IGOR FILLIPPE GOLDSTEIN

CargoAffect: aplicação da computação afetiva na experiência de usuário no contexto de motoristas de caminhões

São Paulo

2018

IGOR FILLIPPE GOLDSTEIN

CargoAffect: aplicação da computação afetiva na experiência de usuário no contexto de motoristas de caminhões

Trabalho de Conclusão de Curso apresentado à Escola Politécnica da Universidade de São Paulo

Orientadora: Prof.ª Dr.ª Lucia Vilela Leite Filgueiras

São Paulo

2018

AGRADECIMENTOS

À professora Lucia Vilela Leite Filgueiras, pela orientação exigente, contínua e propositiva, trazendo materiais de referência e de apoio, ideias, questionamentos e outros estímulos que foram essenciais para a realização deste projeto.

À empresa CargoX, e em especial ao meu gestor, Luiz, que permitiu o desenvolvimento deste projeto paralelamente às atividades do estágio que realizei lá, fornecendo apoio e todo o tempo que precisei para me dedicar a este projeto.

À Escola Politécnica, por ter me preparado ao longo de muitos anos de estudo e experiências, ainda que com algumas desilusões e abdicações no caminho, para a realização deste trabalho.

A minha namorada, Luísa, que foi uma grande fonte de apoio não só afetiva e emocional, mas também trazendo sugestões de melhoria no design do banner do projeto e em diversas outras fases do projeto.

Aos meus pais, por todo o inestimável apoio, amor e carinho ao longo de 24 anos e principalmente pela compreensão e paciência neste ano.

A todos os voluntários da fase de catalogação dos áudios e caminhoneiros e operadores de cargas entrevistados no estudo de usuário, que foram essenciais para implementar o sistema proposto e melhorar sua concepção ao longo do tempo.

A todos meus amigos que colaboraram direta ou indiretamente para a realização do trabalho: os amigos do Grupo de Teatro da Poli, do CoralUSP e dos grupos políticos e de militância, sem os quais eu dificilmente teria chegado ao fim da graduação.

RESUMO

A computação afetiva é uma área da computação que se baseia na captura do estado emocional de usuários, utilizando-o de alguma forma em aplicações interativas. Sabe-se que um dos principais sinais de expressão de emoções é a voz. Os motoristas de caminhões, em sua atividade profissional de transporte de carga, enfrentam situações de stress e de sono. No contexto da experiência de usuário atual de motoristas de caminhões, não há recursos disponíveis que possam ser embarcados no caminhão com a meta de alertar o profissional para situações de alto nível de estresse ou sono, reduzindo a probabilidade de acidentes. O objetivo deste trabalho é produzir uma aplicação com base na computação afetiva que, a partir do processamento do áudio das comunicações entre caminhoneiros e supervisores de transportadoras, identifique condições de estresse e sono e negocie com o caminhoneiro a redução do risco agregado a essas condições. Para tanto, a partir de pesquisa bibliográfica e pesquisa de campo, identificando as condições atuais de interação entre caminhoneiros e supervisores, extrai-se e processa-se um corpus de áudio, possibilitando a identificação das condições de estresse e sono. Constrói-se a aplicação sobre um algoritmo supervisionado e avalia-se a eficácia por testes com usuários. O resultado é obter um produto mínimo viável, testado com usuários e operacional da aplicação, que consiste em um aplicativo de captura de voz para o sistema Android e um serviço web que realiza o processamento de amostras de voz de caminhoneiros com base em um algoritmo supervisionado, tomando ações de acordo com o estado emocional detectado.

Palavras-Chave: Computação afetiva, processamento de áudio

ABSTRACT

Affective computing is a field based on the capture of users' emotional state and its utilization in interactive applications. It is known that one of the main emotion expression signals is the human voice. Truck drivers, in their daily professional activities, face stress and fatigue situations. In the context of truck drivers' current user experience, there are no available resources to be embedded on a truck in order to warn them about high-level stress or fatigue situations, reducing the possibility of accidents. The objective of this work is to design and develop an application based on affective computing that processes audio communications between truck drivers and carrier company supervisors, identifies stress and fatigue conditions and negotiates risk reductions with the drivers. For this purpose, in order to enable the identification of stress and fatigue conditions, an audio corpus is extracted and processed, based on bibliographical and field research and identification of the current interaction conditions between truck drivers and supervisors. The application is developed based on a supervised algorithm and its accuracy is evaluated by tests with users. The result is a minimum viable product, tested with users and operational, that consists on an application that captures voice for Android and a web service that processes the voice samples of truck drivers based on a supervised algorithm, taking actions according to the detected emotional state.

Keywords: Affective computing, audio processing

SUMÁRIO

1 INTRODUÇÃO

Esta monografia relata o projeto do sistema CargoAffect, composto por um aplicativo de captura de voz para o sistema Android e um serviço web que realiza o processamento de amostras de voz de caminhoneiros com base em um algoritmo supervisionado, tomando ações de acordo com o estado emocional detectado.

Nesta introdução, apresentam-se a motivação, o objetivo e a estrutura do trabalho.

1.1 MOTIVAÇÃO

As áreas de Interação Humano-Computador (IHC) e de Design de Interação passaram por grandes evoluções desde seu surgimento. O foco das atividades gradativamente deixou de estar apenas no desenvolvimento e avaliação de aplicações voltadas para o trabalho e incorporou aplicações orientadas ao lazer, como jogos, computação social, arte e ferramentas voltadas à criatividade.

Esse movimento direcionado a atividades culturais, artísticas, de entretenimento e de fruição causou mudanças nas perspectivas da interação humano-computador. A primeira refere-se ao atributo de qualidade buscado nos projetos. Enquanto no século passado, a qualidade almejada no *design* de sistemas interativos era a usabilidade, fortemente influenciada pela utilidade e pelo atendimento às metas do fazer, nas décadas deste século a qualidade é expressa de forma mais ampla, no conceito de experiência de usuário (UX) que engloba as metas do fazer (pragmáticas) e também as metas do ser (hedônicas). Assim, passou a ser necessário considerar as experiências estéticas e como lidar com as emoções dos usuários.

A segunda mudança é que o ser humano visto pelo computador não é mais somente um ponto na tela ou um caractere do teclado, mas ganha imagem, movimento, voz, corpo e emoções.

A computação afetiva é uma área de pesquisa recente dentro do contexto da Interação Humano-Computador, que busca se utilizar do estado emocional dos usuários de alguma maneira no desenvolvimento de sistemas. Esse ramo de estudo, que interliga as emoções à ciência da computação, foi originado em 1995 a partir de um artigo de Rosalind Picard, professora e pesquisadora do MIT.

Os estudos nesse âmbito buscam gerar sistemas e dispositivos que apresentem a capacidade de reconhecer, interpretar, processar ou simular emoções dos usuários. É uma área altamente interdisciplinar, que além da ciência da computação, engloba também a psicologia, a ciência cognitiva, a fenomenologia e a neurologia. Cada uma dessas áreas contribui com concepções diferentes no entendimento de modelos de emoções e dos papeis da emoção na racionalidade e nas relações humanas (HOOK, 2013).

Uma das principais motivações da computação afetiva é a possibilidade da simulação de empatia e emoções a partir dos sistemas desenvolvidos com esse intuito em mente. Os dispositivos e aplicações desenvolvidas passam a apresentar a potencialidade de interpretar os estados emocionais de seus usuários e, se desejado, adaptar seus comportamentos a partir desses estímulos recebidos, respondendo apropriadamente a cada situação e emoção detectada.

Assim, é possível utilizar-se da computação afetiva para resolver inúmeros problemas relacionados a questões emocionais. Sabe-se, por exemplo, que motoristas de transportes urbanos, como os caminhoneiros, estão sujeitos a diversos tipos de estressores e fatores que dificultam a locomoção e o trabalho (no caso de caminhoneiros), como o trânsito, o sono, a falta de sinalização e fiscalização nas vias rodoviárias, o mau estado de conservação das estradas e outros fatores.

A sonolência é uma situação recorrente no trabalho de caminhoneiros, que trabalham em jornadas longas, muitas vezes durante a noite e madrugada, sem horários de trabalho bem definidos. De acordo com Canani e Barreto (2001), os acidentes automobilísticos são, de fato, uma das principais consequências da sonolência excessiva.

Como a sonolência está associada à falta de atenção na direção do veículo, é possível encontrar na literatura alguns trabalhos de desenvolvimento de sistemas para monitoramento da atividade dos motoristas, como o sistema de reconhecimento facial desenvolvido por Dikkers et al (2004), que procura prever expressões de sonolência por meio da atividade dos olhos.

Além disso, segundo Quirino e Villemor-Amaral (2015), a combinação entre estresse e agressividade no trânsito pode desencadear reações comportamentais de risco, que ocasionam acidentes graves. Esse quadro se agrava no caso de caminhoneiros, que devem lidar com a pressão de cumprimento de horários e itinerários a seguir.

A questão do estresse também se mostra problemática nas comunicações realizadas entre operadores ou supervisores de cargas, que trabalham nas transportadoras, e os motoristas de caminhões. As centrais de comunicação com caminhoneiros ainda dependem muito das trocas de mensagens por meio de aplicativos de mensagens instantâneas. Um sistema que realize detecção de estresse e outros estados emocionais que interfiram na comunicação pode tornar a comunicação mais eficiente, ajudando todos os envolvidos.

1.2 OBJETIVO

Com base nos preceitos de computação e interação afetiva, este projeto apresenta como objetivo principal o desenvolvimento de um sistema composto por um aplicativo de captura de voz para o sistema Android e um serviço web que realiza o processamento de amostras de voz de caminhoneiros com base em um algoritmo supervisionado, tomando ações de acordo com o estado emocional detectado, para melhorar a experiência de direção dos motoristas.

A concepção do sistema descrito acima surgiu a partir de um estudo de usuários, que identificou o áudio como a melhor forma de identificar, no caso dos motoristas de caminhões, os estados emocionais de sono e estresse.

Consideram-se objetivos secundários deste projeto:

- Auxiliar os motoristas de caminhões a lidar com os vários fatores que acarretam estresse durante as locomoções realizadas em vias urbanas;
- Aplicar boas práticas de experiência de usuário no desenvolvimento do sistema interativo para que sua utilização seja mais proveitosa e duradoura;
- Investigar os conceitos e aplicações da computação afetiva, com foco no reconhecimento de emoções por meio da voz;
- Compreender de que maneira as emoções afetam o áudio e como elas podem ser detectadas por esse meio;
- Estudar aplicações de ciência de dados com base em algoritmos supervisionados.

1.3 METODOLOGIA

Para cumprir com os objetivos propostos, este trabalho cumpriu as seguintes etapas:

- Pesquisa bibliográfica, para estabelecer o referencial teórico
- Estudo de usuários, para conhecer o contexto da aplicação
- Determinação da solução tecnológica e especificação de requisitos
- Construção do corpus de áudio e treinamento do algoritmo de reconhecimento de emoções em áudio
- Desenvolvimento do sistema

1.3.1 PESQUISA BIBLIOGRÁFICA

O primeiro passo deste estudo foi a realização de uma pesquisa bibliográfica na área de computação afetiva e de reconhecimento de emoções por meio da voz. Para compreender esse processo de reconhecimento e determinar o algoritmo supervisionado a ser utilizado, realizou-se uma revisão sistemática da literatura com a seguinte questão de pesquisa: “qual algoritmo pode ser usado para detecção de emoções em arquivos de áudio de voz?”

A *string* de busca dessa revisão sistemática foi determinada iterativamente, chegando-se ao seguinte resultado: *“voice" "affective computing" "emotion detection" recognition algorithm "machine learning" AI -facial -image -music*. Foram excluídos artigos da área de saúde, sem foco em aspectos computacionais e realizou-se um recorte baseado nos seguintes aspectos: processo de reconhecimento, algoritmos utilizados, vantagens e dificuldades, emoções detectadas, caracterização dos áudios utilizados, aplicações e resultados obtidos.

1.3.2 ESTUDO DE USUÁRIO

Paralelamente à pesquisa bibliográfica, foi efetuada uma pesquisa de campo com caminhoneiros e operadores de cargas da empresa CargoX, com o objetivo de levantar requisitos de usuário e compreender quais são os maiores obstáculos encontrados por ambas as partes durante a realização do transporte de cargas.

1.3.3 DETERMINAÇÃO DA SOLUÇÃO E ESPECIFICAÇÃO DE REQUISITOS

A determinação da solução tecnológica desenvolvida ocorreu a partir dos resultados obtidos com as pesquisas bibliográficas e de campo (o estudo de usuários). Decidiu-se pela criação de um sistema composto por um aplicativo em Android que realiza a captura de áudios de voz de motoristas de caminhões e um serviço Web que realiza o processamento desses trechos de áudio com base em um algoritmo supervisionado treinado, tomando ações a partir do estado emocional detectado.

A partir do projeto da solução tecnológica, gerou-se uma especificação de requisitos, que trata de aspectos funcionais, como a possibilidade de gravar áudios por meio do aplicativo e exibir na tela do celular o resultado do processamento dos áudios gravados, e aspectos não funcionais, como o tempo de duração máximo para realizar o processamento de cada áudio e outras questões de usabilidade do sistema.

1.3.4 CONSTRUÇÃO DE CORPUS DE ÁUDIO E TREINAMENTO DO ALGORITMO

Para cumprir seu papel de auxílio aos motoristas, a aplicação deve realizar o processamento de *snippets* (porções) de áudio de voz de caminhoneiros, identificar condições de sono ou estresse e realizar, de alguma maneira, a negociação de redução de risco com o usuário.

O treinamento do algoritmo supervisionado que realiza esse processamento foi viabilizado pela coleta de um corpus de áudio com voz de motoristas de caminhões. Para a criação desse corpus, gerado a partir de segmentos de mensagens de áudio enviadas por caminhoneiros a operadores de cargas, foram coletadas amostras de áudio obtidas a partir de conversações entre caminhoneiros e operadores da empresa CargoX.

Com o corpus bem definido, partiu-se ao processo de catalogação, com o apoio de voluntários para definir a emoção (com base no protocolo SAM – *Self Assessment Manikin*) de cada amostra de áudio coletada (BRADLEY e LANG, 2007).

Em seguida, realizou-se a extração de parâmetros e elementos das amostras de áudio necessárias para sua posterior classificação, como os MFCC (*Mel-Frequency Cepstral Coefficients*). Esses elementos extraídos foram utilizados como entrada para os algoritmos supervisionados para realização do treinamento e, posteriormente, para detecção dos estados emocionais de amostras de áudio.

As etapas de extração de parâmetros, treinamento e teste do algoritmo foram realizadas utilizando a biblioteca de análise de áudios *pyAudioAnalysis*, desenvolvida por Theodoros Giannakopoulos (2015). Por fim, realizou-se uma análise estatística dos resultados obtidos com os algoritmos supervisionados, com o objetivo de verificar sua precisão e desempenho.

1.3.5 DESENVOLVIMENTO DO SISTEMA

O sistema concebido apresenta dois subsistemas, como mencionado anteriormente: um aplicativo para o sistema operacional Android e um serviço Web que realiza o processamento dos áudios enviados pelo aplicativo com base em um algoritmo supervisionado previamente treinado.

O aplicativo foi desenvolvido utilizando a plataforma Android Studio, na linguagem Java. Já o serviço Web foi desenvolvido na linguagem Python, utilizando o *framework* Flask para disponibilizar a funcionalidade de processamento de áudios via Web.

1.4 ESTRUTURA DO TRABALHO

Esta monografia estrutura-se da seguinte forma:

O Capítulo 1 é esta Introdução.

O Capítulo 2, sobre aspectos conceituais da computação afetiva, contém o estado-da-arte da área de interesse, mostrando os conceitos relevantes e os trabalhos correlatos.

O Capítulo 3, Estudo de Usuário, apresenta os resultados obtidos na pesquisa de campo, por meio de entrevistas com stakeholders (supervisores e operadores de risco de empresa de transportes) e caminhoneiros.

O Capítulo 4 apresenta as tecnologias utilizadas, incluindo o corpus, o algoritmo escolhido e os resultados do treinamento.

O Capítulo 5 apresenta a especificação de requisitos do sistema.

O Capítulo 6 apresenta, com maiores detalhes, os passos seguidos para desenvolver o procedimento de detecção de emoções por meio de áudio.

O Capítulo 7 apresenta os resultados e discussões.

O Capítulo 8 contém as considerações finais deste projeto.

2 ASPECTOS CONCEITUAIS DA COMPUTAÇÃO AFETIVA

Este capítulo apresenta o referencial teórico deste trabalho. Descreve-se brevemente a perspectiva da computação afetiva, partindo-se do entendimento das emoções e de sua manifestação. Considerando-se que a voz é uma importante manifestação dos estados emocionais, trata-se ainda das técnicas de processamento do áudio para a extração das características emocionais.

2.1 PERSPECTIVA EMOCIONAL NA INTERAÇÃO HUMANO-COMPUTADOR

A computação afetiva desenvolve-se a partir do entendimento das emoções humanas e suas manifestações, para em seguida tratar da aquisição de dados e do processamento computacional desses dados. Nesta seção, parte-se da história do estudo das emoções para evidenciar como as emoções se manifestam e podem ser percebidas pelos sistemas computacionais.

