

Análise deslizante com recolha de características do sinal e classificação básica

*2 Trabalho Laboratorial de Audio Computacional, DEEC-FEUP

André de Azevedo Barata

Dep. de Engenharia Eletrotécnica e de Computadores
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
up201907705@edu.fe.up.pt

André Nogueira Soares

Dep. de Engenharia Eletrotécnica e de Computadores
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
up201905318@edu.fe.up.pt

Abstract—Este documento aborda a exploração a utilização do software Matlab e PRAAT para a análise de sinais de áudio de forma variante no tempo. O estudo consiste na criação de programas básicos cujo foco principal é a análise de características temporais como energia, taxa de passagens por zero e frequência fundamental (f_0), assim como a classificação dos segmentos de áudio em categorias como sinal, silêncio, vozeado, não vozeado e misto.

Index Terms—Processamento de Sinal, Análise Temporal, Análise Espetral.

I. INTRODUCTION

Ao longo deste documento, iremos descrever as experiências realizadas de modo a podermos caracterizar um sinal áudio com cerca de 20,8 segundos. Deste modo, iremos utilizar o Software Matlab para traçar o espectrograma do sinal, calcular frequência fundamental da voz presente no áudio e calcular a energia do sinal. Posteriormente, iremos corroborar os resultados obtidos utilizando o software PRAAT. Por fim, iremos mostrar os resultados do nosso algoritmo de classificação de silêncio, fala vozeada, fala não vozeada e fala mista e comparar com um algoritmo dado pelo professor.

II. AUDIÇÃO, TRANSCRIÇÃO E ANÁLISE INICIAL

A. Audição e Transcrição de Conteúdo de Áudio

O áudio em questão está presente na pasta submetida como nome "ASRF24.wav". O áudio tem a duração de 20,85 segundos que quando lido com uma frequência de amostragem de 44100 Hz resulta em 919414 amostras. Após a leitura do áudio foi realizado um *downsample* para 16000 Hz, isto é, a frequência de amostragem é agora de 16 kHz. Na Figura 1 podemos observar o sinal original a azul e o sinal após o *downsample* a verde, verificando são visualmente idênticos (sem ampliação), contudo são audivelmente distintos, o que revela que não há perda significativa de informação com a operação de reamostragem. Pois, com uma frequência de amostragem de 16 kHz apenas é possível reconstruir sinais com uma frequência até 8 kHz o que exclui grande parte do

range de sons audíveis (entre 20 e 20 000 kHz), no entanto, os sons musicais estão compreendidos entre 20 e 8 kHz o que está compreendido nos limites impostos pela nova frequência de amostragem. Com um *zoom*, é possível observar que os picos se suavizam, tornando-se mais ondulados e retos.

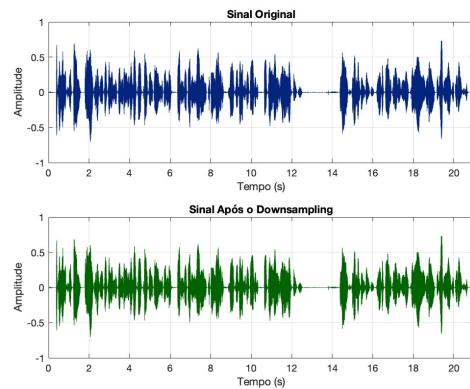


Fig. 1: Comparação entre sinal antes e após o *downsample*

No áudio em analise uma voz masculina recita o seguinte texto: "O problema que se lhe põe a ele que sempre lutou pela justiça e se revoltou contra a injustiça é se esses homens estrangeiros serão aqui bem aproveitados ou pelo menos minimamente pagos e é aí que começa a sua revolta sem papeis como ele um dia como ele são ignobilmente explorados"

III. DESENVOLVIMENTO E AVALIAÇÃO DE SCRIPTS E ALGORITMOS

Nesta secção, iremos descrever a metodologia utilizada para calcular a energia do sinal, as passagens por zero do sinal e a frequência fundamental da voz.

Para conseguirmos realizar as operações no sinal, dividimos o áudio em janelas. Para tal, foram definidos três parâmetros: Duração da janela (t), "window step" (N_H) e o tipo janela. A duração escolhida foi de 30 ms, um valor comum para a fala.

