Relatório Projeto ATD

Autores:

Alexandre Ferreira 2021236702

Eduardo Pereira 2021233890

João Tinoco 2021223708

Introdução/Contextualização

O reconhecimento e interpretação de voz humana tornou-se uma característica comum em dispositivos eletrónicos, como telemóveis e computadores, por meio de comandos de voz. A aquisição e extração de características do sinal áudio são fundamentais nesse processo, visando discriminar diferentes sons, como os dígitos em inglês de 0 a 9.

Este projeto tem como objetivo realizar a análise em frequência de sinais áudio, visando identificar esses dígitos. Isso contribuirá para o desenvolvimento de sistemas avançados de reconhecimento de voz e sua aplicação em assistentes virtuais, automação residencial, segurança, entre outros.

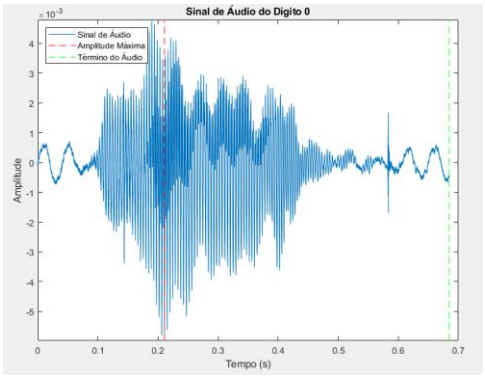
Para construir uma análise abrangente e diversificada, foram utilizados diferentes métodos de processamento de sinais de áudio. Esses métodos incluíram a extração do sinal de cada áudio, o cálculo do espectro de amplitude mediano e a normalização em relação ao número de amostras, além da aplicação da Transformada de Fourier de Tempo Curto (STFT).

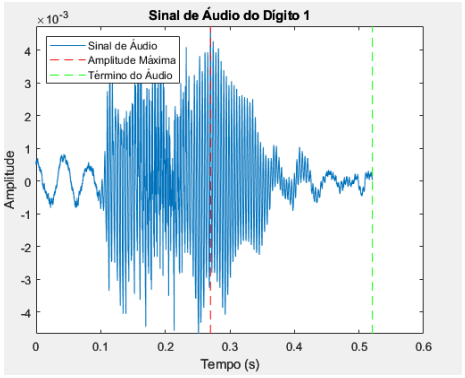
Descrição dos dados

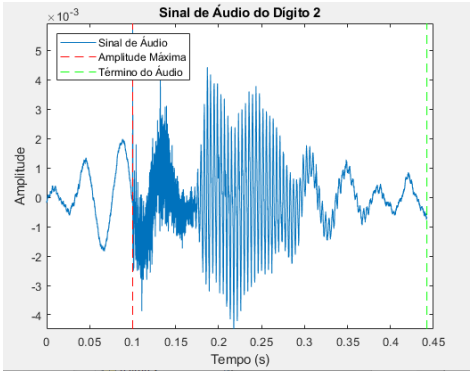
Os dados fornecidos correspondem sinais de voz emitidos por 60 participantes, que repetiram 50 vezes cada um dos dígitos (“0 1 2 3 4 5 6 7 8 9”). Os dados fornecidos consistem em arquivos de áudio em formato .wav, adquiridos a uma taxa de amostragem de 48000 Hz em modo mono-canal.

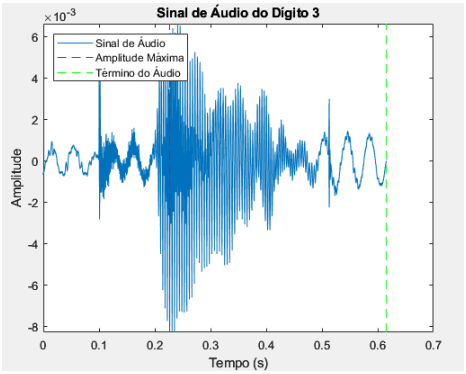
Metodologias implementadas, Resultados

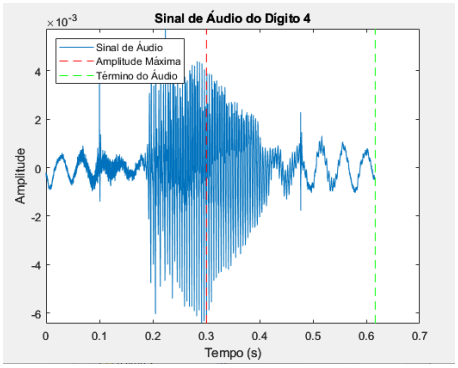
Os seguintes gráficos são o resultado das representações gráficas de um exemplo de cada dígito, a cada par de dígitos. As principais diferenças visuais são o momento terminal do áudio, as amplitudes máximas e o momento onde existe uma energia elevada. Para obtermos estes resultados, inicialmente extraímos o sinal de cada dígito aleatoriamente e verificámos onde ocorre o término do áudio e amplitude máxima.