2.1.1 HISTÓRICO DAS EMOÇÕES

De acordo com Rosenwein e Cristiani (2018), os primeiros estudos e teorizações voltadas às emoções ocorreram na época da Grécia Antiga. Filósofos como Aristóteles e Platão dedicaram parte de suas obras a esse tópico. De acordo com Aristóteles (322 a.e.c.), as emoções são todas as sensações que fazem com que o ser humano mude de opinião em relação a seus julgamentos e são acompanhadas por prazer e dor, como a raiva, a pena e o medo. Ou seja, Aristóteles já reconhece o fator cognitivo que há nas emoções, dado que dependem da avaliação e julgamento de cada indivíduo em uma determinada situação. A emoção e a ação dependem primariamente da razão, na sua visão.

Platão, por sua vez, foi possivelmente o primeiro a dividir a alma (ou mente) em três diferentes partes ou funções: a parte apetitiva ou concupiscente (mais instintiva e irracional, relativa aos desejos), a colérica ou irascível (relativa à proteção do corpo e segurança) e a racional (ligada ao conhecimento e à sabedoria). Cada uma dessas funções é responsável por guiar as ações e sentimentos humanos, sendo que a razão é a mais importante de todas. Essa fórmula tripartite foi muito abordada em estudos psicológicos desde então (LAZARUS, 1999).

Com o tempo, as discussões sobre as emoções tornaram-se mais complexas. René Descartes tratou da separação entre o corpo e a mente, uma dualidade muito repercutida desde então, enquanto John Locke atribuiu emoções, como o amor e a culpa, ao produto da experiência (ROSENWEIN e CRISTIANI, 2018).

Em 1872, Charles Darwin publicou um estudo denominado “A Expressão das Emoções no Homem e nos Animais”, que discorre sobre a expressão de emoções em animais e seres humanos. Desde então, as emoções se tornaram o foco principal de estudo de vários cientistas e psicólogos, como Paul Ekman e William James, ganhando definições diferentes ao longo do tempo.

De acordo com Hook (2013), houve uma nova onda de pesquisa nos anos 1990 focada na discussão do papel das emoções, envolvendo áreas como a psicologia, a neurologia, a medicina e a sociologia. As discussões incitadas a partir desse período colocaram a emoção em um novo patamar na dualidade razão-emoção, já que foi possível perceber que as emoções são uma das principais bases que permitem a racionalidade e as relações mútuas entre seres humanos.

Além disso, foi a partir desse período, com base nos avanços obtidos no estudo das emoções, que ganharam força as pesquisas e inovações tecnológicas considerando estados emocionais dos usuários dos sistemas desenvolvidos, permitindo o surgimento de áreas como a computação afetiva.

2.1.2 MODULAÇÃO SÊNTICA

Manfred Clynes (1977) cunhou o adjetivo sêntico, derivado de “*sentio*” (latim para sentir), para se referir aos estados emocionais, buscando escapar da conotação negativa geralmente associada às emoções.

O trabalho de Clynes desenvolve o fato de que as emoções são inerentemente ligadas ao sistema motor humano. É graças a essa ligação que a comunicação de emoções se torna possível. O estado emocional é expressado por modulações sutis específicas das ações motoras envolvidas, que correspondem precisamente à manifestação do estado sêntico. (CLYNES, 1977)

Clynes formulou princípios para a comunicação sêntica (emocional), que ocorre a partir de estados sênticos, uma descrição dada a estados emocionais (usada para escapar da conotação negativa associada à ideia de emoção). Ele enfatiza, em sua

obra, que emoções modulam nossa comunicação física e o sistema motor atua como um comunicador do estado emocional.

O problema nesse aspecto, e também uma das grandes questões em torno da computação afetiva, se resume ao fato de que não é possível realizar uma medição absolutamente objetiva do estado emocional: ela depende de autoavaliações, que podem ser altamente variáveis e subjetivas. Entretanto, é possível medir respostas fisiológicas (expressões faciais ou amostras de voz, por exemplo) que frequentemente surgem durante a expressão de emoções. Ou seja, pode-se medir fisiologicamente as emoções que já estão manifestadas.

Uma questão que emerge a partir da medição de emoções a partir de respostas fisiológicas é a padronização universal de respostas emocionais. Muitas vezes, indivíduos diferentes exibem respostas fisiológicas diferentes para um mesmo estado emocional.

Embora seria uma grande realização resolver o problema do reconhecimento universal, se o problema pode ser resolvido com base simplesmente na manifestação de um único interlocutor (no caso, o usuário ou grupo de usuários do sistema), então a tarefa de detecção do estado emocional é realizada de modo satisfatório. (PICARD, 1995)

Os experimentos de identificação de estados emocionais a partir de observações de expressões físicas só necessitam demonstrar padronização consistente para um único indivíduo (ou grupo de indivíduos) em um dado contexto perceptível. É possível adquirir informação contextual e de percepção do ambiente (se a pessoa está subindo escadas ou a temperatura mudou, por exemplo) para identificar respostas emocionais condicionadas em fatores perceptíveis não-emocionais. O contexto perceptível pode incluir não só informações físicas, mas também cognitivas (por exemplo, se a pessoa investiu no mercado de ações e pode estar ansiosa em uma época de crise).

Há várias respostas fisiológicas que variam com o tempo e que podem potencialmente ser combinadas para auxiliar no reconhecimento de estados emocionais. Elas incluem a taxa de batimentos cardíacos, a pressão sanguínea, o pulso, a dilatação das pupilas, a respiração, a condutância na pele, a temperatura ou outros sinais fisiológicos.

Em um de seus experimentos, por exemplo, Clynes realizou a medição da pressão nos dedos da mão em milhares de pessoas, verificando a revelação de traços

distintos de “forma emocional” para estados como a ausência de emoção, raiva, ódio, tristeza, amor, alegria, sexo e reverência.

Outra forma muito reconhecida de modulação sêntica é a voz humana. Sabe-se que emoções vocais podem ser entendidas até mesmo por crianças pequenas, antes mesmo que entendam o que está sendo dito, e por cachorros. A comunicação falada é maior que as palavras ditas em si, dado que engloba diversos aspectos da voz e da entonação.

Uma grande variedade de características de discurso é modulada pela emoção. Murray e Arnott (1993) estudam essas características, dividindo-as em três principais categorias: qualidade, timing e tom/timbre da voz. Eles discutem como esses parâmetros podem ser manipulados para dar a computadores a habilidade de falar com emoção.

2.1.3 EXPERIÊNCIA, EXPRESSÃO E ESTADO EMOCIONAL

Picard (1995), em seu artigo sobre a computação afetiva, faz uma diferenciação das terminologias mais utilizadas ao se referir ao estudo e reconhecimento das emoções. As três terminologias que a autora define são:

- Estado emocional/afetivo/sêntico – refere-se ao estado dinâmico do indivíduo ao experienciar determinada emoção;
- Experiência emocional – todos os sinais e sentidos que são conscientemente percebidos em um estado emocional. Essa experiência pode ser considerada como o “sentimento emocional”;
- Expressão emocional – o estado emocional de uma pessoa não pode ser diretamente observado por outra. O que se revela, voluntariamente ou não, é a expressão emocional, ou em outras palavras, “sintomas” emocionais. Essa expressão que passa pelo sistema motor (modulação sêntica) ajuda outras pessoas a deduzir o estado emocional de uma pessoa.

Quando indivíduos são requisitados ou estimulados (a partir de histórias ou filmes, por exemplo) a experienciar um estado emocional particular, eles podem ou não expressar o seu estado emocional de fato. Isso significa que a expressão livre e deliberada é mais útil para inferir um estado emocional.

2.2 COMPUTAÇÃO AFETIVA

De acordo com Picard (1995), a computação afetiva simboliza todo tipo de computação que se relaciona com emoções, surge a partir delas ou as influencia. Esse campo de pesquisa na computação é relativamente recente e teve seu início marcado por resultados no reconhecimento de expressões faciais e síntese de inflexões vocais.

Entretanto, sabe-se hoje que há uma grande variedade de medições fisiológicas disponíveis que podem ajudar a indicar o estado emocional oculto do usuário. Picard propôs alguns modelos possíveis para a identificação do estado emocional de um indivíduo, tratando o reconhecimento como um problema de reconhecimento dinâmico de padrões.

Atualmente, os computadores estão começando a adquirir a habilidade de expressar e reconhecer emoções e podem em breve inclusive ganhar a habilidade de “possuir emoções”. O papel essencial da emoção na cognição e percepção humana, como demonstrado por estudos neurológicos, indica que computadores afetivos podem não só obter melhor performance no auxílio aos seres humanos, mas também podem ter melhor desempenho na tomada de decisões.

2.2.1 RECONHECIMENTO DE PADRÕES DE ESTADOS EMOCIONAIS

Pensamentos e sentimentos são expressados e comunicados por meio da voz, gestos, música e outras formas de expressão, sendo todas elas imperfeitas e limitadas. Apesar de que seja possível atualmente distinguir novas regiões e níveis de atividade no cérebro com a ajuda de novos dispositivos de medição, ainda não é possível acessar diretamente os pensamentos e sentimentos de outras pessoas.

Entretanto, o reconhecimento científico de estados afetivos parece ser possível em muitos casos, por meio da medição da modulação sêntica. Não se mede o estado emocional diretamente, mas sim expressões observáveis desse estado. Essas medições levam a reconhecimento bem-sucedido na maior parte dos casos de expressões voluntárias, mas podem também ser úteis durante expressões involuntárias.

A tarefa de “reconhecer emoções” deve ser interpretada como a medição de observações de comportamentos do sistema motor que correspondem, com alta probabilidade, a uma emoção ou conjunto de emoções subjacente.

Apesar de ser uma tarefa de grande dificuldade, o reconhecimento de estados emocionais expressados parece ser uma tarefa mais fácil que o reconhecimento de pensamentos. No reconhecimento de padrões, a dificuldade do problema cresce com o número de possibilidades variadas. O número de possíveis pensamentos do ser humano é ilimitado e os pensamentos humanos não são facilmente categorizados em conjuntos menores de possibilidades. O reconhecimento de pensamentos, mesmo com o aumento na sofisticação de técnicas de imagem e escaneamento, possivelmente é o maior “problema inverso” imaginável.

Em contraste, para reconhecimento de emoção, um número relativamente pequeno de categorias simplificadoras para emoções é normalmente proposto. As quatro emoções mais comuns que aparecem nas listas de emoções básicas ou prototípicas na literatura são: medo, raiva, tristeza e alegria.

Há também autores que se preocupam mais com determinadas dimensões da emoção, como sua negatividade ou positividade, que com a quantidade de emoções básicas observáveis. Nesse caso, três dimensões aparecem mais frequentemente. Embora seus nomes variem, as duas categorias mais comuns para as dimensões são a excitação (calmo ou relaxado/excitado ou estimulado) e valência (negativo ou triste/positivo ou feliz). A terceira dimensão tende a se chamar “controle” ou “atenção”, relativa à fonte interna ou externa da emoção (desprezo/surpresa).

No trabalho de Bradley e Lang (1994) que define o instrumento de avaliação de emoções SAM (*Self-Assessment Manikin*), utilizam-se as escalas de valência, excitação e dominância (correspondente à dimensão de controle ou atenção mencionada acima) para medição de respostas emocionais a estímulos de qualquer tipo, como pinturas ou sons.

Faz sentido simplificar as possíveis categorias de emoções para computadores, para que eles possam começar em um nível mais básico, reconhecendo emoções mais óbvias. Na computação afetiva, os problemas de reconhecimento e modelagem são simplificados pela suposição de um conjunto pequeno de emoções discretas, ou um número pequeno de dimensões.

O princípio da exclusividade de estados emocionais de Clynes (1977) sugere que não é possível expressar uma emoção quando está se sentindo outra. O autor

enfatizou, em seu artigo, a “pureza” dos estados emocionais básicos e sugeriu que todos os outros estados emocionais são derivados desse pequeno conjunto de estados puros (por exemplo, melancolia seria uma mistura de amor e tristeza).

Dado que o ser humano esteja a cada momento em um estado emocional, como ódio, então certos valores de observações do sistema motor, como uma voz tensa, uma expressão notória ou uma pressão maior dos dedos são mais prováveis. As taxas de respiração e batimentos cardíacos também podem aumentar.

Em contraste, com sentimentos de alegria, a voz pode subir de tom, a face revelar um sorriso e a pressão aplicada nos dedos pode fazer com que aparentem levemente agitados.

2.2.2 MODELOS DE ESTADOS EMOCIONAIS

Para conseguir realizar o reconhecimento de padrões emocionais, como mencionado na seção anterior, a partir da captura e processamento de dados fisiológicos, é necessário criar modelos de estados emocionais.

A figura 1, a seguir, mostra um exemplo de modelo para estados emocionais, o Modelo Oculto de Markov (HMM). O estado emocional de uma pessoa (no caso da figura, interesse, angústia ou alegria) não pode ser observado diretamente, mas pode-se observar manifestações e expressões desse estado. O modelo de Markov (oculto) da imagem caracteriza probabilidades de transições entre três estados “escondidos” (I, Al e An), assim como probabilidades de observações (formas sênticas mensuráveis, como características da inflexão da voz, V) em um dado estado. (PICARD, 1995)

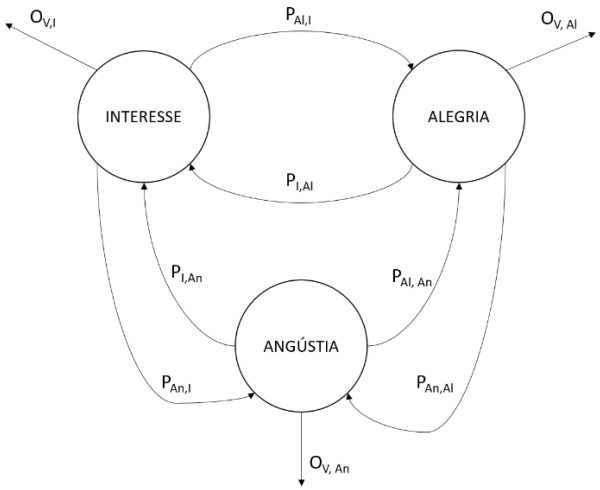

Fig. 1 – Modelo Oculto de Markov de estados emocionais (PICARD, 1995)

O exemplo mostra somente três estados para facilitar a ilustração, mas é possível incluir mais estados, inclusive um estado de “sem emoção”. A ideia básica é que a qualquer instante um indivíduo se encontra em um determinado estado e é possível transitar entre estados com determinadas probabilidades. Por exemplo, no caso da figura, espera-se que a probabilidade de mover de um estado de interesse para alegria seja maior que a de mover de um estado de angústia para alegria.

O HMM é treinado com base em observações, ou seja, qualquer medição de modulação sêntica. HMMs diferentes podem ser treinados para diferentes contextos ou situações. Por isso, as probabilidades e estados podem variar dependendo do contexto da situação (local, pessoas envolvidas).

Como mencionado anteriormente, as manifestações do sistema motor, como a pressão nos dedos, a voz ou até mesmo os movimentos de respiração, são utilizadas para determinar a modulação sêntica envolvida e variam para cada estado do modelo. As medições obtidas a partir dessas manifestações são usadas para treino do algoritmo e posterior reconhecimento.

Uma variedade de técnicas de aprendizagem e reconhecimento de padrões está disponível, porque o reconhecimento de estados emocionais pode ser realizado como um problema de classificação supervisionada, ou seja, em que classes podem ser especificadas a priori.

2.2.3 CATEGORIAS DA COMPUTAÇÃO AFETIVA

Embora a expressão e reconhecimento de emoções sejam importantes na área da interação humano-computador, atribuir emoções ao comportamento motivacional do computador é um problema diferente.

A noção de um computador “ter emoções”, de acordo com Picard (1995), se dá por meio de uma dimensão descritiva, ou seja, por exemplo, chama-se de frustração quando o computador se encontra em um estado no qual recebeu muita informação conflitante. É pouco provável que computadores eletrônicos possuam sentimentos de fato, mas há um paralelo em relação a possibilidade de computadores terem consciência.

Dado isso, Picard apresenta quatro categorias de computação afetiva, com foco no reconhecimento e expressão de emoções:

- Computador não expressa nem reconhece emoção – a maior parte dos computadores se encaixa nessa categoria. São computadores que não são pessoais ou empáticos, sem reconhecimento ou expressão de emoções;
- Computador expressa, mas não reconhece emoção – nessa categoria, se encaixa o desenvolvimento de vozes computadorizadas com entonação natural e faces computadorizadas com expressões naturais;
- Computador reconhece, mas não expressa emoção – essa categoria permite a um computador a percepção do estado emocional do usuário, permitindo a ele ajustar sua resposta de modo que possa, por exemplo, fazê-lo ser um professor ou assistente melhor;
- Computador reconhece e expressa emoção – essa categoria maximiza a comunicação emocional entre humano e computador, providenciando computação verdadeiramente pessoal e "*user-friendly*". Ela não implica que o computador é dirigido por suas emoções.

Além desses, é possível considerar outros casos. Por exemplo, também é possível levar em conta a habilidade de induzir emoções no usuário ou não, ou a habilidade de agir baseado em emoções ou não. Mas essas possibilidades não são abordadas no artigo de Picard devido a implicações éticas e filosóficas que fogem do escopo.

