Desenvolvimento de Software com Reconhecimento de Fala para Registro, Processamento e Análise de Dados de Partículas Suspensas na Atmosfera de Belo Horizonte: Um Projeto em Parceria com a Universidade de Salamanca

Marina B. Diniz[1], Lucas A. Brandão[1], Francisco A. Gonçalves[2], Guilherme Augusto de Oliveira[2], André R. da Cruz[3], Natália C. Batista[3], Sandro R. Dias[3]

[1]Membro PET de Engenharia da Computação (COMPET)
[2]Egresso PET de Engenharia da Computação (COMPET)
[3]Tutor(a) PET de Engenharia da Computação (COMPET)
Departamento de Computação – Centro Federal de Educação Tecnológica de Minas Gerais, Av. Amazonas 7675, Nova Gameleira, Belo Horizonte – MG – Brasil.

{andradelucasbrandao, marinacefetmg, franciscoabreu1408, guilhermeaugustodeoliveira66}@gmail.com, {dacruz,nataliabatista, sandrord}@cefetmg.br

Abstract. *The research on pollen and fungal spores in the atmosphere is a relevant issue in various parts of the world, with implications for public health, urban planning, agriculture, and other areas. Sample collection points need to be established, and with the use of an airborne particle collector, samples are obtained and taken to a laboratory where microscopic analysis is performed for particle identification and counting by experts. In this way, the data recording process is carried out manually, requiring time and human resources. In this context, the project's objective is to develop software to expedite particle identification and analysis, based on speech recognition.*

Resumo. *A pesquisa sobre pólen e esporos fúngicos na atmosfera é uma questão relevante em várias partes do mundo, com implicações na saúde pública, planejamento urbano, agricultura e outras áreas. Pontos de recolhimento de amostras devem ser estabelecidos e, com a utilização de um captador de partículas suspensas no ar, as amostras são obtidas e levadas a um laboratório onde é realizada a análise microscópica para identificação e contagem das partículas por especialistas. Dessa forma, o processo de registro dos dados é realizado manualmente, demandando tempo e recursos humanos. Neste contexto, o objetivo do projeto é desenvolver um software para agilizar a identificação e análise das partículas, baseado em reconhecimento de fala.*

1. Introdução

A Aeropalinologia, ciência que estuda o conteúdo e comportamento atmosférico de pólen e esporos fúngicos, tem amplas aplicações em campos como agronomia (na previsão de colheitas e controle de pragas), medicina (relacionada a alergias), planejamento urbano e outros. Nessa área, os pesquisadores utilizam calendários polínicos, nos quais se pode apreciar a evolução anual no aparecimento dos diferentes elementos aeronavegantes estudados (Antón et al., 2020; Erbas et al, 2015). Na atualidade, existem numerosas estações de amostragem ao redor do mundo (Buters, 2018), possibilitando a análise de partículas na atmosfera mediante o estabelecimento de pontos de recolhimento de amostras localizados em diversas regiões, o que permite estabelecer e comparar os diferentes padrões de comportamento aerobiológico das partículas biológicas.

Para obter as amostras, é necessário um captador de partículas suspensas no ar. O dispositivo possui uma cabeça removível que contém um tambor cilíndrico sobre o qual é colocada uma fita plástica impregnada com uma substância adesiva, que possibilita reter todas as partículas capturadas durante a amostragem. As amostras obtidas são levadas a um laboratório onde é realizada a análise microscópica para identificação e contagem das partículas por especialistas que anotam em folhas de resultados diários o número total de grãos de pólen e esporos (Domínguez et al., 1991; Galán et al., 2007). Dessa forma, o processo de registro dos dados é realizado manualmente, o que demanda tempo e recursos humanos.

Neste contexto, o objetivo deste projeto é desenvolver um sistema para registro, processamento e análise eficiente de dados de partículas suspensas na atmosfera. O sistema contará com um módulo para suporte à leitura e identificação das partículas baseado em reconhecimento de fala, agilizando o trabalho do pesquisador que fará essas leituras e fornecendo uma análise dos dados e apresentação dos resultados. Os dados serão armazenados em um banco de dados, que será integrado ao sistema. Ao final do projeto, o sistema web permitirá o acesso às informações destas análises sazonais e um sistema de suporte à leitura das amostras.