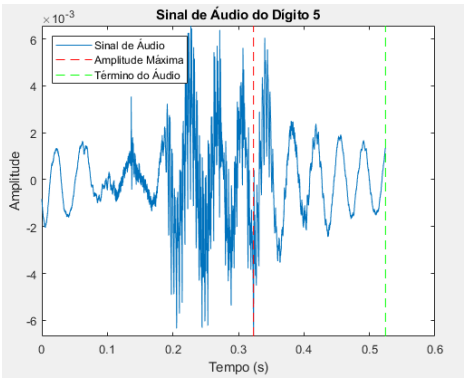


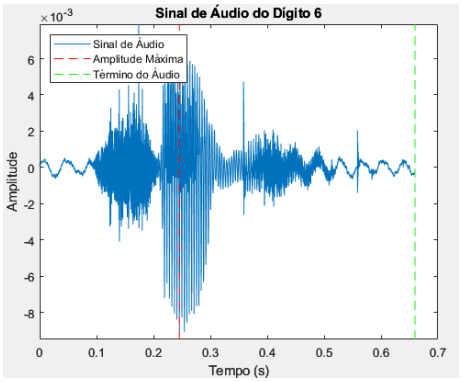


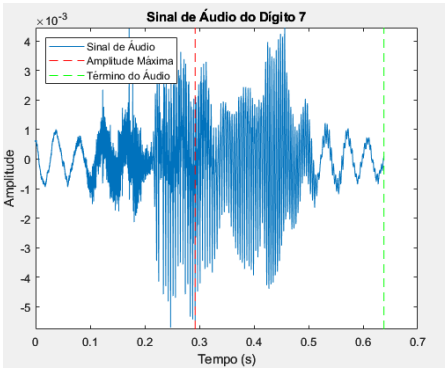


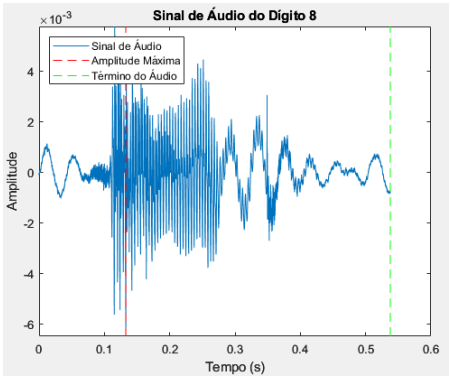


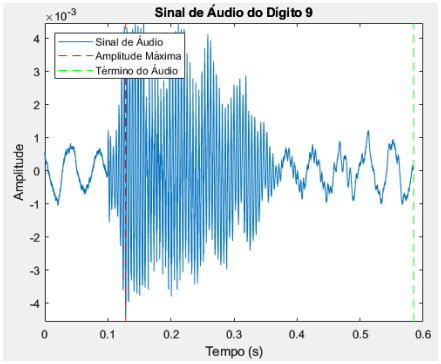




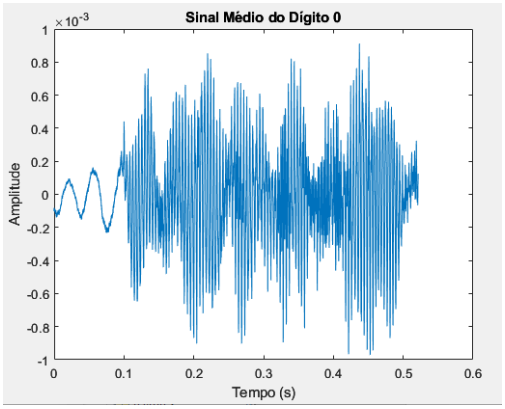


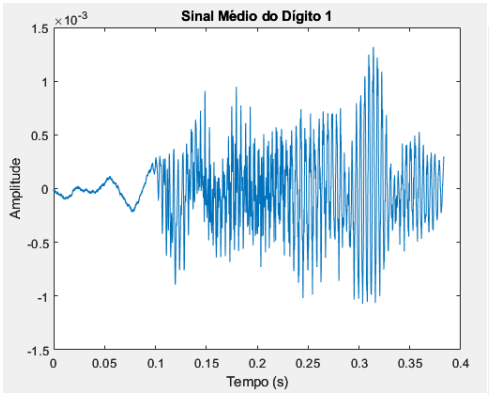


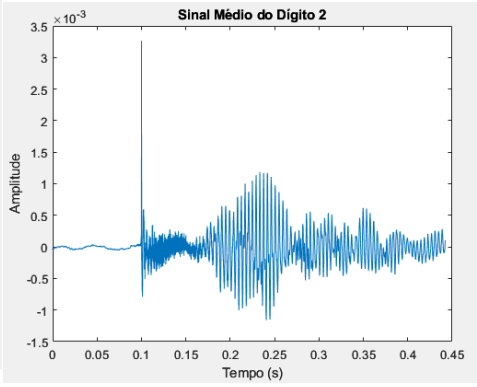


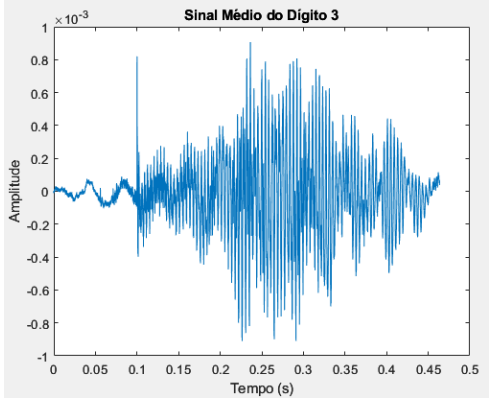


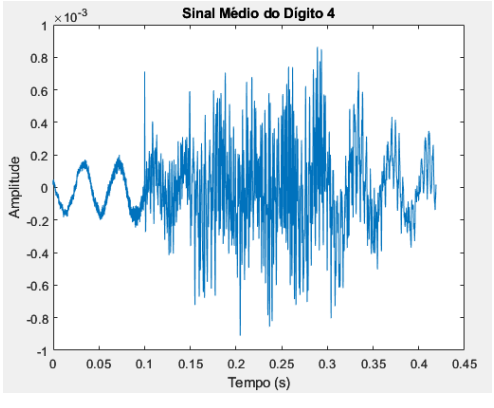
De seguida, foram calculadas as médias dos sinais de cada dígito com o objetivo de compreender o comportamento geral do sinal. Para o realizar, calcula-se o sinal médio para cada dígito, agrupando as amostras de áudio do mesmo dígito e calculando a média dos sinais. Em seguida, ele plota o sinal médio de cada dígito.

