UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ (UTFPR)
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
DEPARTAMENTO ACADÊMICO DE ELETRÔNICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

FERNANDO ARGENTINO DA SILVA
NATAN SCHIECK REGINALDO

DISPOSITIVO DE REALIDADE AUMENTADA PARA VISUALIZAÇÃO DE CONTEXTO SONORO PARA INCLUSÃO DE INDIVÍDUOS COM PERDA AUDITIVA

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2022

FERNANDO ARGENTINO DA SILVA
NATAN SCHIECK REGINALDO

DISPOSITIVO DE REALIDADE AUMENTADA PARA VISUALIZAÇÃO DE CONTEXTO SONORO PARA INCLUSÃO DE INDIVÍDUOS COM PERDA AUDITIVA

Augmented reality sound context visualization device for the hearing impaired

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná (UTFPR), como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Bogdan Tomoyuki Nassu
DAINF - Departamento Acadêmico de Informática - UTFPR

CURITIBA
2022

4.0 Internacional

Esta licença permite *download* e compartilhamento do trabalho desde que sejam atribuídos créditos ao(s) autor(es), sem a possibilidade de alterá-lo ou utilizá-lo para fins comerciais. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

FERNANDO ARGENTINO DA SILVA
NATAN SCHIECK REGINALDO

DISPOSITIVO DE REALIDADE AUMENTADA PARA VISUALIZAÇÃO DE CONTEXTO SONORO PARA INCLUSÃO DE INDIVÍDUOS COM PERDA AUDITIVA

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia da Computação da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 03/Março/2022

_____

Bogdan Tomoyuki Nassu
Doutorado
Universidade Tecnológica Federal do Paraná

_____

João Alberto Fabro
Doutorado
Universidade Tecnológica Federal do Paraná

_____

André Eugênio Lazzaretti
Doutorado
Universidade Tecnológica Federal do Paraná

CURITIBA
2022

AGRADECIMENTOS

Somos gratos ao nosso professor e orientador Bogdan Tomoyuki Nassu por todo o apoio, assistência e disponibilidade, não somente durante o desenvolvimento deste trabalho, mas também ao longo de todo o curso.

Agradecemos a universidade e todos os docentes que nos acompanharam nesta jornada, além de familiares e amigos que sempre nos apoiaram.

The right man in the wrong place can make all the difference in the world.

— G-Man, Half-Life 2

RESUMO

DA SILVA, F. A.; REGINALDO, N. S.. Dispositivo de realidade aumentada para visualização de contexto sonoro para inclusão de indivíduos com perda auditiva. 2022. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná (UTFPR). Curitiba, 2022.

Recentemente diversas tecnologias têm sido propostas para auxiliar a comunicação e melhorar a qualidade de vida de pessoas com perda auditiva. Soluções tradicionais de tecnologia assistiva se baseiam principalmente na transformação de informações de áudio, limitando-se à amplificação e/ou compressão de faixa dinâmica de som. No entanto, esta abordagem tem desempenho reduzido para situações com ruído competitivo, não atendendo às necessidades dos indivíduos que dependem desses dispositivos para atividades de socialização. Consequentemente, métodos alternativos têm sido explorados. Com o objetivo de melhorar as experiências de comunicação interpessoal para indivíduos com perda auditiva em situações complexas, uma plataforma de realidade aumentada (AR) é proposta, combinando várias tecnologias em um sistema que altera a percepção do ambiente, aprimorando-o em tempo real pela adição de informações relevantes. A plataforma foi realizada através da construção de um protótipo similar a um capacete ou óculos que possibilita a visualização de informações no campo de visão do usuário. O trabalho estabelece alguns objetivos principais, destacando-se a transcrição de diálogo em tempo real e determinação e visualização de potenciais fontes de som, promovendo assim uma conscientização sobre o contexto sonoro no qual o indivíduo está inserido. Para atingir as metas definidas, métodos de aprimoramento de áudio e supressão de ruído são empregados, beneficiados ainda pela utilização de elementos de visão computacional para reconhecimento de contexto de diálogo, número de interlocutores e foco da atenção do usuário, através de estimativa de pose. Finalmente, questões de usabilidade são consideradas buscando-se minimizar a intrusividade do dispositivo e permitindo seu uso móvel. Constatou-se que a plataforma construída tem capacidade de processamento adequada para a execução de algoritmos recentes de aprendizado de máquina, que foram aplicados para determinar o número de interlocutores e seu papel no contexto. Especificamente, uma abordagem *bottom-up* para estimativa de pose humana utilizando *Deep Residual Learning* (ResNet) e *Part Affinity Fields* (PAF) é considerada. Sua implementação mostrou-se satisfatória para a aplicação, estimando 17 pontos de interesse de pose humana em situações com múltiplos indivíduos, resiliente até mesmo a situações de oclusão severa, atingindo ainda tempo de inferência suficientemente baixo para permitir o processamento de até 10 quadros por segundo em tempo real na plataforma móvel Nvidia Jetson Nano utilizada. O desenvolvimento e decisões tomadas para a construção de um protótipo funcional são descritas. Nesse sentido, a viabilidade da plataforma desenvolvida é demonstrada, mostrando-se versátil e extensível, tendo potencial para servir como ponto de partida de soluções de tecnologia assistiva auditiva atuais e futuras.

Palavras-chave: Perda Auditiva; Realidade Aumentada; Reconhecimento de Fala; Estimativa de Pose

ABSTRACT

DA SILVA, F. A.; REGINALDO, N. S.. Augmented reality sound context visualization device for the hearing impaired. 2022. 41 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná (UTFPR). Curitiba, 2022.

Recently, many technologies have been proposed to assist communication for people with hearing loss. Assistive technology solutions are mainly based on audio signal transformation, limited to amplification or dynamic range compression. However, this approach has poor effectiveness for situations where competing noise is present, preventing these individuals from socializing in these environments. This leads to alternative approaches being explored to improve these devices. With the intent to improve communication for hearing impaired individuals in complex situations, an Augmented Reality (AR) platform is proposed. This kind of platform combines many technologies in a system that changes the ambient perception, improving it by adding relevant information in real-time. The platform was implemented with the construction of a prototype similar to a helmet or eyeglasses which enables information visualization in the user's field of view. This work establishes the following main objectives: highlighting the dialogue transcription in real-time and determination/visualization of potential sound sources, promoting a better ambient perception for the user. To reach these goals, methods to improve the audio signal and noise suppression have been implemented, aided by computer vision elements to recognize dialogue context, number of speakers, and user attention focus using pose estimation. Lastly, usability concerns have been considered, trying to minimize the device intrusiveness and also considering the need for mobility in order to make it viable to be used in real scenarios. In conclusion, the platform that was built has been verified as having enough processing capability to execute recent machine learning algorithms, which have been applied in computer vision tasks with the intent to determine the number of speakers and their roles in a given context. Specifically, a bottom-up approach has been used for the pose estimation task using Deep Residual Learning (ResNet) and Part Affinity Fields (PAF). Its implementation proved to be acceptable for the application, estimating 17 points of interest of human pose in situations with multiple individuals, being resilient even in circumstances where severe occlusion was present, reaching inference times low enough to allow a processing time of 10 frames per second in real-time in the Nvidia Jetson Nano platform. The development and decisions to build a working prototype are described. The viability of the developed platform is demonstrated, shown to be versatile and extensible, allowing it to be a starting point for future hearing assistance solutions.

Keywords: Hearing Loss; Augmented Reality; Speech Recognition; Pose Estimation

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AEC	*Acoustic Echo Cancellation*, do inglês, Cancelamento Acústico de Eco.
API	*Application Programming Interface*, do inglês, Interface de Programação de Aplicação.
AR	*Augmented Reality*, do inglês, Realidade Aumentada.
ASR	*Automatic Speech Recognition*, do inglês, Reconhecimento Automático de Fala.
DNN	*Deep Neural Network*, do inglês, Rede Neural Profunda.
DOA	*Direction Of Arrival*, do inglês, Direção de Chegada.
DSP	*Digital Signal Processing*, do inglês, Processamento Digital de Sinais.
GCC-PHAT	*Generalized Cross-Correlation with Phase Transform*, do inglês, Correlação Cruzada Generalizada com Transformação de Fase.
HMD	*Head Mounted Device*, do inglês, Dispositivo Montado na Cabeça.
IFFT	*Inverse Fast Fourier Transform*, do inglês, Transformada Rápida de Fourier Inversa.
IID	*Inter-Microphone Intensity Difference*, do inglês, diferença de intensidade entre microfones.
IOT	*Internet Of Things*, do inglês, Internet das coisas.
JSON	*JavaScript Object Notation*, do inglês, Notação de Objeto *JavaScript*.
Libras	Língua Brasileira de Sinais
ML	*Machine Learning*, do inglês, Aprendizado de Máquina.
ODAS	*Open Embedded Audition System*, do inglês, Sistema Embarcado Aberto de Audição.
PAF	*Part Affinity Fields*, do inglês, Campos de Afinidade de Partes.
PcD	Pessoa com Deficiência
PHAT	*Phase Transform*, do inglês, Transformação de Fase.
ResNet	*Deep Residual Learning*, do inglês, Aprendizado Profundo Residual.

SRP	*Steered Response Power*, do inglês, Potência de Resposta Dirigida.
SSL	*Sound Source Localization*, do inglês, Localização de Fonte de Som.
SST	*Sound Source Tracking*, do inglês, Rastreamento de Fonte de Som.
TDOA	*Time Difference Of Arrival*, do inglês, Diferença de Tempo de Chegada.
TOA	*Time Of Arrival*, do inglês, Tempo de Chegada.
VR	*Virtual Reality*, do inglês, Realidade Virtual.
WER	*Word Error Rate*, do inglês, Taxa de Erro de Palavras.

SUMÁRIO

1 INTRODUÇÃO

Tecnologias assistivas buscam promover maior eficiência e autonomia, tornando atividades de interesse mais acessíveis para pessoas com deficiência (PcD). Neste contexto, aspectos de comunicação podem ser uma grande barreira para a convivência e educação. Uma possível abordagem para inclusão se dá pela "utilização de artefatos tecnológicos, os quais funcionam como meios e fins alternativos para adequar os usos e benefícios aos diferentes tipos de deficiência" (CHAIBEN, 2019).

Existem diversos argumentos convincentes sobre o potencial da utilização de realidade aumentada com o objetivo de aperfeiçoar equipamentos de tecnologia assistiva para pessoas com deficiência auditiva Mehra et al. (2020) fazem uma análise teórica exaustiva sobre todos os elementos necessários para realizar um sistema baseado neste conceito e suas ramificações. A conclusão é que, devido à evolução recente de tecnologias de reconhecimento de fala e capacidade de processamento, um dispositivo de realidade aumentada(RA) tem potencial para auxiliar pessoas com perda auditiva além do que é atualmente possível com tecnologias assistivas baseadas apenas em áudio.

Figura 1 – Protótipo desenvolvido e amostra da interface implementada.