O projeto é realizado em parceria com a Universidade de Salamanca, na Espanha, e conta com a participação do grupo COMPET, PET de computação do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e trará como contribuição à comunidade científica um banco de dados das partículas suspensas, mais especificamente pólen e esporos, em uma região de Belo Horizonte, além de análises sazonais das incidências dessas partículas, fomentando outras pesquisas ou análises sanitárias específicas.

Este artigo apresenta o desenvolvimento do sistema, que está em andamento, e os passos futuros. A próxima seção apresenta os conceitos e referências relacionados à metodologia proposta e à sua implementação. A metodologia é apresentada na Seção 3 e

os resultados obtidos até o momento são descritos na Seção 4. Finalmente, as considerações finais são abordadas na Seção 5.

2. Revisão bibliográfica

Reconhecimento de fala (ASR-*Automatic Speech Recognition*) é um recurso computacional que usa técnicas de inteligência artificial e aprendizado de máquina para permitir que um programa processe a fala humana em um formato escrito, possibilitando converter áudio em texto (*speech-to-text*) (Dhanjal; Singh, 2023; Kuligowska et al. 2018; Saksamudre et al., 2015). Atualmente existem no mercado várias aplicações da tecnologia de fala, por exemplo no setor automotivo (sistemas de navegação ativados por voz e recursos de procura em rádios automotivos), segurança (autenticação com base em voz), vendas (central de atendimento), etc.

Existem diversos aplicativos e dispositivos de reconhecimento de fala disponíveis, sendo que alguns oferecem serviços pagos como o IBM Watson Speech to Text (IBM, 2023) e o Google Speech- To-Text API (Google Cloud, 2023) e outros que são de código aberto como Fairseq (Fairseq Developers, 2023), ESPnet (ESPnet Developers, 2023), Mozilla DeepSpeech (DeepSpeech Developers, 2023), Kaldi (Kaldi ASR Developers, 2023), dentre outros. Entretanto, nem todas as soluções permitem personalizar e adaptar a tecnologia a requisitos específicos, por exemplo idioma, pronúncia, sotaque e reconhecimento de termos específicos que não constam no vocabulário de base do idioma, como o nome científico de espécies na biologia. Além disso, um bom modelo deve ser capaz de gerenciar áudios com ruídos de vários ambientes, variações no volume e na densidade.

Os sistemas ASR podem ser avaliados em sua taxa de precisão, ou seja, taxa de erro de palavra (WER) e velocidade, denre outras métricas. Para aumentar a precisão dos sistemas existentes de conversão de fala em texto, pode-se realizar o treinamento do modelo de inteligência artificial com dados novos no domínio específico e utilizar transferência de aprendizado, já que a maioria é baseada em redes neurais profundas. Os modelos de aprendizado profundo (Goodfellow et al., 2016), são um conjunto de algoritmos de aprendizado de máquina inspirados no funcionamento do cérebro humano compostos de redes neurais artificiais com múltiplas camadas. Essas redes, conhecidas como redes neurais profundas, são capazes de aprender representações complexas de dados, tornando-se especialmente eficazes na resolução de tarefas avançadas, como reconhecimento de objetos em imagens e processamento de linguagem natural.

Quando o modelo de um sistema ASR é treinado para uma determinada língua e são apresentadas palavras que não estão em seu vocabulário (OOV - *Out-Of-Vocabulary words)*, o modelo possivelmente não conseguirá reconhecer a palavra corretamente, prejudicando o desempenho do sistema. Essas palavras OOV são

geralmente nomes e locais. Há várias abordagens propostas para resolver este problema (Kitaoka et al., 2021; Qin, 2013), como por exemplo a utilização de um léxico híbrido e um modelo de linguagem híbrida que incorpora palavras e sub-unidades lexicais, que foi a escolha para este trabalho, pois há nomes de espécies em latim em meio a comandos em inglês. Não foram encontrados trabalhos que realizam reconhecimento de fala especificamente no contexto da aerobiologia.