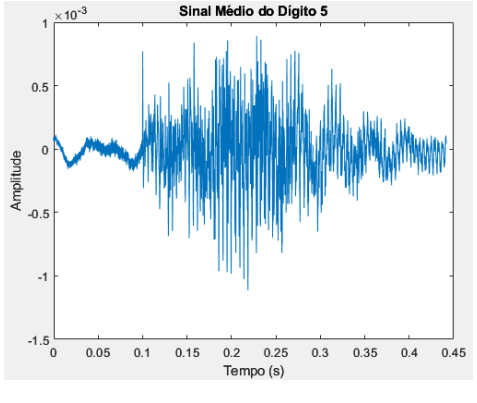


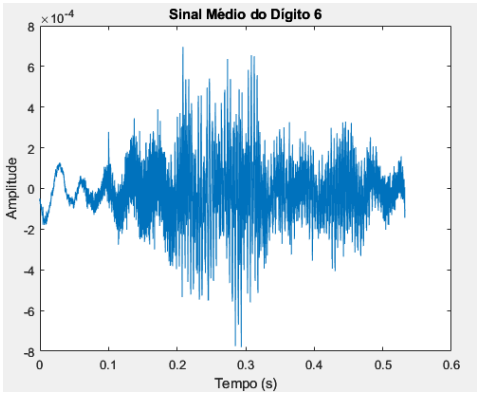


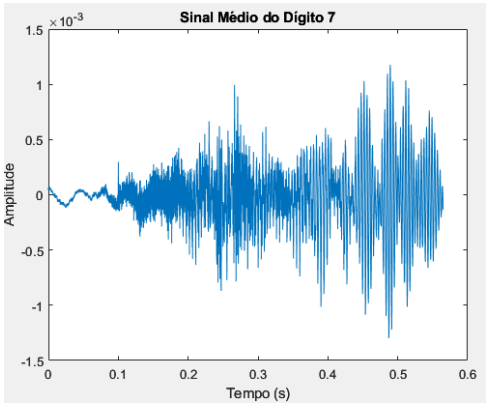


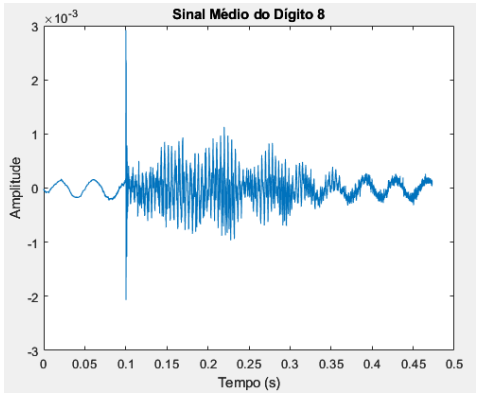


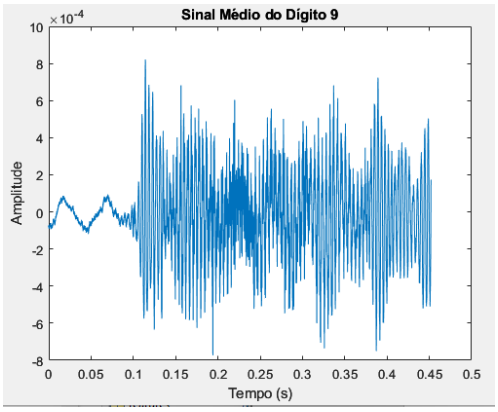






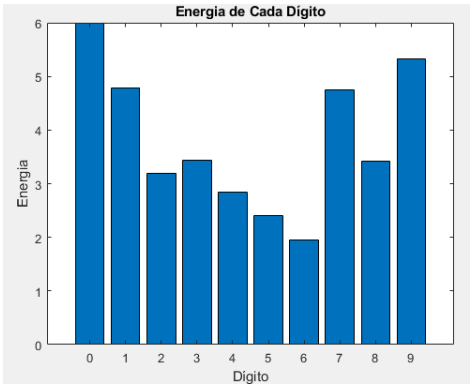


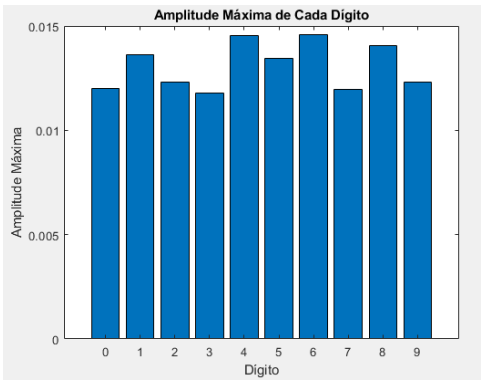


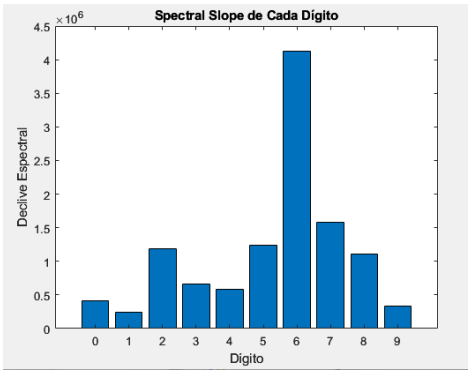


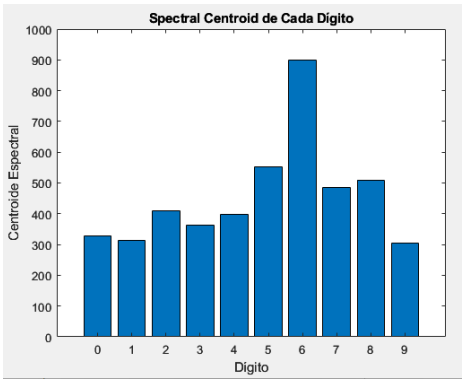
Posteriormente, identificaram-se características temporais que permitissem diferenciar todos os dígitos da forma mais percetível. Para cada dígito existe uma barra que identifica o seu nível de energia, amplitude máxima, Taxa de Cruzamento por Zero, Centroide Espectral, Declive Espectral. O código calcula estas características iterando sobre as amostras de áudio de cada dígito, calculando essas características usando técnicas como a transformada de Fourier.

