Desenvolvimento de um sistema automático para confecção de atas por captação de fala

RAISSON ALMEIDA DE SANTANA[1]

MARCOS BATISTA FIGUEREDO[2]

Resumo

Atas são documentos importantes que registram de forma formal e rigorosa o que foi feito ou acontecido em uma determinada reunião ou evento. Através desses registros, é possível certificar os eventos e garantir a veracidade dos fatos. Este artigo tem como objetivo explorar como o uso da tecnologia da informação pode melhorar o processo de redação de atas de reuniões. O problema é que criar minutas precisas pode ser um desafio, e erros na documentação podem levar a tomadas de decisão incorretas. Portanto, o estudo propõe um sistema baseado em reconhecimento contínuo de fala Continuous Speech Recognition (CSR), para automatizar a tarefa de escrever atas. A tecnologia da informação pode melhorar o processo de documentação técnica das reuniões, melhorando assim o acesso às decisões e aumentando a divulgação das resoluções aplicadas nas reuniões.

Palavras-chave: Reconhecimento Contínuo de Fala. Rede Neural Artificial. Extração de Características.

Abstract

Minutes are important documents that formally and accurately record what was done or happened at a specific meeting or event. Through this document it is possible to verify the events and ensure the reliability of the facts. This article aims to analyze how the use of information technology can optimize the process of preparing minutes in meetings. The problem is that creating accurate drafts can be challenging, and mistakes in documentation can lead to incorrect decision making. Therefore, the study proposes a system based on Continuous Speech Recognition (CSR) to automate the task of writing minutes. Information technology can improve the process of technical documentation of meetings, thus facilitating

[1] Graduando em Sistema de Informação – almeidaraisson@gmail.com

[2] Graduado em Licenciatura em Matemática, Doutor em Modelagem computacional e Tecnologia Industrial e Professor da UNEB –mbfigueredo@uneb.br

access to decisions and expanding the dissemination of resolutions adopted at meetings. The objective of this work was to propose a model for the automatic generation of minutes from the audio transcription of meetings. Despite the difficulties encountered, we were able to elaborate an initial proposal that can serve as a basis for future improvements. However, we present some of the challenges and limitations that must be taken into account for the development of an efficient and reliable system for the production of minutes by voice recognition.

Keywords: Continuous Speech Recognition. Artificial Neural Network. Feature Extraction.

1 Introdução

Na elaboração da ata temos a oportunidade de expressar por escrito de forma formal e rigorosa o que foi feito ou acontecido através da emissão de um documento denominado Registo. Existem também várias expressões que usam esta palavra em latim (por exemplo, ad acta, que significa "ao sujeito"). Para certificar um evento, coloquialmente utiliza-se a expressão “tirar a ata”. A maioria das instituições e organizações possuem diários de bordo. Que geralmente contém o que foi acordado na reunião (as páginas são numeradas para evitar alteração não autorizada do conteúdo deste livro). Logicamente, depois que um protocolo é escrito, as pessoas interagem com sua redação para aprová-lo. Como podemos pode ver, este é um formulário destinado a dar garantias por escrito da veracidade dos fatos. Assim como os sistemas robóticos são baseados em habilidades humanas como movimento, articulação, consciência espacial e expressões faciais, a história da pesquisa no campo do processamento da linguagem é baseada na capacidade humana de falar, ouvir, ler, escrever, reconhecer pela voz, traduzir palavras de um idioma para outro e tantos outros recursos relacionados ao assunto. Com o avanço da tecnologia, o acesso ao computador torna-se essencial em muitas áreas e, com isso, a necessidade de facilitar a interação torna-se maior. (MORAIS, LOPER, 2014). Embora isto pareça simples, mas, é complexo por natureza.

Em qualquer ambiente que a ata se faz necessária, é notório que confeccioná-la requer bastante atenção para garantir que ela não leve rasuras ou equívocos. Os descuidos com a retratação podem levar a tomadas de decisões erradas. Assim como é importante que o responsável pela ata não participe ativamente da reunião. Dessa forma, ela pode se concentrar em sua tarefa de criar o documento. Ainda que esse registro seja designado pelo secretário da mesa da assembleia, é essencial que todos os envolvidos saibam como desenvolvê-la. Devido a essas problemáticas, notou-se a pertinência de, por meio da tecnologia da informação, confiar tarefas aos sistemas informáticos e não aos recursos humanos, melhorando a produtividade e permitindo que os participantes se concentrem em tarefas de maior valor agregado. Para isto, como o registro técnico de assembleias pode ser melhorado através da Tecnologia da Informação? Com objetivo de responder esta pergunta, este artigo abordou os recursos atuais disponíveis e buscou propor um modelo que melhor cumprisse este objetivo. Atrelado a isso, o interesse por esse estudo surgiu com a participação do Programa Institucional de Iniciação Científica (IC), da Universidade do Estado da Bahia (UNEB), e devido às pesquisas relacionadas a área de Modelagem Computacional, voltada para o Reconhecimento Automático de Fala. O que tornou possível observar a necessidade de um sistema automático que produza atas de reuniões ordinárias, por meio da captação de fala. Desta forma, este artigo propõe uma modernização administrativa.