3. Metodologia

A metodologia deste projeto consiste em seis etapas principais. A primeira etapa abrange a construção da base de dados para obter um conjunto de áudios representativo para treinamento do sistema de reconhecimento de voz. Os áudios foram gravados por diversas pessoas e contém os nomes científicos de grãos de pólen e comandos que serão usados no sistema. A base de dados é composta por 43 nomes de partículas biológicas, em latim, e 5 comandos a serem dados ao sistema, em inglês.

Para a transcrição fonética em inglês, foi utilizado um dicionário online que oferece a transcrição fonética das palavras em inglês de acordo com o IPA (*International Phonetic Alphabet*), padrão internacional[1]. Já para a transcrição fonética em latim, foi necessário utilizar um transcritor fonético *online* (OpenIPA, 2023). Além disso, foram fornecidos áudios gravados pela pesquisadora da Universidade de Salamanca (Professora Estefanía S. Reyes) com a pronúncia correta dos nomes dos pólen e esporos fúngicos. Com isso, foi possível elaborar roteiros e um fornecer um manual de pronúncia para os voluntários e usuários do sistema. Esse roteiro foi montado por um script na linguagem python. Ao todo, cada voluntário gravou 48 arquivos de áudio no formato wav (43 nomes e 5 comandos), um áudio para cada palavra.

Neste projeto, optou-se por utilizar a biblioteca Kaldi, que é um kit de ferramentas contendo quase todos os algoritmos atualmente usados em sistemas ASR. Ele também contém receitas para treinar seus próprios modelos acústicos em corpora de fala comumente usados, como o Wall Street Journal Corpus, TIMIT e outros. Essas receitas também podem servir como modelo para treinar modelos acústicos em seus próprios dados de fala. Sua arquitetura modular permite aos desenvolvedores personalizar cada etapa do processo de reconhecimento de fala, o que facilita a adaptação do sistema para atender às necessidades específicas de diferentes aplicações, como é o caso deste projeto que envolve vocabulário relativo à aeropalinologia, especificamente nomes científicos em latim. O Kaldi fornece ferramentas para o treinamento de modelos de reconhecimento de voz, incluindo o uso de Modelos Ocultos de Markov (HMMs) e Redes Neurais Profundas (DNNs) (Kaldi ASR Developers, 2023), e é usado principalmente em reconhecimento de fala para assistentes virtuais e biometria de voz, como também em transcrições.

1 https://www.dictionary.com/browse/biology. Acesso em: 13 out. 2023.

O Kaldi requer vários formatos das transcrições para treinamento de modelos acústicos. Além disso, é preciso os horários de início e término de cada elocução, o ID do locutor de cada elocução e uma lista de todas as palavras e fonemas presentes na transcrição. É fundamental destacar que, mesmo quando a pronúncia está correta, cada indivíduo possui uma maneira única de pronunciar cada palavra, o que resulta em diferenças nas amplitudes e frequências das ondas sonoras. Essa diversidade contribui para a robustez do modelo, pois um modelo ideal deve ser capaz de reconhecer e adaptar-se a diversas formas de pronúncia e variações de timbre de voz.

A segunda etapa refere-se à obtenção de um modelo de reconhecimento de fala. Como mencionado anteriormente, optou-se pela ferramenta Kaldi, uma vez que a estrutura da base de dados requisitada é utilizada em outras ferramentas como a Vosk (Alpha Cephei, 2023), e a ESPnet. Para criar um sistema ASR simples no kit de ferramentas Kaldi usando o próprio conjunto de dados é necessário possuir o sistema operacional Linux e utilizar as linguagens Python, C++, Shell e Perl. Os parâmetros do modelo acústico são estimados em etapas de treinamento acústico. No entanto, o processo pode ser melhor otimizado passando pelas fases de treinamento e alinhamento. Isso também é conhecido como treinamento de Viterbi (procedimentos relacionados, mas computacionalmente mais caros, incluem o algoritmo Forward-Backward e Expectation Maximization). Ao alinhar o áudio à transcrição de referência com o modelo acústico mais atual, algoritmos de treinamento adicionais podem usar essa saída para melhorar ou refinar os parâmetros do modelo. Portanto, cada etapa do treinamento será seguida de uma etapa de alinhamento onde o áudio e o texto poderão ser realinhados.

