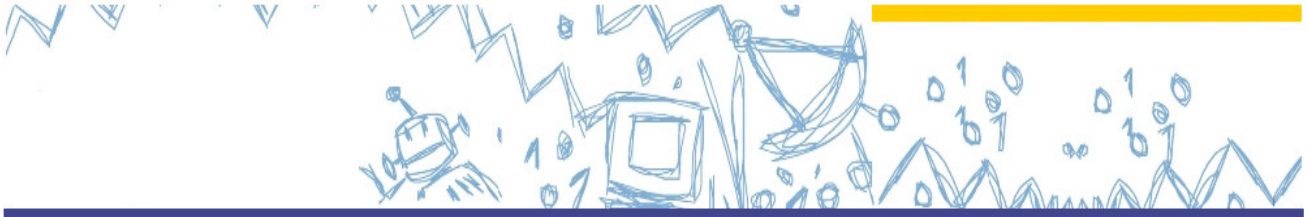

Projeto 2024

Identificação de dígitos através de características extraídas de sinais de áudio

Enquadramento: *Na atualidade, o reconhecimento e interpretação de voz humana está disponível em praticamente todos os dispositivos. Por exemplo, é comum os telemóveis/computadores realizarem tarefas ativadas por comandos de voz. A fase inicial de qualquer tarefa de reconhecimento passa pela aquisição e extração de características do sinal áudio. O objetivo principal é que estas características permitam discriminar diferentes sons, correspondentes a diferentes comandos de voz.*

Objetivo: *Este projeto visa a análise nos domínios da frequência e do tempo de sinais áudio com o objetivo de identificar os dígitos em inglês entre 0 e 9.*

Introdução[1]: A Figura 1 mostra uma ilustração do sistema de produção da fala humana. Os principais componentes anatómicos deste sistema são os pulmões, traqueia, laringe (órgão de produção da fala), cavidade faríngea (garganta), cavidade bucal e cavidade nasal. Existem também outros componentes anatómicos que contribuem para a produção da fala, como as pregas vocais (ou cordas), língua, lábios, dentes e mandíbula. Estes últimos são referidos como articuladores e movem-se para diferentes posições a fim de produzir vários sons da fala como vogais, consoantes ou outras unidades fonológicas de uma língua.

O processo de produção da fala humana pode ser resumido da seguinte forma. Durante a fala, o ar expelido dos pulmões segue para a traqueia e sobe para a glote, onde o fluxo é periodicamente interrompido pelo movimento das cordas vocais. A tensão das cordas vocais é ajustada pela laringe causando as vibrações das cordas. Dentro do trato vocal, também são produzidas constrições à passagem do ar através da língua, lábios e nariz. Estas mudanças nas configurações do sistema de fala humana, irão fazer com que as vibrações emitidas através dos lábios para o ar circundante (i.e., o som) possuam componentes de frequência diferentes.

A produção da fala pode ser vista como uma operação de filtragem, na qual as três cavidades principais do sistema de produção da fala constituem o filtro acústico principal.

[1] https://link.springer.com/referenceworkentry/10.1007/978-0-387-73003-5_199

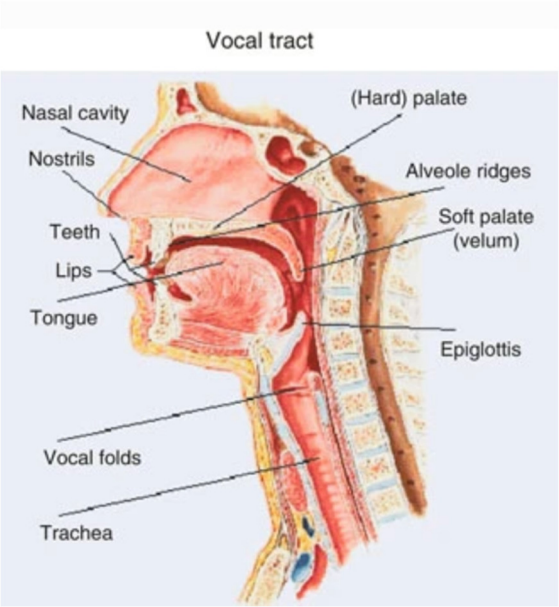

Fig 1: Componentes principais do sistema de produção de fala humana.
(Ref: https://link.springer.com/referenceworkentry/10.1007/978-0-387-73003-5_199)

Linguagem de Programação: **MATLAB (alternativamente Python).**

Organização: **Grupos de dois alunos (sempre que possível)**

Dados: https://www.kaggle.com/datasets/sripaadsrinivasan/audio-mnist/download?datasetVersionNumber=1
(Deve descompactar o ficheiro e considerar somente os sinais em bruto disponíveis na pasta *data*).