Fonte: autoria própria.

Este trabalho descreve o desenvolvimento de um protótipo funcional, utilizando uma câmera atuando em conjunto com uma matriz de microfones para captar dados do

ambiente, os quais são processados e mostrados ao usuário, criando assim uma plataforma de Realidade Aumentada móvel, comprovando a viabilidade de um sistema desse tipo. A Figura 1 mostra o dispositivo que foi construído e a visualização da interface de AR implementada.

Para a captação de áudio e transcrição de fala, a matriz de microfones ficou centralizada no topo do protótipo, desta forma, com ele equipado, a matriz está sempre alinhada com a direção em que o usuário está apontando. O som é captado e enviado à plataforma móvel *Raspberry Pi*, onde este sinal é então enviado para duas rotinas separadas, uma para a transcrição de fala, a qual utiliza o *Google Cloud Speech API* e outra para a identificação e localização de potenciais fontes de som, utilizando o framework *Open Embedded Audition System* (ODAS).

Para a estimativa de pose foi utilizada uma câmera posicionada no topo do dispositivo, conectada à plataforma móvel Nvidia Jetson Nano, sendo ela responsável pelo processamento da imagem utilizando o kit de desenvolvimento de software *TensorRT Pose Estimation* (NVIDIA-AI-IOT, 2022). A abordagem conseguiu estimar 17 diferentes pontos de interesse, além da inferência da direção do olhar, o que nos permitiu determinar se um ou mais indivíduos detectados estão olhando na direção do usuário do protótipo ou não.

Juntando os dados coletados através de comunicação via *WebSockets*, eles são enviados para um programa centralizador que sincroniza as informações recebidas de cada sensor utilizado e então envia pacotes periodicamente para a plataforma de renderização de interface criada utilizando o motor gráfico *Unity*. Com estes pacotes, os dados são processados e as decisões em relação ao que mostrar no *display* são feitas, determinando se a transcrição irá aparecer ou não, mostrando os potenciais de som e localização de indivíduos detectados, também revelando se estão ou não olhando na direção do usuário.

1.1 Objetivos

De acordo com o Decreto n. 5.626, de 22 de Dezembro de 2005 que dispõe sobre a Língua Brasileira de Sinais (LIBRAS), uma pessoa surda é aquela que “por ter perda auditiva, compreende e interage com o mundo por meio de experiências visuais, manifestando sua cultura principalmente pelo uso da Libras”. Partindo dessa definição, este trabalho tem como objetivo expandir a compreensão de contexto sonoro de maneira visual, auxiliando pessoas com perda auditiva total ou parcial através de transcrição de áudio para Português e pela visualização de outros indicadores relevantes.

É preciso esclarecer que esse processo não torna acessíveis interações para esse grupo como um todo, já que, principalmente para pessoas que nasceram com deficiência auditiva, Libras é a sua língua materna. Portanto as legendas e transcrições funcionam exclusivamente para aqueles que, além da Libras, entendem o Português, ou seja, são bilíngues. Já para o grupo denominado surdos oralizados, aqueles que perderam a audição

após a aquisição de linguagem, o desenvolvimento deste trabalho pode ser de grande relevância, visto que costumam depender exclusivamente de leitura labial e/ou de aparelhos auditivos tradicionais.

1.2 Estrutura da monografia

O restante deste trabalho se divide como se segue. No capítulo 2, são apresentados trabalhos relacionados e uma visão geral da fundamentação do tema proposto. Em seguida, no capítulo 3, é apresentada a metodologia utilizada para cada área explorada, detalhando algoritmos e decisões tomadas para a definição de como implementar o protótipo proposto. Então, no capítulo 4 é demonstrada como foi feita esta implementação, apresentando e discutindo os resultados obtidos. Por fim, no capítulo 5 são apresentados as considerações finais e propostas de trabalhos que possam dar continuidade a este.

2 REVISÃO DE LITERATURA

Alguns tópicos precisam ser introduzidos para que o desenvolvimento do trabalho seja melhor compreendido. A seção 2.1 apresenta o Problema do coquetel, definição que representa uma síntese das questões principais abordadas pelo trabalho. A seção 2.2 aborda a transcrição de fala para texto, elemento central do desenvolvimento. No item 2.3 é introduzida a ideia de que o ato de conversação não depende de elementos puramente sonoros, o que leva a uma breve discussão em 2.4 sobre quais elementos visuais seriam relevantes neste caso. A seção 2.5 ilustra o estado atual de algoritmos para localização e rastreamento de fonte de som, essenciais para o desenvolvimento que segue. Em 2.6 uma fundamentação é apresentada a respeito da plataforma de realidade aumentada proposta, exemplificada pelos trabalhos relacionados citados no subsequente item 2.7.

2.1 Problema do coquetel

O Problema do Coquetel, ou "*Cocktail-Party Problem*", descreve um conjunto de problemas que surgem em situações de interação social devido à simultaneidade de eventos e fontes de som. Essa confluência de fatores que prejudicam a inteligibilidade da comunicação é de especial interesse para estudos relacionados com a perda auditiva, já que trata das dificuldades reais encontradas por estes indivíduos, como a segregação de tons complexos e outros fatores como co-modulação e origem (BEE; MICHEYL, 2008). Enquanto os aparelhos auditivos e os implantes cocleares melhoram a percepção da fala em ambientes silenciosos, eles normalmente oferecem muito menos benefícios aos usuários em situações barulhentas do mundo real (MOORE; PETERS; STONE, 1999) (OXENHAM; KREFT, 2014). Nesse sentido, uma solução de tecnologia assistiva para pessoas com perda auditiva melhor serviria seu propósito caso abordasse os problemas assim definidos, destacando-se: separação e isolamento de interlocutores, detecção de intenção do ouvinte, supressão de ruídos externos e um sistema de aprimoramento dos sinais de interesse. É com base nestes objetivos que a solução apresentada neste trabalho se baseia.

2.2 Reconhecimento automático de fala (ASR) / Transcrição de texto

Novas tecnologias de auxílio auditivo são possibilitadas pelo recente aumento de dados disponíveis e sua utilização em conjunto com técnicas de Aprendizado de Máquina (ML) utilizando redes neurais profundas (DNNs). Essas redes são aproximações de funções universais, encontrando as palavras, por exemplo, que melhor explicam os dados de áudio. A maior parte das tecnologias de reconhecimento automático de fala atuais são baseadas em modelos de interpretação dependentes no tempo em que uma DNN prevê o próximo

símbolo, por vezes partes de uma palavra, a partir do áudio de entrada. Isso é implementado como uma forma de regressão não linear em que a rede converte símbolos de entrada, áudio em um espaço de espectrograma de alta dimensão, em uma probabilidade de milhares de partes de palavras diferentes (SLANEY et al., 2020).

Word Error Rate (WER) é a métrica mais popular para avaliação de sistemas de ASR, e é baseada na estatística de substituições, inserções e remoções de palavras considerando-se o número total de palavras processadas. Já não é mais considerada ideal devido a possíveis avaliações inconclusivas ou que não consideram todos os parâmetros, mas ainda é amplamente utilizada, e é útil para uma comparação geral de desempenho entre diferentes sistemas (ERRATTAHI; HANNANI; OUAHMANE, 2018).

Um exemplo de implementação utilizando DNNs é descrito por Chiu et al. (2017), com uma rede de 100 milhões de parâmetros atingindo um WER de 5.8% quando treinado utilizando 12500 horas de frases em Inglês. Um número cada vez maior de sistemas utilizados em nosso dia-a-dia são implementados aplicando esta abordagem e são considerados atualmente estado da arte (SLANEY et al., 2020).

É possível argumentar que apenas recentemente, ao atingirem taxas de erro relativamente aceitáveis e serem tão amplamente disponibilizadas, tecnologias de transcrição de voz em tempo-real tornaram-se uma alternativa ou adição viável para utilização no contexto de tecnologias assistivas.

2.3 Reconhecimento do contexto e intenção do ouvinte

A disciplina de psicolinguística estabelece decisivamente a natureza complementar entre aspectos verbais e não verbais da expressão humana, concluindo que diferentes modalidades de comunicação se completam. Havendo ambiguidade na estrutura de uma modalidade a outra pode fornecer evidências para uma correta interpretação (QUEK et al., 2000).

Uma demostração da utilização dessa característica é apresentada por Ephrat et al. (2018), que desenvolveu um modelo baseado em DNNs incorporando ambos sinais visuais e auditivos para resolver os problemas apresentados em uma situação de coquetel. O método demonstrou uma clara vantagem sobre sistemas de separação de fala que utilizam somente informações de áudio.

Assim, em situações de discurso e conversas, a avaliação e reconhecimento de características do contexto como um todo é fundamental, considerando informações como o número de pessoas envolvidas, a direção do olhar do ouvinte e do interlocutor, ou ainda mudanças no ambiente e de carga cognitiva. Muitas dessas características podem ser estimadas relacionando-se informações de áudio e vídeo.

2.4 Estimativa de pose e olhar

Para buscar compreender as intenções sociais de um ou mais indivíduos, a linguagem corporal é de grande relevância. Neste sentido, a estimativa de pose se torna importante ao determinar a posição dos membros e junções, estimando também membros que podem estar oclusos, dependendo da distância e ângulo dos indivíduos em relação ao observador.

Pelo lado visual, o rastreamento do olhar revela uma boa indicação acerca da informação que o usuário busca (HAKKANI-TÜR et al., 2014). Tal rastreamento permite inferir a origem na qual ele está procurando estes dados, seja pela busca de objetos ou até mesmo para interações sociais.

2.5 Localização e rastreamento de fonte de som

O rastreamento da fonte de som trata da determinação da localização de onde o sinal de áudio é originado em relação a um elemento receptor, usualmente uma matriz de microfones. Os métodos mais comuns e eficazes para localização de som incluem aqueles baseados em energia, Tempo De Chegada (TOA), Diferença de Tempo de Chegada (TDOA), Direção de Chegada (DOA), *beamforming*, *Inter-Microphone Intensity Difference*(IID) e *Steered Response Power*(SRP) (LIAQUAT et al., 2021). Cada método apresenta vantagens e desvantagens, além de requisitos próprios para implementação, apresentados no Quadro 1.

Para nosso trabalho a característica mais relevante é o requisito de robustez em ambiente ruidosos, o que levou a uma análise mais profunda do método *Steered Response Power*(SRP).

2.6 Plataformas de realidade aumentada (AR) para apoio a PcDs

Em termos gerais, realidade aumentada é a tecnologia que faz a conexão entre o mundo real e o virtual através de interações síncronas. Em outras palavras, é uma ponte entre os dois mundos feita através de um sistema multissensorial, o qual está constantemente monitorando a realidade, seja por som, por imagem ou pelos mais variados tipos de sensores, muitas vezes atuando em conjunto. Estes dados são coletados e enviados a um programa que irá processá-los, conectando o mundo virtual ao mundo real, onde interações no digital resultam em ações reais.