O "window step" 10 ms, de modo a representar um terço da duração da janela (valor comum em processamento de fala). A janela utilizada foi a janela de Hamming, já que fornece uma ideia mais precisa do espectro de frequência do sinal original, evitando a contaminação de qualquer componente do espectro por componentes mais distantes. A partir destes parâmetros também é possível calcular o tamanho da janela (N_f) e a sobreposição das janelas (θ), equações 1 e 2, respetivamente, nas quais f_s representa a frequência de amostragem.

$$N_f = t \times f_s \quad (1)$$

$$\theta = (t - N_H) \times f_s \quad (2)$$

Como foi realizada um *downsample*, o valor de f_s a ser considerado é 16 kHz. Os valores dos parâmetros utilizados estão representados na tabela III.

t	N_H	N_f	θ	f_s	janela
30ms	10ms	480	320	16kHz	Hamming

Na Figura 2 a) encontra-se uma captura de ecrã do código para importar o áudio e definir os parâmetros e na Figura 2 b) o ciclo que processa a imagem janela a janela. Efetivamente, a definição dos parâmetros é implementada de acordo com as equações 1 e 2 e o dentro do ciclo os a considerar do vetor que contém a informação do áudio são atualizados (ou seja, a janela é redefinida).

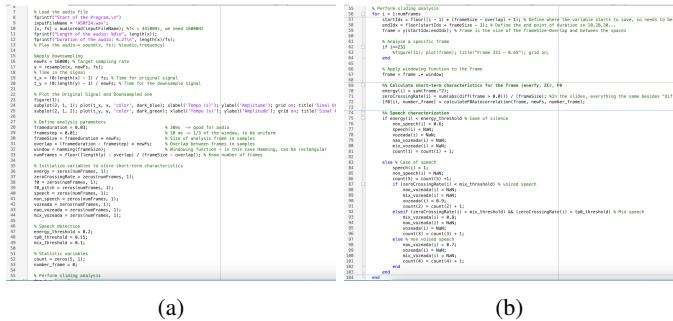


Fig. 2: Segmento do código MATLAB utilizado no processamento

A. Análise de Energia do sinal

A energia de um dada janela do sinal é obtida através da equação 3. Na Figura 2 b) podemos ver a implementação deste equação que para janela, calcula a soma da amplitude de cada amostra (linha 70 do código).

$$E(m) = \sum_{n=m-N+1}^m s^2(n) \quad (3)$$

Por fim, traçamos o gráfico da energia em cada janela (Figura 3). Comparando com o gráfico da Figura 1 verificamos que as zonas com menor amplitude do sinal correspondem às janelas com menor energia, como por exemplo entre os segundo 12 e 14. Para além disso, estas secções de menor

intensidade/energia também estão relacionadas com as zonas de silêncio. Este fenómeno será analisado na secção III-D.

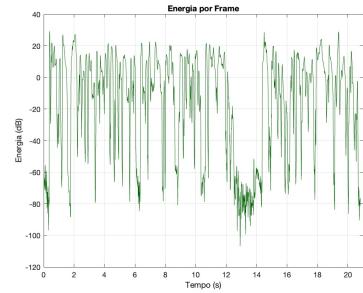


Fig. 3: Energia no Sinal

B. Análise da Taxa de Passagens por Zero do sinal

A taxa de passagens por zeros expressa a quantidade de vezes que um dado sinal passa por zero, por outras palavras quantas vezes passa de positivo por negativo e vice-versa. Tratando-se de um sinal digital, este valor será calculado pelo cociente entre o número de amostras consecutivas que têm sinais diferentes e o número total de amostras. A equação 4 representa a forma mais comum de calcular a taxa de passagem por zero, no entanto, de modo a reduzir a complexidade computacional, adotamos as equações 5a e 5b que consiste em verificar se uma dada amostra é positiva ou negativa e atribuir um valor lógico de 1 ou 0, respetivamente, que depois é realizada a diferença entre estes valores lógicos, resultado em no valor 1 caso exista passagem por zero ou em 0 caso não exista, por fim estes valores são somados, resultando em todas as passagens por zero que por sua vez é dividido pelo tamanho da janela. A implementação das equações 5a e 5b está representada na Figura 2 b na linha 71 do código, de notar que esta linha está incluída dentro de um ciclo que resolve as equações para cada janela, guardando os resultados num vetor.