2.2.4 SIMETRIA AFETIVA

Computadores que leem emoções (ou seja, deduzem estados emocionais escondidos baseando-se em observações psicológicas e comportamentais) também devem mostrar o que estão lendo ao usuário. Em outras palavras, a interação afetiva com computadores pode facilmente dar um feedback direto, que é geralmente ausente na interação humana. (PICARD, 1995)

Os modelos de "estados ocultos" propostos acima podem revelar seus estados ao usuário, indicando que estado emocional o computador reconheceu. Esse feedback ajuda não só a depurar o desenvolvimento desses sistemas, mas é também útil para alguém que sente que outras pessoas confundem suas expressões. Esses indivíduos podem nunca obter um feedback preciso de pessoas para saber como

melhorar suas habilidades de comunicação, mas, por outro lado, o computador pode ceder feedback pessoal contínuo.

2.3 RECONHECIMENTO DE EMOÇÕES POR ÁUDIO

A determinação do procedimento de reconhecimento foi realizada a partir de uma revisão bibliográfica, que revelou uma sequência de passos bem definida e recorrente em quase todos os artigos analisados, com poucas alterações entre cada um. O procedimento é descrito a seguir, passo a passo.

A primeira etapa ao se desenvolver qualquer aplicação baseada em reconhecimento de emoções com base em voz é a coleta e construção de um corpus de áudio, com amostras de voz que serão utilizadas para treinar e testar o algoritmo supervisionado escolhido.

Lubis et al (2014), por exemplo, realizaram a construção de um corpus baseado em amostras de áudio obtidas de gravações de programas de *talk show* da Indonésia, que apresentam discussões sobre diversos temas, o que favorece o surgimento de manifestações emocionais de diversos tipos. Entretanto, é possível também utilizar corpora de áudio já existentes previamente.

Após a coleta de amostras de áudio para criação do corpus, deve-se realizar a etapa de tratamento dos arquivos obtidos. Pode-se segmentar as amostras em pequenos pedaços de áudio, de 5 a 10 segundos, para facilitar a classificação de cada amostra de acordo com um estado emocional determinado. A remoção de ruído por meio de *softwares* de edição de áudio também é uma prática recomendada.

Com as amostras de áudio tratadas, pode-se proceder à etapa de catalogação. Para realizar essa etapa, deve-se contar com o apoio voluntário de pessoas que serão responsáveis por catalogar cada amostra de áudio de acordo com a emoção que acredita ser correspondente. É recomendável realizar essa etapa com ao menos três pessoas, de gêneros diferentes, em sessões padronizadas. (BURKHARDT et al, 2009)

Em seguida, deve-se realizar a extração de elementos e características das amostras de áudio que serão utilizadas como entradas pelo algoritmo supervisionado para a classificação. De acordo com Garg e Sehgal (2015), é possível classificar essas características em alguns grupos. Há as características acústicas, geralmente mais utilizadas nessa situação, que estão relacionadas a aspectos como volume, tom,

duração e frequências máximas, mínimas e médias. Outras classificações menos utilizadas incluem as características prosódicas (relativas ao ritmo de fala, ênfases e entonação) e as características paralinguísticas (combinação de aspectos espectrais e prosódicos, dinâmica da fala em termos de tom e ritmo).

Uma das características de áudio mais frequentemente usadas no reconhecimento de emoções são os MFCC (*Mel-Frequency Cepstral Coefficients*, ou coeficientes Mel-cepstrais). Esses coeficientes compõem, em conjunto, um cepstro de frequência Mel, que corresponde a uma representação do espectro de potência de um som, baseado na transformada inversa de Fourier do espectro do sinal em escala logarítmica. Esse espectro de potência pode ser usado como um vetor de características para representar a voz humana e serve de entrada para os algoritmos supervisionados que realizam o reconhecimento de emoção pela voz.

Após extrair as características de cada amostra de áudio, elas devem ser utilizadas no algoritmo supervisionado classificador para realizar o reconhecimento automático da emoção. Recomenda-se separar as amostras de áudio em dois grupos: um grupo de treinamento do algoritmo e um grupo de teste, para verificar se o algoritmo realiza a classificação de acordo com o esperado com base na etapa de catalogação.

Segundo Garg e Sergal (2015), um dos classificadores mais importantes e utilizados para essa tarefa é a máquina de vetores de suporte (*Support Vector Machines* – SVM). As SVMs consistem em modelos algorítmicos de aprendizagem supervisionada que analisam dados e reconhecem padrões, sendo frequentemente utilizados para classificação e análise de regressão.

Basicamente, a SVM, assim como outros algoritmos de classificação, recebe um conjunto de dados de entrada e prevê, para cada conjunto, qual de duas ou mais possíveis classes representa a saída.

Há trabalhos como o de Deshpande et al (2017), por outro lado, que realizam o processamento das amostras de áudio com mais de um classificador, para poder comparar o desempenho de cada algoritmo ao final do processo. No caso desse trabalho, os autores utilizaram, além das máquinas de vetores de suporte, os algoritmos de K vizinhos mais próximos (*k-nearest neighbors* – kNN) e floresta aleatória (*Random Forest*).

Após realizar o processamento das amostras de áudio com base em um algoritmo supervisionado, a última etapa do processo é a análise estatística dos resultados obtidos.

3 ESTUDO DE USUÁRIO

Neste capítulo, descreve-se o estudo de usuário no contexto da aplicação, realizado como parte da disciplina PCS3573 (Interação Humano-Computador), da Escola Politécnica, ministrada pela Prof.ª Dr.ª Lucia Vilela Leite Filgueiras.

Estudos de usuário são relevantes para prover imersão no contexto do problema e identificar requisitos de usuário para a solução. No caso deste projeto, foi importante para realizar decisões de projeto e determinar qual seria o sinal fisiológico utilizado para realizar a detecção do estado emocional dos motoristas de caminhões.

3.1 *STAKEHOLDERS*, PAPEIS E VARIÁVEIS DE PERFIL

No contexto em que esta aplicação é desenvolvida, foram identificados os seguintes *stakeholders*: os motoristas de caminhões, como usuários primários; os operadores e supervisores de risco de empresas de transporte, que se comunicam com os motoristas, como usuários secundários; o desenvolvedor da aplicação e a diretoria das empresas de transporte, como demais interessados.

Os caminhoneiros foram escolhidos como usuários primários e alvo do estudo de usuário, dado que o sistema a ser desenvolvido apresenta como principal objetivo a obtenção de melhorias na sua experiência de direção, trazendo maior segurança e reduzindo a ocorrência de situações de risco.

Na aplicação, os motoristas de caminhões são atores: o nível de estresse e sonolência percebido por meio de amostras captadas de voz dos caminhoneiros são entradas do sistema. Além disso, os caminhoneiros devem receber um feedback da aplicação, após processamento dos arquivos de áudio, de acordo com os estados emocionais detectados.

Para consolidar o perfil da população de caminhoneiros, algumas variáveis relevantes para a interação com o sistema são:

- faixa etária - no cruzamento dos resultados obtidos, pode-se verificar se a idade dos motoristas de caminhões influencia na aceitação do sistema, na habilidade com o uso de smartphones ou na frequência das situações de estresse e sonolência;

- nível de escolaridade - da mesma maneira que a faixa etária, é um fator relevante que pode influenciar na aceitação do sistema e se relacionar com outros resultados obtidos em campo;
- anos de experiência como caminhoneiro - com esse dado, pode-se verificar se a experiência dos motoristas tem alguma relação com a frequência das situações de estresse e cansaço em seu cotidiano;
- quantidade de horas de sono - dado importante para analisar os resultados obtidos em relação às situações de cansaço e sonolência relatadas pelos caminhoneiros; e
- posse de smartphone com câmera - assumindo que o sistema a ser desenvolvido inclua um aplicativo para envio e análise de mensagens de áudio, é importante garantir que a população de foco majoritariamente tenha acesso ao sistema.

Os operadores e supervisores de risco de empresas de transporte também apresentam um papel importante no contexto da aplicação, dado que são os responsáveis por garantir que os caminhoneiros realizem a entrega das cargas sem contratempos e, se possível, sem riscos ocasionados por oscilações negativas de estados emocionais.

Eles entram em contato com os caminhoneiros com grande frequência em sua jornada de trabalho, para supervisionar os transportes de carga e garantir que a entrega seja bem-sucedida.

O contato ocorre antes da saída da carga, para verificar se os caminhoneiros receberam todos os documentos necessários para deixar o pátio e um adiantamento do valor total que devem receber pela viagem. Além disso, ocorre também durante a viagem, para fins de monitoramento, e ao fim da viagem, para confirmação da entrega e verificação do recebimento do restante do valor da viagem (saldo) por parte do caminhoneiro.

No contexto do sistema a ser desenvolvido, os operadores e supervisores devem receber um *feedback* da aplicação, após o processamento dos áudios enviados pelos caminhoneiros durante o transporte de cargas.

3.2 NECESSIDADES

Após a identificação dos papeis primários e secundários da aplicação a ser desenvolvida, é possível delimitar as principais necessidades e obstáculos de cada um dos envolvidos, sabendo que o sistema irá atender a algumas dessas necessidades e diminuir os obstáculos, com foco na experiência de usuário do papel primário, o motorista de caminhão.

As necessidades de caminhoneiros detectadas estão relacionadas aos seguintes tópicos:

- seus maiores problemas enfrentados antes, durante e depois do transporte de cargas;
- como fazem para se manter suficientemente descansados entre uma viagem e outra;
- o que fazem em situações de estresse, durante o trabalho;
- com que frequência enfrentam situações de estresse;
- o que fazem em situações de cansaço ou sonolência, durante o trabalho;
- com que frequência enfrentam situações de cansaço ou sonolência; e
- que outros fatores adversos atrapalham suas viagens ou comunicações com operadores e supervisores.

Já no caso dos operadores e supervisores de empresas de transporte, as necessidades se resumem aos seguintes pontos:

- como melhorar a comunicação com os caminhoneiros;
- como procedem quando percebem que o caminhoneiro está irritado ou cansado;
- com que frequência entram em contato com caminhoneiros;
- que meio mais utilizam para falar com caminhoneiros; e
- que fator atrapalha mais a comunicação entre eles e os caminhoneiros – estresse ou cansaço.

3.3 INSTRUMENTOS E AMOSTRA

Para a realização da fase de pesquisa de campo, foram escolhidas duas técnicas de estudo de usuário, com o objetivo de garantir resultados mais satisfatórios tanto em termos qualitativos como em termos quantitativos.

Uma das técnicas é a entrevista em profundidade, um instrumento de cunho reflexivo e qualitativo, para obter mais informações com os operadores e supervisores de risco de uma empresa de transporte. Seis pessoas, que trabalham na empresa de tecnologia e transportes CargoX, foram entrevistadas por meio dessa técnica. São colaboradores que entram em contato diário com os caminhoneiros, para realizar monitoramento e garantir que o transporte de cargas ocorra sem maiores problemas.

A escolha dessa técnica se deu pelo fato de permitir que a entrevista tenha um caráter exploratório e mais aberto. Para consolidar uma lista de requisitos para o sistema que realmente cumpre com os objetivos iniciais, é necessário realizar uma imersão no contexto dos caminhoneiros, operadores e a comunicação que ocorre entre eles. Assim, a entrevista em profundidade é adequada para conseguir aprofundar detalhes mais específicos e dificuldades que encontram cotidianamente no trabalho.

A outra técnica escolhida foi a entrevista semiestruturada, para obter mais informações com motoristas de caminhões. A entrevista semiestruturada é um modelo mais fechado de entrevista, que se assemelha a um questionário por ter questões fechadas (múltipla escolha), mas também permite a realização de questões abertas, de resposta livre. Assim, caracteriza-se por ser uma técnica de cunho qualitativa e quantitativa ao mesmo tempo, além de também possuir um caráter reflexivo.

Esse modelo foi escolhido para conseguir obter uma boa quantidade de informações de cada caminhoneiro, em um intervalo de tempo mais curto. Assim, não se ocupa muito tempo de cada motorista e torna-se viável a realização da entrevista com uma quantidade maior de caminhoneiros, para conseguir obter uma amostra razoável para realizar uma análise quantitativa dos resultados.

Para a realização da entrevista em profundidade com operadores e supervisores de risco, obteve-se uma amostra de seis pessoas entrevistadas, colaboradores da empresa CargoX.

Foram também realizadas dezesseis entrevistas semiestruturadas com motoristas de caminhões.

3.4 RESULTADOS DA PESQUISA DE USUÁRIOS

Nesta seção, apresentam-se os resultados da pesquisa feita com os usuários, que descreve o contexto do trabalho dos operadores e dos caminhoneiros.

3.4.1 ENTREVISTA EM PROFUNDIDADE (OPERADORES)

Por meio da entrevista em profundidade realizada, foi possível verificar que o meio de comunicação principal entre caminhoneiros e operadores é, unanimemente, o aplicativo de mensagens instantâneas WhatsApp (exceto por algumas ocasiões de caráter mais emergencial, em que se realizam ligações telefônicas). Geralmente, a comunicação entre eles se dá por mensagens de áudio.

Além disso, o contato com caminhoneiros em situação de estresse é muito frequente, por conta de condições de trabalho extenuantes, atrasos no recebimento de documentos necessários para transportar a carga e atraso para receber um adiantamento em dinheiro, que os caminhoneiros costumam precisar para conseguir colocar o combustível necessário para a viagem em seus veículos.

O estresse dos caminhoneiros afeta negativamente o transporte das cargas, aumentando os riscos da ocorrência de acidentes, e, principalmente, a comunicação entre eles e os supervisores, já que o estresse gera confusões na troca de informações necessárias para a fluidez do transporte e ocasiona desconfortos aos operadores que recebem as mensagens.

O cansaço e sonolência dos caminhoneiros também é percebido pelos operadores e supervisores, mas com menor frequência que os casos de alto nível de estresse. Além disso, o cansaço não afeta a comunicação entre caminhoneiros e operadores tanto quanto o estresse.

Por fim, alguns dos supervisores já monitoraram caminhoneiros que sofreram acidente de trânsito, mas é uma ocorrência mais incomum e não necessariamente se relaciona aos fatores de risco discorridos anteriormente. De fato, todos os operadores não souberam dizer os motivos que levaram à ocorrência dos acidentes.

3.4.1 ENTREVISTA SEMIESTRUTURADA (CAMINHONEIROS)

Foram entrevistados, até o presente momento, dezesseis motoristas de caminhões, alguns por telefone e outros, pessoalmente. Essas entrevistas permitiram a definição inicial do perfil desse público.

Metade dos caminhoneiros entrevistados apresenta entre 30 e 39 anos, com outros 37,5% na faixa de idade entre 40 e 49 anos. Ou seja, em média, pode-se considerar que a faixa etária foco se encontra entre 35 e 45 anos.

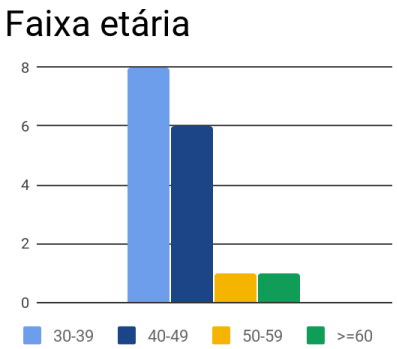

Fig. 2 – Faixa etária dos entrevistados

Além disso, metade dos caminhoneiros apresenta o nível médio de escolaridade completo (2º grau) e 37,5% não chegou ao nível médio, parando no nível fundamental (incompleto ou completo).

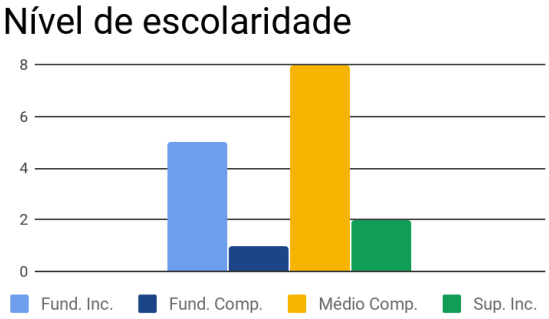

Fig. 3 – Nível de escolaridade dos entrevistados

A maioria dos caminhoneiros apresenta mais de dez anos de experiência nesse trabalho (56%), o que pode indicar que é uma profissão muitas vezes herdada entre gerações e que não costuma ser trocada (possivelmente pela quantia inicial investida no caminhão).

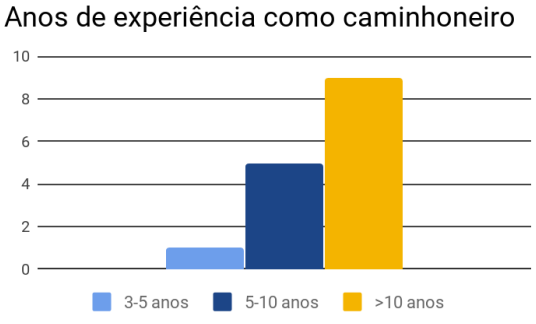

Fig. 4 – Quantidade de anos de experiência dos entrevistados

A maior parte dos motoristas entrevistados (68%) enfrenta situações de estresse diariamente em seu trabalho. Os motivos relatados para essas situações incluem os atrasos ao carregar e descarregar cargas, a falta de cargas para retornar de uma viagem realizada, os altos custos de sobrevivência, o preço muitas vezes baixo pago pelos fretes aos motoristas, os problemas nas estradas, a inflação dos combustíveis e situações de desrespeito por parte dos clientes e embarcadores de cargas, resultando muitas vezes em trocas de ofensas e mais atrasos.