Uma plataforma AR é uma classe de tecnologias que nos permite criar estímulos virtuais que podem ser mesclados com o nosso mundo real. Isto contrasta com o termo *Virtual Reality* (VR), onde o estímulo virtual substitui completamente a realidade, uma discussão mais profunda sobre isto pode ser vista em Hohmann et al. (2020).

Quadro 1 – Comparação de métodos para localização de fontes sonoras.

Método	Sincronização	Requisitos	Vantagens	Desvantagens
Energy-based methods	Direção de chegada, arranjos de microfones ad-hoc, rede de sensores acústicos sem fio, classificação de sinal de áudio, estimativa de localização.	Dispositivos de captura e transmissão mais simples.	Energia eficiente. Menor suscetibilidade à perturbação. Robusto. Largura de banda baixa.	A calibração de ganho é necessária nos nós para alta relação de energia
Beamforming	Localização da fonte, direção de chegada, trilateração, estimativa de atraso de tempo, calibração de posição.	A medição simultânea de dados é um requisito.	Os resultados têm boa resolução espacial. Rápida velocidade de análise.	Funciona com frequências acima de 1000 Hz. Existe uma troca entre alcance e precisão.
TDOA	Arrays de microfones ad-hoc, algoritmo ESPRIT, aprimoramento de fala, localização de fonte acústica, calibração de posição.	Funções de custo lineares são usadas para superar a dificuldade de não linearidade nas medições.	Custo computacional moderado. Largura de banda desejável. Menor potência de transmissão.	Propenso a erros como ruído e interferência.
Steered Response Power	Máxima verossimilhança, processamento de sinal, algoritmo MUSIC, beamforming, redes de sensores sem fio.	Mapas de potência SRP são necessários.	Robusto em ambiente ruidoso	Unidades de processamento gráfico (GPUs) são necessárias para a implementação.
TOA	Sincronização precisa.	Hardware de temporização precisa é necessário	Usando suposições razoáveis, maior precisão junto com tempo de execução reduzido podem ser alcançados.	Atrasos internos desconhecidos que precisam ser tratados com ajuste de dados.
DOA	Funciona facilmente com entradas não sincronizadas em baixas taxas de aquisição.	Associação de dados precisa ser feita para alarmes falsos.	Baixo uso de largura de banda.	Computação complexa.
Inter-microphone Intensity Difference	No domínio da frequência, existe correlação.	Incorporação do mapeamento baseado na aprendizagem.	Robusto contra interferências.	Funciona apenas para matrizes de 2 microfones.

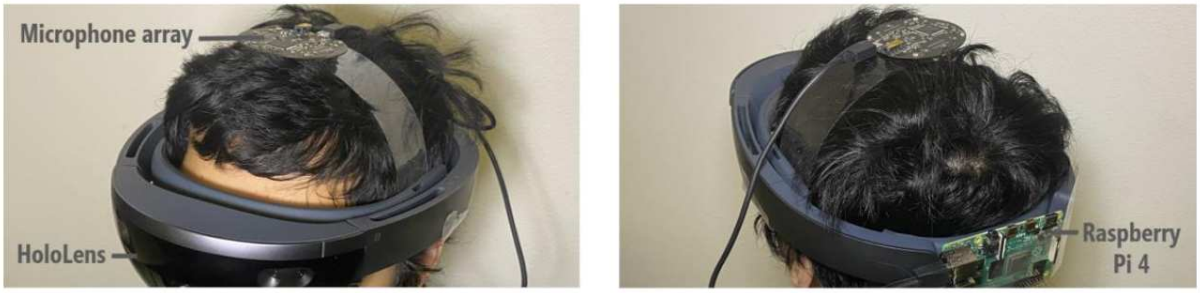
Fonte: (LIAQUAT et al., 2021), traduzido.

Seguindo o contexto de apoio a pessoas com deficiências (PcDs), diversas plataformas baseadas em AR foram propostas para facilitar a vida dessas pessoas. Como exemplo podem-se citar detectores de obstáculos e objetos, interpretadores de linguagem de sinais, transcritores de fala, entre outros, exemplificados na seção a seguir.

2.7 Trabalhos relacionados

Diversos trabalhos apresentam uma abordagem prática com o objetivo de desenvolvimento de um dispositivo de tecnologia assistiva para pessoas com perda auditiva, mas apresentam diferentes níveis de complexidade e resultados, por vezes focando em problemas específicos da implementação, como considerações sobre a interface de usuário (SCHIPPER; BRINKMAN, 2017) e acurácia de transcrição de fala (DABRAN et al., 2017), ou ainda uma visão mais geral dos componentes mínimos para um protótipo viável (GUO et al., 2020). O dispositivo mais profundamente explorado foi desenvolvido por Moraru (2018) em sua tese de mestrado, envolvendo a implementação e estudos de caso de sua utilização, incluindo questionários e dados sobre a utilização do dispositivo por um grupo de pessoas com perda auditiva.

Figura 2 – Protótipo do dispositivo *HoloSound*.



Fonte: (GUO et al., 2020)

Nas implementações realizadas por Guo et al. (2020) e Moraru (2018) a plataforma de realidade aumentada HoloLens da Microsoft foi utilizada. Esse dispositivo de visualização similar a um capacete, que pode ser visto na Figura 2 e 3, tem a vantagem de já ser disponibilizado comercialmente como um único componente, diferente da montagem da interface de realidade aumentada proposta neste trabalho, com a desvantagem de seu alto custo.

Figura 3 – Dispositivo criado por Moraru.

Fonte: (MORARU, 2018)

3 METODOLOGIA

Neste capítulo, são descritos os algoritmos e técnicas utilizadas para a construção do protótipo proposto, incluindo os passos para processamento de áudio em todas as suas etapas: a realização da supressão de ruídos, cancelamento de eco e reverberações, ganho automático, a determinação da localização das fontes de emissão de som e o reconhecimento de fala através da transcrição. Também é descrita a parte de visão computacional, que inclui estimativa de pose e a detecção da direção de olhar. Até, por fim, apresentar a metodologia utilizada para realizar cada uma destas tarefas, considerando as limitações e dificuldades encontradas para aplicar um projeto deste tipo a um sistema embarcado.

3.1 Plataforma de realidade aumentada

Por definição, a interface de realidade aumentada é a visão direta ou indireta em tempo real de um ambiente físico do mundo real que foi aprimorado/melhorado pela adição de informações geradas por computador (CARMIGNIANI; FURHT, 2011).

Um dispositivo de realidade aumentada é inerentemente multissensor, ou seja, deve utilizar e agregar informações de várias fontes. No caso da construção de uma plataforma para aprimoramento da inteligibilidade de áudio, dispositivos como câmeras, microfones e processadores de áudio devem ser utilizados, já que a compreensão de fala não é um fenômeno puramente acústico.

Em nosso protótipo, devido à necessidade de realizar a localização das fontes de áudio, funcionalidade que é explicada na seção 3.3.1, optamos por uma matriz com quatro microfones. Para a câmera, não se fez necessário um critério muito rigoroso, já que consideramos suficiente um dispositivo com capacidade aceitável para termos imagens nítidas, permitindo a estimativa de pose e determinação de direção do olhar em um ambiente razoavelmente iluminado.

3.2 Interface

Para a aplicação em um dispositivo de tecnologia assistiva com o objetivo de aprimoramento de contexto sonoro, ressaltando o foco em conversações, uma interface não intrusiva deve ser considerada. São diversas as opções de visualização em realidade aumentada, Carmigniani e Furht (2011) categorizam como possibilidades dispositivos montados na cabeça (HMD), *handheld* ou espaciais.

A decisão quanto à utilização de um HMD é justificada pelo fato de que um *handheld* requer um desvio de atenção do usuário incompatível com o contexto de diálogo, e um sistema de interface espacial é atualmente impraticável para utilização móvel. Ainda

assim é necessário ressaltar que a utilização de um sistema desse tipo pode causar considerável inconveniência, apesar da continuada miniaturização dos componentes necessários como unidades de processamento e bateria.

Outra característica de interface a ser considerada é o método de visualização que de acordo com Carmigniani e Furht (2011) pode ser separada como visualização por vídeo, visualização óptica ou aprimoramento direto (projeções). Cada método tem vantagens e desvantagens mas como o objetivo levantado anteriormente é a não intrusividade, a visualização óptica foi considerada mais adequada, já que minimiza o impacto na percepção natural dos ambientes.

3.3 Processamento de áudio

Os requisitos de processamento de áudio que possibilitam o desenvolvimento da solução, com foco em sinais de conversação, são relacionados principalmente à aplicação de algoritmos de supressão de ruídos externos, enriquecimento de sinal e localização/separação de fontes de som. Para estes fins são inúmeros os avanços e soluções atualmente disponíveis. Apesar de ser tema central do desenvolvimento deste trabalho, não é aqui proposta nenhuma solução inovadora para tratamento de áudio, escolhendo-se aplicar soluções bem estabelecidas como redução de eco e ganho automático.

3.3.1 Localização e rastreamento de fontes de áudio

Estimar a direção da fonte de som foi considerado um ponto chave para o funcionamento do sistema já que a separação de interlocutores é necessária para a visualização e compreensão correta de diálogo. Nesse sentido, abordagens que exploram a utilização de múltiplos microfones e as relações entre os sinais são a solução mais comumente utilizada, destacando-se a decisão da utilização de uma matriz circular de microfones que reduz ambiguidades quando comparada a matrizes lineares (PAVLIDI et al., 2013).

A estratégia utilizada para localização de fontes de som (SSL) foi descrita por Grondin e Michaud (2018) e implementada como uma ferramenta chamada de *Open embeddeD Audition System* (ODAS) por Grondin et al. (2021). Notadamente, implementa o método chamado *Steered Response Power with Phase Transform* (SRP-PHAT) que é geralmente calculado utilizando *Generalized Cross-Correlation with Phase Transform* (GCC-PHAT) para cada par em uma matriz de microfones, utilizando o conceito de *Time Difference of Arrival* (TDOA). Para o rastreamento de potenciais fontes um filtro Kalman 3D modificado é empregado.

