Processamento e Transcrição de Voz em Língua Portuguesa voltado para Assistente Inteligente*

Acácio Torres de Andrade[1], **Shayenne Moura**[1], **Alfredo Goldman**[1]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)

{acaciotda,shayenne.moura}@usp.br, gold@ime.usp.br

Abstract. *Inserted in intelligent assistants context and part of the Advanced Distributed Assistant (ADA) project, this work proposes the adaptation of a speech recognition system that allows user interaction through voice commands, transcribing their commands, given in Portuguese, to text format, based on open-source software and available systems. A HMM-based system architecture is adopted. The preliminary system registered a 44.3% word error rate. Posteriorly, the system's acoustic model development will include a hyperparameter optimization stage and experiments with more complex approaches, an assistant specific language model will be introduced.*

Resumo. *Inserido no contexto de assistentes inteligentes, e parte do projeto Assistente Distribuída Avançada (ADA), este trabalho propõe a adaptação de um sistema de reconhecimento de fala que permita a interação do usuário por meio de comandos de voz, transcrevendo seus comandos para texto em língua portuguesa. É adotada uma arquitetura de sistema baseada em modelos ocultos de Markov. O sistema preliminar registrou taxa de palavras erradas de 44.3%. Posteriormente, este sistema terá uma etapa de otimização de hiper-parâmetros do modelo acústico e experimentos com abordagens mais complexas, assim como a introdução de um modelo de linguagem específico para assistentes.*

1. Introdução

O projeto ADA [1] (Assistente Distribuída Avançada) propõe a criação de uma assistente pessoal distribuída inteligente, isto é, uma agente virtual capaz de interagir com o usuário a partir de um ecossistema de dispositivos, como os IoT (Internet das Coisas), por meio de comandos de voz em português. Esses comandos serão transcritos para programas cujas operações básicas serão disponibilizadas pelos aparelhos administrados pela assistente. Neste contexto, entre os objetivos deste projeto de iniciação científica está o desenvolvimento de um módulo de processamento de sinal do áudio capturado e de transcrição dos comandos de voz em linguagem natural. Adotado a priori como motivador, o componente de Reconhecimento de Fala Automático da plataforma Snips [Coucke et al. 2018] norteou o desenvolvimento do sistema, [Gales et al. 2008] fornecendo a base teórica.

A tarefa de reconhecimento de fala pode ser vista como uma questão de otimização estatística, trata-se de encontrar a mais provável sequência de palavras que tenha sido responsável por gerar um sinal de áudio observado. Sistemas de reconhecimento de fala

*Este projeto foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pelo grupo de extensão CodeLab uclab.xyz/site

[1]www.uclab.xyz/ada

baseados em modelos ocultos de Markov – Figura 1 – são classicamente adotados para estimar essas sequências. Sua performance é avaliada, principalmente, em função da taxa de palavras erradas (WER), computada usando a distância de edição entre as strings esperadas e as decodificadas.

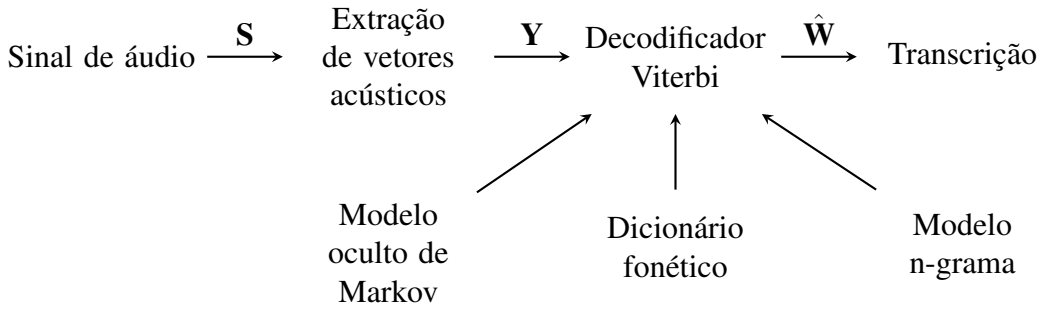

Figura 1. Arquitetura de um sistema de reconhecimento de fala baseado em Modelos ocultos de Markov.

O projeto Kaldi [Povey et al. 2011] provê o ferramental necessário para o desenvolvimento integral de um sistema de reconhecimento de fala. Os softwares Sequitur G2P [Bisani and Ney 2008] e SRILM [Stolcke 2002] fornecem componentes miscelâneos não disponíveis pelo projeto, como um conversor de grafemas para fonemas, e um construtor de modelos de linguagem, respectivamente.

2. Requisitos

Modelos acústicos, como o empregado pelo modelo oculto de Markov, têm seus parâmetros estimados através de um conjunto de dados de treinamento constituído de clipes de áudio e transcrição ortográfica da fala contida em cada clipe. Em se tratando de língua portuguesa, grande dificuldade no desenvolvimento de tais modelos reside na inexistência de datasets públicos apropriados.

A medida adotada para contornar esse cenário foi a construção de um dataset a partir de dados extraídos de vídeos compartilhados na plataforma YouTube. Vídeos de conferências TEDx foram utilizados como referência porque além de contarem com legenda disponível e revisada, tratam-se de vídeos em que a fala é dada de maneira natural, em oposição, por exemplo, à fala oriunda da leitura. Essa característica é importante pois simula a forma como a fala será recebida pelo sistema em um contexto real. Para fins de experimentos iniciais foram extraídos aproximadamente 29 mil clipes de áudio, armazenados em arquivos WAV com 16 kHz/16 bits, em um único canal. As legendas, assim como informações a respeito dos palestrantes, foram armazenadas em arquivos de texto. No total, 27 horas de áudio contendo a fala de 146 locutores diferentes, sendo 77 do sexo feminino e 69 do sexo masculino, compuseram o corpus de áudio.