À medida que o algoritmo de reconhecimento de fala é treinado e validado, entra-se na terceira etapa, de planejamento e execução de experimentos abrangentes. Esses experimentos são projetados para testar e aprimorar ainda mais os modelos gerados. Esta etapa envolverá a utilização do sistema por potenciais usuários. Nesta etapa, serão utilizados dados reais provenientes das análises de amostras coletadas em Belo Horizonte, pois está prevista a aquisição e instalação de um captador de partículas suspensas no ar no Campus Nova Suíça do CEFET-MG, proveniente de recursos aprovados na Demanda Universal da Fapemig 2022. Considerando que não há sistema similar em Minas Gerais (há no Brasil apenas 4 equipamentos no estado de São Paulo e um no Amazonas), a instalação deste captador e a avaliação dos dados trará benefícios para a cidade e para a comunidade científica que disporá dessas informações para desenvolver pesquisas voltadas a estratégias sanitárias ou relacionadas ao impacto dessa dispersão de partículas aéreas.

A quarta etapa consiste na integração do sistema de reconhecimento de fala a um banco de dados, permitindo aos pesquisadores inserir e consultar informações de maneira eficiente e intuitiva. A implementação deste sistema pode auxiliar significativamente na análise de grandes volumes de dados, contribuindo para uma

avaliação das análises sazonais, diárias e anuais da dispersão das partículas. armazenamento e recuperação de dados. Utilizou-se o Firebase Firestore, que é um banco de dados NoSQL que permite armazenar, sincronizar e consultar dados em escala global. Essa foi a escolha para o projeto tendo em vista a necessidade de sincronizar as informações de grãos de polens e esporos com diferentes pesquisadores. Além disso, a biblioteca *ChartJS* está sendo utilizada para a geração de gráficos, permitindo que os dados coletados por meio da captura de voz sejam estudados através de representações visuais.

A quinta etapa refere-se ao desenvolvimento do sistema web que permitirá o acesso às informações armazenadas no banco de dados, incluindo a construção dos gráficos que permitirá uma visualização clara e objetiva dos dados coletados. Este sistema se baseia em tecnologias como o React e as ferramentas do Firebase, incluindo o Auth e o Firestore. React é uma biblioteca JavaScript para a construção de interfaces de usuário. Desenvolvido pelo Facebook, React é amplamente adotado por sua eficiência e flexibilidade, permitindo a criação de componentes reutilizáveis, que podem ser compostos para formar interfaces de usuário complexas. Além disso, o React utiliza um modelo de programação declarativa, o que facilita o rastreamento de como a interface do usuário deve ser atualizada em resposta a mudanças de estado (Abramov, 2019). Firebase é uma plataforma de desenvolvimento de aplicativos, desenvolvida pelo Google, que fornece uma série de serviços *serverless*, ou seja, serviços que não exigem a gestão direta de servidores por parte dos desenvolvedores. Isso permite que os desenvolvedores se concentrem na lógica do aplicativo, enquanto o Firebase cuida de aspectos como autenticação de usuários (Firebase Auth), armazenamento e recuperação de dados (Firebase Firestore), entre outros.

Na sexta etapa será realizada a análise dos resultados obtidos, examinando-se as saídas e comparando-as com os resultados esperados, a fim de detectar falhas e possíveis melhorias, bem como desafios ou gargalos.