O ODAS usa a geometria da matriz de microfones para realizar a localização, definida na inicialização em um arquivo de configuração. Além da posição, a orientação de cada microfone também fornece informações úteis quando os microfones estão em uma configuração de matriz fechada (por exemplo, quando instalados na cabeça ou tronco de um

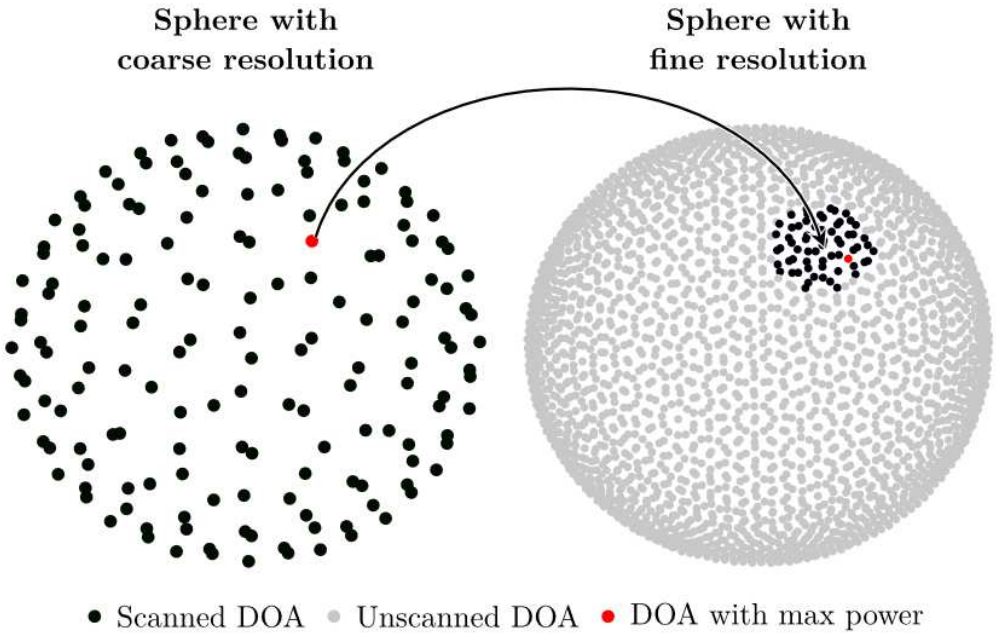
robô, ou seja, em posições relativamente próximas). Na prática microfones geralmente são omnidirecionais, mas podem ser parcialmente ocultos por algumas superfícies, o que torna sua orientação relevante. A localização baseia-se no método *Generalized Cross-Correlation with Phase Transform* (GCC-PHAT), calculado para cada par de microfones. ODAS usa a Transformada Rápida de Fourier Inversa (IFFT) para calcular a correlação cruzada de forma eficiente a partir dos sinais no domínio da frequência. Ao lidar com matrizes construídas com pequeno número de microfones o ODAS também pode interpolar o sinal de correlação cruzada para melhorar a precisão da localização e lidar com os artefatos de discretização TDOA introduzido pelo IFFT. Além disso, o framework também explora a diretividade dos microfones para computar apenas GCC-PHAT entre pares de microfones que podem ser excitados simultaneamente por uma fonte sonora.

A implementação calcula o Steered-Response Power with Phase Transform (SRP-PHAT) para todos os DOAs que se encontram em uma esfera unitária discretizada com 2562 pontos. Para cada DOA, calcula a potência SRP-PHAT somando o valor da correlação cruzada associada à diferença de tempo de chegada correspondente (TDOA) obtida com GCC-PHAT para cada par de microfones e retorna o DOA associado à maior potência. Como pode haver mais de uma fonte de som ativa por vez, os TDOAs correspondentes são zerados e a varredura é realizada novamente para recuperar o próximo DOA com a potência mais alta. Essas varreduras sucessivas geralmente são repetidas para gerar até quatro DOAs potenciais, número escolhido arbitrariamente para reduzir a carga computacional total. No entanto, a varredura de cada ponto na esfera unitária envolve vários acessos à memória que tornam o processamento bastante lento. Para acelerá-lo, o ODAS usa duas esferas unitárias, uma com resolução grosseira (162 pontos discretos) e outra com resolução mais fina (2562 pontos discretos). O ODAS primeiro varre todos os DOAs na esfera grosseira, encontra aquelas associadas às potências máximas e, em seguida, refina a busca em uma pequena região ao redor desta DOA na esfera fina.

A Figura 4 ilustra esse processo, que reduz consideravelmente o número de acessos à memória enquanto fornece uma precisão de estimativa de DOA semelhante. Por exemplo, ao executar o ODAS em um Raspberry Pi 3, essa estratégia reduz o uso da CPU para realizar a localização usando um único núcleo para uma matriz de 8 microfones em um fator de 3 (de um uso de núcleo único de 38% para 14%).

A tarefa de rastreamento de fontes de som é desafiadora devido à natureza das fontes sonoras que são não-estacionárias e esporádicas no domínio do tempo. Um algoritmo desse tipo deve ainda levar em consideração curtos períodos de silêncio, surgimento e inatividade das fontes e acompanhamento de trajetória durante o tempo. A localização da fonte sonora descrita anteriormente fornece uma ou várias DOAs potenciais, e o rastreamento mapeia e relaciona cada observação para uma fonte previamente rastreada, para uma nova fonte ou para uma detecção falsa. Para lidar com fontes de som estáticas e em movimento um filtro de partículas pode ser considerado, modelando a dinâmica

Figura 4 – Esferas unitárias da busca de potenciais fontes de som (ODAS).



Fonte: (GRONDIN et al., 2021)

de cada fonte (GRONDIN et al., 2013). As partículas de cada filtro estão associadas a três possíveis estados: estática, em movimento com velocidade constante, em aceleração. Essa abordagem, no entanto, envolve uma quantidade significativa de cálculos, pois o filtro geralmente é feito de milhares de partículas, sendo que cada uma delas precisa ser atualizada individualmente. Em vez disso, o ODAS usa o filtro Kalman para cada fonte rastreada. Os resultados apresentados por Grondin et al. (2021) demonstram desempenho de rastreamento semelhante, mas com uma redução significativa na carga computacional (por exemplo, por um fator de 14, de um único uso de núcleo de 98% até 7% ao rastrear quatro fontes com um Raspberry Pi 3).

3.3.2 Supressão de ruídos

Além do processamento pela aplicação de algoritmos e tratamento de som em software foi realizada a decisão de utilizar soluções disponíveis em hardware, já que estas diminuem a carga computacional e são amplamente difundidas. A supressão de ruídos foi executada através da utilização do processador de fala XVF3000 da XMOS que foi configurado para aplicar diferentes algoritmos. A implementação destes é abstraída pelo fabricante, descritos apenas como algoritmos avançados executados na plataforma de processamento de sinal digital.

3.3.3 Reconhecimento de fala

A etapa de reconhecimento de fala ocorre após o pré-processamento de áudio obtido da matriz de microfones. Os sistemas de Reconhecimento Automático de Fala (ASR) fornecem uma maneira eficiente de extrair uma transcrição textual dos sinais de fala, implementando várias abordagens de extração de recursos, bem como empregando diferentes tipos de métodos de classificação (ANGGRAINI et al., 2018).

Como discutido na seção 2.2, algoritmos de ASR de melhor performance são atualmente implementados por DNNs treinadas com extensas bases de conversação e extensa otimização de hiperparâmetros. Assim, grandes empresas oferecem esse tipo solução como serviço, já que ao manter suas bases de dados e parâmetros privadas podem maximizar sua rentabilidade. Esse modelo é ofertado pela IBM, Microsoft e Google, por exemplo. A solução da Google, *gcloud Speech-to-Text* foi escolhida pela extensa documentação de API, boa acurácia para a linguagem Português Brasileiro (HERCHONVICZ; FRANCO; JASINSKI, 2019) e principalmente pela oferta, no tempo da implementação deste projeto, de US$300 em créditos para novos usuários que gostariam de avaliar seus sistemas. Pelo fato de ser uma solução comercial existem poucos detalhes a respeito de seu funcionamento interno.

3.4 Visão computacional

Como discutido na seção 2.3 aspectos visuais são parte importante para o contexto de conversação, tornando-se necessária a aplicação de técnicas de visão computacional. Entre os elementos definidos relevantes para a aplicação destacam-se o levantamento do número de indivíduos inseridos no contexto e a tentativa de estimação de seu papel na conversação, ou seja, das pessoas presentes no campo de visão do usuário, quais estão próximas e/ou mantendo contato visual, como se movimentam e sua linguagem corporal. Entre as tarefas mais comumente aplicadas no campo de visão computacional a detecção de rostos e estimativa de pose foram elencadas como o mínimo para viabilizar os objetivos destacados, já que permitem o acompanhamento de pontos de referência nos indivíduos presentes na cena.

3.4.1 Estimativa de pose

A estimativa de pose, similarmente a outros tipos de tarefa de visão computacional, é um desafio normalmente tratado utilizando *Deep Learning*. Normalmente, quanto maior a profundidade deste tipo de rede maior é a dificuldade de treiná-la, pois em maior profundidade a acurácia é saturada e, de forma contraintuitiva, acrescentar mais camadas leva a um maior erro de treinamento. Para solucionar este problema, é utilizado então o modelo de *Deep Residual Learning* (ResNet), o qual foi inicialmente apresentado no artigo *Deep Residual Learning for Image Recognition* (HE et al., 2015). Este foi o modelo utilizado

para realizar a estimativa de pose empregado pelo TensorRT Pose (projeto em que se baseia a nossa implementação), com uma ResNet de 224x224, o qual foi implementado em nosso protótipo para a estimativa de pose.

Existem duas maneiras de realizar a estimativa de pose. A mais comum e mais simples, é o modo *top-down*, ou seja, primeiro são colocados *bounding boxes* nos humanos detectados e então as partes do corpo são localizadas para cada *bounding box*. A outra maneira é através do modo *bottom-up*, abordagem com foco invertido desta primeira, detectando-se inicialmente as partes do corpo e então agrupando as partes pertencentes a cada pessoa específica. Esta última é a abordagem empregada no TRT Pose, devido à sua maior velocidade para contextos com mais pessoas. A razão pela qual ela é mais rápida é resultado da forma como as imagens são enviadas para serem processadas pela rede. Ao contrário da maneira *top-down*, a imagem é alimentada apenas uma vez de forma integral na rede neural, levando a uma menor resolução, enquanto a outra precisa enviar à rede neural cada recorte de humanos detectados, tendo maior resolução, porém necessitando de mais passos, de acordo com a quantidade de indivíduos presentes.

O TensorRT Pose segue o método de *bottom-up*, realizando 6 passos até o objetivo final de obter as partes interconectadas para cada humano detectado.

O primeiro passo, chamado de inferência, consiste na obtenção de mapas de calor, para gerar os *Part Affinity Fields* (PAF) a partir do modelo.

O segundo passo trata de extrair as partes do corpo das áreas de maior confiança, isto é feito localizando máximos locais usando Supressão de Não-Máximos (NMS).