$$Z(m) = \frac{1}{2N} \sum_{n=m-N+1}^m [sign[s(n)] - [sign[s(n-1)]] \quad (4)$$

$$V(n) = \begin{cases} 1 & \text{if } s(n) > 0 \\ 0 & \text{if } s(n) < 0 \end{cases} \quad (5a)$$

$$Z(m) = \frac{1}{N} \sum_{n=m-N+1}^m V(n) - V(n-1) \quad (5b)$$

A taxa de passagens por zero do sinal e análise está representada na Figura 4 por janela, ou seja, cada valor representa o número de passagens de uma dada janela. O eixo X foi colocado em segundos para que a relação entre passagens por zero, energia e momentos de silêncio que será abordada na secção III-D seja mais evidente.

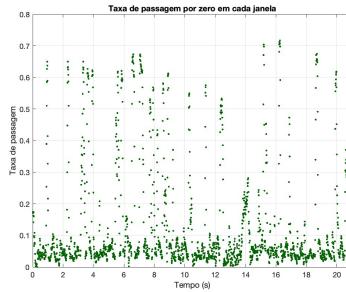


Fig. 4: Passagens por zero em cada janela

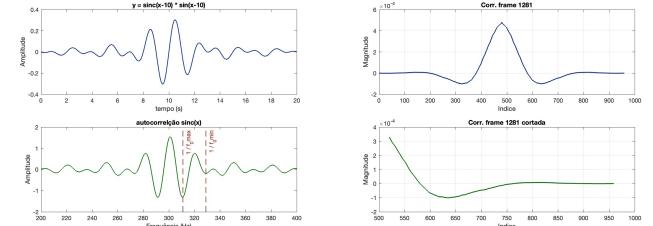


Fig. 5: Limites da voz humana numa auto correlação

C. Análise da Frequênci Fundamental do Sinal

A frequência fundamental (f_0) é a frequência mais baixa à qual uma onda periódica oscila, esta representa o tom principal (em inglês, *main pitch*) de um dado sinal áudio. O pitch foi calculado pela através da auto correlação para cada janela definida em III e encontrando o valor da frequência que corresponde ao segundo pico mais alto da correlação. Escolhemos o segundo pico, uma vez que o valor mais alto da correlação é sempre em zero, já que quando as janelas estão sobrepostas a auto correlação é um.

A voz humana está compreendida entre 70 e 250Hz, logo a frequência fundamental está dentro deste limite. Contudo, os momentos de silêncio e de maior ruído terão valores de F_0 bastante distintos daqueles oriundos da voz e que não devem ser ignorados. Assim, procuramos valores da frequência fundamental entre os valores de $1/f_{0,max}$ e o resto do sinal. Tratando-se de uma correlação, o sinal terá o dobro do tamanho do original e será simétrico, podendo ser apenas considerada a parte direita do sinal. Na Figura 5 está uma explicação gráfica do método de definição destes valores limites, de facto, o sinal verde representa a auto correlação do sinal a azul. Na Figura 6 encontra-se a função implementada em Matlab que realiza a auto correlação de uma dada janela: corta a janela pelos limites definidos acima e encontra o máximo da função dentro desse limite. Em 5 b) temos uma situação real que corresponde à janela 1281, cujo sinal azul é a auto correlação e a verde após o corte. De referir que, como estamos lidar com um sinal digital é necessário converter os valores de amostras para frequência, daí as consecutivas aparições de f_s no código. Esta função é chamada dentro de um ciclo for que percorre cada janela do sinal.

```

161 % We define a function to calculate PB using autocorrelation
162 function [fb, number_frame] = calculatePB(frame, fs, number_frame)
163
164
165
166
167
168
169
170
171 % We need to search the second peak, because the first peak is always 0
172 % The voice of man is between 70 and 250 Hz, so we need to find the peak in this interval
173 f0_max = 250;
174 f0_min = 70;
175
176 % Calculate autocorrelation
177 autocorr2 = autocorr(length(frame)) + filfilt(fs/f0_max); end;
178
179 % Find the index of the maximum peak in the autocorrelation function
180 [fb, number_frame] = findpeaks(autocorr2);
181
182 % Calculate F0 using the index
183
184 fb = fs / (fs * filfilt(fs/f0_max));
185
186 % Plot the autocorrelation function
187 if number_frame>=2000
188 subplot(1,2,1); plot(autocorr2, 'color', 'darkblue', 'LineWidth', 1.5); xlabel('Frame'); ylabel('Magnitude');
189 subplot(1,2,2); plot(autocorr2, 'color', 'darkblue', 'LineWidth', 1.5); xlabel('Frame'); ylabel('Magnitude');
190 end