A unidade básica do som com que modelos acústicos trabalham são os fonemas, esses representam qualquer som elementar da linguagem articulada. A palavra “deslumbre", por exemplo, tem transcrição fonética dada por “d ʒ i z l ũ b ɾ i". O mapeamento de uma sequência de palavras para uma sequência de fonemas e vice-versa é feito com auxílio de um dicionário composto por todo vocabulário de palavras que deseja-se que o sistema conheça, assim como suas respectivas transcrições fonéticas. O dicionário fonético

LUPo [Ashby et al. 2012], desenvolvido em Lisboa pelo então Instituto de Linguística Teórica e Computacional foi incorporado e utilizado para realização de tal mapeamento.

Foram acrescentadas ao dicionário todas as palavras presentes no corpus de texto proveniente das transcrições dos áudios contidos em um conjunto pré-determinado para o desenvolvimento dos modelos acústicos. A incorporação foi feita usando um conversor Sequitur G2P treinado usando dados do dicionário LUPo. Sua capacidade de generalização foi avaliada por meio de um esquema de validação cruzada com cinco *folds*, a taxa de fonemas errados média foi de 1.6%. O dicionário fonético final possui o mapeamento de 63357 palavras.

Outro componente prévio necessário para o desenvolvimento do sistema, o modelo n-grama de linguagem utilizado é disponibilizado pelo FalaBrasil [Batista et al. 2018], grupo de pesquisa criado pelo Laboratório de Processamento de Sinais (LaPS) da Universidade Federal do Pará (UFPA). O modelo é um n-grama de ordem 3 treinado em um corpus de 1.6 milhões de frases extraídas do jornal Folha de São Paulo, seu valor de perplexidade é 170.

3. Modelo Acústico

O modelo acústico é fruto do refinamento sequencial de quatro modelos intermediários, cada um dependente do anterior, porém, em geral, resultado de abordagens mais sofisticadas. Todos baseados em modelos de misturas de Gaussianas [Povey et al. 2011], os modelos diferem entre si em relação a representação de cada estado, a composição dos vetores acústico utilizados como atributos para distinguir as classes, ao número máximo de estados da cadeia, o número máximo de Gaussianas utilizadas para construção dos modelos de misturas que definem a distribuição de probabilidade em cada estado, e, por fim, o volume de dados usado em seu treinamento. A Tabela 1 ilustra essas diferenças e apresenta a performance do sistema quando cada um dos modelos é integrado.

Tabela 1. Performance do sistema avaliada através de experimentos.

Modelo	Estados	Gaussianas	Clipes	WER(%)
Monofone (MFCCs)	145	989	2000	78.4
Trifone (MFCCs + Δ + $\Delta\Delta$)	1464	10025	5000	60.7
Trifone (LDA-MLLT)	1960	15031	10000	54.9
Trifone (LDA-MLLT + fMLLR)	1976	15018	10000	47.0
Trifone (LDA-MLLT + fMLLR)	3408	40044	28794	44.3

4. Resultados

Para avaliação do sistema, foi selecionado um conjunto de teste, disjunto do de treinamento também em termos de falantes presentes, contendo 1175 clipes, estratégia adotada devido a limitações computacionais para realização de uma avaliação mais robusta, como Cross-validation. A taxa de palavras erradas entre as sentenças desse conjunto decodificadas pelo sistema foi de 44.3%, a taxa de sentenças erradas, 87.2%.

Para um certo clipe presente no conjunto de teste, foi decodificada a sentença "cada um de nós hoje é que está cortando vários objetos". Sendo "cada um de nós hoje aqui está portando vários objetos" a sentença esperada, esse exemplo ajuda a ilustrar

dois possíveis tipos de erro: a palavra “portando” foi decodificada como “cortando”, isso ocorre porque a palavra “portando” não está presente no dicionário fonético, portanto, esse é um erro esperado. Já a palavra “aqui” foi decodificada como “é que”, verificando que a palavra “aqui” está presente no dicionário fonético, conclui-se que esse erro é uma falha do modelo acústico. O alto grau de complexidade do dicionário fonético e do modelo de linguagem na etapa de desenvolvimento do modelo acústico, permite se chegar a esses tipos de conclusões.

5. Próximos Passos

Para a conclusão do desenvolvimento do modelo acústico, deve ser incluída uma etapa de otimização de hiper-parâmetros em cada estágio de treinamento. Além disso, mais um estágio pode ser acrescentado, incorporando ao modelo acústico redes neurais profundas para determinar a distribuição de probabilidade associada aos vetores acústicos. Idealmente, mais dados serão obtidos.

Levando em conta que as transcrições devem ser boas o suficiente para que se possam extrair as intenções do locutor, um corpus de texto contendo diversos exemplos de comandos pode ser usado para o desenvolvimento de um modelo de linguagem específico para essa tarefa. Espera-se que isso reduza o espaço de busca e aumente a velocidade de processamento, não ponderada até então.

Será incluído no conjunto de teste o corpus de áudio “LaPS Benchmark” usado em [Batista et al. 2018], possibilitando uma comparação entre os dois sistemas. Além disso, experimentos integrados ao projeto A.D.A. serão conduzidos pelos membros do projeto.

Referências

Ashby, S., Barbosa, S., Brandão, S., Ferreira, J. P., Janssen, M., Silva, C., and Viaro, M. E. (2012). A rule based pronunciation generator and regional accent databank for portuguese. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Batista, C. T., Dias, A. L., and Neto, N. C. S. (2018). Baseline acoustic models for brazilian portuguese using kaldi tools. In *IberSPEECH*, pages 77–81.

Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al. (2018). Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Gales, M., Young, S., et al. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.