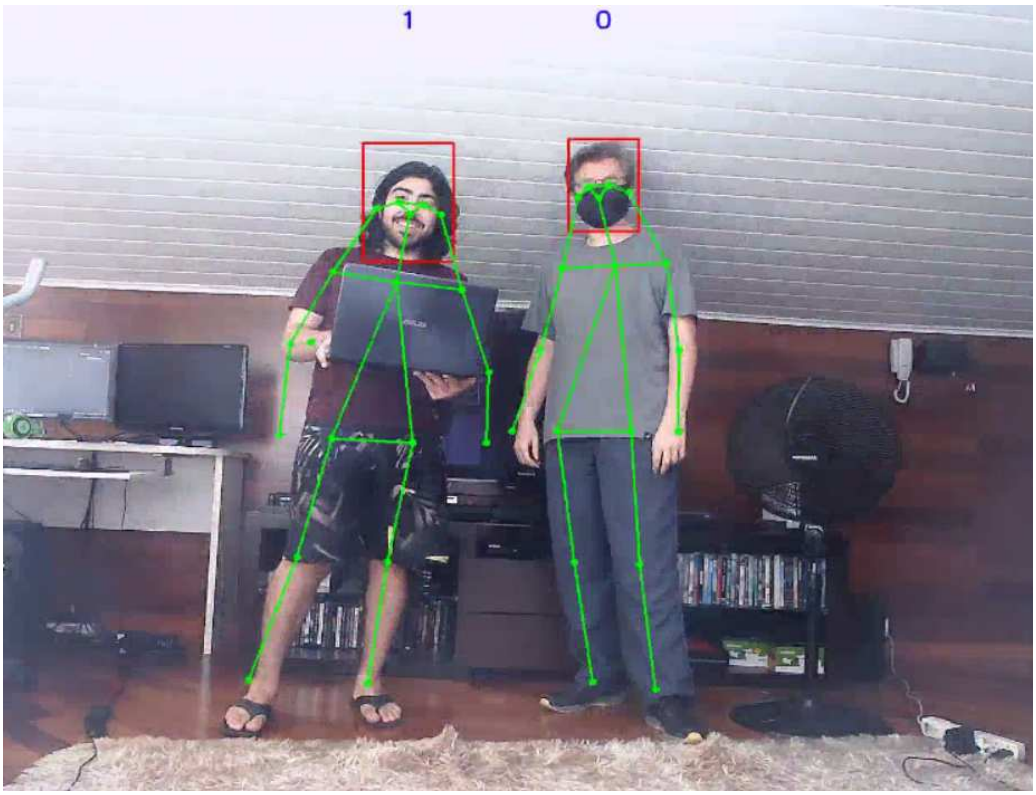
No terceiro e quarto passos, é criado um grafo bipartido para associar partes do corpo a uma pessoa única. Isto é útil principalmente para os casos em que há mais indivíduos na mesma imagem, permitindo identificar as ligações entre as partes do corpo para cada pessoa separadamente. Para isso, as arestas do grafo recebem pesos, para cada par de candidatos é computada a linha integral ao longo do vetor conectando os pares. Então normaliza-se o vetor e finalmente é computada a linha integral do produto dos componentes dos PAF nas direções X e Y com os correspondentes X e Y do vetor anteriormente computado.

O quinto passo resolve então este grafo utilizando o método Húngaro, eliminando conexões que não fazem sentido para o que buscamos, restando apenas aquelas que possuem significado.

No sexto e último passo, é finalmente realizado o desenho do esqueleto humano, conectando os pares obtidos anteriormente até não restarem mais pares a serem conectados, diferenciando partes iguais como sendo de pessoas diferentes.

Ao final destes passos, temos a estimativa de pose completa com os pontos de interesse marcados e conectados, como pode ser visto na Figura 5.

Figura 5 – Pontos de interesse da estimativa de pose conectados.



Fonte: autoria própria.

3.4.2 Detecção de olhar

A partir da estimativa de pose, de uma forma mais simples, obtemos a direção de olhar considerando os pontos de interesse faciais. Para isto, é calculada a distância entre o olho esquerdo e orelha esquerda, e da mesma forma para o lado direito, a distância entre o olho direito e a orelha direita. Comparando-se a proporção adimensional entre estas duas distâncias, se apresentar valor próximo de unitário consideramos que a pessoa está voltada para o usuário, portanto possivelmente a direção de olhar também está.

3.5 Processamento de informações

O processamento de informações no contexto do projeto apresenta alguns requisitos essenciais para a viabilização do dispositivo proposto, principalmente no que se refere à minimização do tempo entre aquisição de dados pelos sensores e visualização final das informações pelo usuário, intervalo de tempo comumente chamado de latência em sistemas de informação. Existe também o aspecto móvel do protótipo, característica que define restrições de dimensão e consumo de energia que devem ser respeitadas, impactando diretamente a capacidade total de processamento disponível.

Como sugerido anteriormente, o sistema apresenta duas classes principais de

tarefas de processamento, aquelas relacionadas ao processamento de áudio/voz e aquelas relacionadas à visão computacional.

A tarefa de visão computacional deve ser realizada localmente no próprio dispositivo, isso porque a transmissão contínua do vídeo capturado e subsequente processamento remoto produziriam um atraso incompatível com as necessidades de visualização das informações obtidas. Assim foi considerada a utilização de um módulo de computação embarcada capaz de executar esse tipo de tarefa.

A plataforma NVIDIA Jetson Nano alcança um balanço entre baixo consumo de energia e desempenho na execução de tarefas de visão computacional (MITTAL, 2019) e foi escolhida para este propósito em nosso projeto.

Inicialmente a utilização de um segundo computador de placa única (SBC), além do Jetson Nano, foi considerada pela restrição de conectividade da matriz de microfones, mais especificamente, a *ReSpeaker 6-Mic Circular Array* que foi elaborada inicialmente para funcionamento com a plataforma Raspberry Pi. Posteriormente essa ideia foi suplantada pela escolha da matriz de microfones *ReSpeaker USB Mic Array*, permitindo funcionamento com um número maior de plataformas, inclusive NVIDIA Jetson Nano. Entretanto, foi observada uma saturação da utilização de recursos no Jetson Nano pelas tarefas de visão computacional e renderização de interface, portanto justificando ainda a utilização de um segundo módulo de processamento na forma de uma placa Raspberry Pi que fica responsável pelas tarefas relacionadas ao processamento de áudio e coordenação de tarefas.

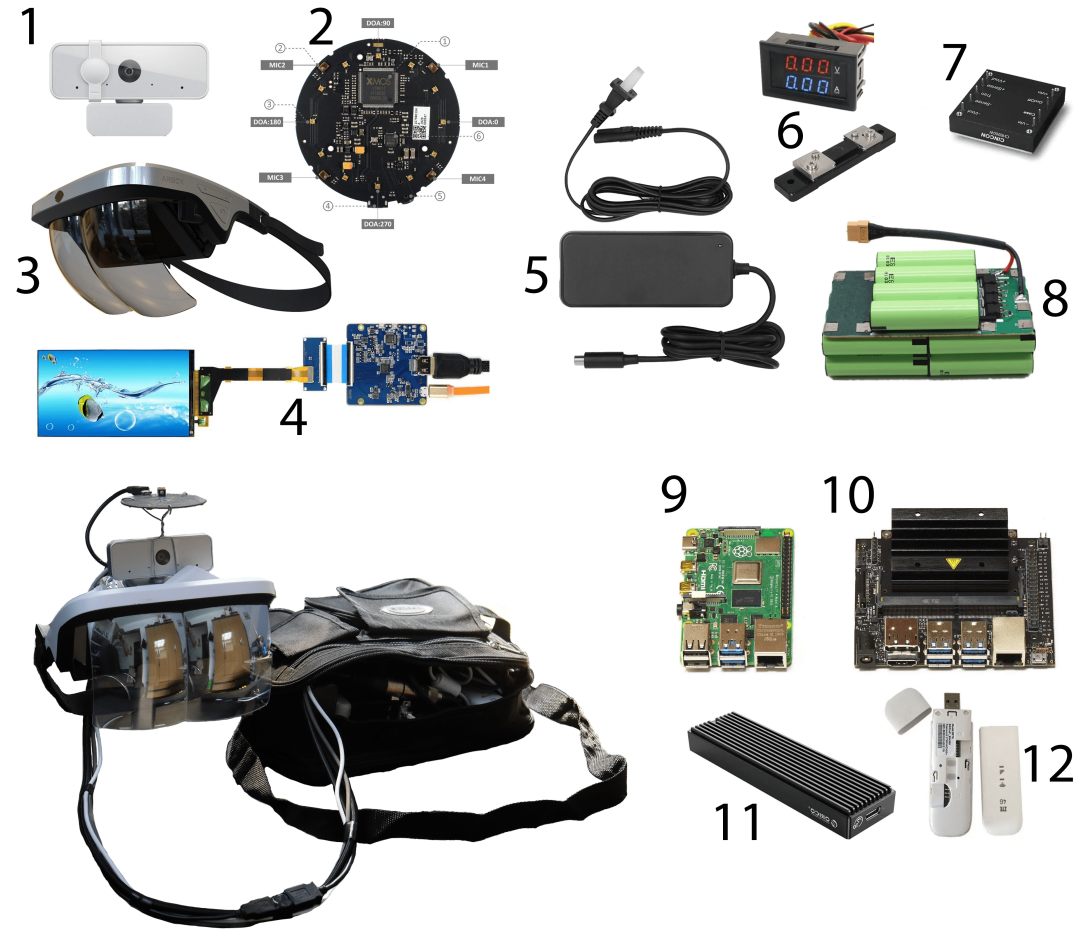
4 IMPLEMENTAÇÃO E RESULTADOS

Neste capítulo, detalhamos os passos seguidos para a construção de um protótipo capaz de captar áudios de conversação e transcrevê-los em tempo real. Além disto, apresentamos e discutimos os resultados obtidos, com ênfase no reconhecimento de fala, demonstrando a viabilidade do projeto como um todo.

4.1 Plataforma de desenvolvimento

Para a realização do protótipo, diversos componentes foram utilizados em conjunto, apresentados separadamente na lista a seguir, podendo ser visualizados na Figura 6:

Figura 6 – Componentes utilizados no protótipo.



Fonte: autoria própria.

1. Webcam Lenovo 300 FHD DFOV 95°
2. *ReSpeaker USB Mic Array, 4 mic, XMOS XVF-3000 DSP*
3. *Augmented Reality Headset* (baseado em reflexão óptica)

4. *6.0"1440x2560 LCD Display*
5. Carregador Adaptador AC/DC 42V 71W
6. Voltímetro Digital com Amperímetro 50A 100VDC + Resistor Shunt
7. Conversor DC/DC 24-72V para 5V 20A
8. Bateria 36v 4,4Ah 10s2p 18650 + Sistema de gerenciamento de bateria (BMS)
9. *Raspberry Pi 4 Model B 4GB*
10. *Jetson Nano 2GB Developer Kit*
11. *128GB NVME SSD*
12. Modem 4G Wi-Fi ZTE MF79U

O elemento principal da construção do dispositivo é a estrutura análoga a um capacete ou óculos, indicada como item 3, moldada em plástico com compartimento para uma tela ou *smartphone* e a lente curva reflexiva que permite o efeito óptico para realidade aumentada, similar ao efeito fantasma de Pepper (ANGELI; O'NEILL, 2015). A câmera (item 1), matriz de microfones (item 2) e *display* (item 4) são acoplados a essa estrutura (item 3), formando assim o dispositivo montado na cabeça (HMD, de *Head Mounted Device*) É esta a interface entre o sistema e o usuário, que vê em seu campo de visão as informações a respeito de contexto sonoro obtidas através das informações visuais e auditivas capturadas e processadas. Nas seções subsequentes cada componente utilizado no protótipo é explicado em maiores detalhes.