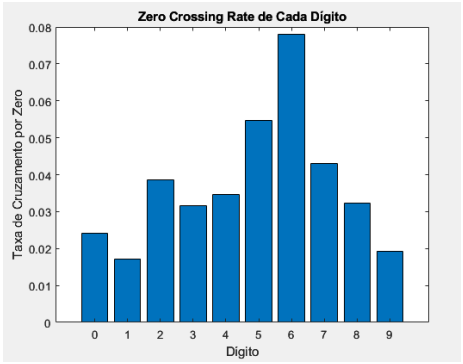
O código gera gráficos para visualizar as características calculadas. Ele plota a energia e a amplitude máxima para cada dígito, bem como a taxa de cruzamento por zero, o centroide espectral e o declive espectral. Também é gerado um gráfico 3D que representa a energia, a amplitude máxima e o centroide espectral dos dígitos, permitindo visualizar as superfícies de divisão entre os dígitos.

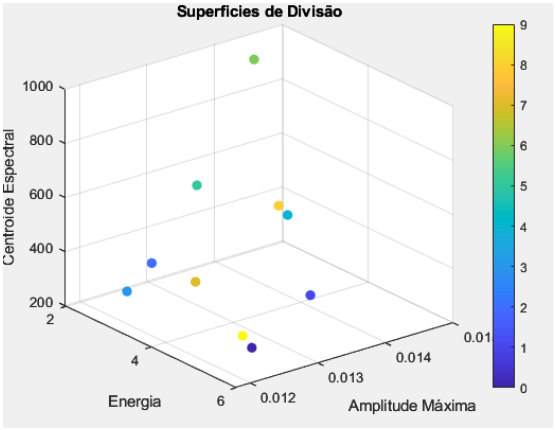












Verifica-se que na última figura, ou seja, um espaço de três dimensões que é possível criar regras de decisão.

Se o valor do Spectral Centroid for o mais baixo de todos (ou próximo disso), então é o dígito 6. Caso o valor da energia for o mais baixo de todos (ou próximo disso), então é o dígito 5. Se o valor da amplitude máxima for o mais baixo de todos (ou próximo disso), então é o dígito 3. Se o valor do Spectral Centroid for o mais alto de todos (ou próximo disso), então é o dígito 6. Já se o valor da energia for o mais alto de todos (ou próximo disso), então é o dígito 0. Se o valor da amplitude máxima for o mais alto de todos (ou próximo disso), então é o dígito 6. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for alta, então é o dígito 8. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for baixa, então é o dígito 1. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for moderada, então é o dígito 4. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for alta, então é o dígito 7. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for moderada, então é o dígito 2. Se os valores do Spectral Centroid e da energia forem intermediários e a amplitude máxima for baixa, então é o dígito 9.

As características usadas para estas regras foram: a Spectral Centroid, a energia e a amplitude máxima.

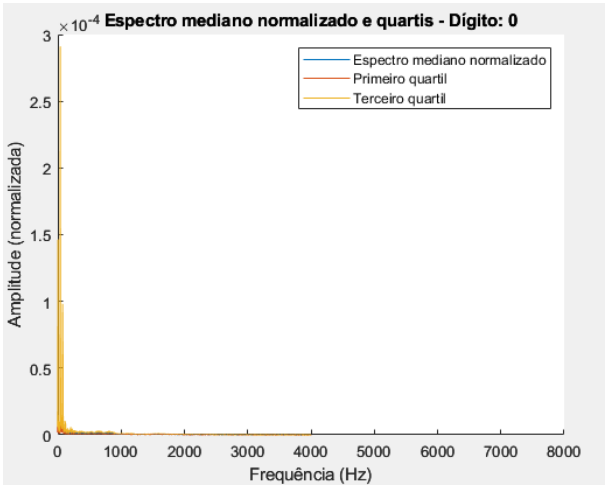
Spectral Centroid (Centroide Espectral): É uma medida estatística que indica a posição central do espectro de frequência de um sinal. Representa o centro de gravidade do espectro e fornece uma estimativa de onde a maior parte da energia espectral está concentrada. O cálculo do Spectral Centroid envolve a ponderação das frequências pelos valores de magnitude espectral.