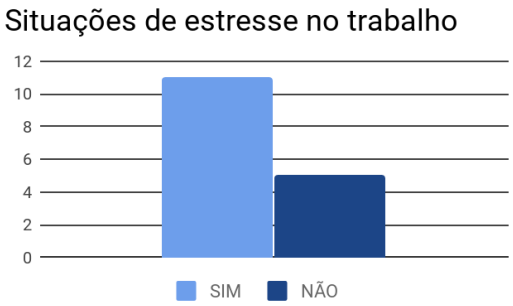

Fig. 5 – Quantidade de entrevistados que sofrem com estresse no trabalho

Da mesma maneira, a maioria dos caminhoneiros entrevistados (68%) passa por situações de cansaço ou sonolência ao longo de suas viagens. Os motivos para isso são as jornadas noturnas recorrentes, os atrasos no carregamento e descarregamento de cargas, a quantidade pequena de horas de sono diárias e a necessidade de esticar a jornada de trabalho para conseguir dinheiro suficiente para sobrevivência.

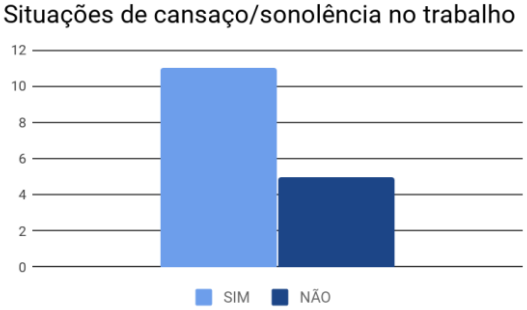

Fig. 6 – Quantidade de entrevistados que sofrem com cansaço no trabalho

Os motoristas também relataram diversas outras situações adversas que atrapalham o transporte de cargas, além do sono e do estresse. Eles costumam sofrer

com problemas de saúde, muitas vezes em articulações das pernas e braços. Há também problemas de falta de segurança, com roubos de cargas frequentes. Outros problemas incluem o preço baixo pago aos caminhoneiros pelos fretes, as situações de maus tratos por parte de clientes, embarcadores e seguradoras, a fome durante o trabalho e a inconstância de horários.

A maior parte dos caminhoneiros entrevistados (60%) relata não conhecer ninguém que tenha sofrido acidentes de trânsito por conta de sono ou estresse. Porém, vale dizer que os motoristas que relatam conhecer dizem que esses acidentes são frequentes, e segundo muitos dos motoristas relataram, geralmente fatais.

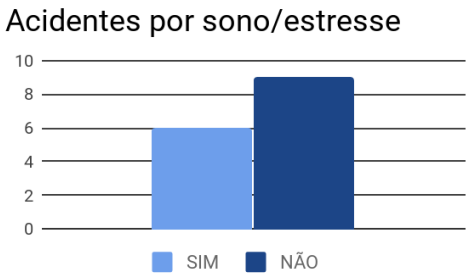

Fig. 7 – Entrevistados que conhecem quem sofreu acidente por sono/estresse

Ao perguntar qual dos fatores, entre sono e estresse, atrapalhava mais os motoristas durante sua jornada de trabalho, o resultado se mostrou dividido: oito caminhoneiros são mais afetados pelo estresse e sete deles são mais afetados pelo sono. Isso significa que o projeto CargoAffect deve, de fato, procurar lidar com ambas as situações para auxiliar uma quantidade maior de motoristas.

A maioria dos motoristas (56%) relatou que dorme menos de seis horas de sono por dia, o que é um dado alarmante e ilustra a rotina cansativa e arriscada da profissão. Somente três dos caminhoneiros entrevistados conseguem dormir mais de oito horas diárias.

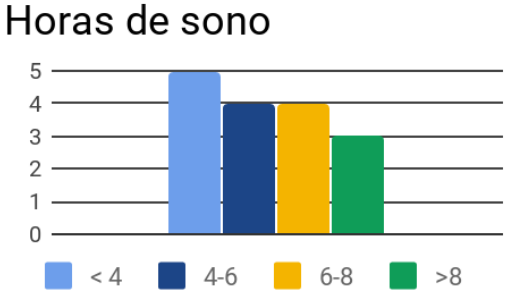

Fig. 8 – Quantidade de horas de sono de cada entrevistado

Todos os caminhoneiros entrevistados possuem smartphones com câmera e utilizam aplicativos de trocas de mensagens, como o WhatsApp, no mínimo uma ou duas vezes por dia. Isso indica que os caminhoneiros possuem certa familiaridade com a tecnologia e conseguem lidar ao menos com funções básicas dos smartphones.

Fig. 9 – Frequência de utilização de WhatsApp de cada entrevistado

Porém, o resultado mais significativo foi em relação à última pergunta da entrevista, que buscou verificar se os caminhoneiros se mostravam interessados no produto inicial: um aplicativo de mensagem que detectasse, pelas mensagens de voz enviadas a supervisores de risco e operadores, situações como estresse e cansaço, e os ajudasse a resolver essas situações, negociando o risco envolvido.

Mais de 80% dos motoristas se mostrou disposto a conhecer ou testar a ideia, mas não houve grande receptividade ou entusiasmo. Uma das maiores resistências encontradas é em relação à dominância do WhatsApp como aplicativo principal de troca de mensagens. Muitos não substituiriam o WhatsApp por um novo aplicativo, mesmo que ele incluísse a funcionalidade de reconhecimento emocional.

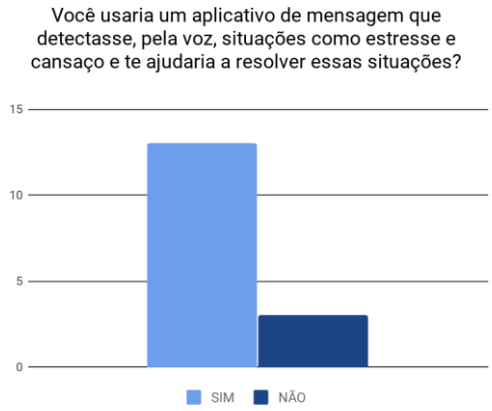

Fig. 10 – Disposição dos entrevistados a usar o sistema proposto

Além disso, muitos se mostraram temerosos em relação a questões de privacidade dos dados obtidos, por não ficar claro à primeira vista o que seria feito das mensagens de áudio enviadas.

4 TECNOLOGIAS UTILIZADAS

Este capítulo detalha as decisões de projeto tomadas para desenvolver o sistema, o protótipo inicial desenvolvido e a arquitetura final do CargoAffect.

4.1 DECISÕES DE PROJETO

Para conseguir desenvolver o sistema baseado em computação afetiva, um dos requisitos mais importantes é a capacidade de realizar a detecção do estado emocional do usuário no momento em que ele utiliza o sistema, para que ele possa (ou não) responder adequadamente de acordo com o estímulo recebido.

Preliminarmente, foram identificadas três possibilidades de detecção de emoções dos usuários para utilização no sistema a ser desenvolvido: processamento de vídeo em tempo real, por meio do reconhecimento das emoções transmitidas por expressões faciais; processamento de mensagens de áudio enviadas pelos caminhoneiros aos supervisores de transportadoras para comunicar a situação do transporte de cargas; processamento de mensagens de texto (linguagem natural) enviadas pelos caminhoneiros aos supervisores, com o mesmo intuito das mensagens de áudio.

Para o desenvolvimento da aplicação, optou-se por realizar o processamento de arquivos de áudio, que será abordado nessa seção, dadas as dificuldades relacionadas ao processamento de texto e de vídeo, tanto em termos da realização de processamento como, no caso do vídeo, por dificuldades logísticas, já que seria custoso manter uma câmera operando o tempo todo, direcionada à face do motorista, em termos de processamento de imagens e do equipamento necessário para realizar essa tarefa). A necessidade de utilizar o celular dos motoristas para captura de vídeo aumenta a complexidade de detecção e desenvolvimento do sistema e surgem empecilhos como o alto consumo de bateria e impossibilidade de realizar a captura em segundo plano.

4.2 PROTÓTIPO INICIAL DESCARTADO

Com base nos resultados iniciais obtidos com o estudo de usuários, havia sido realizado um protótipo inicial, de baixa fidelidade, da aplicação que seria desenvolvida:

um aplicativo de troca de mensagens entre caminhoneiros e supervisores. As imagens do protótipo encontram-se a seguir:

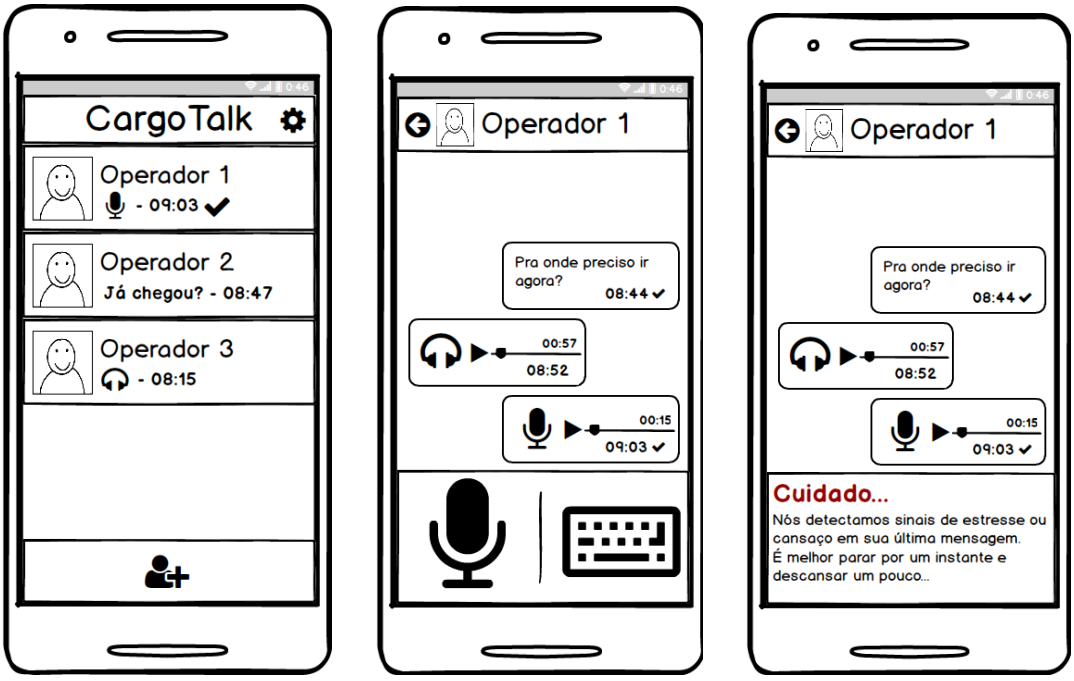

Fig. 2 – Telas do protótipo inicial

Entretanto, como consta na seção anterior, a pesquisa de campo com caminhoneiros e operadores apontou que muitos entrevistados são resistentes à ideia de substituir o aplicativo de troca de mensagens WhatsApp por outro aplicativo de chat, por questões de costume.

Além disso, verificou-se que o projeto de um novo aplicativo de chat para comunicação que também realizasse a captura e processamento de mensagens de áudio representava um escopo muito grande, o que potencialmente dificultaria sua implementação posterior.

Com essas questões em mente, optou-se por descartar o protótipo acima e simplificar o projeto da aplicação de computação afetiva a um aplicativo que realiza somente a captura e processamento de amostras da voz dos caminhoneiros em segundo plano, com feedback contínuo do estado emocional detectado.

4.3 ARQUITETURA DO PROJETO

O sistema CargoAffect, concebido para cumprir com os objetivos propostos, com base no estudo de usuários realizado, apresenta dois subsistemas: um aplicativo

para o sistema operacional Android que realiza gravação de áudios e envio das gravações para processamento e um serviço Web (uma API) que realiza o processamento dos áudios enviados pelo aplicativo com base em um algoritmo supervisionado previamente treinado.

A API, neste caso, tem o objetivo de encapsular a funcionalidade de processamento e disponibilizá-la para o aplicativo utilizar, possibilitando o reuso da lógica de detecção de emoções.

A imagem abaixo ilustra os componentes do sistema e o fluxo de dados que ocorre entre eles.

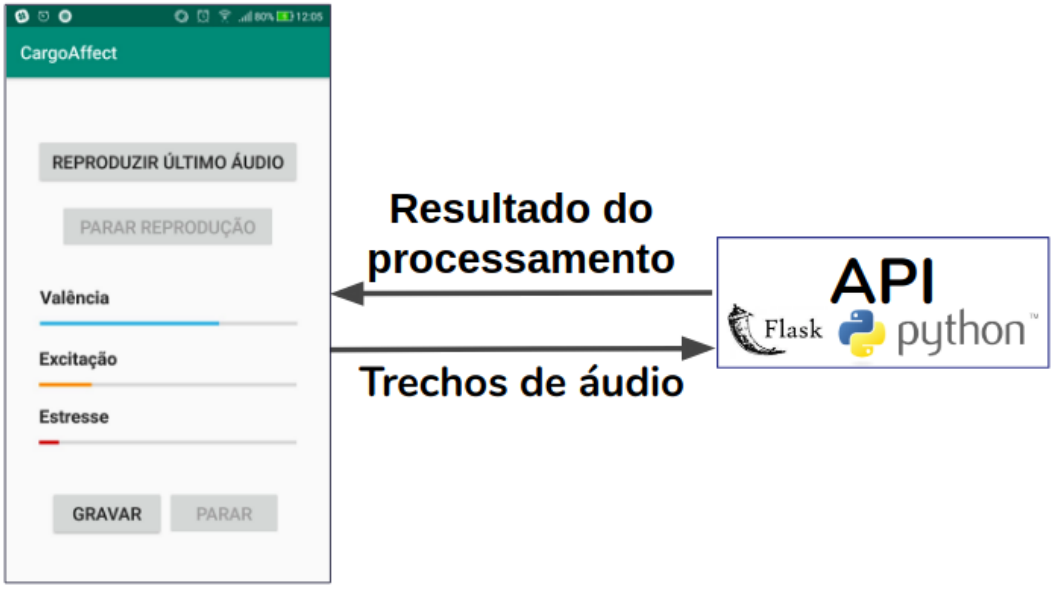

Fig. 3 – Esquema ilustrativo do funcionamento do sistema

4.3.1 O APLICATIVO

O aplicativo foi desenvolvido utilizando a plataforma Android Studio, na linguagem Java.

O aplicativo apresenta, basicamente, duas funcionalidades: a gravação de áudios e a reprodução do último áudio gravado. Há outros dois botões, que servem para interromper a gravação do áudio após seu início e a de interromper a reprodução do último áudio gravado. Além disso, o aplicativo possui três barras de progresso, que correspondem às escalas emocionais utilizadas neste projeto: valência, excitação e estresse.

Após a gravação de um áudio, o aplicativo o envia imediatamente ao serviço Web para que seja realizado seu processamento e reconhecimento de emoções. Isso ocorre por meio de uma requisição POST (HTTP) contendo o arquivo de áudio, que parte do aplicativo para o serviço Web.

4.3.2 O SERVIÇO DE PROCESSAMENTO DE ÁUDIO

O serviço Web foi desenvolvido na linguagem Python, utilizando o *framework* Flask para disponibilizar a funcionalidade de processamento de áudios via Web.

Uma vez recebido o arquivo, o serviço se encarrega das atividades iniciais de separar o áudio do restante dos dados que vem pela requisição POST, salvar o áudio temporariamente no servidor e convertê-lo para o formato .wav, já que o aplicativo cria áudios no formato .3gp e o algoritmo classificador requer o formato .wav para realizar as tarefas de extração de parâmetros e reconhecimento de emoções.

Após essas atividades iniciais, o arquivo de áudio convertido passa pelos procedimentos de extração de parâmetros e classificação nas escalas emocionais adotadas, por meio de módulos da biblioteca *pyAudioAnalysis*, de Theodoros Giannakopoulos. As etapas de extração de parâmetros, classificação nas escalas emocionais, treinamento e teste do algoritmo supervisionado serão descritas com maior nível de detalhe no capítulo 6, que trata do procedimento completo de reconhecimento de emoções desde a obtenção de um corpus de áudio com trechos de voz até o teste do algoritmo supervisionado classificador.

Após o processamento do áudio, a função do serviço Web retorna os resultados da classificação nas três escalas emocionais no formato JSON. Esses resultados são recebidos novamente pelo aplicativo como resposta pela requisição POST realizada anteriormente.

Então, por fim, o aplicativo processa o arquivo JSON recebido e disponibiliza os valores calculados para cada escala de emoção nas barras de progresso correspondentes.

5 ESPECIFICAÇÃO DE REQUISITOS

5.1 REQUISITOS FUNCIONAIS

R1 – O aplicativo deve permitir ao usuário a gravação de áudios, de qualquer duração.

R2 – O aplicativo deve fornecer a funcionalidade de interrupção da gravação a qualquer momento após ter sido iniciada.

R3 – O aplicativo deve permitir ao usuário a reprodução do último áudio gravado.

R4 – O aplicativo deve fornecer a funcionalidade de interrupção da reprodução do último áudio gravado a qualquer momento após ter sido iniciada.

R5 – O aplicativo deve mostrar, por meio de barras de progresso, o resultado do processamento do áudio em função das escalas emocionais de valência, excitação e estresse.

R6 – O aplicativo deve enviar ao serviço Web o arquivo de áudio imediatamente após sua gravação.

R7 – O serviço Web deve realizar a conversão do arquivo de áudio recebido para o formato .wav.

R8 – O serviço Web deve executar o processamento do áudio, incluindo a extração de parâmetros e a classificação nas escalas emocionais pelo algoritmo supervisionado.

R9 – O serviço Web deve enviar o resultado do processamento do arquivo de áudio, que consiste na classificação com base nas três escalas emocionais, ao aplicativo.