4.1.1 Processamento de informações

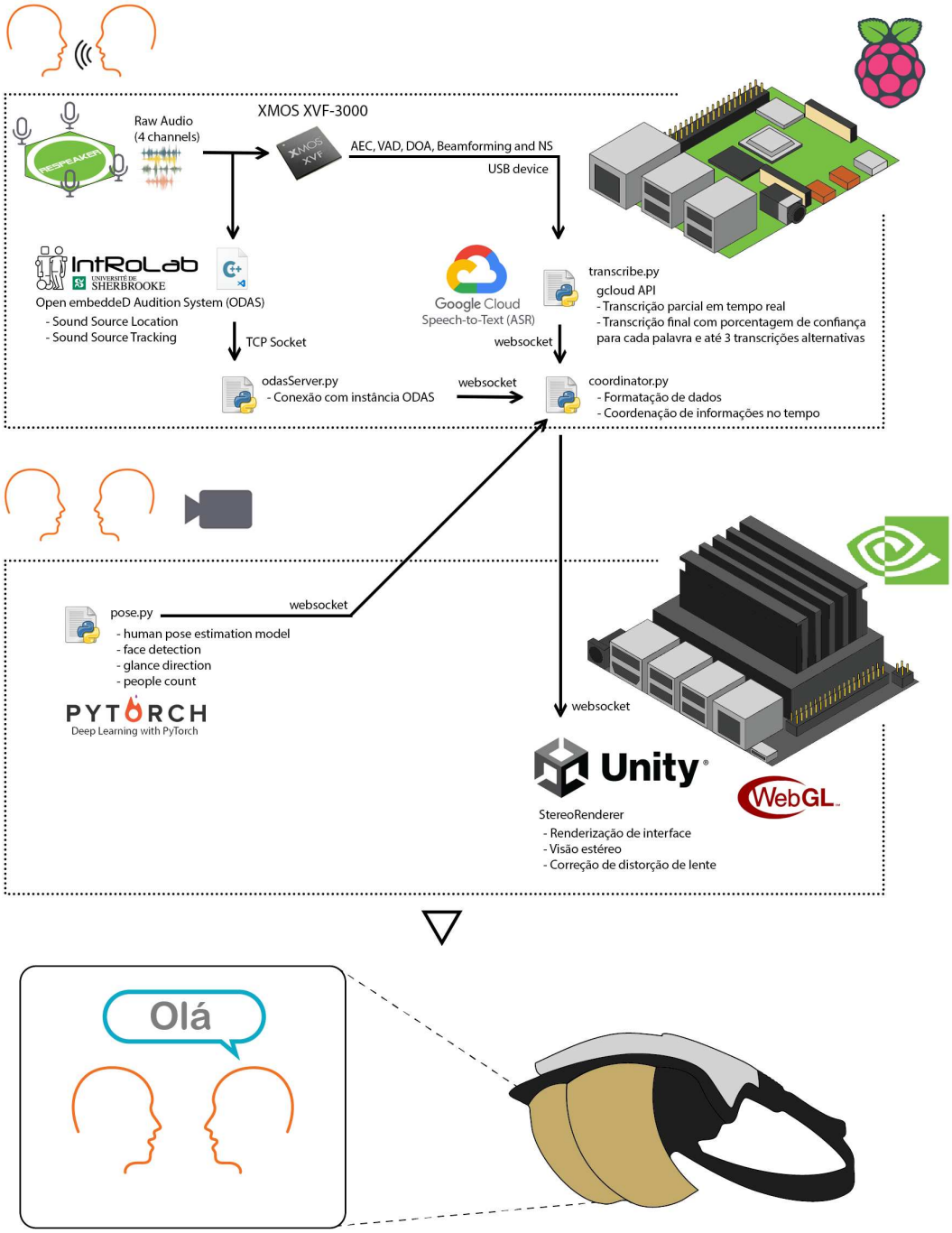
Como discutido na seção 3.5 temos duas vertentes principais para o processamento de informações, o processamento de áudio/voz e a parte relacionada à visão computacional, sendo a primeira realizada na plataforma NVIDIA Jetson Nano, enquanto a segunda na plataforma Raspberry Pi, como esquematizado na Figura 7.

Escolhemos realizar a separação desta maneira, deixando o processamento das imagens na primeira plataforma, devido a dois principais fatores, primeiramente por realizarmos a *Pose Estimation* utilizando *Tensor RT*, uma tecnologia da própria NVIDIA, e pelo fato de possuir maior capacidade computacional. Desta forma nossa necessidade foi muito bem atendida, nos permitindo fazer a estimativa de pose e a detecção de olhar em tempo real.

Já para o processamento de áudio, além da matriz de microfones utilizada possuir um *chipset* que é capaz de realizar um processamento a nível de hardware, o sinal é captado e enviado para duas ferramentas distintas: Para a ferramenta denominada *Open embeddeD Audition System* (ODAS), onde é feita a localização das fontes de som a partir dos potenciais, e para a ferramenta baseada em nuvem de transcrição de áudio, o *Google Cloud Speech-To-Text*.

Para cada um destes diferentes serviços geramos pacotes no formato *JavaScript Object Notation* (JSON), os quais possuem os dados e a informação do momento em que

Figura 7 – Diagrama de funcionamento.



Fonte: autoria própria.

foram gerados. Então, através de comunicação via *Websocket*, estes pacotes são enviados para um programa centralizador, o qual é responsável por sincronizá-los e gerar um pacote único que é enviado à *Unity Engine* da mesma forma. Estes pacotes recebidos são, por fim, continuamente desempacotados e os dados são processados, o que nos permite ter com garantia dados do mesmo instante sempre em cada pacote. Com isso conseguimos

realizar as decisões de fazer a transcrição ou não, com base na direção de chegada do áudio, em conjunto com os dados de estimativa de pose e pela detecção do olhar. Também através da *Unity Engine* é feito a comunicação com a interface, enviando ao *display* o texto de transcrição, a informação da localização de potenciais detectados e a localização de indivíduos detectados, junto à informação se estão olhando na direção do usuário.

4.1.2 Matrizes de microfones de alta ordem

Como discutido na seção 3.3, matrizes de microfones possibilitam a utilização de algoritmos para localização e separação de som. Nesse contexto o ReSpeaker Mic Array 2.0, dispositivo disponibilizado comercialmente e com esquemático ofertado abertamente, apresenta diversas vantagens e excelente performance (MISCHIE; NXAĂRESC, 2020). Além de disponibilizar os canais de áudio de seus 4 microfones, utiliza ainda o XMOS XVF-3000, um chipset que integra algoritmos de processamento digital de sinais (DSP) avançados incluindo cancelamento de eco acústico (AEC), formação de feixe, redução de reverberação, supressão de ruído e controle automático de ganho. Estas características são voltadas especificamente ao processamento para áudio de fala, que é o foco de nossa aplicação.

4.1.3 Conectividade

Como discutido posteriormente na seção 4.2.1, soluções estado da arte para *Automatic Speech Recognition* (ASR) são implementadas atualmente no modelo de serviço online disponibilizado por corporações que executam tecnologias proprietárias em seus próprios servidores. Assim, para atingir a menor taxa de erro de transcrição de fala para texto se faz necessária uma conexão contínua entre o protótipo desenvolvido e os serviços em *Cloud*.

A utilização do Modem 4G Wi-Fi ZTE MF79U neste caso não só permite a conexão móvel mas, devido à implementação de um ponto de acesso Wi-Fi para compartilhamento de rede, acaba simplificando a construção do protótipo, isso porque ao conectar o modem ao Jetson Nano através de uma porta USB e o Raspberry Pi à rede Wi-Fi subsequentemente criada, uma rede local é estabelecida, mecanismo utilizado para a comunicação entre os dois dispositivos.

4.1.4 Portabilidade e gerenciamento de energia

Os componentes escolhidos para o projeto sempre foram considerados primariamente pela perspectiva de portabilidade, buscando-se minimizar tamanho e peso de suas partes. Outro aspecto crítico para portabilidade está relacionado à fonte de energia utilizada pelo dispositivo. Nesse sentido decidiu-se pela utilização de uma bateria (Figura 6, item 8) que utiliza um conjunto de células 18650 em configuração 10s2p, totalizando

uma tensão nominal de saída de 36V e capacidade de 4,4Ah. Conjuntos de células desse tipo são amplamente utilizados, têm características de funcionamento comprovadas, e sua escolha se deve principalmente à alta densidade de energia de baterias lítio-íon, que permitem uma capacidade total e peso aceitáveis (MANTHIRAM, 2017). Ainda sobre a alimentação do sistema, o mesmo pode ser energizado e/ou ter sua bateria carregada por um adaptador AC/DC 42V (Figura 6, item 5). Devido à variação de tensão de entrada do sistema, que pode ir de 42V (tensão de carga na etapa de tensão constante) até 32V (tensão mínima de descarga das células gerenciada pelo BMS), é necessário o uso de um conversor DC/DC (Figura 6, item 7) de ampla faixa de tensão de entrada e que apresente potência de saída suficiente para o funcionamento de todo o sistema (entre 8 e 25W medido, todos os componentes são alimentados em 5V). Para uma estimativa de quantidade de carga restante da bateria e consumo do sistema um voltímetro/amperímetro (Figura 6, item 6) é posicionado junto à saída da bateria.

Considerando-se a potência de pico do sistema em 25W, medida com o voltímetro/amperímetro logo após a bateria, ou seja, levando-se em consideração a eficiência do conversor DC/DC, é possível estimar que o sistema permite 6h20min de uso contínuo (Equação 1, pior caso).

$$t_{min} = \frac{V_{bat} * Capacidade}{P_{pico}} = \frac{36V * 4.4Ah}{25W} = 6.336\ h \tag{1}$$

Na prática, foi possível utilizar o dispositivo por até 10 horas continuamente.

Figura 8 – Dispositivo Montado na Cabeça (HMD) e Bolsa.

Fonte: autoria própria.

Outro ponto importante para a mobilidade é relacionado às dimensões e invólucro do dispositivo. No total o sistema pesa 2,3kg, divididos como 300 gramas no HMD (Figura 8, esquerda), contendo câmera, *display* e matriz de microfones, e 2kg na bolsa (Figura 8, direita) que contém a bateria, sistema de gerenciamento de energia, elementos de processamento e comunicação. O conjunto de cabos que liga o HMD à bolsa tem exatamente 1 metro de comprimento.

4.2 Algoritmos de processamento de áudio

Como citado na seção 3.3, apesar da implementação dos algoritmos de áudio ser fundamental para a realização deste trabalho, e a grande oportunização de exploração desse tema pela aquisição da matriz de microfones de alta ordem, optou-se pela busca de soluções já bem estabelecidas para viabilizar a implementação do trabalho.