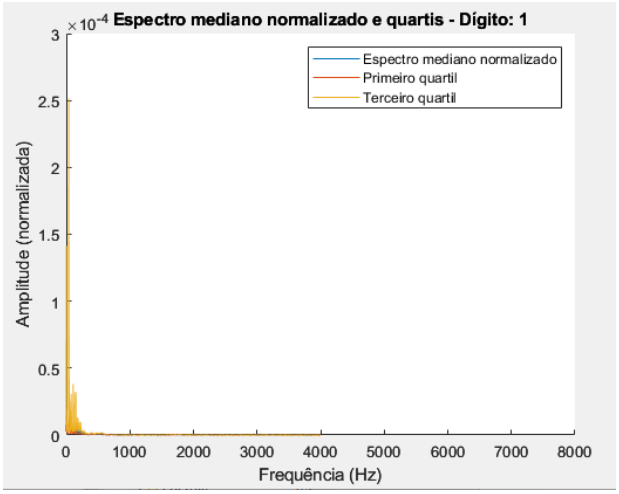
A energia de um sinal de áudio refere-se à quantidade total de energia contida nesse sinal. É uma medida da potência ou intensidade do sinal ao longo do tempo. A energia é calculada somando-se o quadrado dos valores do sinal em cada ponto no tempo. No contexto do gráfico tridimensional mencionado, a energia é representada pelo eixo dos x.

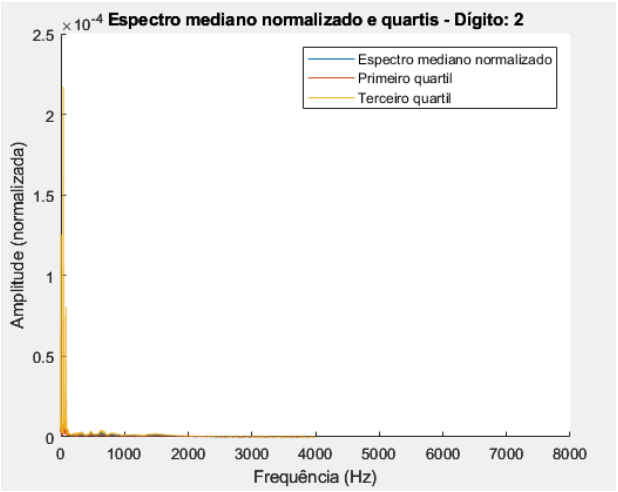
A amplitude máxima de um sinal de áudio é o valor máximo alcançado pela onda sonora em termos de deslocamento máximo da posição de equilíbrio. Ela indica o valor mais alto do sinal em termos de sua magnitude.

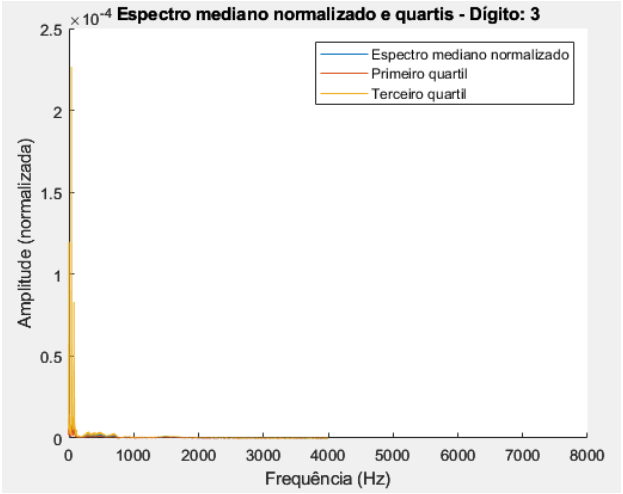
De seguida, são apresentados os gráficos obtidos através do espectro de amplitude mediano e normalizado pelo número de amostras para cada dígito. O código normaliza os sinais de áudio preenchendo-os com zeros para terem o mesmo tamanho. Em seguida, os sinais são somados e a média é calculada.