A elaboração de atas, em sua maioria, causa um desconforto para seus responsáveis por ser considerado meticuloso. O uso de sistemas voltados para essa área busca solucionar, com praticidade, essas problemáticas tanto para as pessoas que produzem as atas diretamente como para as que utilizam de forma indireta. Desta forma a Tecnologia da Informação melhora o processo de registro técnico documental de reuniões, assembleias e afins, pois é a partir disto que se pode democratizar o acesso às decisões tomadas e aos assuntos abordados. Aprimorar esse processo de registro é essencial para a melhora da formalização e divulgação de resoluções aplicadas em reuniões.

2 Fundamentação teórica

Neste capitulo iremos abordar as principais características do Processamento de linguagem natural e as principais características de uma Rede neural, posteriormente, o processo de reconhecimento de fala.

2.1 Reconhecimento de fala

O reconhecimento de fala é a conversão da fala humana em texto ou sinal de controle por meio de algoritmos inteligentes(JOLAD; KHANAI, 2019). Já o Processamento de linguagem natural(PLN), também conhecida como linguística computacional, pertence aos campos da ciência da computação e da linguística como um subcampo da inteligência artificial (IA) MARTINEZ e ZEROUAL; LAKHOUAJA. O processo de fala humana possui basicamente duas etapas, a primeira o falante converte a informação que pretende transmitir em símbolos de uma estrutura linguística. Na segunda etapa, é feita a materialização desses símbolos em unidades acústicas, acionando os músculos necessários à geração de fluxo de ar proveniente dos pulmões e para haver uma modelação desse fluxo através da geometria das cordas vocais e do trato vocal. Assim produzindo uma onda de pressão acústica, notando-se que esse sinal é devolvido ao próprio falante via seu aparelho auditivo, permitindo avaliar e controlar o processo de produção de fala. No entanto, durante a transmissão, os sinais de fala são frequentemente afetados pelos mais variados tipos de ruído (por exemplo, vozes de outras pessoas) e pela distorção do canal (por exemplo, o corte em altas frequências superiores a 4 Khz). Finalmente, ao receber a fala, o ouvinte capta a onda de pressão sonora através do ouvido e tenta extrair a informação nela contida usando um processo inverso ao processo de fabricação.

O reconhecimento de fala é uma tarefa difícil devido às diferenças nos dispositivos de gravação, alto-falantes, contexto e ambiente. No mundo real o reconhecimento da fala humana quase sempre envolve ouvir com ruído de fundo. O impacto desse ruído nos sinais de fala e no desempenho da inteligibilidade aumenta com a separação do ouvinte do falante, (MEYER; DENTEL; MEUNIER, 2013). O problema com o reconhecimento de fala é determinar o que você diz com base no valor medido da onda de pressão sonora, que é geralmente uma versão amostrada e quantizada de um sinal de microfone. Segundo CORO et al., atualmente os reconhecedores automáticos de fala (ASR) comerciais alcançam alto desempenho no reconhecimento de fala limpa, mas suas abordagens são pouco relacionadas ao reconhecimento de fala humana. É realmente uma questão de modelação de um sistema inverso. Para resolver esse problema, a análise deve usar técnicas que apliquem o conhecimento das tensões que ocorrem no mecanismo que produz a fala. Tal conhecimento pode estimar os parâmetros

expressivos que ocorrem durante a geração da sentença analisada e, em última análise, permitir a reconstrução do evento de fala original.

O processo de reconhecimento se dá em três etapas: pré-processamento de fala, extração de características e classificação de fala, (JOLAD; KHANAI, 2019). No préprocessamento de fala busca-se remover o ruído do sinal de fala de entrada, segmentando o sinal de fala do ruído de fundo e segmentar as palavras da fala. No que diz respeito a técnica de extração de recursos, há variadas técnicas difundidas como: métodos estatísticos„ métodos espectrais, métodos baseados em modelos, métodos baseados em transformers e métodos de reconhecimento de padrões. Neste aplicação, foi utilizado o Coeficiente Cepstral de Frequência do Mel (MFCC), que converte a escala de frequência linear para a escala de frequência Mel logarítmica reduzida. Esta técnica sofre com o ruido de fundo presente no sinal de fala. E seu desempenho depende do número de banco de filtros Mel(VIJAYAN et al., 2017).