4.2.1 Reconhecimento de fala

Como ilustrado na Figura 7 uma implementação em Python foi escrita para o acesso à *gcloud API*, realizando a transmissão contínua do áudio pré-processado, que é transcrito na plataforma externa e retorna em "tempo real" os dados referentes às palavras obtidas pelo reconhecimento de fala na forma de resultados parciais e finais. Adicionalmente o serviço foi configurado para disponibilizar a confiança sobre o resultado obtido, apresentado como variável de "probabilidade" no intervalo 0,0 a 1,0 que indica a qualidade da transcrição de cada palavra, em outras palavras, o grau de precisão da resposta, para o trecho de diálogo considerado. É uma métrica útil para decisão quanto à exibição de resultados alternativos, recurso configurado em nossa aplicação para sugerir até três possíveis frases potencialmente intercambiáveis no contexto. Mais detalhes da utilização destes dados são descritos na seção 4.4.

Pensando sobre a viabilidade da implantação do protótipo, custos de manutenção deste serviço devem ser considerados. Uma possível estimativa poderia considerar a utilização do dispositivo por 8 horas diariamente, todos os dias, com atividade de voz em 50% do tempo. Considerando os valores na data de realização de nossos testes, Fevereiro de 2022, o custo total mensal seria entre US$115 e US$260, dependendo da qualidade do modelo selecionado e permissão do usuário quanto à reutilização dos dados de áudio obtidos (o que também implica em considerações sobre privacidade).

Apesar de custo considerável e a exigência de conexão móvel para a modalidade online de ASR, ainda escolhemos implementa-la já que a qualidade da transcrição obtida e tempo de inferência são características críticas para o funcionamento satisfatório do sistema (GAZETIć, 2017). Em trabalhos futuros uma estratégia *offline* e/ou de código aberto pode ser considerada para diminuir a complexidade e custos de manutenção do sistema.

4.2.2 Localização e rastreamento de fonte de som

Como discutido em 3.3.1 a ferramenta *Open embeddeD Audition System* (ODAS) desenvolvida por Grondin et al. (2021) foi utilizada para localização e rastreamento de potenciais fontes de som. A Figura 7 mostra como a implementação se conecta ao resto do sistema. Especificamente, uma rotina escrita em Python foi criada para servir como destino dos dois *Sockets TCP* relacionados às tarefas de Localização de Fonte de Som

(SSL) e Rastreamento de Fonte de Som (SST) provenientes da implementação de Grondin e configuradas para levar em consideração detalhes específicos da matriz de microfones utilizada, já que o algoritmo depende de informações como distância e posicionamento dos microfones.

4.3 Estimativa de pose e olhar

O mecanismo de inferência para o algoritmo apresentado em 3.4.1 foi baseado na implementação disponível em Nvidia-Ai-Iot (2022). Executado no Jetson Nano, resulta na obtenção de até 17 pontos de interesse da pose humana para múltiplos indivíduos simultaneamente. Os pontos são listados a seguir: nariz, olho esquerdo, olho direito, orelha esquerda, orelha direita, ombro esquerdo, ombro direito, cotovelo esquerdo, cotovelo direito, pulso esquerdo, pulso direito, quadril esquerdo, quadril direito, joelho esquerdo, joelho direito, tornozelo esquerdo, tornozelo direito e pescoço.

Dados os pontos obtidos pela inferência, tarefas de contagem de pessoas, detecção de rosto e estimativa de direção de olhar tornam-se triviais, já que um resultado aproximado pode ser deduzido a partir da estimativa inicial de pose humana. Algoritmos especializados para estas tarefas de visão computacional foram considerados mas em última análise julgou-se suficiente os resultados obtidos utilizando apenas o algoritmo inicialmente proposto. Em trabalhos futuros a utilização de implementações específicas pode ser melhor explorada, dada a versatilidade de desenvolvimento da plataforma.

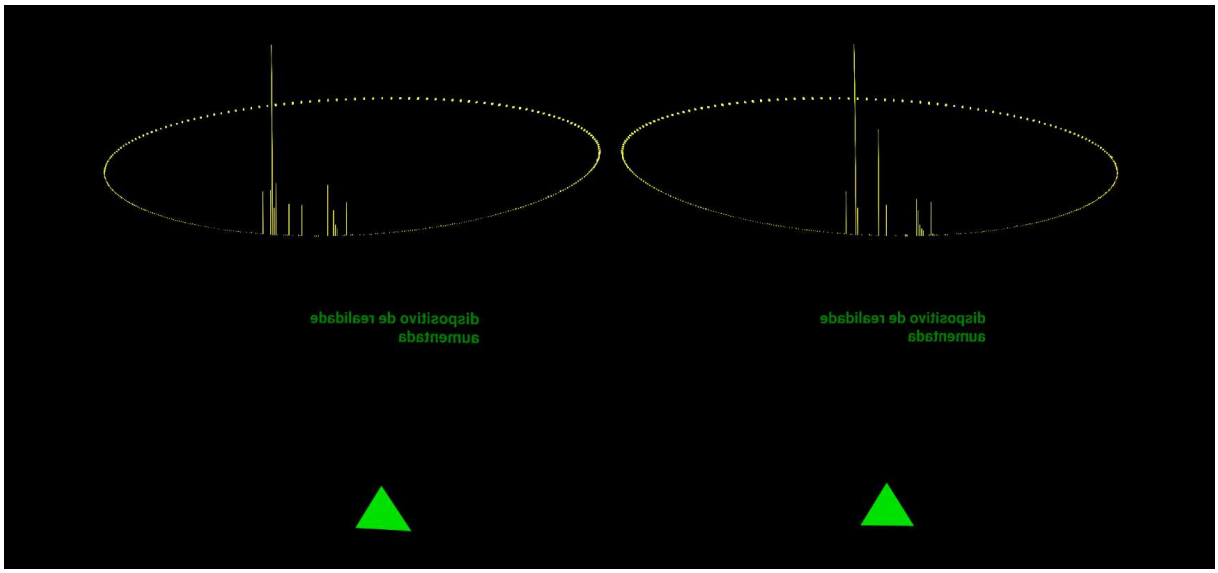
4.4 Interface de Realidade Aumentada

Para a implementação da interface de Realidade Aumentada um mecanismo de visualização capaz de se integrar ao campo de visão do usuário precisou ser considerado. Para que isso ocorresse de forma a minimizar a intrusividade do dispositivo, como discutido na seção 3.2, um dispositivo montado na cabeça (HMD) aplicando método de visualização óptica foi utilizado. Consequentemente, os objetos que fazem parte da Realidade Aumentada foram renderizados para que se adequassem a esta forma de visualização. Na prática isso resulta na implementação de visão estereoscópica e compensação de distorção de lente. Partindo destes requisitos básicos ferramentas de computação gráfica foram consideradas, levando ainda em consideração a plataforma na qual seriam executadas, mais especificamente um processador ARM Cortex-A57 de arquitetura aarch64 executando o sistema operacional Ubuntu 18.04.5 (Jetson Nano) com quantidade limitada de recursos. Deliberou-se a utilização da ferramenta e motor gráfico Unity Engine, devido à versatilidade para execução em diferentes plataformas e disponibilidade de mecanismos para implementação da interface com as características citadas anteriormente. Um detalhe importante é que o Unity não possibilita atualmente a compilação para execução nativa em ARM Linux, isso foi solucionado direcionando a execução para que fosse “agnóstica” à

plataforma utilizando o recurso de execução empregando WebGL, ou seja, voltada para execução em navegadores como Google Chrome e Firefox. Esta "tradução" tem um custo computacional considerável mas que se mostrou suficientemente otimizada para oferecer um desempenho adequado à proposta. A interface é executada utilizando o navegador *Chromium* através do acesso a um *webserver* local que disponibiliza os arquivos necessários gerados pela Unity Engine.

Os dados gerados por outras partes do sistema são recebidos pela aplicação de interface através de um *websocket*, esquematizado na Figura 7, para que sejam agregados e finalmente visualizados pelo usuário. A interface desenvolvida pode ser dividida em três elementos principais, relacionados à transcrição de texto em tempo real, visualização de potenciais fontes de som e indicação de contexto de diálogo.

Figura 9 – Renderização da interface.



Fonte: autoria própria.

O elemento amarelo similar a um compasso na Figura 9 representa o potencial acumulado durante os últimos segundos para uma dada direção, separada em 360 componentes para representação de 360 graus. Essa informação é derivada dos mecanismos de separação e rastreamento de fontes de som discutidas em 3.3.1 e tem como objetivo informar ao usuário sobre o contexto sonoro, alterações e possíveis pontos de interesse que de outra forma passariam despercebidos.

O próximo elemento visual é a transcrição em tempo real do diálogo obtido por meio do serviço descrito em 4.2.1, localizado em posição central com o objetivo de evitar a perda de contato visual e/ou foco pelo usuário em situações de comunicação. Para transcrições parciais, ou seja, que são recebidas antes mesmo da consideração de um trecho de diálogo como um todo e que estão sujeitas a mudanças, uma cor azul clara é utilizada, sendo disponibilizada ao usuário imediatamente devido às características dinâmicas de

conversação. Dado um trecho de transcrição final, normalmente obtida em momentos de pausa do discurso, uma visualização em cores verde, no caso de uma confiabilidade auto-relatada de palavra superior a um limiar estipulado em 70% e amarela caso seja inferior, é utilizada. Possíveis alternativas são obtidas para palavras com baixa probabilidade de acerto, mas no momento de execução deste trabalho não foi encontrado um mecanismo considerado aceitável para visualização destas palavras, já que durante os testes causaram uma perda de legibilidade das sentenças.

Por último, uma indicação simplificada de interlocutores identificados é disponibilizada, mostrando sua posição horizontal no campo de visão do usuário. A informação relacionada à detecção de direção de olhar também é exemplificada alterando-se a cor do indicador caso o usuário e dialogador estejam frente-a-frente, ou seja, em uma situação de provável discurso dirigido.

4.5 Discussão

Uma breve discussão a respeito das conclusões obtidas pela realização do trabalho segue, com considerações sobre seu escopo e validação.

4.5.1 Métricas para avaliação

É possível argumentar que métricas quantitativas para cada tarefa e suas potenciais alternativas não foram, admitidamente, exploradas com a devida profundidade, entretanto isso pode ser possivelmente justificado dado o escopo do projeto executado. A maior parte dos resultados obtidos se refere à exequibilidade e funcionamento subjetivamente adequado aos objetivos propostos. Detalhes quanto à iteração da construção do protótipo e de implementações inviáveis de algoritmos para a aplicação foram omitidos deste trabalho.

4.5.2 Validação dos resultados

A validação dos resultados obtidos se deu em uma capacidade bastante limitada, notando-se ainda a omissão do processo de testes e retorno que poderia ser obtido com a discussão do desenvolvimento e resultados através da comunicação com o público-alvo. Esse processo incluiria um levantamento de requisitos e outros aspectos referentes à disciplina de design de interação (ELLWANGER; ROCHA; SILVA, 2015).

