriormente transita para o local

onde o cursor se encontrava no início do ditado. Contudo, e tal como referido, existe sempre a possibilidade da solução de RIS e LIS integrar o Invox Medical Dictation.

Quais são as perspetivas de evolução do Invox Medical Dictation nos próximos anos em Portugal?

O Invox Medical Dictation, propriedade da Vócali, está inserido num grupo de soluções de reconhecimento de voz. À data de hoje a Vócali tem, também, uma solução de transcrição de consulta, chamada Invox Medical Genesis, esta já foi uma evolução do Invox Medical Dictation. O Genesis tem como objetivo otimizar o processo de registo de informação clínica em âmbito de consulta. O sistema tem um módulo de Inteligência Artificial que analisa aquilo que foi dito na consulta, entre o médico e o paciente, e com base numa estrutura de relatório previamente configurada, extrai a informação relevante para inserir no processo clínico do utente. A Vócali continuará a evoluir e a melhorar as suas soluções, otimizando os modelos de linguagem e de IA a si associados e, consequentemente, a transcrição do ditado/reconhecimento de voz.

Anexo E – Respostas aos inquéritos concretizados no Serviço de Anatomia Patológica do Hospital Curry Cabral

	Utilizador 1	Utilizador 2	Utilizador 3	Utilizador 4	Utilizador 5	Utilizador 6	Utilizador 7
Categoria profissional	TSDT	TSDT	Patologista	Patologista	Patologista	TSDT	TSDT
Tempo de utilização	1-6 meses	> 1 ano	1-6 meses	1-6 meses	< 1 mês	> 1 ano	> 1 ano
O tempo necessário para elaborar os relatórios diminuiu	Concordo totalmente	Concordo totalmente	Concordo totalmente	Neutro	Neutro	Concordo totalmente	Concordo
O *software* contribui para a celeridade da emissão de relatórios	Concordo	Concordo totalmente	Concordo totalmente	Neutro	Concordo	Neutro	Concordo totalmente
O *software* é fácil de utilizar	Concordo	Concordo totalmente	Concordo	Concordo	Neutro	Concordo totalmente	Concordo totalmente
Encontra frequentemente erros de reconhecimento de voz	Ocasionalmente	Ocasionalmente	Ocasionalmente	Ocasionalmente	Frequentemente	Frequentemente	Frequentemente
A interface é intuitiva e adequada à sua prática	Sim	Sim	Sim	Sim	Parcialmente	Sim	Sim
A transcrição apresenta elevada precisão em termos médicos	Concordo	Concordo totalmente	Concordo	Concordo	Concordo	Concordo totalmente	Concordo
Necessita de realizar muitas correções após transcrição	Frequentemente	Raramente	Raramente	Frequentemente	Frequentemente	Frequentemente	Frequentemente
Confia na fiabilidade da transcrição criada pelo *software*	Parcialmente	Sim	Parcialmente	Sim	Não	Parcialmente	Sim
Está satisfeito com o desempenho geral do *software*	Satisfeito	Muito satisfeito	Muito satisfeito	Muito satisfeito	Indiferente	Satisfeito	Satisfeito
Recomendaria o uso a outros profissionais	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Período de adaptação	Muito curto	Curto	Adequado	Muito curto	Ainda não me adaptei	Curto	Curto
Recebeu formação para utilizar o *software*	Sim	Sim	Sim	Não	Parcialmente	Não	Sim
O *software* está bem integrado com o sistema informático	Discordo	Discordo	Discordo totalmente	Discordo	Discordo	Discordo totalmente	Discordo

Cálculos detalhados com base nos 7 inquiridos:

Escala utilizada	
Discordo totalmente	1
Discordo	2
Neutro	3
Concordo	4
Concordo totalmente	5

Escala utilizada	
Insatisfeito	1
Indiferente	2
Satisfeito	3
Muito satisfeito	4

Escala utilizada	
Não	1
Parcialmente	2
Sim	3

- Tempo de elaboração dos relatórios diminuiu

57% Concordo totalmente	4 respostas
14% Concordo	1 resposta
29% Neutro	2 respostas
Média	4,3/5

- Contribuição para a celeridade

43% Concordo totalmente	3 respostas
29% Concordo	2 respostas
29% Neutro	2 respostas
Média	4,1/5

- Facilidade de utilização

43% Concordo totalmente	3 respostas
43% Concordo	3 respostas
14% Neutro	1 resposta
Média	4,3/5

- Interface intuitiva

86% Sim	6 respostas
14% Parcialmente	1 resposta
Média	2,9/3

- Precisão termos médicos

71% Concordo	5 respostas
29% Concordo totalmente	2 resposta
Média	4,3/5

- Confiança na transcrição

43% Parcialmente	3 respostas
43% Sim	3 respostas
14% Não	1 resposta
Média	2,3/3

- Satisfação geral

43% Muito satisfeito	3 respostas
43% Satisfeito	3 respostas
14% Indiferente	1 resposta
Média	3,3/4

- Formação

57% Sim	4 respostas
29% Não	2 respostas
14% Parcialmente	1 resposta
Média	2,3/3

- Integração com ANAPAT

71% Discordo	5 respostas
29% Discordo totalmente	2 resposta
Média	1,7/5

Anexo F – Resumo das características dos dois métodos propostos para personalização de sistemas de reconhecimento de voz ao utilizador (*software one-to-one*).

Características	***i-vectors***	***x-vectors***
Base	Modelos estatísticos (GMM-UBM)	Redes Neuronais Profundas
Treino	Não supervisionado	Supervisionado (classificação do orador)
Informação captada	Orador + Canal (condições acústicas)	Orador (mais discriminativo)
Dependência do tempo	Estatísticas acumuladas (perde detalhes temporais)	Modelação explícita de contexto temporal
Robustez	Bom desempenho com pouco treino e dados limitados	Bom desempenho com grandes volumes de dados e treino supervisionado
Aplicação moderna	Utilizado sobretudo quando existe baixa disponibilidade de dados	Padrão atual para sistemas de alto desempenho