2.2 Captação de fala

A captura de sinais de áudio é uma parte fundamental do desenvolvimento de sistemas de reconhecimento de fala. Há bancos de dados que fornecem arquivos de áudio, útil projetos em que a captura de áudio não é vital, portanto, os bancos de dados podem ser usados para testes que não requerem aquisição de sinal de áudio. E este é o nosso caso. Para este trabalho optamos por utilizar gravações de voz de bancos disponíveis, como o Common Voice — Mozilla Foundation. De acordo com a própria fundação, o objetivo do projeto é construir um conjunto de dados de fala multilíngue e de código aberto que qualquer pessoa possa usar para treinar aplicativos de fala.

2.3 Rede neural artificial

Se tratando de Rede neural artificial (ANN) é importante saber que, este modelo possui três camadas: camada de entrada, camada oculta e camada de saída. Durante o seu treinamento ajusta os pesos das camadas ocultas para predição do rótulo correto da classe, esta é uma uma rede que pré-treina o modelo antes de testar (KUMAR; KUMAR; RAJAN, 2009), As redes neurais até podem resolver tarefas cognitivas mais complexas, mas não funcionam tão bem quanto o Hidden Markov Model (HMM) ao lidar com grandes vocabulários (SAKSAMUDRE; SHRISHRIMAL; DESHMUKH, 2015);

Atualmente, a Rede neural recorrente (RNN) tem se mostrado uma ferramenta eficaz em modelos de reconhecimento de fala. Por exemplo no projeto Deepspeech, HANNUN et al., que utiliza em seu núcleo uma RNN treinada para receber espectogramas de fala e retornar transcrições em inglês. Contendo em sua arquitetura quatro partes, sendo elas a camada de convolução, Unidade linear retificada, Camada de pool e camada totalmente conectada. Considerada como rede neural profunda, as RNN são bastante difundidas para fins de processamentos de dados temporais, como sequencias de texto ou áudio.

2.4 Classificadores

Tratando-se de classificadores em reconhecimento de fala, há diferentes classificadores capazes de reconhecer a palavra falada corretamente comparando-a com dados pré-treinados. O Classificador K-Vizinho Mais Próximo (kNN) é um método de reconhecimento de padrões bastante difundido, como um dos melhores algoritmos de classificação de texto e mais

simples,porém, é um algoritmo preguiçoso. as amostras de treinamento são escaneadas e as distâncias são calculadas sobrecarregando a classificação da amostra de teste. A Máquina de vetores de suporte (SVM), outro classificador não linear. O SVM é um classificador discriminativo eficaz com várias propriedades salientes (SOLERA-UREñA et al., 2007), Para banco de dados maior e maior conjunto de recursos, o treinamento SVM torna-se custoso (DONG et al., 2010). Já o classificador Naïve Bayes, usado no reconhecimento de fala multiclasse, é baseado na teoria Bayesiana de classificação de probabilidade. Bayes é simples de implementar, porém o seu funcionamento só é melhor em pequenos conjuntos de dados (SANCHIS; JUAN; VIDAL, 2012) (BHAKRE; BANG,
2016).

É fundamental para um modelo de ASR ter um classificar eficiente, pois é ele o responsável por identificar os padrões de fala e atribuir as transcrições corretas às entradas de fala. Um bom classificador pode melhorar significativamente a precisão do modelo e reduzir o número de erros.

3 Abordagem Metodológica

Para GERHARDT; SILVEIRA, "metodologia é o estudo da organização, dos caminhos a serem percorridos, para se realizar uma pesquisa ou um estudo, ou para se fazer ciência". Esta pesquisa fundamentou-se na abordagem de natureza qualitativa, ou seja, "não se preocupa com representatividade numérica, mas, sim, com o aprofundamento da compreensão de um grupo social, de uma organização, etc." (GERHARDT; SILVEIRA, 2009), p.31).

Quanto aos objetivos, esta pesquisa pode ser considerada como exploratória. "Este tipo de pesquisa tem como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses" (GERHARDT; SILVEIRA, 2009).

Para o desenvolvimento desse projeto podemos utilizar o Deepspeech, que é uma biblioteca Python de código aberto que nos permite construir sistemas automáticos de reconhecimento de fala (HANNUN et al., 2014). Em conjunto com o banco de vozes aberto Mozilla CommonVoice, na sua versão Delta Segment 13.0, contendo 132 vozes e 12 horas validadas, utilizado para treinar e testar inicialmente o modelo proposto. Posteriormente planejou-se a obtenção de um novo conjunto de dados para avaliação do modelo treinado, averiguando-se de forma comparativa sua acurácia resultados de conversão de áudio para textos.