5 CONSIDERAÇÕES FINAIS

Este projeto foi proposto com o objetivo de delinear o desenvolvimento de um dispositivo de tecnologia assistiva voltado a pessoas com perda auditiva. Apresenta-se uma revisão teórica sobre os elementos fundamentais necessários para o início do projeto e seu desenvolvimento. Uma metodologia é proposta para atingir os objetivos levantados e sua implementação descrita. Por fim, o dispositivo finalizado é apresentado e resultados provenientes de sua utilização são descritos, com ressalvas a respeito de sua usabilidade. Resulta do trabalho um protótipo funcional que apesar das limitações demostra a viabilidade da plataforma proposta.

5.1 Trabalhos futuros

Como indicado em 4.5.1, a realização de uma validação quantitativa de cada componente da arquitetura implementada é recomendada para trabalhos futuros, com a definição de métricas importantes para este tipo de dispositivo, permitindo assim uma comparação adequada de seu desempenho relativo a outras plataformas.

Como prova de conceito o protótipo foi construído integrando diferentes soluções comerciais de hardware. Entretanto, para reduzir as dimensões e complexidade do dispositivo a unificação de todos os componentes em uma plataforma única seria desejável.

Em trabalhos futuros uma estratégia *offline* e/ou de código aberto pode ser considerada para a implementação do algoritmo de Reconhecimento Automático de Fala (ASR), já que a solução comercial utilizada incorre em custos significativos de operação.

Como sugerido na seção 4.5.2 a interface desenvolvida, e o projeto como um todo, precisariam ser considerados de um ponto de vista de usabilidade, o que exigiria contato próximo com PcDs que auxiliariam no processo iterativo de avaliação e testes. Técnicas de Design de Interação seriam seguidas para obter resultados relevantes ao contexto em que o dispositivo seria utilizado, levando em consideração as experiências do público alvo e suas reflexões a respeito das funcionalidades e finalidade da plataforma desenvolvida. Essa análise demanda um planejamento regrado e cuidadoso já que envolve ainda questões de ética e de acessibilidade, com suas possíveis etapas e processos bem estabelecidos na literatura (LÖWGREN; STOLTERMAN, 2007) (ANTONA; STEPHANIDIS, 2017).

Referências

ANGELI, D. D.; O'NEILL, E. J. Development of an Inexpensive Augmented Reality (AR) Headset. In: **CHI EA '15: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2015. p. 971–976. ISBN 978-1-45033146-3. Citado na página 30.

ANGGRAINI, N. et al. Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API. **Telkomnika (Telecommunication Computing Electronics and Control)**, Universitas Ahmad Dahlan, v. 16, n. 6, p. 2733–2739, Dec 2018. ISSN 1693-6930. Citado na página 25.

ANTONA, M.; STEPHANIDIS, C. **Universal Access in Human–Computer Interaction. Design and Development Approaches and Methods: 11th International Conference, UAHCI 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I**. Heidelberg, Germany: Springer Verlag, 2017. v. 10277. ISBN 978-3-319-58705-9. Citado na página 38.

BEE, M. A.; MICHEYL, C. The Cocktail Party Problem: What Is It? How Can It Be Solved? And Why Should Animal Behaviorists Study It? **Journal of comparative psychology (Washington, D.C. : 1983)**, American Psychological Association, v. 122, n. 3, p. 235–51, Sep 2008. ISSN 0735-7036. Citado na página 16.

CARMIGNIANI, J.; FURHT, B. Augmented Reality: An Overview. In: **Handbook of Augmented Reality**. New York, NY, USA: Springer, New York, NY, 2011. p. 3–46. ISBN 978-1-4614-0063-9. Citado 2 vezes nas páginas 21 e 22.

CHAIBEN, G. H. **Políticas públicas para discentes com deficiência: a UTFPR.** Tese (Doutorado) — Universidade Tecnológica Federal do Paraná, Dec 2019. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/4742. Citado na página 13.

CHIU, C.-C. et al. An online sequence-to-sequence model for noisy speech recognition. **arXiv**, Jun 2017. Citado na página 17.

DABRAN, I. et al. Augmented reality speech recognition for the hearing impaired. In: **2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)**. [S.l.]: IEEE, 2017. p. 1–4. Citado na página 19.

ELLWANGER, C.; ROCHA, R. A. da; SILVA, R. P. da. Design de Interação, Design Experiencial e Design Thinking: a triângulação da Interação Humano-Computador. **Revista de Ciências da Administração**, Universidade Federal de Santa Catarina, v. 1, n. 1, p. 26, Dec 2015. ISSN 1516-3865. Citado na página 37.

EPHRAT, A. et al. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. **ACM Trans. Graphics**, Association for Computing Machinery, v. 37, n. 4, Apr 2018. ISSN 0730-0301. Citado na página 17.

ERRATTAHI, R.; HANNANI, A. E.; OUAHMANE, H. Automatic Speech Recognition Errors Detection and Correction: A Review. **Procedia Comput. Sci.**, Elsevier, v. 128, p. 32–37, Jan 2018. ISSN 1877-0509. Citado na página 17.

GAZETIć, E. **Comparison Between Cloud-based and Offline Sp eech Recognition Systems**. Dissertação (Masterarbeit) — Technische Universität München, 2017. Citado na página 34.

GRONDIN, F. et al. The ManyEars open framework. **Auton. Robots**, Kluwer Academic Publishers, v. 34, n. 3, p. 217–232, Apr 2013. ISSN 0929-5593. Citado na página 24.

GRONDIN, F. et al. ODAS: Open embeddeD Audition System. **arXiv**, Mar 2021. Disponível em: https://arxiv.org/abs/2103.03954v1. Citado 3 vezes nas páginas 22, 24 e 34.

GRONDIN, F.; MICHAUD, F. Lightweight and Optimized Sound Source Localization and Tracking Methods for Open and Closed Microphone Array Configurations. **arXiv**, Dec 2018. Disponível em: https://arxiv.org/abs/1812.00115v1. Citado na página 22.

GUO, R. et al. HoloSound: Combining Speech and Sound Identification for Deaf or Hard of Hearing Users on a Head-mounted Display. In: **ASSETS '20: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility**. New York, NY, USA: Association for Computing Machinery, 2020. p. 1–4. ISBN 978-1-45037103-2. Citado 2 vezes nas páginas 19 e 20.

HAKKANI-TÜR, D. et al. Eye Gaze for Spoken Language Understanding in Multi-modal Conversational Interactions. In: **ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction**. New York, NY, USA: Association for Computing Machinery, 2014. p. 263–266. ISBN 978-1-45032885-2. Citado na página 18.

HE, K. et al. Deep Residual Learning for Image Recognition. **arXiv**, Dec 2015. Disponível em: https://arxiv.org/abs/1512.03385v1. Citado na página 25.

HERCHONVICZ, A. L.; FRANCO, C. R.; JASINSKI, M. G. A comparison of cloud-based speech recognition engines. **Anais do Computer on the Beach**, v. 0, n. 0, p. 366–375, May 2019. ISSN 2358-0852. Citado na página 25.

HOHMANN, V. et al. The Virtual Reality Lab: Realization and Application of Virtual Sound Environments. **Ear Hear.**, Lippincott, Williams & Wilkins, v. 41, n. Supplement 1, p. 31S–38S, Nov 2020. ISSN 0196-0202. Citado na página 18.

LIAQUAT, M. U. et al. Localization of Sound Sources: A Systematic Review. **Energies**, MDPI, v. 14, n. 13, p. 3910, Jun 2021. ISSN 1996-1073. Citado 2 vezes nas páginas 18 e 19.

LÖWGREN, J.; STOLTERMAN, E. **Thoughtful Interaction Design: A Design Perspective on Information Technology**. [S.l.: s.n.], 2007. ISBN 978-0-26225657-5. Citado na página 38.

MANTHIRAM, A. An Outlook on Lithium Ion Battery Technology. **ACS Cent. Sci.**, American Chemical Society, v. 3, n. 10, p. 1063–1069, Oct 2017. ISSN 2374-7943. Citado na página 33.

MEHRA, R. et al. Potential of Augmented Reality Platforms to Improve Individual Hearing Aids and to Support More Ecologically Valid Research. **Ear Hear.**, Wolters Kluwer, v. 41, p. 140S–146S, Oct 2020. ISSN 1538-4667. Citado na página 13.

MISCHIE, S.; NXAĂRESC, G. Găxn-p. On Using ReSpeaker Mic Array 2.0 for speech processing algorithms. In: **2020 International Symposium on Electronics and Telecommunications (ISETC)**. [S.l.]: IEEE, 2020. p. 1–4. Citado na página 32.

MITTAL, S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. **J. Syst. Archit.**, North-Holland, v. 97, p. 428–442, Aug 2019. ISSN 1383-7621. Citado na página 28.

MOORE, B. C. J.; PETERS, R. W.; STONE, M. A. Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. **J. Acoust. Soc. Am.**, Acoustical Society of America, v. 105, n. 1, p. 400–11, Feb 1999. ISSN 0001-4966. Citado na página 16.

MORARU, O.-A. **Real-time subtitle for the hearing impaired in augmented reality**. Tese (Doutorado) — Wien, Wien, Austria, Nov 2018. Citado 2 vezes nas páginas 19 e 20.

NVIDIA-AI-IOT. **trt_pose**. 2022. [Online; accessed 16. Feb. 2022]. Disponível em: https: //github.com/NVIDIA-AI-IOT/trt_pose. Citado 2 vezes nas páginas 14 e 35.

OXENHAM, A. J.; KREFT, H. A. Speech Perception in Tones and Noise via Cochlear Implants Reveals Influence of Spectral Resolution on Temporal Processing. **Trends in hearing**, SAGE Publications, v. 18, May 2014. ISSN 2331-2165. Citado na página 16.

PAVLIDI, D. et al. Real-Time Multiple Sound Source Localization and Counting Using a Circular Microphone Array. **IEEE Trans. Audio Speech Lang. Process.**, IEEE, v. 21, n. 10, p. 2193–2206, Jul 2013. ISSN 1558-7924. Citado na página 22.

QUEK, F. et al. **Gesture, speech, and gaze cues for discourse segmentation**. [S.l.]: Institute of Electrical and Electronics Engineers, 2000. v. 2. ISBN 978-0-7695-0662. Citado na página 17.

SCHIPPER, C.; BRINKMAN, B. Caption Placement on an Augmented Reality Head Worn Device. In: **ASSETS '17: Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility**. New York, NY, USA: Association for Computing Machinery, 2017. p. 365–366. ISBN 978-1-45034926-0. Citado na página 19.

SLANEY, M. et al. Auditory Measures for the Next Billion Users. **Ear Hear.**, Wolters Kluwer, v. 41, p. 131S–139S, Oct 2020. ISSN 1538-4667. Citado na página 17.