Sinais e Sistemas (ES413) – Cin/UFPE

Projeto: Transformada de Fourier em Reconhecimento de Fala

Os humanos se comunicam preferencialmente por meio da fala, usando o mesmo idioma. O reconhecimento de fala pode ser entendido como a capacidade de compreender as palavras e as sentenças. O reconhecimento automático de fala (ASR) (Malik et al., 2021) diz respeito ao reconhecimento da fala humana e sua transcrição em texto pelo computador. Este campo de pesquisa ganhou muito destaque nas últimas décadas. Os primeiros métodos consistiam na extração manual de características para representar os sinais de áudio e posterior aplicação de técnicas convencionais para reconhecimento, como Modelos de Mistura Gaussiana (GMM), o algoritmo *Dynamic Time Warping* (DTW) e Modelos Escondidos de Markov (HMM). Há alguns anos, há uma tendência de utilização de redes neurais (MLP, RNNs, CNNs, *transformers*) para realizar o reconhecimento.

A Transformada de Fourier (FT) é uma ferramenta útil em sistemas ASR. Ela serve para analisar os componentes de frequência de um sinal de fala e extrair características relevantes que podem ser usados para reconhecer palavras faladas. Em suma, a FT é empregada para decompor um sinal de fala em suas frequências constituintes, um conjunto de características empregado no reconhecimento de fala.

Um sistema ASR que empregue FT é tipicamente formulado com as seguintes etapas:

- Pré-processamento de sinal de áudio para remover ruído e melhorar a relação sinal-ruído;
- Transformação do sinal pré-processado para o domínio da frequência usando a FT;
- Extração de características (componentes de frequência) de um sinal;
- Reconhecimento das palavras empregando algoritmos de reconhecimento de padrões.

A FT apresenta fatores importantes para seu uso com sistemas ASR:

- Precisão na análise dos componentes de frequência do sinal de fala;
- Robustez ao ruído, perturbação frequente na aquisição de sinais de áudio;
- Aplicabilidade para sistemas de tempo real por ser velozmente calculada com algoritmos rápidos, como a Transformada Rápida de Fourier (FFT) (Duhamel & Vetterli, 1990; Liu et al., 2019; Rajaby & Sayedi, 2022; Shchekotov et al., 2022; Sorensen et al., 1987).

Neste projeto há interesse em usar a FT como parte obrigatória de um extrator de características (Becoulet & Verguet, 2021), utilizar extratores de características adicionais para representação dos áudios (Borandağ, 2019; Polur et al., 2005), empregar modelos convencionais ou baseados em aprendizagem de máquinas para fazer o reconhecimento do áudio (Abdusalomov et al., 2022; Borandağ, 2019; Polur et al., 2005; Li et al., 2020) e usar bases de dados públicas para testar os sistemas propostos (Fendji et al., 2022).

Objetivo:

Construir um modelo para reconhecimento de fala atendendo às especificações a seguir:

1. Os sinais de áudio serão coletados (recomendação forte) de bases de dados públicas;
2. Cada grupo deve utilizar ao menos um método de pré-processamento;
3. Cada grupo deve selecionar ao menos um outro outro extrator de características para atuar em conjunto com aquele baseado em FT;
4. Cada grupo deve definir uma opção para algoritmo de reconhecimento de padrões.

Equipes:

1. Cada equipe terá até 5 estudantes sendo um deles o líder.
    a. Será permitida uma equipe com 3 ou 4 alunos ou até 2 equipes com 6 alunos no caso do número total de alunos que fará o trabalho não ser múltiplo de 5.
2. Cada aluno selecionará sua equipe que deve ser informado no *classroom* da disciplina.
3. O(s) monitor(es) ajudará(ão) as equipes no seu projeto.

Avaliação:

1. O projeto valerá 50% da segunda nota, os outros 50% será da segunda prova.
2. Os projetos estarão competindo entre si. Eles serão avaliados nos quesitos de formalização, criatividade e funcionalidades.
3. A participação de cada aluno é opcional. Porém, se o aluno decidir participar do projeto ele receberá a nota por sua participação. Em caso de desistência receberá nota 0,0 (zero).
4. Os alunos terão suas notas de forma decrescente, sendo a(s) equipe(s) ganhadora(s) com pontuação máxima. É possível que haja empates de notas.
5. Cada equipe será acompanhada de perto pelo(s) monitore(s) e componentes que não estejam cooperando serão dispensados da equipe e receberão nota zero.
6. Adicionar no escopo do projeto funcionalidades coerentes garante a equipe boa pontuação.
7. Atrasos nas entregas do projeto gera perda da nota (10% ao dia). A etapa final não pode ser entregue com atraso.
8. Não pode haver dois grupos com a mesma combinação de componentes nos seus modelos.
9. A entrega do relatório do projeto se dará em 26/07/2024, quando ocorrerá a apresentação dos projetos.
10. Cada aluno será avaliado por sua apresentação (individual) e pelo conteúdo oral e escrito do projeto (em grupo).
11. O tempo de cada aluno na apresentação deve ser o mais equânime possível em cada grupo. Variações significativas no tempo (para cima ou para baixo) prejudicam a nota.