4 Considerações finais

Embora o desafio de desenvolver um modelo de criação de atas não tenha se concretizado totalmente, pudemos obter um importante ponto de partida para tal. Com isso, listamos alguns possíveis problemas a serem considerados para a eficácia do modelo de um sistema automático para confecção de atas por captação de fala.

Primeiro, o problema de decisão de quais dados acústicos serão estimados no processamento. É fundamental encontrar uma representação que reduza a complexidade do modelo, enquanto preserva-se a informação linguística, sem embargo da influência do falante, canal ou características ambientais. Segundo, o problema de modelagem, ou seja, como deve ser

calculado cada evidencia acústica. Muitas vezes, são necessários vários modelos para a caracterização de como os falantes pronuncia cada palavra. De forma ampla, os modelos são altamente dependentes do tipo de aplicação como por exemplo: comandos, fala fluente ou ditado. E por diversas vezes, restringidos para se tornarem computacionalmente viáveis. Terceiro, o problema de modelagem linguística, ou seja, qual melhor ou mais provável calculo que representa uma sequência de palavras. O mais popular modelo se baseia na suposição markoviana de que uma palavra em uma sentença é condicionada apenas pelas palavras N-1 anteriores. Quarto e último, O problema de busca, qual seria a melhor transcrição de palavras para a evidencia acústicas, frente aos modelos acústicos e linguagem. Sabendo que é impraticável pesquisar minunciosamente todas as sequencias de palavras possíveis. Alguns métodos já foram pensados justamente para reduzir os requisitos computacionais

Com o avanço da tecnologia, novas abordagens, como o aprendizado profundo, estão sendo usadas para melhorar ainda mais os sistemas de fala. Essas abordagens permitem que os sistemas aprendam diretamente com os dados, sem a necessidade de recursos de entrada especializados ou ajustes manuais por especialistas no assunto. Isso pode levar a sistemas de fala mais precisos e eficientes.

Referências

BHAKRE, S.; BANG, A. Emotion recognition on the basis of audio signal using naive bayes classifier. In: . [S.l.: s.n.], 2016. p. 2363–2367. Citado na página 5.

CORO, G. et al. Psycho-acoustics inspired automatic speech recognition. *Computers Electrical Engineering*, v. 93, 06 2021. Citado na página 3.

DONG, F. et al. Speech emotion recognition based on multi-output gmm and svm. In: . [S.l.: s.n.], 2010. p. 1 – 4. Citado na página 5.

GERHARDT, T.; SILVEIRA, D. *Métodos de pesquisa*. [S.l.: s.n.], 2009. 120 p. ISBN 9788538600718. Citado na página 5.

HANNUN, A. et al. *Deep Speech: Scaling up end-to-end speech recognition*. 2014. Citado 2 vezes nas páginas 4 e 5.

JOLAD, B.; KHANAI, R. An art of speech recognition: A review. In: . [S.l.: s.n.], 2019. p. 31–35. Citado 2 vezes nas páginas 3 e 4.

KUMAR, T.; KUMAR, T.; RAJAN, K. Speech recognition using neural networks. In: . [S.l.: s.n.], 2009. p. 248 – 252. Citado na página 4.

MARTINEZ, A. R. Natural language processing. *WIREs Computational Statistics*, v. 2, n. 3, p. 352–357, 2010. Disponível em: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.76>. Citado na página 3.

MEYER, J.; DENTEL, L.; MEUNIER, F. Speech recognition in natural background noise.

PloS one, Public Library of Science, United States, v. 8, n. 11, p. e79279, 2013. ISSN 1932-6203. Citado na página 3.

SAKSAMUDRE, S.; SHRISHRIMAL, P.; DESHMUKH, R. A review on different approaches for speech recognition system. *International Journal of Computer Applications*, v. 115, p. 23–28, 04 2015. Citado na página 4.

SANCHIS, A.; JUAN, A.; VIDAL, E. A word-based naive bayes classifier for confidence estimation in speech recognition. *IEEE Transactions on Audio, Speech Language Processing*, v. 20, p. 565–574, 02 2012. Citado na página 5.

SOLERA-UREñA, R. et al. Svms for automatic speech recognition: A survey. In: . [S.l.: s.n.], 2007. p. 190–216. ISBN 978-3-540-71503-0. Citado na página 5.

VIJAYAN, A. et al. Throat microphone speech recognition using mfcc. In: . [S.l.: s.n.], 2017. p. 392–395. Citado na página 4.

ZEROUAL, I.; LAKHOUAJA, A. Data science in light of natural language processing: An overview. *Procedia Computer Science*, v. 127, p. 82–91, 01 2018. Citado na página 3.