Descrição dos dados: Os dados fornecidos correspondem a sinais de voz emitidos por 60 participantes (cada uma das subdiretorias na pasta *data* contém os sinais correspondentes a 1 participante). Cada participante repetiu 50 vezes cada um dos dígitos, ou seja, cada uma das 60 pastas contém 500 sinais de áudio em formato *.wav*. Cada sinal de áudio foi adquirido a uma taxa de amostragem de 48000 Hz em modo mono-canal. Mais detalhes sobre os dados podem ser obtidos em:
https://github.com/soerenab/AudioMNIST
https://www.kaggle.com/datasets/sripaadsrinivasan/audio-mnist
https://arxiv.org/abs/1807.03418.

Método de avaliação:

O trabalho foi dividido em quatro metas, cujas datas estipuladas para conclusão de cada uma delas são:
Meta 1: 03/03/2024, 23:59
Meta 2: 08/04/2024, 23:59
Meta 3: 05/05/2024, 23:59
Meta 4: 19/05/2024, 23:59

Até à data estipulada para conclusão de cada meta, o grupo deverá submeter o trabalho correspondente no inforestudante.

Se o grupo não conseguir cumprir com a data de submissão estipulada, poderá incluir o trabalho na submissão da meta seguinte, com as seguintes penalizações nas notas das metas correspondentes:
Atraso de 1 *deadline*: 25 %
Atraso de 2 *deadlines*: 50 %
Atraso de 3 *deadlines*: 75 %

Materiais a entregar: É altamente recomendável entregar de forma integrada **código**, **figuras** e **texto** (questões colocadas ao longo dos exercícios). Para isto, poderá utilizar ferramentas como o Júpiter Notebook (para MATLAB ou Python), ou o Live Script do MATLAB.

Defesas:

- Defesa Final: a decorrerem na semana de **20 a 24 de maio**, em horário das PLs ou em horários a combinar com o professor da turma.
- Serão realizadas avaliações nas aulas PLs após a entrega de cada meta.
- A nota da defesa e as presenças nas PLs irão multiplicar pela nota obtida no trabalho implementado.

$$\text{Nota do Projeto} = (30\%\ \text{Relatório} + 70\%\ \text{Código}) \times \text{Defesa}\ (0-100\%) \times \text{Presenças}\ (0-100\%)$$

Registo:

1. Registe o seu grupo em:
https://docs.google.com/spreadsheets/d/10BxMIxyZZ7qSLq_LlJsht0FSEJftXqN6KCkraBVngS4/edit#gid=0

2. Escolha o participante/pasta que o seu grupo irá trabalhar e indique no formulário.

Meta 1: 03/Março/2024

1. Desenvolva código para importação dos sinais de áudio.
   Nota: pode utilizar a função **audioread** do MATLAB.

2. Reproduza e represente graficamente um exemplo dos sinais importados, identificando o dígito a que cada um corresponde, e indicando o eixo horizontal com o tempo em segundos, como apresentado no exemplo da Figura 2.