4. Resultados

As etapas explicadas na Seção 4 encontram-se em andamento, porém alguns resultados iniciais mostram o potencial do sistema para a aplicação descrita. Os resultados do treinamento do sistema de reconhecimento de fala, realizado com 189 amostras de áudio, mostraram que uma baixa frequência de repetição das palavras leva a resultados insatisfatórios, sendo necessário ampliar a base de dados. Essa situação levou à necessidade de reformular a base de dados, a fim de obter uma frequência mais elevada de cada palavra. O tamanho mínimo recomendado para esse modelo, conforme os requisitos, é de aproximadamente 50 vezes para cada palavra[2]. Dessa forma, foi

2 De acordo com as respostas no fórum da comunidade Kaldi: https://groups.google.com/g/kaldi-help/.

necessário convocar mais voluntários para popular a base e reforçar o modelo.

Em relação à interface web, foram desenvolvidas neste projeto algumas funcionalidades básicas. Como pode ser visto na Figura 1, o cadastro pode ser feito com a autenticação do Google, ou com email e senha. Com o usuário devidamente logado, é possível ter acesso ao menu de seleção para a área em que foi feita a coleta da amostra que será implantada no banco de dados e, após a seleção, o usuário tem acesso à área de reconhecimento de voz e cadastro dos dados no sistema (Figura 2). Nessa área, o usuário pode falar o nome do pólen e a quantidade observada, bem como falar comandos como *minus* (o sistema subtrai a quantidade falada após o comando do tipo de pólen escolhido) e *read* (o sistema informa a quantidade já registrada para o tipo de pólen escolhido). A Figura 3 mostra a planilha após a inserção de alguns tipos de pólen e sua quantidade.

O dados cadastrados podem ser filtrados e analisados de diferentes formas, como por meio da análise do comportamento intra-diário e do estudo da variação sazonal. Para tal, são representadas graficamente as concentrações médias diárias registradas ao longo do período de estudo, utilizando uma média móvel de 5 dias, bem como as concentrações mensais totais. A Figura 4 mostra a página que permite gerar o gráfico da média móvel a partir dos dados coletados, no período selecionado e a Figura 5 mostra um exemplo de gráfico gerado a partir dos dados coletados.

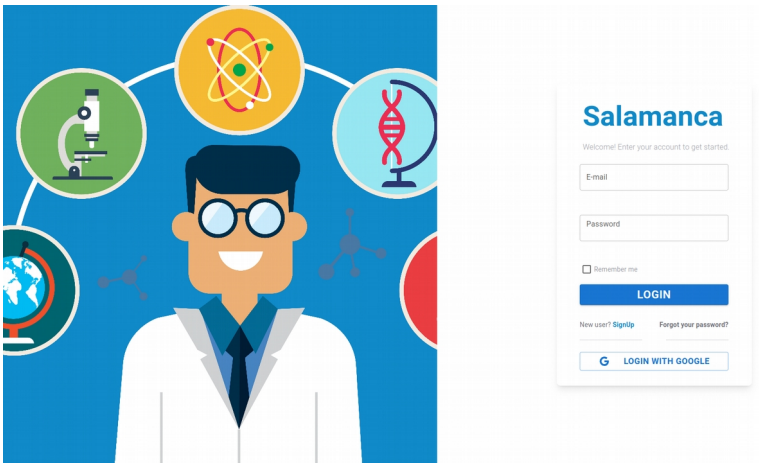

Figura 1. Página de login do sistema.

Acesso em: 13 out. 2023.