R10 – O aplicativo deve exibir o resultado do processamento nas barras de progresso correspondentes a cada escala emocional, imediatamente após recebe-lo do serviço Web.

5.2 REQUISITOS NÃO-FUNCIONAIS

R11 – O sistema requer conexão a uma rede (Wi-Fi ou móvel) para realizar o envio do arquivo de áudio ao serviço Web e para que o resultado de seu processamento seja enviado de volta ao aplicativo.

R12 – O aplicativo deve ser compatível com o sistema operacional Android.

R13 – O envio de um arquivo de áudio ao serviço Web deve ocorrer imediatamente após sua gravação.

R14 – O resultado do processamento do arquivo de áudio deve ser enviado ao aplicativo logo que esteja disponível.

R15 – O aplicativo deve exibir o resultado do processamento nas barras de progresso imediatamente após recebê-lo.

6 DESENVOLVIMENTO DO PROCEDIMENTO DE RECONHECIMENTO

Este capítulo se aprofunda nos passos que foram realizados para a implementação do sistema que realiza a detecção e o reconhecimento de emoções por meio de áudios de voz humana. É importante ressaltar que o sistema CargoAffect se baseia na constatação que a voz humana é uma forma reconhecida de modulação sêntica, conforme exposto no capítulo 2. Além disso, como abordado no mesmo capítulo, a aplicação desenvolvida tende a se encaixar na terceira categoria da computação afetiva (sistema reconhece, mas não expressa emoções), dado que é uma ferramenta que realiza o processamento de amostras de áudio para detecção e reconhecimento de emoção, mas não apresenta a capacidade de expressar emoções por si só.

Além disso, optou-se por realizar o processamento de somente amostras de áudio devido às dificuldades relacionadas ao processamento de mensagens de texto majoritariamente em linguagem informal ou coloquial e aos obstáculos relacionados à utilização de vídeo, tanto em termos de seu custoso processamento em tempo real como em relação à dificuldade de viabilizar um cenário real em que uma câmera monitore continuamente a expressão facial de caminhoneiros.

Para tanto, a sequência de ações realizadas encontra-se ilustrada no diagrama abaixo.

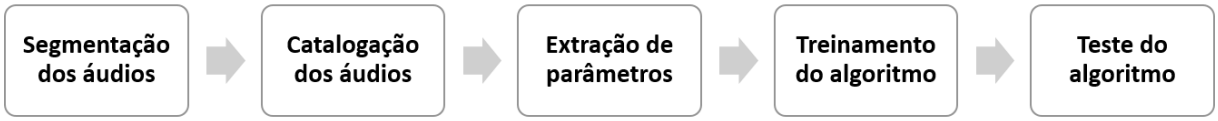

Fig. 3 – Sequência de ações do procedimento de reconhecimento de emoções

6.1 SEGMENTAÇÃO E CLASSIFICAÇÃO (CATALOGAÇÃO) DOS ÁUDIOS

Verificou-se, por meio da revisão bibliográfica, que vários pesquisadores da área de detecção de emoções em áudios de voz humana (BURKHARDT et al, 2009; DESHPANDE et al, 2017; LUBIS et al, 2014) iniciam o processo pela segmentação e classificação (ou catalogação) dos áudios.

Nessa fase, os áudios coletados são segmentados em trechos mais curtos (de no máximo cinco segundos, geralmente) e, se necessário, passam por tratamento de redução de ruídos e outros sinais indesejados. Em seguida, os novos trechos de áudio segmentados são expostos a pessoas voluntárias, que realizam a catalogação dos áudios segundo determinados aspectos emocionais.

No caso deste projeto, inicialmente foram coletados 68 áudios, enviados por mais de quinze caminhoneiros a operadores de cargas da empresa CargoX por meio do aplicativo de mensagens instantâneas WhatsApp. Realizou-se a segmentação e tratamento desses áudios para reduzir ruídos utilizando a ferramenta de edição de áudio Audacity. Todos os áudios foram segmentados de modo a terem duração de cinco segundos, sendo que alguns deles foram preservados de acordo com a gravação original, sem redução de ruído.

Ao final do processo definido acima, foram obtidos cem fragmentos de áudios definitivos, que foram utilizados nas fases seguintes do processo de detecção de emoções.

A fase de catalogação dos áudios foi realizada seguindo o instrumento de avaliação SAM (*Self-Assessment Manikin*), definido por Margaret M. Bradley e Peter J. Lang. Em um relatório publicado em 2007, estes autores abordam a criação de um corpus de áudio denominado IADS (*International Affective Digitized Sounds*), que tinha como objetivo o desenvolvimento de um grande conjunto de estímulos sonoros padronizado e acessível internacionalmente, e a utilização do SAM para realizar a classificação dos áudios de acordo com três parâmetros: valência (prazer ou desprazer), excitação (calmo ou agitado) e dominância (submissão ou dominância).

Nesse modelo de experimento, são utilizadas imagens ilustrativas de valores dos parâmetros mencionados anteriormente, em escalas de 1 a 9, para que os voluntários do experimento classifiquem cada fragmento de áudio de acordo com os sentimentos evocados a partir da escuta de cada trecho, em termos emocionais. As imagens padronizadas definidas neste método encontram-se a seguir:

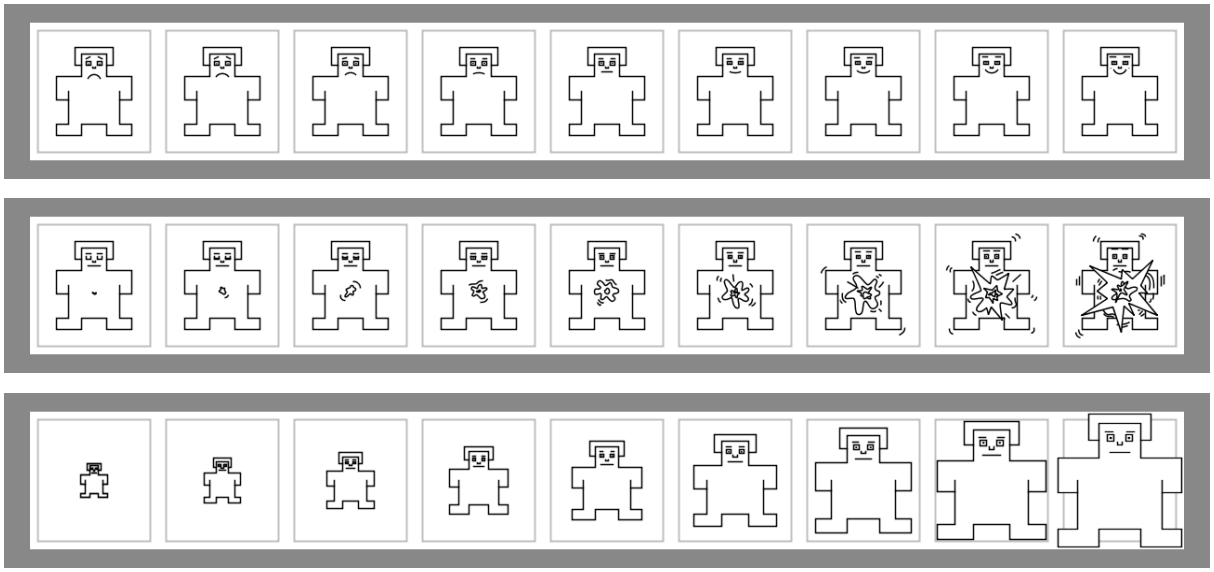

Fig. 4 - Escalas de valência, excitação e dominância

No relatório de Bradley e Lang, descreve-se também como idealmente deve ocorrer o experimento de catalogação dos áudios. Cada conjunto de sons a ser rotulado por cada pessoa deve conter por volta de 60 áudios, que variam nas dimensões de valência, excitação e dominância. Os participantes devem contemplar diferentes faixas etárias e gêneros. Antes de iniciar a classificação dos áudios, realiza-se um procedimento de teste, incluindo três áudios de treino, para que os participantes realizem a rotulação e se familiarizem com o processo.

Cada rodada de catalogação de áudio se inicia com um som de preparação com um intervalo de cinco segundos, seguido pelo áudio propriamente dito (com seis segundos de duração), terminando-se com um intervalo de quinze segundos para a rotulação do áudio nas dimensões de valência, excitação e dominância. Essa rodada se repete para todos os fragmentos de áudio a serem classificados.

Na catalogação dos áudios de caminhoneiros coletados neste projeto, realizou-se uma abordagem muito semelhante, com algumas adaptações. O experimento de rotulação foi realizado com 15 pessoas (8 mulheres e 7 homens), sendo que 7 delas realizam contato diário com motoristas, o que foi considerado como um aspecto primordial para a escolha dos voluntários, já que são pessoas que estão habituadas a ouvir motoristas e potencialmente reconhecem melhor seus estados emocionais pela voz.

Cada um dos 15 voluntários realizou a catalogação de 40 áudios em escalas de 1 a 9, com base em três dimensões: valência, excitação e estresse. O parâmetro de dominância foi descartado por não ser uma dimensão primária de variância em avaliações emocionais (BRADLEY e LANG, 2007) e o parâmetro do estresse foi adicionado por ser um aspecto muito importante na detecção de estados emocionais que representam riscos aos motoristas no caso deste projeto.

É importante destacar que a escala de estresse é baseada na percepção de cada pessoa em relação a esse estado emocional, sem embasamento em alguma teoria ou definição formal. Para manter o padrão de uso de imagens ilustrativas para as escalas emocionais, criou-se uma escala de cores para representar a intensidade do estresse, apresentada na imagem abaixo:

Fig. 5 - Escala de estresse

No início do experimento, os participantes assinaram um termo de consentimento livre e esclarecido (TCLE), cujo modelo se encontra nos anexos desta monografia. Em seguida, cada voluntário recebeu um link de acesso a um formulário criado na ferramenta Google Forms, que continha uma descrição do procedimento e uma seção destinada à rotulação de cada áudio. A descrição foi também narrada, para que o participante pudesse verificar se o volume do som nos fones de ouvido estava adequado.

O experimento realizado apresentou duas fases: na primeira, o formulário contendo o instrumento de avaliação dos áudios baseado no SAM foi apresentado ao participante e realizou-se um treino, para que os participantes realizassem a catalogação, preenchendo este formulário, de cinco áudios de teste, retirados do IADS. Na segunda fase, foram apresentados aos voluntários os 40 trechos de áudio contendo vozes de motoristas de caminhões, para que classificassem esses áudios de acordo com as emoções que sentiam ao escutá-los. Foram criadas cinco possibilidades de experimento com os 100 áudios coletados, de modo que cada áudio

fosse avaliado por exatamente seis pessoas. Cada experimento durou cerca de 25 minutos.

Na descrição anterior à catalogação dos áudios, buscou-se enfatizar que as classificações deveriam ser realizadas com base no que cada participante sentia ao escutar cada fragmento, evitando realizar um julgamento mais aprofundado do suposto estado emocional do caminhoneiro no momento em que o áudio foi gravado. Além disso, deu-se ênfase ao fato de que, por ser uma análise subjetiva, o experimento não apresenta respostas certas ou erradas. A descrição completa do experimento utilizada encontra-se nos anexos desta monografia.

Na fase de rotulação dos áudios, seguiu-se o modelo de rodada de catalogação descrito por Bradley e Lang: cinco segundos de preparação para o áudio a ser escutado, com um alerta auditivo indicando o número do próximo áudio a ser classificado. Em seguida, o áudio de cinco segundos era reproduzido e, logo depois, o participante tinha quinze segundos para classificar o áudio nas três escalas (valência, excitação e estresse), após outro alerta auditivo (“Classifique o áudio nas três escalas”). A imagem a seguir mostra um trecho do formulário, para ilustrar sua estrutura:

Fig. 6 - Estrutura do formulário de catalogação de áudios

Os resultados da catalogação dos áudios foram compilados em uma planilha, para reunir todas as seis avaliações de valência, excitação e estresse para cada áudio e para poder obter um número médio de cada dimensão (V para valência, E para excitação e S para estresse) para cada áudio, como ilustrado na figura abaixo:

Áudio	Experimento 1									Experimento 2									Experimento 3									Experimento 4									Experimento 5									Média		
Pessoa	1			6			11			2			7			12			3			8			13			4			9			14			5			10			15					
Escala	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S	V	E	S
1	2	9	9	1	9	9	1	8	9																			5	9	9	2	4	9	3	8	8												2,333	7,833	8,833
2										5	7	4	4	4	3	5	5	8																			9	6	5	6	4	4	7	3	3	6,000	4,833	4,500

Fig. 7 – Excerto da planilha com compilação das avaliações dos áudios

Esses valores médios de valência, excitação e estresse de cada áudio foram utilizados nas fases seguintes do processo de detecção de emoções nos áudios: a extração de parâmetros (*features*) dos áudios e a realização do treinamento do algoritmo supervisionado.

6.2 EXTRAÇÃO DE PARÂMETROS (FEATURES)

A etapa de obtenção de parâmetros dos áudios rotulados é a fase seguinte do processo de detecção de emoções em áudio. É uma etapa crucial neste processo, dado que os parâmetros extraídos de cada arquivo de áudio, em conjunto com os valores de classificação obtidos na etapa anterior, serão os dados de entrada para treinamento do algoritmo não supervisionado. Assim, é importante garantir que todos os parâmetros sejam relevantes, evitando a utilização de valores atípicos.

Os parâmetros a serem extraídos podem ser divididos em três categorias: parâmetros no domínio do tempo, extraídos diretamente de amostras do sinal bruto; no domínio da frequência, cuja extração é baseada na magnitude da Transformada Discreta de Fourier (DFT) do áudio; e parâmetros no domínio cepstral, que surgem da aplicação da Transformada Discreta de Fourier inversa (IFT) no espectro logarítmico.

Nessa última categoria, obtém-se o inverso da frequência ao realizar a transformada inversa, o que se convenciona chamar de “quefrência”. Entretanto, a “quefrência” volta a ser uma medida de tempo (já que se aplicou uma transformada inversa em cima do resultado de uma transformada direta), embora não no sentido de um sinal no domínio do tempo, como no primeiro caso. As imagens a seguir, obtidas por meio da ferramenta Audacity, ilustram essas três categorias no caso de um dos áudios de caminhoneiros utilizados no projeto.

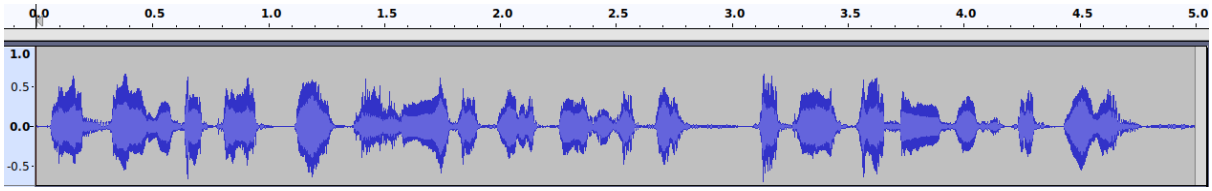

Fig. 8 - Forma de onda no domínio do tempo, com amplitude no eixo vertical

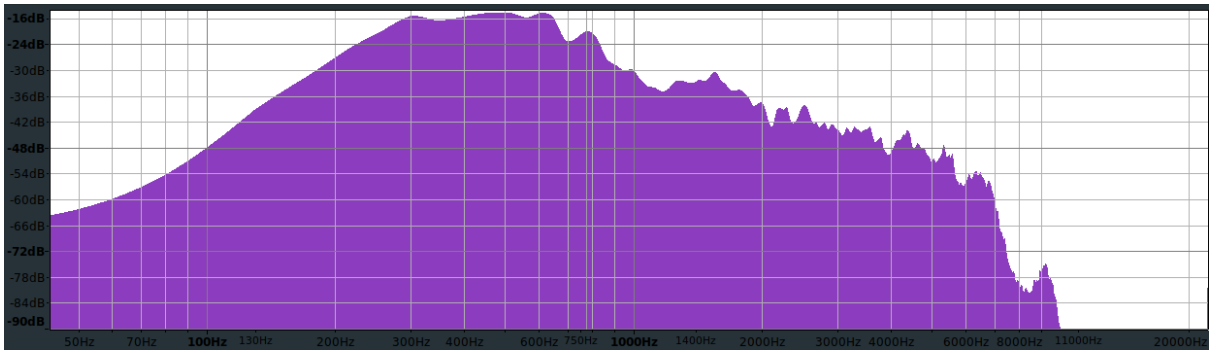

Fig. 9 - Espectro de potência no domínio da frequência, amplitude no eixo vertical

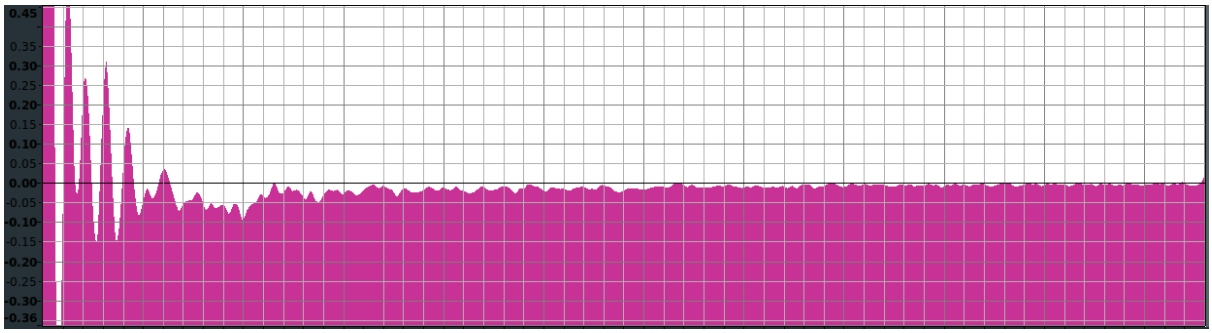

Fig. 10 - Cepstro no domínio da “quefrência” (unidade de tempo)

Além disso, a extração dos parâmetros pode ocorrer por meio de uma análise de curto prazo ou de médio prazo. Na análise de curto prazo, o sinal de áudio é dividido em janelas (denominadas como quadros) curtas de tempo e todos os parâmetros são calculados para cada janela de tempo, resultando em uma sequência de vetores de parâmetros. Geralmente, as janelas de tempo da análise de curto prazo apresentam de 20 a 100 milissegundos. O processo pode ser feito de maneira sobreposta (os quadros se sobrepõem no tempo) ou não sobreposta (um quadro imediatamente após o outro).