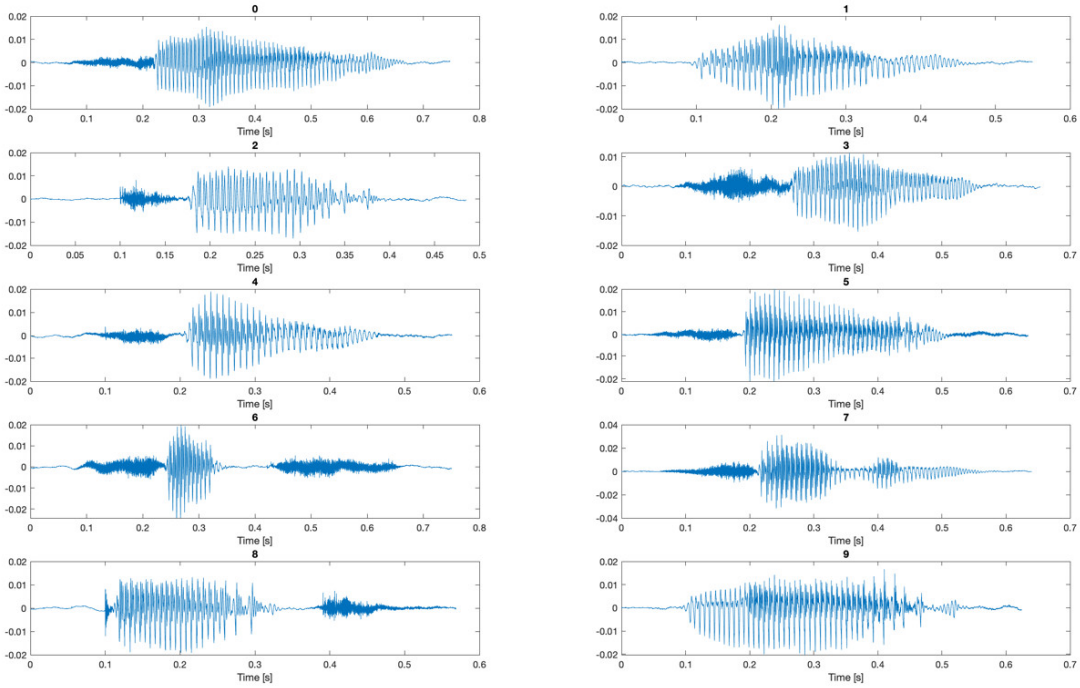

Fig. 2: Sinais de áudio correspondentes aos dígitos de 0 a 9.

3. Identifique visualmente e implemente o cálculo de possíveis características (*features*) temporais do sinal que permitam a diferenciação entre dígitos, como energia (total, por intervalos temporais, diferenças entre intervalos, etc.), amplitude (máxima, mínima, razão de amplitudes, desvio padrão, etc.) ou outras medidas temporais que considere relevantes.

   Nota: Uma estratégia que pode melhorar a diferenciação entre dígitos é fazer uma etapa de pré-processamento para garantir que todos os ficheiros têm a mesma duração e intervalo de amplitude. Para isso pode recorrer às seguintes estratégias:
   - Retirar o "silêncio" inicial dos sinais, de forma a garantir que todos começam exatamente ao mesmo tempo. Isto pode ser implementado através da análise da Energia em janelas de tempo.
   - Normalizar a amplitude com base na amplitude máxima e mínima das amostras. Isto garante que os problemas usuais da gravação do som (por exemplo, a distância da pessoa ao microfone) não interferem na análise.
   - Adicionar (ou retirar) silêncio no fim dos ficheiros para garantir que todos têm a mesma duração total.

4. Recorrendo à representação gráfica das características temporais extraídas no ponto anterior, identifique as três características que permitem uma melhor discriminação dos dígitos. Para isto pode utilizar *boxplots*, gráficos 2D, 3D, etc. Alguns exemplos são apresentados na Figura 3. **Indique as características escolhidas**.