12. Haverá entregas intermediárias obrigatórias de cada projeto.

Datas importantes:

28/06/2024: Data limite para definição dos grupos. Cada grupo deve ser postado no mural da disciplina e seu líder indicado até 12 horas (meio-dia).

03/07/2024: Data limite para postagem nas atividades do *classroom* da escolha do algoritmo reconhecedor de padrões (deve ter uma justificativa baseada em trabalhos anteriores). O documento deve ter até 2 (duas) páginas.

05/07/2024: Data limite para postagem nas atividades do *classroom* da escolha dos algoritmos para extração de características (deve ter uma justificativa baseada em referências) e dos dados a serem empregados para testes, dados adquiridos pelo grupo ou conjunto de dados de domínio público (neste caso deve informar o repositório dela). É preciso informar o que irá usar para pré-processamento também. Vale 1 (um) ponto.

26/07/2024: Apresentação do projeto.

Contato da monitora:

Lucas Inoja Mosendes da Silva <lims@cin.ufpe.br>
Pedro Nascimento Coelho <pnc2@cin.ufpe.br>

Referências:

Abdusalomov, A. B., Safarov, F., Rakhimov, M., Turaev, B., & Whangbo, T. K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. *Sensors*, 22(21).

Becoulet, A., & Verguet, A. (2021). A depth-first iterative algorithm for the conjugate pair fast Fourier transform. *IEEE Transactions on Signal Processing*, 69: 1537-1547.

Borandağ, E. (2019). Markov model based real time speaker recognition using k-means, fast fourier transform and mel frequency cepstral coefficients. *Celal Bayar University Journal of Science*, 15(3): 287-292.

Duhamel, P., & Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art. *Signal Processing*, 19(4): 259-299.

Fendji, J. L. K. E., Tala, D. C., Yenke, B. O., & Atemkeng, M. (2022). Automatic speech recognition using limited vocabulary: A survey. *Applied Artificial Intelligence*, 36(1).

Li, S., Xue, K., Zhu, B., Ding, C., Gao, X., Wei, D., & Wan, T. (2020). Falcon: A Fourier transform based approach for fast and secure convolutional neural network predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8705-8714.

Liu, W., Liao, Q., Qiao, F., Xia, W., Wang, C., & Lombardi, F. (2019). Approximate designs for fast Fourier transform (FFT) with application to speech recognition. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 66(12): 4727-4739.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80: 9411-9457.

Polur, P. D., & Miller, G. E. (2005). Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(4): 558-561.

Rajaby, E., & Sayedi, S. M. (2022). A structured review of sparse fast Fourier transform algorithms. *Digital Signal Processing*, 123.

Shchekotov, I., Andreev, P., Ivanov, O., Alanov, A., & Vetrov, D. (2022). Ffc-se: Fast Fourier convolution for speech enhancement. *arXiv preprint arXiv*:2204.03042.

Sorensen, H. V., Jones, D., Heideman, M., & Burrus, C. (1987). Real-valued fast Fourier transform algorithms. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(6): 849-863.

Links interessantes:
https://www.researchgate.net/publication/267450621_AUTOMATIC_SPEECH_RECOGNITION_USING_FOURIER_TRANSFORM_AND_NEURAL_NETWORK
https://www.ijert.org/research/speech-recognition-using-fast-fourier-transform-algorithm-IJERTCONV10IS08007.pdf
https://developer.nvidia.com/blog/essential-guide-to-automatic-speech-recognition-technology/
https://www.scaler.com/topics/nlp/architecture-of-automatic-speech-recognition/
https://realpython.com/python-scipy-fft/

Vídeos interessantes:
https://www.youtube.com/watch?v=BXghmsH-mKY&list=PL4K9r9dYCOoqmykdiyCq2jyAb0zwO0p-b
https://www.youtube.com/watch?v=DpchUWUsYs0&list=PLASpGWv0ToUHyTZ6FL0K2rhTkVaV9GPND
https://www.youtube.com/watch?v=E8HeD-MUrjY
https://www.youtube.com/watch?v=iCwMQJnKk2c&list=PL-wATfeyAMNqIee7cH3q1bh4QJFAaeNv0
https://www.youtube.com/watch?v=gMQyGASOZO0&list=PLuWx2S0SyaDd_eMm68ep0XEDUphnw0HcI
https://www.youtube.com/watch?v=q67z7PTGRi8&list=PLpCZr5mhfo86H0eRtTGuDSs-FYsHcTzk9