Já na análise de médio prazo, o sinal de áudio é dividido em janelas (denominadas como segmentos) médias de tempo, que podem se sobrepor ou não. Para cada segmento, ocorre o processo de análise de curto prazo mencionado anteriormente (o segmento é dividido em quadros menores e os parâmetros são calculados com base nos quadros) e, então, a sequência de parâmetros de cada segmento de médio prazo é usada em análises estatísticas (para obter, por exemplo, o valor médio de um determinado parâmetro no segmento). Ou seja, cada segmento de médio termo é representado por um conjunto de valores estatísticos, calculados a

partir da análise de curto prazo em quadros do segmento. Geralmente, cada segmento apresenta de 1 a 10 segundos.

Para a etapa de extração de parâmetros, optou-se pela utilização de uma biblioteca de análise de áudios na linguagem Python criada por Theodoros Giannakopoulos, chamada *pyAudioAnalysis*. Por meio da utilização desta biblioteca, é possível obter 34 parâmetros de cada arquivo de áudio, cujas definições encontram-se na tabela abaixo (GIANNAKOPOULOS, 2015 e ESQUERRA; DIMATTIA, 2008):

Índice	**Nome**	**Descrição**
1	Taxa de cruzamento por zero (ZCR)	Taxa de mudanças de sinal (de positivo para negativo ou vice-versa) no quadro.
2	Energia	Soma dos quadrados dos valores do sinal, normalizada pelo tamanho do quadro.
3	Entropia de energia	Entropia das energias normalizadas dos subquadros. Funciona como um medidor de mudanças abruptas.
4	Centroide espectral	Centro de gravidade do espectro logarítmico. Apresenta forte relação com o timbre do som.
5	Espalhamento espectral	Segundo momento central do espectro. Ajuda a diferenciar ruídos de sons tonais.
6	Entropia espectral	Entropia das energias espectrais normalizadas para um conjunto de subquadros.
7	Fluxo espectral	Diferença quadrática entre as magnitudes normalizadas dos espectros de dois quadros sucessivos.
8	*Roll-off* espectral	Frequência abaixo da qual 90% da distribuição de magnitude do espectro está concentrada.
9-21	MFCCs (*Mel-Frequency Cepstral Coefficients*)	Formam uma representação cepstral na qual as bandas de frequência não são lineares, mas sim distribuídas de acordo com a escala Mel (uma escala baseada no mapeamento entre as frequências reais e as percebidas pelo sistema auditivo humano).

22-33	Vetor cromático	Representação de 12 elementos da energia espectral, que representam os 12 semitons igualmente espaçados da música ocidental.
34	Desvio cromático	O desvio padrão dos 12 coeficientes cromáticos

Os parâmetros de 1 a 3 se encaixam na categoria do domínio do tempo. Os parâmetros de 4 a 8 e de 22 a 34 se encontram no domínio da frequência. Já os parâmetros 9 a 21, que consistem nos MFCCs, são obtidos a partir do domínio cepstral. Estes últimos merecem destaque especial, dado que são amplamente utilizados por pesquisadores em detecção de emoção em áudios.

No trabalho desenvolvido por DESHPANDE et al. (2017), por exemplo, são utilizados como parâmetros de curto prazo, além de 13 MFCCs, o valor eficaz (valor quadrático médio) da energia, a entropia (de energia e espectral) e o tom dos áudios analisados, com o objetivo de realizar avaliações da precisão de classificações de emoção para trechos de voz em ambientes de trabalho (não atuados, estimulados ou simulados). O trabalho de BURKHARDT et al. (2009) também utiliza os MFCCs como parâmetros, extraindo a média de 12 coeficientes cepstrais para introduzir no algoritmo de classificação.

Para demonstrar a fase de extração de parâmetros nesta monografia, mostra-se essa extração com base em um dos áudios de caminhoneiros coletados. O primeiro passo é realizar a conversão do arquivo de áudio para o formato WAVE (*.wav), que é um formato padrão para armazenamento de áudio em PCs e armazena *bitstreams* (fluxos de *bits*) de áudio em blocos (*chunks*). Para realizar essa conversão, pode-se utilizar o programa ffmpeg, que faz a transcodificação necessária para gerar um arquivo de extensão WAV.

Após a conversão do arquivo, utiliza-se uma função (*featureExtractionFile*) de um dos módulos da biblioteca *pyAudioAnalysis* (*audioAnalysis.py*) para extrair os parâmetros, que gera como saída dois arquivos .csv (*comma-separated values*, ou seja, valores separados por vírgulas) com os valores dos parâmetros: um para os valores de médio prazo e outro para os valores de curto prazo.

Em ambos os casos, cada sequência de parâmetros é armazenada em uma coluna e as linhas correspondem às janelas de tempo. Assim, um trecho de áudio de cinco segundos utilizado como entrada nesse processo gera um arquivo .csv de 34

colunas (correspondentes aos 34 parâmetros extraídos, listados na tabela acima) e 100 linhas (dado que a janela e o passo sejam de 0,05 segundos) para o caso de parâmetros de curto prazo e outro arquivo .csv de 68 colunas (correspondentes ao valor médio e ao desvio padrão de cada parâmetro de curto prazo) e 5 linhas (dado que a janela e o passo sejam de 1 segundo) para o caso de parâmetros de longo prazo. As imagens abaixo ilustram os arquivos gerados:

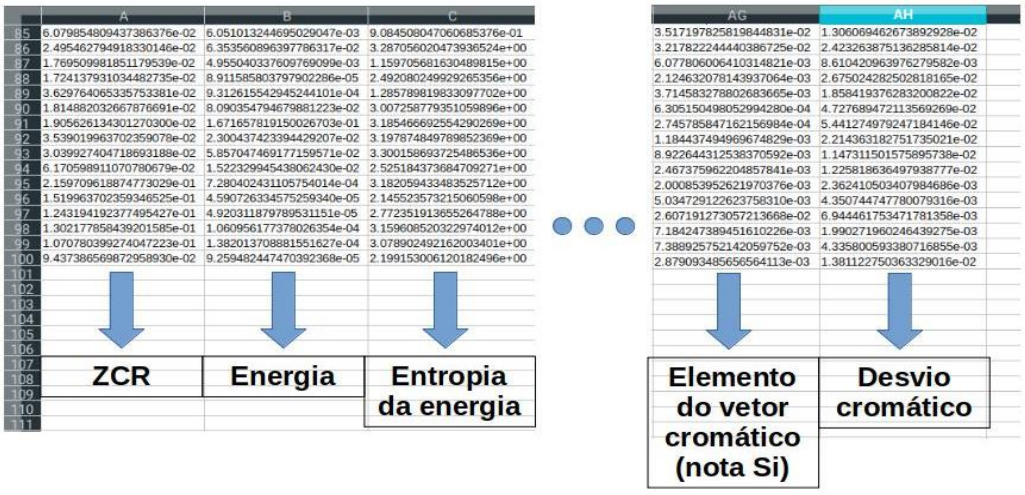

Fig. 11 - Arquivo .csv com os parâmetros de curto prazo

Fig. 12 - Arquivo .csv com os parâmetros de longo prazo

No caso do processo de reconhecimento de emoções, não são gerados arquivos .csv com os parâmetros extraídos de cada áudio. Os parâmetros serão utilizados como entrada para os algoritmos supervisionados reconhecedores, como treinamento ou como teste, como será descrito na seção seguinte.

6.3 TREINAMENTO E TESTE DO ALGORITMO SUPERVISIONADO

Após a extração dos parâmetros de um trecho de áudio, os vetores contendo estes parâmetros são utilizados como entrada para algoritmos supervisionados classificadores, com o intuito de realizar, enfim, a classificação do áudio em termos de estado emocional.

Optou-se pela utilização de um algoritmo supervisionado para processamento das amostras de áudio devido aos fatores mencionados na seção 2.2.2 (Modelos de Estados Emocionais).

Na literatura, são mencionados dezenas de possibilidades de algoritmos e abordagens, com suas vantagens e desvantagens para cada caso. GARG e SEHGAL (2015) mencionam alguns dos algoritmos mais utilizados com o objetivo de compará-los. O algoritmo k-NN (*k-Nearest Neighbor*, ou k-ésimo vizinho mais próximo) compara a instância (o vetor, neste caso) inserida com as k instâncias de treino que sejam mais similares ou próximas dela. Geralmente, utiliza-se a distância euclidiana para medir essa similaridade.

As redes neurais artificiais (ANN) também são comumente usadas para classificação emocional de áudios de voz. Elas apresentam bons resultados para aplicações que envolvem reconhecimento de padrões e classificação de dados por meio de um processo de aprendizado.

Como mencionado no capítulo de computação afetiva, também são utilizados com frequência os modelos ocultos de Markov (HMM), que são variantes próximas de máquinas de estado finito, embora não sejam determinísticos, mas sim estocásticos (transições determinadas por modelos probabilísticos).

Entretanto, o algoritmo escolhido para realizar a classificação dos trechos de áudio foi o SVM (*Support Vector Machine*, ou máquina de vetores de suporte). Optou-se pela utilização deste classificador por ser um dos mais utilizados e frequentemente citados na literatura (GARG e SEHGAL, 2015; GUVEN e BOCK, 2010; LUBIS et al., 2014), com altos índices de precisão na classificação de emoções.

As máquinas de vetores de suporte são modelos com algoritmos de aprendizado supervisionado que analisam dados e reconhecem padrões, sendo utilizados para análises de classificação e regressão. Em seu caso mais simples, esse tipo de máquina recebe como entrada um conjunto de dados e prevê, para cada

entrada, a qual de duas possíveis classes corresponde a saída, sendo assim um classificador não-probabilístico binário e linear.

O processo de classificação de uma SVM ocorre da seguinte forma: a máquina constrói, em um espaço com muitas ou até mesmo infinitas dimensões, um hiperplano (subespaço com uma dimensão a menos que o espaço de variáveis sob análise) ou um conjunto de hiperplanos entre dados de classes diferentes, que são utilizados nas tarefas de classificação, regressão ou busca de pontos anômalos (outliers). O algoritmo busca criar hiperplanos que maximizem a distância entre os dados mais próximos em relação a cada uma das classes, para evitar erros de generalização.

Para realizar o treinamento e, posteriormente, o teste das instâncias de áudio com a voz de caminhoneiros neste trabalho, optou-se por utilizar novamente a biblioteca *pyAudioAnalysis* (GIANNAKOPOULOS, 2015). Os módulos dessa biblioteca implementam o algoritmo kNN e trazem implementações de outras bibliotecas dos algoritmos SVM (com função *kernel* linear ou RBF - *Radial Basis Function*, ou função com base radial), árvores de decisão aleatórias e outros algoritmos.

Dentre os algoritmos que a biblioteca oferece, optou-se por utilizar o SVM linear, pelos motivos mencionados acima em relação às vantagens de se utilizar máquinas de vetores de suporte na tarefa de classificação e regressão de dados. A regressão dos dados é particularmente importante no caso de reconhecimento de emoções em áudio, já que o objetivo é determinar o estado emocional não simplesmente por meio de uma classe discreta (contém ou não contém estresse, por exemplo), mas por uma medição real estimada dos valores de valência, excitação e estresse dos áudios de teste a partir dos valores dos áudios usados no treinamento do algoritmo.

Utilizou-se um módulo da biblioteca em Python que lê os arquivos de áudio contidos em um diretório com seus respectivos valores de valência, excitação e estresse contidos em três arquivos .csv, realiza a extração de parâmetros de todos os áudios e retorna um modelo de regressão, com base nos dados extraídos dos áudios e nas escalas emocionais contidas nos arquivos .csv. A figura abaixo ilustra a organização dos arquivos no diretório e os arquivos com valores dos parâmetros emocionais de cada áudio.

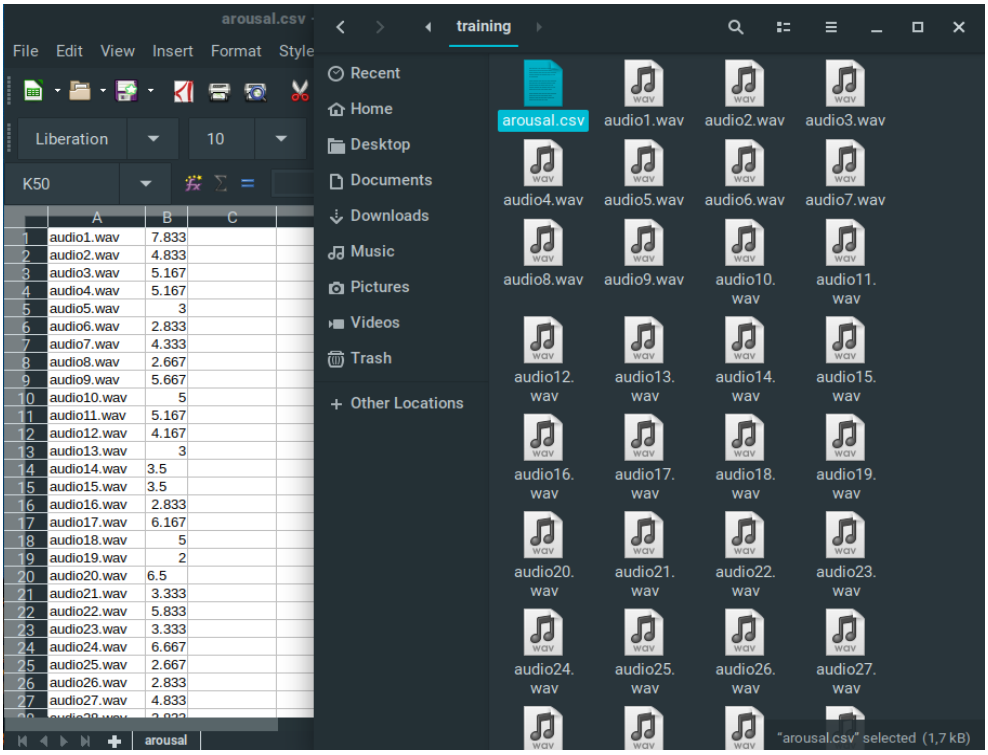

Fig. 13 - Organização dos arquivos, planilha com os valores de excitação dos áudios

Para realizar o treinamento do SVM, optou-se pela subdivisão de 85:15 entre conjunto de treinamento e conjunto de teste, assim como descrito no trabalho de LUBIS et al. (2014). Ou seja, dos 100 áudios de caminhoneiros coletados, 85 foram usados no treinamento do algoritmo e 15 foram utilizados para realizar o teste do algoritmo. Os áudios para treinamento e para teste foram selecionados de modo a contemplar o máximo possível de valores de valência, excitação e estresse. Ou seja, os áudios de teste possuem valores que vão dos mais baixos aos mais altos nas três escalas.

O treinamento foi feito a partir da execução de um script, com o seguinte comando: **python3 audioAnalysis.py trainRegression -i training/ --method svm -o training/**. Esse comando executa a função *featureAndTrainRegression()* do módulo *audioAnalysis.py* da biblioteca *pyAudioAnalysis*, utilizando como entrada o diretório training, que contém todos os arquivos de áudio a serem utilizados como treinamento e as planilhas com os valores de cada escala emocional para cada áudio. A saída do processo também tem como destino o mesmo diretório. Durante a execução do comando, são exibidas algumas mensagens de progresso, como ilustrado na figura a seguir.