```

Fig. 6: Código MATLAB - Calculo de F0 de uma dada janela

A Figura 7 apresenta as diversas frequências fundamentais calculadas em cada janela no sinal. 7 b) é o mesmo sinal que 7 a), mas após aplicado um filtro que consiste em calcular a mediada dos valores de f_0 de todas as janelas e aplicar um *threshold* de 50 Hz entre a diferença desse valor e de f_0 calculada em cada janela, de modo a remover outliers. Em 7 b) é claro um segmento entre os 50 e os 200 Hz e algum ruído a frequências bastante altas. Na secção III-D iremos elaborar a análise destes resultados.

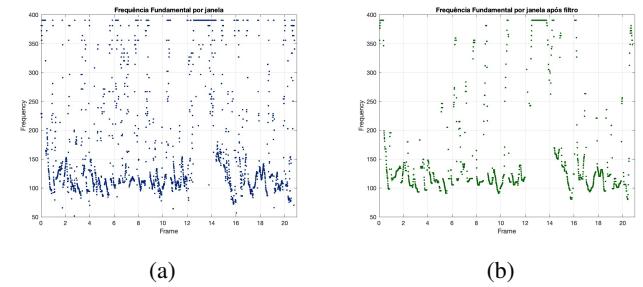


Fig. 7: Frequência fundamental de cada janela

Por fim, calculamos o valor médio das frequências fundamentais de cada janela, tendo apenas considerado os limites da voz humana, obtendo um valor de 112.36 Hz que corresponde à frequência fundamental da voz presente no áudio. Este valor é corroborado pelo histograma apresentado na Figura 8, já que está compreendido nos limites mais frequentes.

D. Desenvolvimento e Avaliação do Script de Classificação

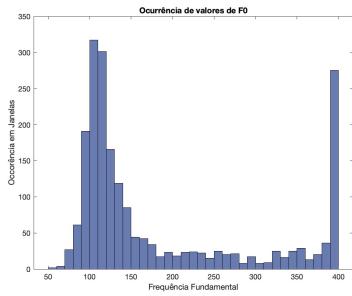


Fig. 8: Histograma de f_0 ao longo do sinal audio

Para distinguirmos entre silêncio e sinal, utilizamos a magnitude da energia. Através de um limiar (*threshold*), quando a energia é inferior a um valor específico, consideramos que se trata de silêncio; quando é superior, identificamos como voz.

Para verificar a correção do nosso programa, analisamos o segmento entre 12 e 14 segundos, Figura 9, uma vez que neste encontra-se uma pausa no discurso do orador. Assim sendo devia ser detetado um silêncio.

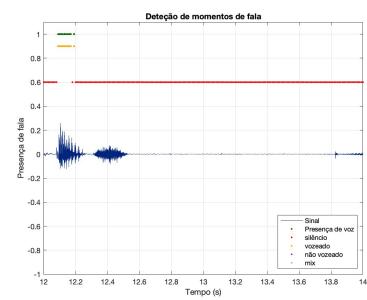


Fig. 9: Verificação do sinal Fala e Silêncio - exemplo 1

Como se pode verificar no gráfico anterior, um pouco depois dos 12 segundos enquanto termina a palavra 'pagos', nota-se a deteção de discurso, seguida de uma deteção de silêncio quando ocorre uma pausa por parte do orador.

Outro exemplo, que iremos analisar posteriormente em outros casos, situa-se entre os 7 e 9 segundos, Figura 10.

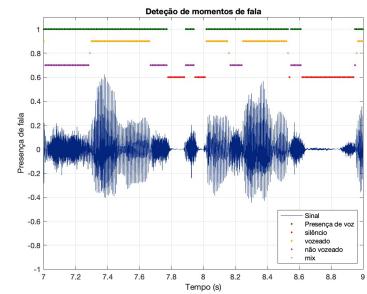


Fig. 10: Verificação do sinal Fala e Silêncio - exemplo 2

Neste excerto de fala é dito 'esses homens estrangeiros se' ou, tendo em conta a forma como é pronunciado, 'esseshomens es trangeiros se'. É possível observar claramente essas divisões entre fala e silêncio. Entre alguns momentos tem pequenas falhas, mas nada que degenera o sinal, obtido.