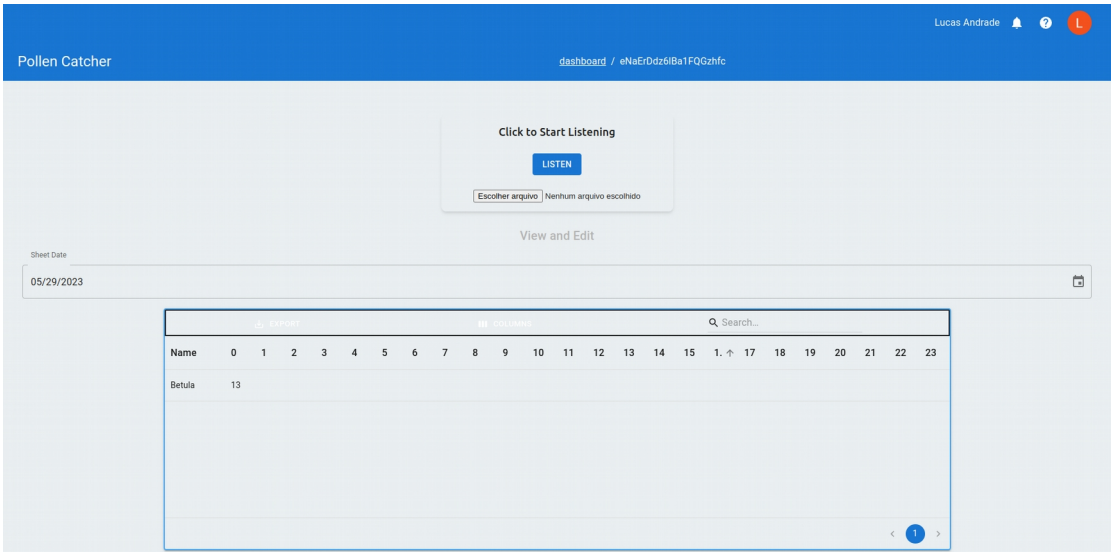

Figura 2. Página que permite o reconhecimento de voz e coleta de dados.

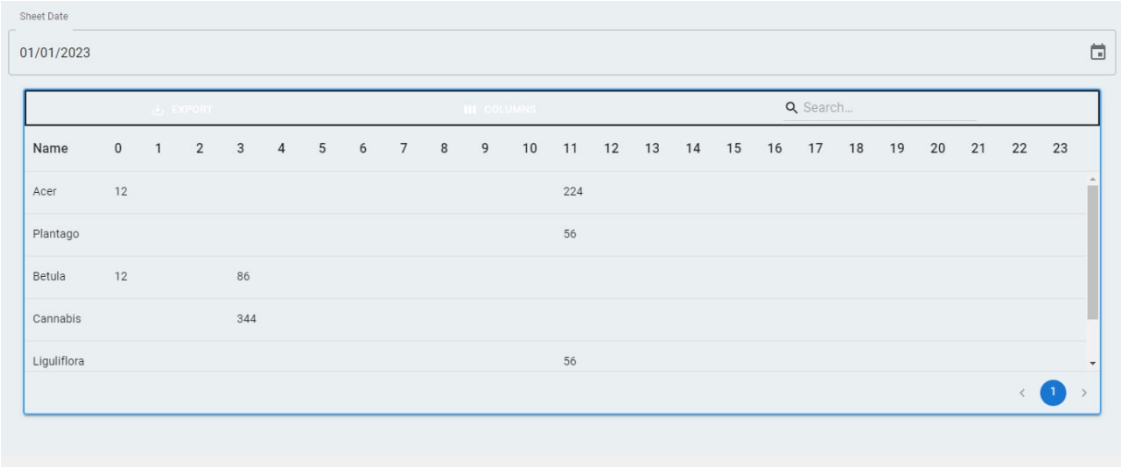

Figura 3. Página que mostra a planilha após a inserção de alguns tipos de pólen e sua quantidade.