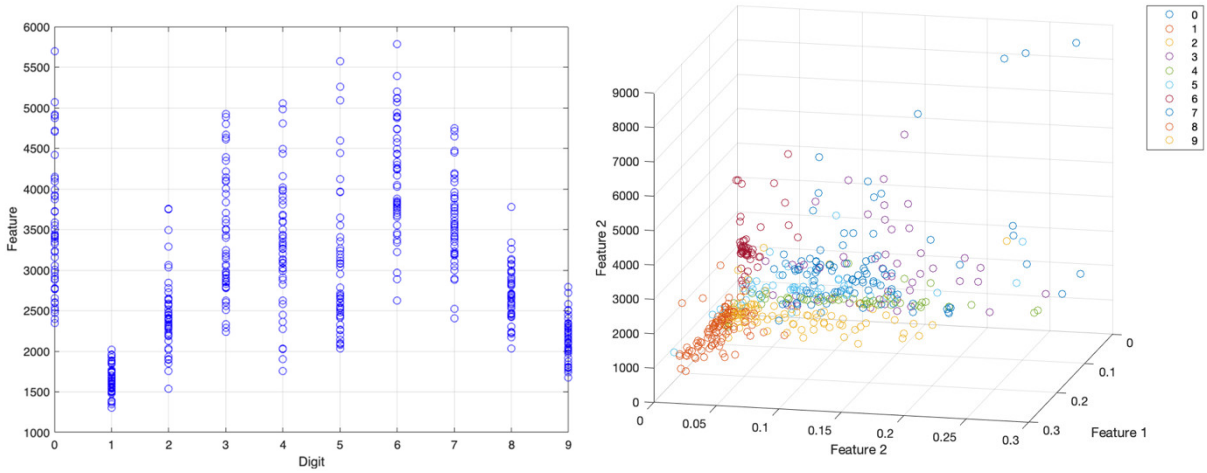

Fig. 3. Representação gráfica das características extraídas.

Meta 2: 8/Abril/2024

5. Calcule para cada dígito, o espectro de amplitude mediano, normalizado pelo número de amostras (i.e., equivalente ao módulo dos coeficientes da série complexa de Fourier), e somente para frequências positivas. Calcule também o primeiro quartil (25%) e terceiro quartil (75%). A Figura 4 mostra um exemplo.
   Nota: pode utilizar a função **quantile** do MATLAB.

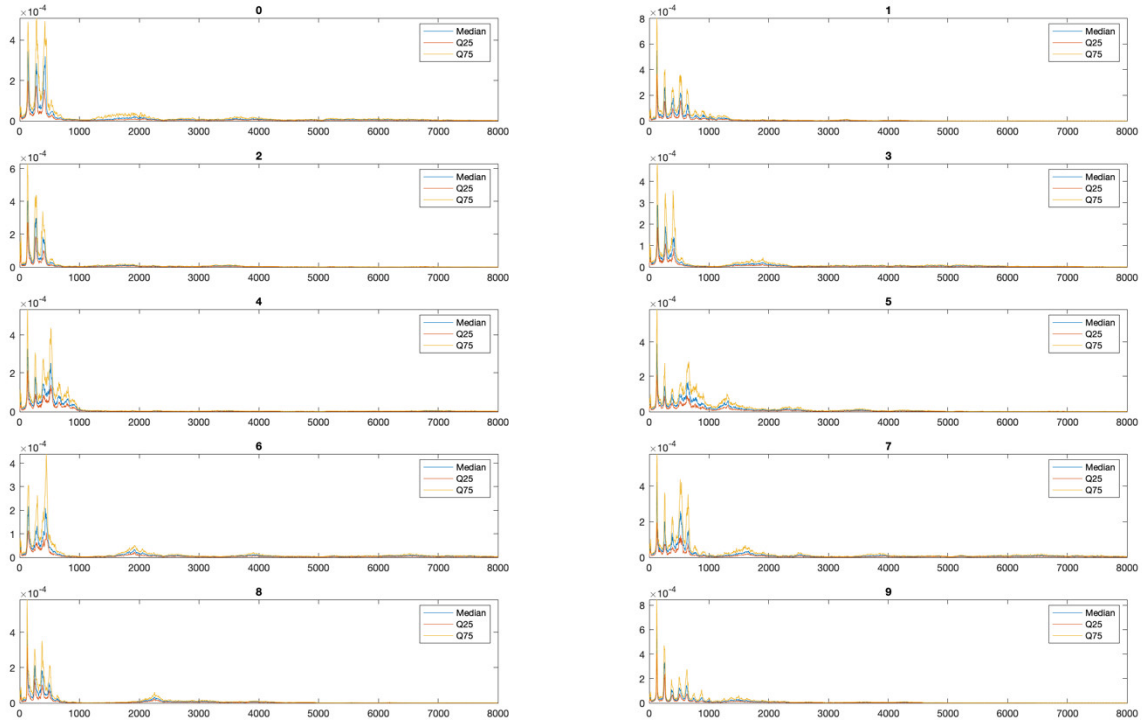

Fig. 4: Espectro de amplitude mediano normalizado, primeiro quartil e terceiro quartil, para os dígitos de 0-9. Nota: para efeitos de uma melhor visualização não é apresentado o eixo de frequências completo.

6. Compare três tipos de janela diferentes. Qual é o efeito das diferentes janelas? **Comente os resultados**.

7. Identifique possíveis características espectrais que permitam a diferenciação entre dígitos como: máximos espectrais (posição e amplitude), médias espectrais, *spectral edge frequency*, etc.