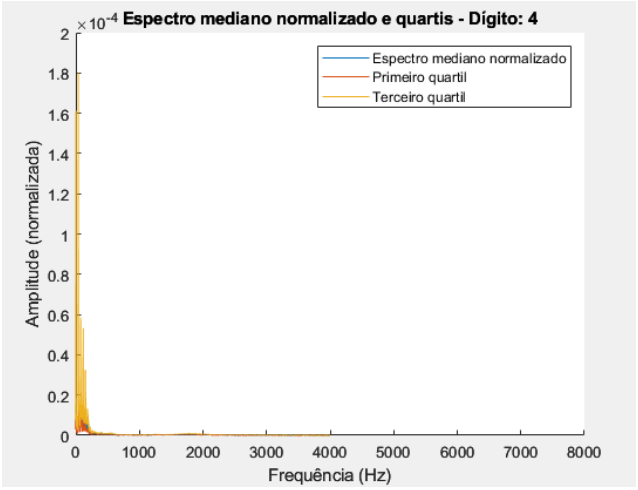
Calcula-se o espectro de amplitude para cada sinal de áudio selecionado, usando a Transformada Rápida de Fourier (FFT). O espectro mediano, o primeiro quartil e o terceiro quartil são calculados para cada frequência. O espectro de amplitude é normalizado pelo tamanho do sinal. Por fim são gerados gráficos para visualizar os resultados. É plotado o espectro mediano normalizado juntamente com o primeiro e terceiro quartis. Além disso, são gerados gráficos comparando diferentes janelas de análise (retangular, Hann, Hamming e Blackman) e seus respectivos espectros medianos normalizados.