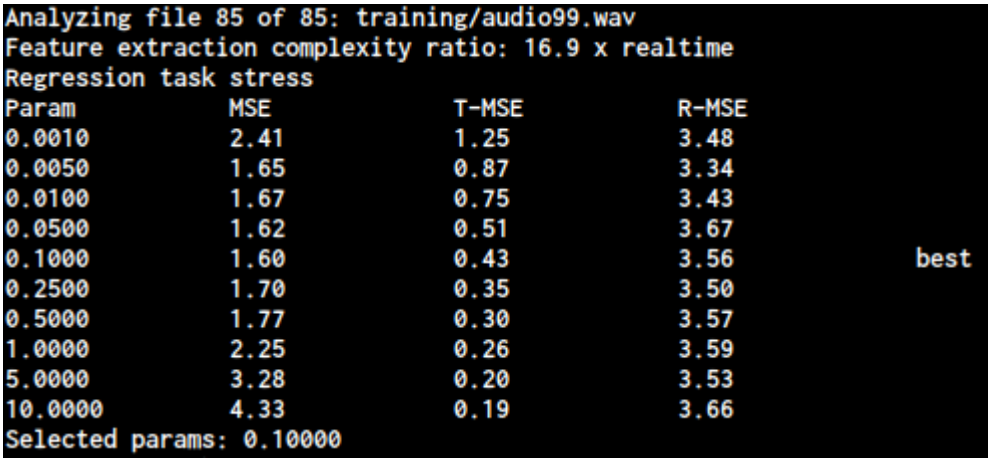

Fig. 14 - Mensagens exibidas durante o treinamento do algoritmo

Como é possível ver pela figura, há um passo intermediário na tarefa de treinamento, que é a seleção dos parâmetros C (*soft margin*) do SVM para cada escala de emoção. Esse parâmetro é um indicativo à máquina de vetores de suporte do quanto se deseja escapar de erros de classificação para cada arquivo utilizado no treinamento. Assim, para valores maiores de C, a otimização escolhe hiperplanos com margens menores, dado que façam a tarefa de classificar os áudios corretamente. Por outro lado, para valores menores do parâmetro, a otimização busca hiperplanos de separação com margens maiores, mesmo que isso gere erros de classificação para alguns pontos.

Ao lado de cada parâmetro, há os valores de erros quadráticos médios. O algoritmo de treinamento se encarrega de escolher o parâmetro para o qual o valor do erro quadrático médio (MSE) é o menor possível, fazendo o mesmo para as três escalas de emoção adotadas neste trabalho.

Após o treinamento do algoritmo SVM, a próxima e última fase do processo de reconhecimento de emoções por áudio é a realização do teste do algoritmo, para verificar se os valores processados pelo algoritmo para cada áudio de teste correspondem aos valores esperados (ou seja, a média dos valores obtidos na fase de classificação dos áudios por pessoas voluntárias).

Para realizar este teste, também utilizando a biblioteca *pyAudioAnalysis* (GIANNAKOPOULOS, 2015), criou-se uma API (interface de programação de aplicação) em Python, utilizando o framework web Flask, para criar uma rota que permite a realização dos métodos de requisição POST e GET do HTTP. Essa rota recebe, por meio do método POST, um arquivo de áudio. O arquivo de áudio recebido

passa inicialmente por um processo de conversão para .wav, utilizando o mesmo programa de conversão de arquivos de áudio mencionado anteriormente (ffmpeg).

Após a conversão do arquivo de áudio (quando necessário, caso o arquivo de entrada não esteja no formato .wav), ele é utilizado como entrada para a função *fileRegression()* do módulo *audioTrainTest* da biblioteca, que vai realizar a classificação deste arquivo nas três escalas de emoção com base no treinamento do algoritmo SVM realizado anteriormente. A API retorna, como resultado, um arquivo no formato JSON, com os respectivos valores de valência, excitação e estresse calculados pelo algoritmo a partir dos parâmetros extraídos do áudio.

A rota criada na API também é utilizada pelo aplicativo criado em Android, dado que os áudios gravados por meio do aplicativo são classificados por meio do envio à essa rota da API, por meio de uma requisição com o método POST contendo o arquivo de áudio gravado.

Os resultados obtidos na fase de testes do algoritmo serão abordados na próxima seção desta monografia, a de Resultados e Discussões.

7 RESULTADOS E DISCUSSÕES

Na fase de segmentação e classificação (rotulação) de áudios do projeto, foi possível perceber correspondências entre as escalas de emoção adotadas a partir das classificações médias obtidas para cada trecho de áudio. Por exemplo, há uma tendência de que áudios com alto nível de estresse possuam também alto nível de excitação e baixo nível de valência, e vice-versa. Essas relações entre as escalas, ilustradas pelos gráficos abaixo, podem ajudar a delimitar a definição de estresse como sendo uma reação normalmente caracterizada por níveis mais altos de excitação e mais baixos de valência.

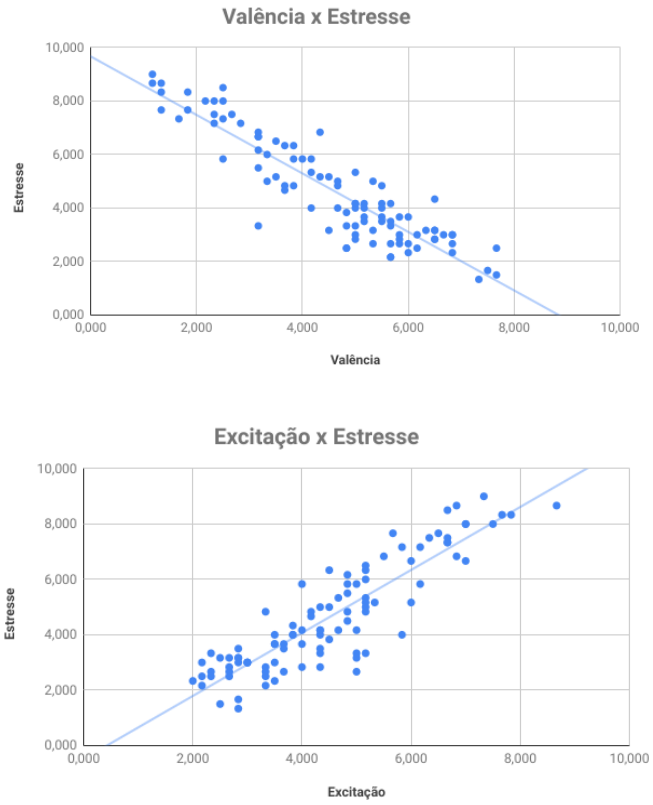

Fig. 15 - Gráficos que demonstram a relação entre as escalas empregadas

Essa constatação em relação às escalas de emoção adotadas no projeto ajudou a estabelecer expectativas mais sólidas em relação ao resultado da utilização do algoritmo de classificação. Espera-se que o algoritmo, ao detectar valores de estresse mais altos em áudios, também retorne valores de excitação mais altos e valores de valência mais baixos.

No teste inicial realizado após o treinamento do algoritmo SVM linear, verificou-se a confirmação dessa tendência para a maior parte dos arquivos de áudio utilizados

como teste. Entretanto, algumas das classificações mostraram erros razoavelmente grandes em relação aos valores previstos na rotulação por voluntários, conforme pode-se verificar na figura a seguir.

Teste do algoritmo SVM Linear									
Áudio testado	Média das Classificações			Valor de retorno do algoritmo			RMSE		
	V	E	S	V	E	S	V	E	S
Áudio 1	2,333	7,833	8,833	2,820	6,588	8,725	-0,487	1,245	0,108
Áudio 2	6,000	4,833	4,500	5,865	2,886	2,198	0,135	1,947	2,302
Áudio 3	4,333	5,167	5,167	5,886	3,536	3,361	-1,553	1,631	1,806
Áudio 5	6,833	3,000	3,000	5,210	3,847	3,981	1,623	-0,847	-0,981
Áudio 19	6,000	2,000	2,333	5,396	3,338	3,746	0,604	-1,338	-1,413
Áudio 29	1,833	7,667	8,333	3,152	5,966	6,093	-1,319	1,701	2,240
Áudio 34	5,000	4,000	4,167	4,615	4,103	4,127	0,385	-0,103	0,040
Áudio 39	7,667	2,500	1,500	5,138	4,167	4,355	2,529	-1,667	-2,855
Áudio 46	2,167	7,500	8,000	2,732	6,531	9,514	-0,565	0,969	-1,514
Áudio 48	3,167	4,833	5,500	5,199	3,281	2,646	-2,032	1,552	2,854
Áudio 67	1,167	8,667	8,667	3,700	5,448	6,989	-2,533	3,219	1,678
Áudio 76	4,500	2,500	3,167	5,822	3,263	2,856	-1,322	-0,763	0,311
Áudio 77	5,333	5,167	5,000	5,136	3,507	3,244	0,197	1,660	1,756
Áudio 87	2,333	6,667	7,500	3,297	6,122	6,783	-0,964	0,545	0,717
Áudio 100	7,500	2,833	1,667	5,721	3,170	2,260	1,779	-0,337	-0,593
							1,430	**1,498**	**1,673**

Fig. 16 - Valores de classificação e de retorno do algoritmo

Para a realização da análise estatística dos valores obtidos, optou-se por utilizar o erro médio quadrático (RMSE). Quanto menor for o valor obtido, maior é a correspondência entre os valores obtidos e os valores esperados. Pela imagem, pode-se verificar que os valores de erro médio quadrático ficaram próximos de 1,5.

Apesar da ocorrência de valores significativos de erro, pode-se dizer que o algoritmo detectou relativamente bem as ocorrências de estresse: nos áudios com valores de estresse previstos acima de 6, o algoritmo retornou valores para estresse também superiores a esse valor, não tendo retornado valores superiores a 6 para os outros áudios com nível de estresse inferior. Se o valor de 6 for considerado como a referência de estresse para o algoritmo, ele foi preciso em 100% dos casos testados, sem a ocorrência de falsos positivos ou falsos negativos. Seria importante, contudo, testar o algoritmo treinado com mais áudios de caminhoneiros, além dos coletados ao longo do projeto, para contestar sua eficácia real na detecção de estresse.

Os valores significativos dos erros médios quadráticos podem ter sido ocasionados por diversos fatores: em LUBIS et al. (2014), os valores dos parâmetros

extraídos de cada áudio foram normalizados para uma escala de [-1, 1]. Esta etapa de normalização dos parâmetros pode, de acordo com os pesquisadores, evitar que parâmetros com valores maiores sejam mais significativos e relevantes que os menores, amenizando eventuais dificuldades numéricas na computação realizada pela SVM. A normalização também poderia ser realizada nos valores das escalas emocionais de valência, excitação e estresse dos áudios de treinamento e teste do algoritmo utilizado.

Além disso, a biblioteca *pyAudioAnalysis* realiza a extração de 34 parâmetros, com seus valores médios e de desvio padrão, como explicado no capítulo anterior. É possível que alguns dos parâmetros não sejam significativos para a detecção do estado emocional neste caso, ou que sejam valores aberrantes em comparação com os demais. Um dos próximos passos mais imediatos na continuação deste projeto é avaliar qual o papel de cada um dos parâmetros extraídos dos áudios e sua relevância no objetivo de detectar emoções.

Outra possível fonte de erro está nos próprios dados considerados como os esperados no procedimento de teste do algoritmo, que são os dados provenientes da fase de classificação (catalogação) dos trechos de áudio por voluntários. Como cada áudio foi classificado por 6 pessoas nas escalas de valência, excitação e estresse, é possível que alguns valores tenham sido atípicos em relação aos demais. Dada essa questão, seria válido realizar uma análise estatística qualitativa para detectar os dados mais discrepantes e possivelmente descartá-los, ou utilizar outra medida estatística que não a média das classificações, o que poderia resultar em valores mais significativos para realizar o treinamento e o teste do algoritmo.

Na continuação deste projeto, também seria importante testar outros algoritmos que a biblioteca *pyAudioAnalysis* oferece, como o SVM com função *kernel* RBF, o kNN e as florestas aleatórias, ou até mesmo buscar outros algoritmos de outras bibliotecas, utilizando alternativas como redes neurais artificiais. O algoritmo SVM, como abordado no capítulo anterior, apresenta diversas vantagens e já foi utilizado em muitos artigos na literatura para realização de detecção de emoções por áudio, mas também apresenta alguns obstáculos em sua utilização e outros algoritmos podem ter melhor desempenho nestes casos.

No capítulo 2, aborda-se a possibilidade de uma modelagem de estados emocionais utilizando o modelo oculto de Markov (HMM). Em trabalhos futuros, a utilização de um algoritmo que implementa esse tipo de modelagem, para comparação

com os resultados obtidos pelo SVM, seria um passo importante para verificar qual algoritmo retorna os resultados mais precisos para a detecção de emoções por meio da voz de motoristas de caminhões.

Em relação ao processo de detecção de emoções em si, além de todas as possibilidades já mencionadas neste capítulo, há algumas mudanças e melhorias mais drásticas que poderiam ser postas em prática para aprimorar a classificação emocional dos áudios testados.

Neste projeto, a detecção e classificação de emoção é realizada somente com base em parâmetros extraídos dos áudios, no domínio do tempo, frequência e cepstral. Entretanto, o discurso dos áudios também poderia ser utilizado para esta tarefa, dado que é um indicativo a mais da possibilidade de estresse: há ocasiões em que o estresse é mais perceptível pelo que é dito pelo emissor, não pelo seu tom de voz. A análise de discurso e reconhecimento de fala (*speech-to-text*), aliada à análise de parâmetros dos sinais dos áudios, já realizada neste projeto, aprimoraria o processo de detecção de emoções por voz humana.

Além disso, para simplificar todo o procedimento, optou-se por restringir a detecção de estado emocional por meio das escalas de valência, excitação e estresse, excluindo o cansaço e a sonolência. Em trabalhos futuros, a inclusão da possibilidade de detectar se um motorista está cansado ou com sono, além do estresse, pode representar um grande avanço, dado que os motoristas entrevistados apontaram o sono e o cansaço como fatores que atrapalham sua rotina diária de trabalho tanto quanto o estresse. Também pode ser válido verificar, na continuação do projeto, se o ruído dos áudios gravados pelos motoristas influencia os resultados obtidos ao final do processamento.

O único retorno que o sistema apresenta ao usuário, na arquitetura implementada, é a classificação do áudio gravado nas três escalas emocionais. Na continuação do projeto, seria importante verificar uma maneira eficaz e não invasiva de alertar o usuário em relação a condições emocionais adversas (como o sono ou o estresse).

Por conta de restrições de tempo, não foi possível realizar um teste de aceitação com os motoristas de caminhões, o principal público-alvo do sistema. Para trabalhos futuros, seria importante realizar esse tipo de teste, com o objetivo de verificar se o sistema cumpre com o objetivo de auxiliar os motoristas a lidar com situações de estresse que o atrapalham em sua rotina diária de trabalho. Também

seria possível verificar se a idade ou o nível de escolaridade influenciam na aceitação do sistema.

Em relação ao sistema desenvolvido, também é possível realizar melhorias em trabalhos futuros. A arquitetura inicial proposta previa a utilização de um gerenciador de tarefas (*WorkManage*r) para lidar com tarefas executadas em segundo plano. Com isso, o motorista não precisaria ter acesso a uma rede móvel ou Wi-Fi para utilizar o sistema: o gerenciador organiza as tarefas que requerem envio ou recebimento de dados pela rede para serem executadas somente quando o dispositivo conseguir conexão.

Além disso, outra melhoria seria o uso do Google Firebase para realizar as tarefas de processamento dos arquivos de áudio, enviando-os para armazenamento na nuvem em um *bucket* do serviço web Amazon S3. O Firebase disponibiliza uma ferramenta (*ML Kit*) que permite o processamento com aprendizado de máquina por meio da inserção de um algoritmo supervisionado treinado. Caso a utilização dessa ferramenta fosse bem-sucedida, provavelmente seria possível realizar o processamento do áudio sem precisar enviá-lo a um serviço web e, portanto, sem necessitar da conexão a uma rede móvel ou Wi-Fi. Como os motoristas costumam trafegar muitas vezes em zonas sem cobertura de redes móveis, essa melhoria permitiria a eles utilizar o sistema sem que o dispositivo estivesse conectado, obtendo *feedbacks* contínuos de seu estado emocional.

Na arquitetura final do projeto, também não há um histórico disponível para que o motorista consiga verificar o progresso dos níveis detectados ao longo do tempo nas escalas emocionais. Isso poderia ser implementado por meio da utilização de um banco de dados como o PostgreSQL, para que os motoristas e também os supervisores e operadores de carga das transportadores pudessem acompanhar alterações no estado emocional dos caminhoneiros em algum intervalo de tempo.

Em trabalhos futuros, pode ser possível verificar se essa estrutura funciona melhor, é mais robusta e veloz do que a que foi implementada. Ela está esquematizada no diagrama abaixo.

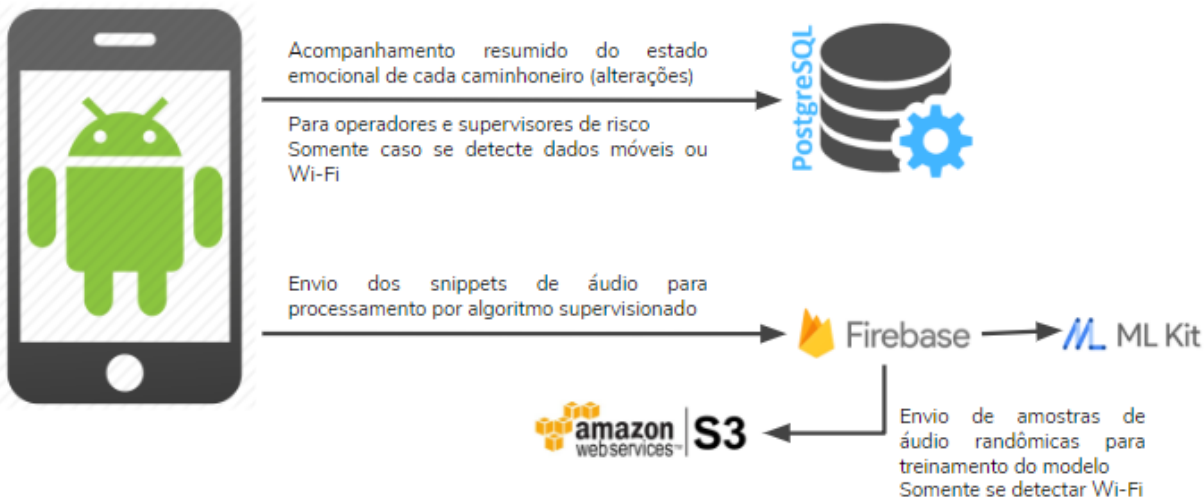

Fig. 17 – Proposta de melhorias na arquitetura do projeto CargoAffect

Para exemplificar uma situação em que a arquitetura acima seria um avanço em relação ao sistema implementado, pode-se pensar no momento imediatamente anterior ao início de um transporte de carga. Para que um caminhoneiro possa sair com a carga do pátio para realizar a entrega, é necessário que ele porte alguns documentos relativos à carga impressos, sob o risco de ser penalizado em uma fiscalização caso não os possua. Porém, a emissão destes documentos pode atrasar em algumas ocasiões, o que pode gerar uma situação de alto nível de estresse para o caminhoneiro.