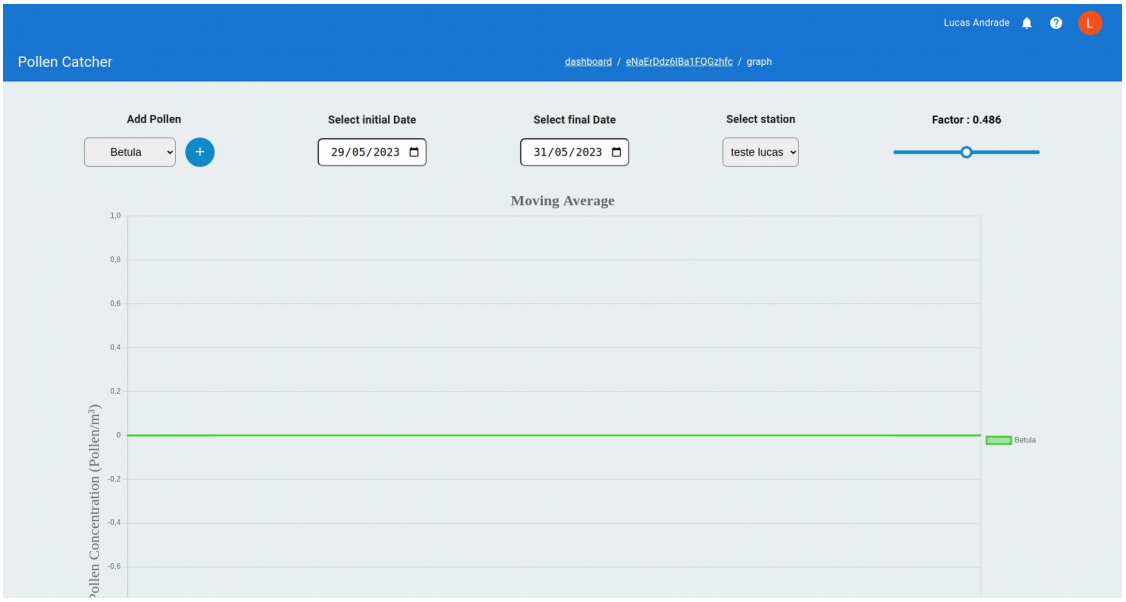

Figura 4. Página para gerar o gráfico da média móvel a partir dos dados coletados, no período selecionado.

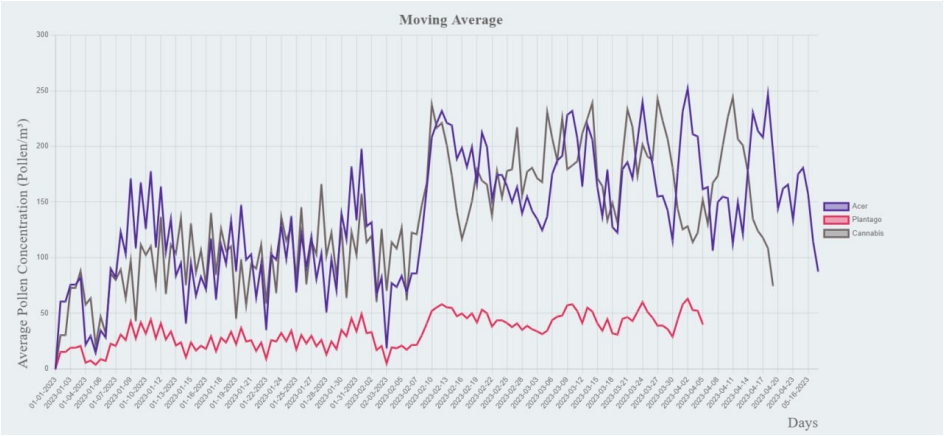

Figura 5. Página que mostra um exemplo de gráfico gerado a partir dos dados coletados.

5. Conclusão

O projeto descrito neste artigo, que é realizado pelo COMPET em parceria com a Universidade de Salamanca, possui uma equipe multidisciplinar dedicada a desenvolver um software que realiza o reconhecimento de fala para registro, processamento e análise de dados de partículas suspensas na atmosfera. Os experimentos iniciais com a ferramenta Kaldi revelaram seu potencial para a criação e treinamento de modelos de fala para texto, porém mostraram desafios em relação à sensibilidade dessa ferramenta ao tamanho da base de dados, evidenciando a necessidade de um número adequado de exemplos para cada palavra e consequentemente levando a busca por uma reformulação

da base de dados.

Além disso, está sendo implementado um banco de dados para armazenamento dessas informações e geração das variadas análises sazonais, diárias e anuais da dispersão das partículas. Com a utilização do banco de dados, o cruzamento desses dados com os dados meteorológicos e sanitários da cidade permitirão conclusões específicas sobre a concentração de partículas específicas na atmosfera e seu espalhamento na região. Isso configurará a criação de um calendário da dispersão polínica na região analisada. Ao final do projeto teremos um sistema web que permitirá o acesso às informações destas análises sazonais e um sistema de suporte à leitura das amostras. Esta proposta representa um avanço para Belo Horizonte e Minas Gerais, que não possuem este tipo de equipamento instalado ou estes dados disponibilizados, o que pode ser comprovado dentro da numerosa comunidade científica internacional dedicada aos estudos aeropalinológicos.