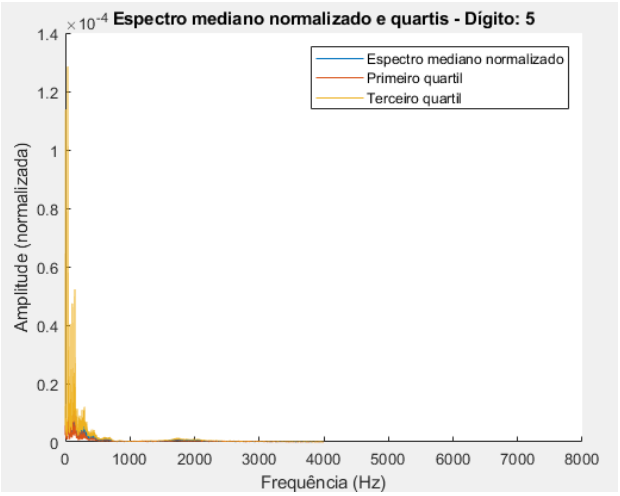


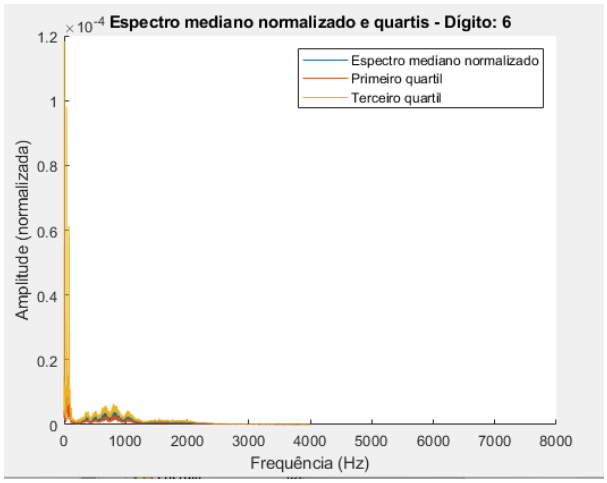


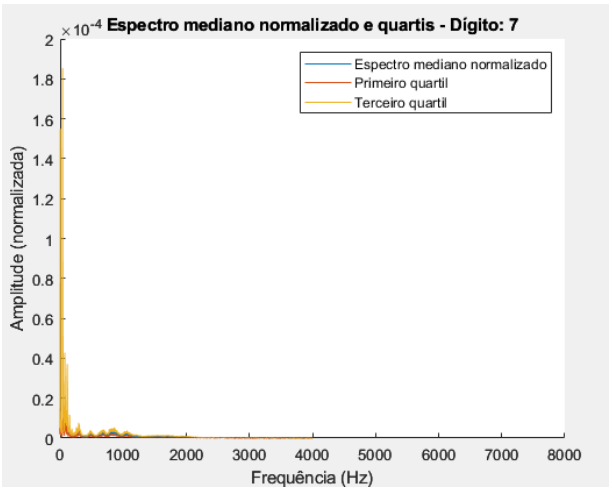


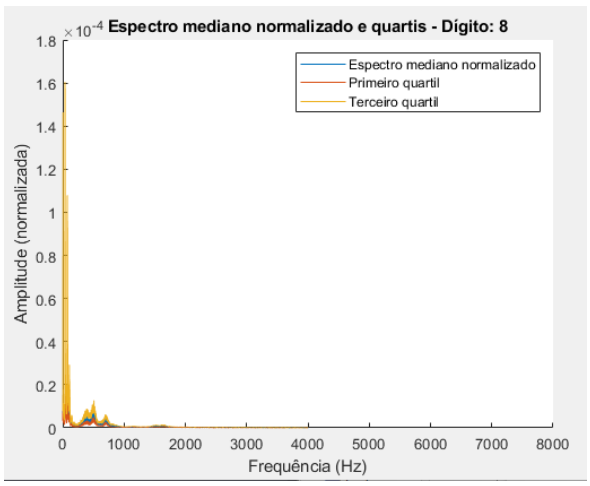