Para obter uma visão das percentagens de cada componente, criamos um histograma, Figura 11, que representa a taxa de ocorrência de cada estado. Este histograma já inclui outras categorias que posteriormente irão ser abordadas. É possível observar que 60% do discurso corresponde à fala, enquanto 40% representa o silêncio.

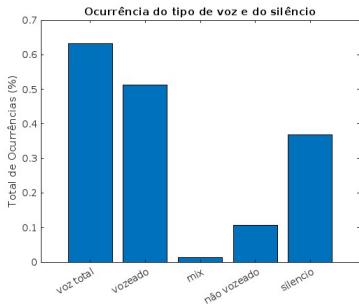


Fig. 11: Histograma com as taxas de ocorrência de cada estado

Agora, considerando para outras categorias, como fala vozzeada, fala não vozzeada e fala mista, observamos também o período entre 7 e 9 segundos, Figura 12, em que o orador pronuncia 'esses homens estrangeiros se'. Para este caso, além de considerarmos a energia, levamos em consideração a taxa de cruzamento por zero, em que um valor mais alto de zero crossing rate é classificado como fala não vozzeada, um valor intermediário como fala mista e um valor mais baixo como fala vozzeada.

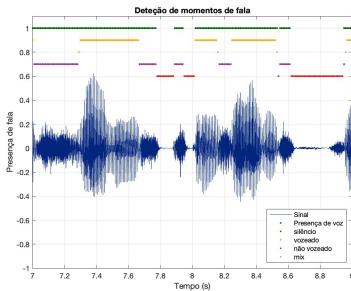


Fig. 12: Verificação do sinal entre 7 e 9 segundos

No inicio temos fala não vozzeada, pois temos o sô de 'ss', posteriormente mista que é no intervalo entre ambos e 'home' que é vozzeada e por fim não vozzeada outra vez com 'ns'. No segundo excerto do corte do áudio a palavra 'estrangeiros' também é possível perceber que obedece sempre conforme a fala não vozzeada e fala vozzeada.

Tendo em conta que estavamos a obter resultados satisfatórios, obtivemos a Figura 13, com o sinal completo e a sua classificação.

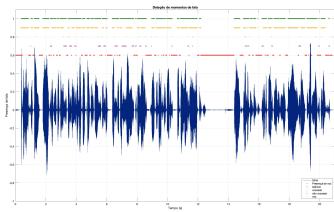


Fig. 13: Sinal completo com classificações

E observamos novamente o histograma, Figura 14, mas neste caso tendo em atenção as restantes categorias e o silêncio.

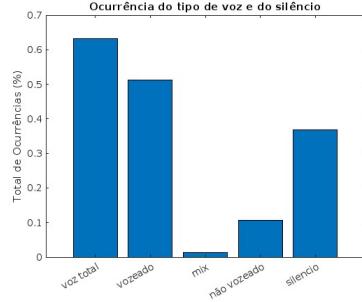


Fig. 14: Histograma com as taxas de ocorrência de cada estado

Computacionalmente, a classificação fala, fala mista, fala vozzeada, fala não vozzeada e silêncio é realizada por uma série de condições *if* apresentados entre as linhas 74 e 103 da Figura 2 b), sendo que os limites (*thresholds*) estão definidos nas linha 47, 48 e 49 da Figura 2 a). Deste modo, mediante a energia calculada para a janela em análise, esta é classificada como silêncio ou discurso. Caso seja discurso, é classificado como vozzeado, não vozzeado ou mix. As ocorrências são guardadas na variável *count*.

E. Algoritmo de Classificação do professor João Paulo Teixeira

Para que possamos ter uma base de comparação com o nosso código, o professor forneceu-nos um código desenvolvido por João Paulo Teixeira. Após efetuarmos as alterações necessárias para atender às nossas necessidades, conseguimos gerar gráficos que nos permitiram fazer comparações com o nosso próprio trabalho.

Quatro Categorias: Indefinido, Silêncio, Ruído e Fala

Neste primeiro caso, o sinal é dividido em quatro categorias: indefinido, silêncio, ruído e voz. O ruído no nosso caso é contabilizado como fala. Para isso, geramos o gráfico com o sinal completo e com as devidas classificações, Figura 15.