6. Agradecimentos

Gostaríamos de expressar nossa sincera gratidão à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) e ao CEFET-MG (Centro Federal de Educação Tecnológica de Minas Gerais) por sua valiosa contribuição e apoio que permitiram a continuidade desta pesquisa. Gostaríamos de agradecer também à Professora Estefanía Sánchez Reyes, da Universidade de Salamanca, por todo seu apoio.

Referências

Abramov, D. (2019). The Road to React. Leanpub.

Alpha Cephei, “Vosk Speech Recognition Toolkit,” https://alphacephei. com/vosk/. Acesso em: 19 set. 2023.

Antón, S. et al. Urban atmospheric levels of allergenic pollen: comparison of two locations in Salamanca, Central-Western Spain. Environmental Monitoring and Assessment, v. 192, n. 414, 2020.

Buters, J. et al. (2018), “Pollen and spore monitoring in the world”.,Clin. Transl. Allergy, 8: 9.

Dhanjal, A. e Singh, W. A comprehensive survey on automatic speech recognition using neural networks. Multimedia Tools and Applications, 2023.

DeepSpeech Developers, “DeepSpeech Documentation,” https:// deepspeech.readthedocs.io/en/r0.9/. Acesso em: 19 set. 2023.

Dempster, A. et al. (1977) “Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Stat. Soc. Ser. B (Methodological) 39, p. 1–38.

Domínguez, E. et al. Manejo y evaluación de los datos obtenidos en los muestreos aerobiológicos. 1991. 18 p. Monografías REA/EAN, 1991.

Erbas, B. et al. Do human rhinovirus infections and food allergy modify grass pollen-induced asthma hospital admissions in children?. The Journal of allergy and clinical immunology, vol. 136, n.4, p. 1118-1120.e2, 2015.

ESPnet Developers, “ESPnet Documentation,” https://espnet.github.io/ espnet/. Acesso em: 19 set. 2023.

Fairseq Developers, “Fairseq Documentation,” https://fairseq. readthedocs.io/en/latest/. Acesso em: 19 set. 2023.

Firebase. (2020), “Firebase Documentation”, https://firebase.google.com/docs. Acesso em: 19 set. 2023.

Galán, C.et al. Manual de Calidad y Gestión de la Red Española de Aerobiología. Servicio de Publicaciones de la Universidad de Córdoba, 2007. 39 p.

Goodfellow, I et al. (2016), Deep Learning, MIT Press, Cambridge, vol. 1.

Google Cloud. (2023) “Google Cloud Speech-to-Text”, https://cloud.google.com/speech-to-text. Acesso em: 19 set. 2023.

IBM Brasil. (2023) https://www.ibm.com/br-pt. Acesso em: 01 set. 2023.

Kaldi ASR Developers, “Kaldi ASR Documentation,” http://kaldi-asr.org/doc/index.html. Acesso em: 19 set. 2023.

Kitaoka, N. et al. Dynamic out-of-vocabulary word registration to language model for speech recognition. EURASIP Journal on Audio, Speech, and Music, vol. 4, 2021.

Kuligowska et al. Managing Development of Speech Recognition Systems: Performance Issues. Annales Universitatis Mariae Curie-Skłodowska, sectio H – Oeconomia, v. 52, n. 2, p. 71-78, 2018.

OpenIPA. (s.d.), “Transcription of Latin”, https://www.openipa.org/transcription/latin. Acesso em: 19 set. 2023.

Qin, L. Learning out-of-vocabulary words in automatic speech recognition. 2013. 111 f. Thesis (Doctor of Philosophy) - Carnegie Mellon University, Pittsburgh, 2013.

Saksamudre et al. A review on different approaches for speech recognition system. International Journal of Computer Applications, v. 115, n. 22, 2015.

Schmidhuber, J. (2015), “Deep learning in neural networks: An overview,” Neural Networks, 61, p. 85–117.