Quando ocorre esse tipo de atraso, o motorista costuma entrar em contato com a transportadora para verificar o motivo. Como esse contato ocorre muitas vezes por meio de envios de áudios em aplicativos de mensagens instantâneas nos smartphones, seria possível coletar esses áudios e processá-los, com o intuito de comprovar uma condição de estresse. Então, assim que verificado o nível de estresse do usuário, seria possível alertar os operadores de cargas das transportadoras caso isso represente um risco na direção.

8 CONSIDERAÇÕES FINAIS

O projeto CargoAffect teve seu objetivo principal cumprido: o desenvolvimento de um sistema composto por um aplicativo de captura de voz para o sistema Android e um serviço web que realiza o processamento de amostras de voz de caminhoneiros com base em um algoritmo supervisionado, tomando ações de acordo com o estado emocional detectado.

Durante a fase de catalogação do corpus de áudio por voluntários, foi possível perceber a correlação existente entre as escalas de valência, excitação e estresse, o que foi um passo importante no avanço da compreensão de estados emocionais e a relação entre eles.

Ao final do projeto, obteve-se um sistema que realiza gravações de voz por meio de um aplicativo para o sistema Android e as envia a um serviço web que faz o processamento dos arquivos de áudio. O resultado do processamento é, então, exibido na tela do dispositivo móvel, com base nas escalas emocionais adotadas. Com isso, pode-se dizer que os conceitos de computação afetiva foram implementados de modo a criar um sistema que realiza o reconhecimento de emoções por meio da voz e tem o potencial de auxiliar os motoristas de caminhões a lidar com fatores que acarretam estresse durante seu trabalho.

Este trabalho contribuiu com a compreensão de conceitos de computação afetiva, a relação entre escalas de medição de estados emocionais e a própria concepção do que significa o estresse em termos dessas escalas, a detecção de emoções em áudio, a utilização do instrumento de avaliação SAM (*Self-Assessment Manikin*) adaptado para catalogar o estado emocional de áudios com base em escalas pré-definidas e a construção de uma aplicação utilizando um aplicativo para dispositivo móvel e um sistema web que realiza a detecção de estados emocionais a partir de áudios de voz e retorna o resultado ao usuário, como previsto na concepção do projeto.

A detecção do estresse nos áudios utilizados como teste do algoritmo supervisionado SVM foi relativamente bem-sucedida. Para os áudios que, na fase de catalogação, os voluntários apontaram altos níveis de estresse, o algoritmo classificador retornou níveis de estresse acima de seis, o que pode ser determinado como o limite aceitável de estresse e pode gerar alertas ao usuário caso o valor detectado esteja acima desse nível.

Porém, o procedimento de detecção de emoções com base nas escalas utilizadas ainda não apresenta bons níveis de precisão em relação aos níveis previstos para cada escala na fase de catalogação dos áudios de teste. Com isso, o *feedback* obtido ainda não possui um bom nível de confiança, sendo necessário verificar os fatores que diminuem a precisão em trabalhos futuros, como descrito no capítulo anterior.

Na continuação do projeto, pode-se realizar a comparação do desempenho do SVM com outros algoritmos, obter mais áudios para o corpus construído para treinar e testar o algoritmo, realizar a catalogação dos áudios com mais voluntários e utilizar análises estatísticas para determinar como utilizar melhor os dados obtidos pela catalogação.

Em trabalhos futuros, também pode-se verificar se todos os parâmetros extraídos de cada áudio neste projeto são, de fato, relevantes e significativos na detecção do estado emocional, realizar a detecção de estados de sono e cansaço além da já implementada detecção de estresse e realizar um teste de aceitação do sistema com motoristas de caminhões.

A maneira de fornecer o retorno ao motorista também pode ser revista na continuação do projeto. Como foi implementado, o sistema retorna os níveis de valência, excitação e estresse para cada áudio, mas não alerta o usuário caso seja detectado um alto nível de estresse, como previsto inicialmente.

Também é possível, em projetos futuros, avaliar se a análise do discurso de cada áudio (por meio da realização de transcrição automática dos áudios, *speech-to-text*) contribuiria na obtenção de resultados mais precisos em relação aos estados emocionais detectados. Há diversas maneiras possíveis de trazer melhorias à arquitetura implementada, que estão também descritas no capítulo anterior.

REFERÊNCIAS

[1] BRADLEY, M. M.; LANG, P. J. **Measuring emotion: The self-assessment manikin and the semantic differential**. Journal of Behavioral Therapy and Experimental Psychiatry 25, p. 49 a 59, 1994.

[2] BRADLEY, M. M.; LANG, P. J. **The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual**. Technical report B-3. University of Florida, 2007.

[3] BURKHARDT F. et al. **Emotion detection in dialog systems: applications, strategies and challenges**. Proceedings of the ACII, Amsterdam, 2009. p. 684-689.

[4] CANANI, S. F.; BARRETO, S. S. M. **Sonolência e acidentes automobilísticos**. Jornal Brasileiro de Pneumologia, v. 27, p. 94 a 96, mar./abr. 2001.

[5] CLYNES, D. M. **Sentics: The Touch of the Emotions**. Anchor Press/Doubleday, 1977.

[6] CYTOWIC, R. E. **The Man Who Tasted Shapes**. New York: G. P. Putnam's Sons, 1993.

[7] DESHPANDE, G. et al. **Empirical evaluation of emotion classification accuracy for non-acted speech**. IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), 2017.

[8] DIKKERS, H. J. et al. **Facial Recognition System for Driver Vigilance Monitoring**. IEEE International Conference on Systems, Man and Cybernetics, 2004.

[9] DIMATTIA, G. **An Automatic Audio Classification System for Radio Newscast**. 2008. Tese (Master in Telecommunication Engineering) – UPC Terrassa, Terrassa.

[10] GARG, P.; SEHGAL S. **Comparison of Emotion Recognition Models in Spoken Dialogs.** International Journal of Software & Hardware Research in Engineering, Vol. 3, Issue 3, Mar. 2015. p.48-54.

[11] GIANNAKOPOULOS, T. **pyAudioAnalysis: An open-source python library for audio signal analysis**. PLoS ONE, vol. 10, n. 12, 2015.

[12] GUVEN, E.; BOCK, P. **Speech emotion recognition using a backward context**. Proceedings of the IEEE 39th Applied Imagery Pattern Recognition Workshop, Washington, D.C., USA, Out. 2010, p. 1–5.

[13] HOOK, K. **Affective Computing**. In**: The Encyclopedia of Human-Computer Interaction.** 2ª Edição. Interaction Design Foundation, 2013. Disponível em: <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/affective-computing>. Acesso em 05 Maio 2018.

[14] LAZARUS, R. S. The Cognition-Emotion Debate: A Bit of History. In: DALGLEISH, T.; POWER, M. **Handbook of Cognition and Emotion**. Chichester: John Wiley & Sons, 1999. Cap. 1, p. 3-20.

[15] LUBIS, N. et al. **Emotion recognition on Indonesian television talk shows**. Proceedings of Spoken Language Technology Workshop (SLT), 2014. p. 466-471.

[16] MURRAY, I. R.; ARNOTT, J. L. **Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion**. Journal of the Acoustic Society of America, vol. 93, 1993. p. 1097-1108.

[17] PICARD, R. **Affective Computing**. MIT Press, 1997.

[18] PICARD, R. **Perceptual user interfaces: affective perception**. Communications of the ACM, v. 43, n. 3, p. 50-51, 2000.

[19] QUIRINO, G. de S.; VILLEMOR-AMARAL, A. E. de. **Relação entre estresse e agressividade em motoristas profissionais**. Rev. Psicol. Saúde, Campo Grande, v. 7, n. 2, p. 125-132, dez. 2015. Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S2177-093X2015000200006&lng=pt&nrm=iso>. Acesso em 20 Ago. 2018.

[20] ROSENWEIN, B. H.; CRISTIANI, R. **What is the history of emotions?** Polity Press, 2018.

APÊNDICE A – PESQUISA DE CAMPO

Roteiro para a entrevista em profundidade, com operadores e supervisores:

1. Faixa etária
    - De 18 a 29 anos
    - De 30 a 39 anos
    - De 40 a 49 anos
    - De 50 a 59 anos
    - Acima de 60 anos
2. Nível de escolaridade
    - Fundamental incompleto
    - Fundamental completo
    - Médio incompleto
    - Médio completo
    - Superior incompleto
    - Superior completo
3. Há quanto tempo você trabalha com caminhoneiros?
4. Com que frequência você entra em contato com caminhoneiros durante seu trabalho?
5. Qual é o meio que você mais utiliza para se comunicar com caminhoneiros?
6. Você já entrou em contato com caminhoneiros que estavam em situação de stress? Se sim, com que frequência e por qual motivo? Isso afetou negativamente o transporte do motorista ou a comunicação entre vocês?
7. Você já entrou em contato com caminhoneiros que relataram ou pareciam estar cansados ou com sono? Se sim, com que frequência? Isso afetou negativamente o transporte do motorista ou a comunicação entre vocês?
8. Você já notou outras situações adversas que colocaram em risco o trabalho dos caminhoneiros ou afetaram negativamente a comunicação entre você e eles? Se sim, quais? Com que frequência e por qual motivo?
9. Você já passou por algum caso de acidente de trânsito com um caminhoneiro que você tenha entrado em contato ou monitorado? Se sim, por qual motivo? Com que frequência?

10. Um aplicativo que conseguisse detectar situações adversas aos caminhoneiros durante o transporte de cargas, como cansaço, sonolência ou estresse, seria útil para seu trabalho?

Roteiro para a entrevista semiestruturada, com caminhoneiros:

1. Faixa etária:
    - De 18 a 29 anos
    - De 30 a 39 anos
    - De 40 a 49 anos
    - De 50 a 59 anos
    - Acima de 60 anos
2. Nível de escolaridade:
    - Fundamental incompleto
    - Fundamental completo
    - Médio incompleto
    - Médio completo
    - Superior incompleto
    - Superior completo
3. Anos de experiência como caminhoneiro(a):
    - Menos de um ano
    - De um a três anos
    - De três a cinco anos
    - De cinco a dez anos
    - Acima de dez anos
4. Você já percebeu ou soube de um colega de trabalho que passou por situações de estresse antes, durante ou depois do transporte de cargas?
    - Sim
    - Não
5. Se sim, com qual frequência? Por quais motivos?
6. Você já percebeu ou soube de um colega de trabalho que sentiu cansaço e/ou sonolência antes, durante ou depois do transporte de cargas?
    - Sim
    - Não

7. Se sim, com qual frequência? Por quais motivos?
8. Você já percebeu ou soube de outras situações parecidas que atrapalharam o trabalho dos seus colegas? Se sim, quais?
9. Você conhece algum caminhoneiro que já sofreu algum acidente por conta de estresse ou cansaço?
    - Sim
    - Não
10. No seu trabalho, o que te afeta mais negativamente e com mais frequência: sono ou estresse?
    - Sono
    - Estresse
11. Com qual frequência? Por quais motivos?
12. Geralmente, em períodos de trabalho, quantas horas de sono você tem por dia?
    - Abaixo de 4 horas
    - De 4 a 6 horas
    - De 6 a 8 horas
    - Acima de 8 horas
13. Você possui um smartphone com câmera?
    - Sim
    - Não
14. Se sim, com que frequência utiliza aplicativos de mensagens (como o WhatsApp) em seu smartphone?
    - Várias vezes por dia
    - Uma ou duas vezes por dia
    - Uma vez a cada três dias
    - Uma vez por semana
    - Não uso
15. Você usaria um aplicativo que detectasse condições adversas durante o seu trabalho, como sono ou estresse?
    - Sim
    - Não

APÊNDICE B – TERMO DE CONSENTIMENTO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Você está sendo convidado(a) a participar de uma pesquisa intitulada: “Aplicação da computação afetiva na experiência de usuário no contexto de motoristas de caminhões”, um trabalho de conclusão de curso coordenado e orientado pela Profª. Drª. Lucia Vilela Leite Filgueiras, da Universidade de São Paulo.

A sua participação não é obrigatória sendo que, a qualquer momento da pesquisa, você poderá desistir e retirar seu consentimento. Sua recusa não trará nenhum prejuízo para sua relação com o pesquisador, com a Universidade de São Paulo ou com as empresas CargoX.

O objetivo desta pesquisa é investigar a possibilidade de aplicar conceitos de computação afetiva no desenvolvimento de um sistema direcionado a motoristas de caminhões, melhorando seu trabalho, garantindo maior segurança em suas viagens e reduzindo a ocorrência de comportamentos de risco, como estresse e sono.

Caso você decida aceitar o convite, será submetido(a) aos seguintes procedimentos: entrevista individual e gravação de vídeo e áudio, para utilização posterior como parte do objeto de pesquisa. As entrevistas e gravações serão utilizadas para a análise dos dados referentes ao objeto de estudo, não sendo utilizados nome ou qualquer dado de identificação dos entrevistados.

O tempo previsto para a sua participação é de aproximadamente 10 minutos.

Os riscos relacionados com sua participação são a comunicação e divulgação de informações relativas a suas opiniões e percepções e serão minimizados pelos seguintes procedimentos: o pesquisador se compromete, no presente termo, a não utilizar o nome ou dados de identificação dos entrevistados. Se o entrevistado achar que determinadas perguntas incomodam, porque as informações coletadas são sobre suas experiências pessoais, podem escolher não responder, ou seja, o entrevistado pode deixar de responder quaisquer perguntas que o façam sentir incomodado.

Os benefícios relacionados com a sua participação serão: sua entrevista ajudará na elaboração do relatório final de pesquisa e na criação de um sistema que possa auxiliar motoristas de caminhões a melhorar seu trabalho, com mais segurança em suas viagens e menos chance de ocorrência de comportamentos de risco, como estresse e sono. Além disso, a entrevista ajuda na produção de conhecimentos na

área de computação afetiva e em relação aos riscos associados ao transporte rodoviário de cargas.

Os resultados desta pesquisa poderão ser apresentados em seminários, congressos e similares, entretanto, os dados/informações obtidos por meio da sua participação serão confidenciais e sigilosos, não possibilitando sua identificação.

A sua participação bem como a de todas as partes envolvidas será voluntária, não havendo remuneração para tal.

Após ser esclarecido (a) sobre as informações do projeto, se você aceitar em participar deste estudo, assine o consentimento de participação, que está em duas vias. Uma delas é sua e a outra é do pesquisador responsável. Em caso de recusa, você não será penalizado. Este consentimento possui mais de uma página, portanto, solicitamos sua assinatura (rubrica) em todas elas.

A qualquer momento, você poderá entrar em contato com o pesquisador principal, podendo tirar suas dúvidas sobre o projeto e sobre sua participação.

Pesquisador Responsável_____

Endereço_____

Telefone _____

Assinatura_____

CONSENTIMENTO DE PARTICIPAÇÃO

Eu, _____, abaixo assinado, concordo em participar do presente estudo como participante e declaro que fui devidamente informado e esclarecido sobre a pesquisa e os procedimentos nela envolvidos, bem como os riscos e benefícios da mesma e aceito o convite para participar. Autorizo a publicação dos resultados da pesquisa, a qual garante o anonimato e o sigilo referente à minha participação.

_____

Assinatura do participante ou Responsável legal

Telefone do participante para contato: _____

APÊNDICE C – REQUISITOS DO SISTEMA INICIAL DESCARTADO

R1 – Os usuários devem se cadastrar antes de utilizar o sistema.

R2 – Os usuários devem realizar login após o cadastro para ter acesso à lista de operadores e ao chat.

R3 – O sistema deve realizar amostragem e processamento de amostras de áudio em segundo plano.

R4 – O sistema deve alertar o usuário caso detecte um estado emocional adverso (sono ou estresse).

R5 – A tela principal deve fornecer ao usuário um histórico de seus estados emocionais.

R6 – O usuário deve conseguir desativar o envio de snippets de áudio e dados detectados de estado emocional.

R7 – O usuário deve ter a possibilidade de configurar o envio de dados somente caso o dispositivo encontre um sinal Wi-Fi.

R8 – O sistema deve enviar amostras de áudio randômicas para um bucker de armazenamento em nuvem da Amazon S3, em uma proporção de 1:10.

R9 – O sistema deve enviar uma identificação dos áudios processados, junto de tags que indicam o estado emocional detectado, a um banco de dados PostgreSQL.

R10 – Apenas usuários autenticados podem ter acesso aos recursos oferecidos pelo sistema.

R11 – O sistema é compatível com o sistema operacional Android.

R12 – O sistema não depende de redes móveis para funcionar, exceto ao realizar cadastro e login ao iniciar o aplicativo.