8. Recorrendo à representação gráfica das características espectrais extraídas no ponto anterior, identifique as três características que permitem uma melhor discriminação dos dígitos. Para isto pode utilizar *boxplots*, gráficos 2D, 3D, etc. **Indique as características escolhidas**.

Meta 3: 5/Maio/2024

9. Calcule, para cada dígito, a STFT, como exemplificado na Figura 5. Use diferentes parametrizações, por exemplo tamanho da janela, sobreposição, número de pontos para cálculo da FFT, etc., e identifique os parâmetros que lhe parecem mais adequados para o objetivo do trabalho. **Comente as observações**.
   Nota: utilize a função **spectrogram** do MATLAB (consulte a documentação associada).

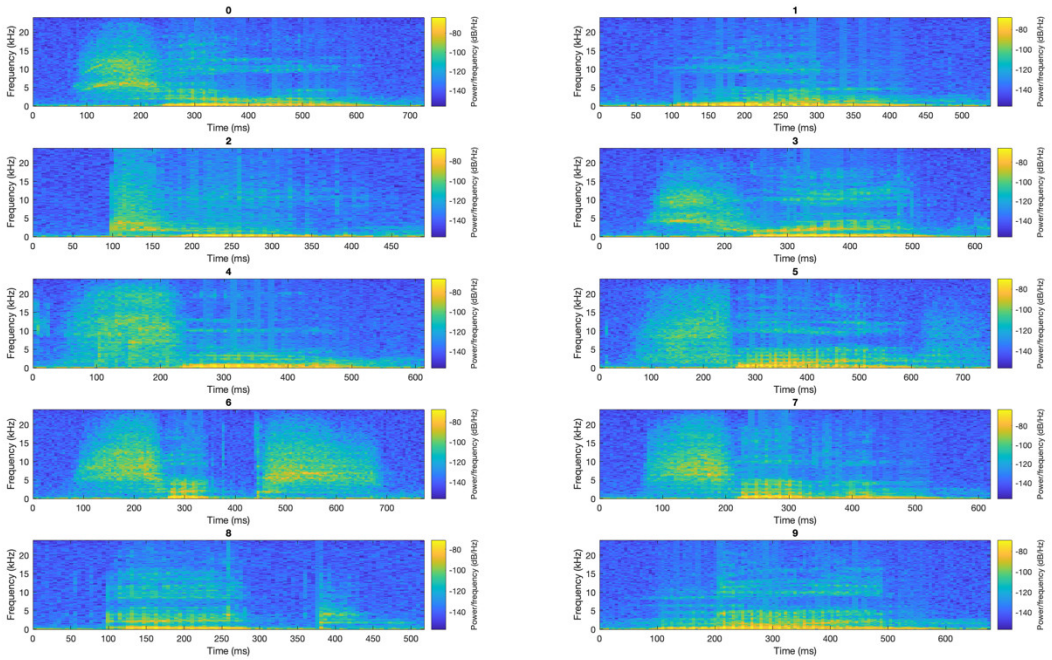

Fig. 5: STFT para os dígitos de 0-9.

10. Extraia características de potência em diferentes janelas tempo-frequência que achar mais relevante para a diferenciação dos dígitos.

11. Recorrendo à representação gráfica das características tempo-frequência extraídas no ponto anterior, identifique as três características que permitem uma melhor discriminação dos dígitos. Para isto pode utilizar *boxplots*, gráficos 2D, 3D, etc. **Indique as características escolhidas**.

Meta 4: 19/Maio/2024

12. Considerando as melhores características extraídas dos sinais no tempo, frequência e tempo-frequência nos pontos 4, 8 e 11, respetivamente, defina possíveis regras de decisão (do tipo *if ... else*) que permitam uma melhor separação dos dígitos. Para a definição do conjunto de regras pode utilizar de 5 a 9 características diferentes.

13. Compare os dígitos atribuídos a cada áudio a partir das regras definidas no ponto anterior, com os dígitos reais e defina a percentagem de acertos como:

$$\text{Acertos} = \frac{\text{Dígitos identificados corretamente}}{\text{Total de dígitos}} x\ 100$$

14. Do conjunto total de características, escolha as três que apresentaram melhor resultado na diferenciação de dígitos e realize um gráfico 3D dos resultados obtidos, como o apresentado na Figura 3. **Comente as observações**.