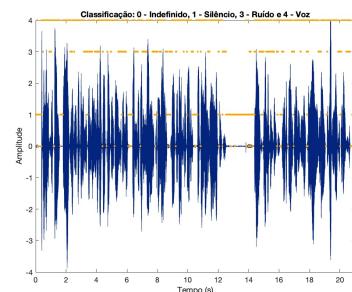


Fig. 15: Sinal total com classificações (professor)

Em termos gerais e considerando o ruído como fala é complicado perceber as diferenças entre o nosso e o do

professor. Assim sendo para obtermos uma visualização de forma a que se conseguisse comparar com o nosso caso, cortamos o sinal num segmento de 2 segundos sendo assim mais fácil de comparar com o obtido por nós, escolhendo para tal o intervalo de 7 a 9 segundos, Figura 16.

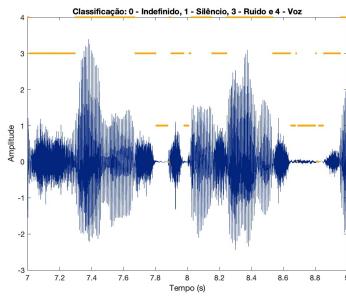


Fig. 16: Sinal com classificações (professor) (7-9 segundos)

Como se pode observar neste gráfico, e tendo em conta o gráfico anterior obtido por nós, podemos notar que, apesar de serem bastante semelhantes e de obtermos resultados praticamente idênticos se considerarmos o ruído como fala, o nosso gráfico é mais sensível à amplitude da energia do que o do professor. Entre os segundos 8.6 e 8.9, consideramos silêncio, enquanto o professor considera ruído. Concluímos, portanto, que se o ruído fosse excluído e o limite entre fala e silêncio fosse ajustado, poderíamos obter o mesmo resultado. Ouvindo este segmento de som, é possível perceber que o orador está em pausa, não falando, o que nos leva a considerar este período como silêncio. Portanto, a nossa abordagem parece mais correta neste contexto.

Para fazer uma comparação com o nosso sinal como um todo, criamos um histograma, Figura 17, que contabiliza a taxa total de cada categoria. Neste caso, consideramos o ruído como fala e o indefinido como silêncio, contudo como já foi observado nem sempre isso é verdade, podendo o ruído ser contabilizado como silêncio e o indefinido como fala, assim sendo foi utilizado apenas como referência:



Fig. 17: Número de ocorrências de Indefinido, Silêncio, Ruido e Voz

Comparando os valores verificamos que contando apenas a voz os valores são bastante próximos, cerca de 60% tal como tínhamos obtido, sendo então claro que parte do ruído é considerado como silêncio por nós e parte como fala.

Seis Categorias: Indefinido, Silêncio, Ruído, Fala Não Vozeada, Fala Mista e Fala Vozeada

Neste segundo caso, o sinal é dividido em seis categorias: indefinido, silêncio, ruído, fala não vozeada, mix (fala mista) e fala vozeada. Assim sendo obtivemos a Figura 18.

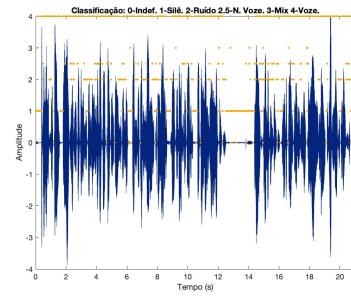


Fig. 18: Sinal completo com seis categorias (professor)

Para conseguirmos comparar melhor com o nosso sinal, cortamos novamente o sinal num segmento de 2 segundos, escolhendo para tal o intervalo de 7 a 9 segundos, Figura 19, novamente, para gerar um gráfico fácil de comparar:

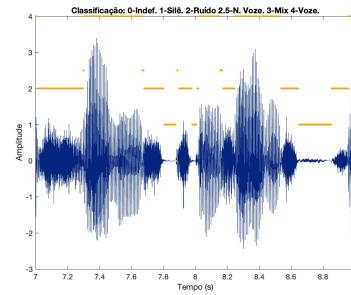


Fig. 19: Sinal com seis categorias (professor) (7-9 segundos)

Como se pode observar visualmente, são bastante parecidos, contudo no do professor podemos considerar que houve incoerência na classificação. Pois, observa-se que quando nós consideramos não vozeado o professor considera ruído. Por isso, pensamos que neste sentido o do professor esteja mal.

Para fazer uma comparação com o nosso sinal, criamos tal como anteriormente um histograma, Figura 20, que contabiliza o total de cada categoria. Neste caso, como referido anteriormente, o professor considerou ruído para não vozeado, por isso consideramos o ruído como não vozeada neste histograma para comparar com o nosso.



Fig. 20: Número de ocorrências de Indefinido, Silêncio, Ruido, N. Vozeada, Mix e Vozeada

Comparando com os nossos valores pode-se observar que neste caso o ruído está dividido entre fala não vozeada e silêncio, contudo comparando os valores de fala vozeada também estão bastante idênticos aos nossos. Assim sendo, podemos concluir que o ruído é sempre dividido entre duas categorias nossas e se isso acontecesse iríamos obter valores bastante idênticos, logo podemos concluir que os nossos valores estão conforme o esperado.

IV. COMPARAÇÃO DE RESULTADOS USANDO A FERRAMENTA PRAAT

A. Inspecção do espetrograma, Frequência Fundamental (F_0) e Energia e a sua comparação

Através do uso do PRAAT, pudemos analisar várias características do nosso sinal. Ao longo deste subcapítulo, examinaremos cada uma delas e, posteriormente, efetuaremos uma comparação entre a frequência fundamental obtida e a energia em relação ao que foi obtido no Matlab.

Espetograma

O espetrograma é uma ferramenta fundamental na análise de sinais complexos, oferecendo uma representação visual minuciosa das características temporais e de frequência de um sinal. Esta representação gráfica é particularmente valiosa em campos como processamento de áudio, fala, música e telecomunicações. Ao observar este espetrograma específico, Figura 21, torna-se possível identificar os momentos de pausa e silêncio no discurso, destacados em branco, bem como discernir quais frequências predominam ao longo do tempo, representadas pelas áreas mais escuras.



Fig. 21: Espetograma do sinal observado no PRAAT

Energia

Utilizando o espetrograma, o software PRAAT tem a capacidade de analisar e definir a energia presente no sinal. Esta análise detalhada do espetrograma permite ao PRAAT

identificar e quantificar a energia em diferentes partes do sinal, fornecendo uma visão abrangente das variações de energia ao longo do tempo, Figura 22, gráfico verde.

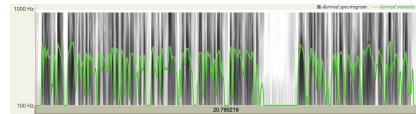


Fig. 22: Energia do sinal observado no PRAAT

Ao comparar a análise de energia realizada pelo PRAAT com aquela obtida através do MATLAB, Figura 3, observa-se uma notável proximidade entre os resultados. Os momentos em que o som atinge maiores níveis de intensidade correspondem aos momentos de maior energia. No espectrograma, esses momentos são representados como tons mais escuros, que se manifestam como picos na Energia.

Frequência Fundamental (F_0)

Através da análise realizada pelo PRAAT, é possível obter estimativas da frequência fundamental com precisão. A frequência fundamental, frequentemente referida como a frequência de pitch, desempenha um papel fundamental na caracterização do conteúdo sonoro. O PRAAT fornece uma representação confiável da frequência fundamental, o que é essencial para a compreensão das propriedades tonais do sinal de áudio, Figura 23.



Fig. 23: Frequência fundamental (F_0) do sinal observado no PRAAT

Ao comparar os resultados com o que foi obtido no MATLAB, conforme ilustrado na Figura 7, é evidente que os dois são notavelmente semelhantes. No entanto, no caso do PRAAT, observa-se a aplicação de um filtro mais robusto que resulta em uma maior coerência no sinal. Isso realça a eficácia do PRAAT em aprimorar a qualidade e a estabilidade das medições, contribuindo para uma análise mais consistente e precisa. O valor médio de frequência fundamental obtido no PRAAT foi de 114.03 Hz que é ligeiramente superior àquele obtido em Matlab na secção III-C.

REFERENCES

- [1] Rossing, Thomas D., Moore, F. Richard, Wheeler, Paul A, *The Science of Sound*, Pearson Education, 2022.