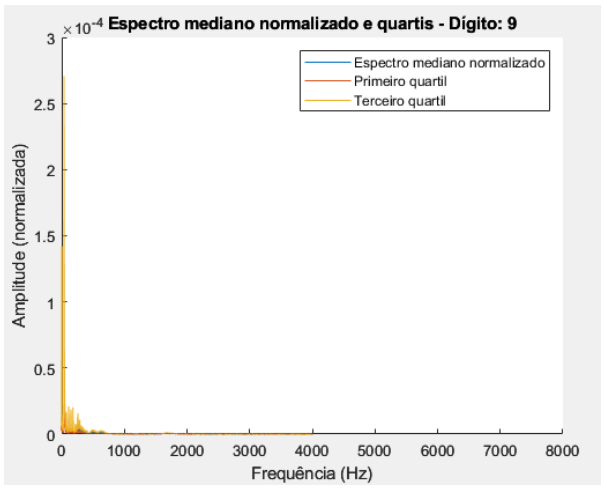




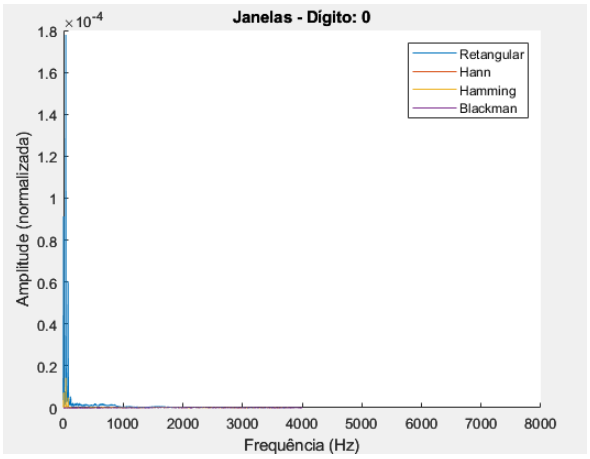


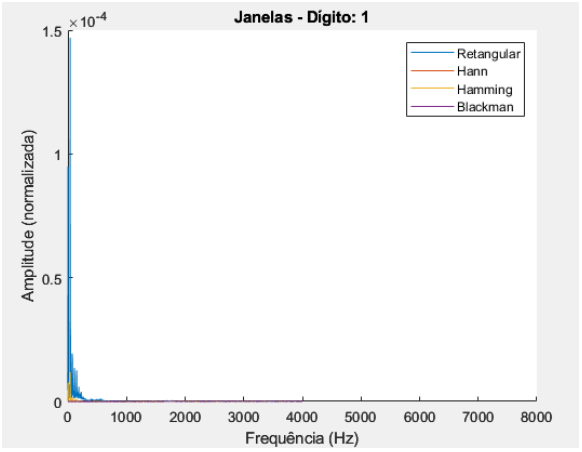


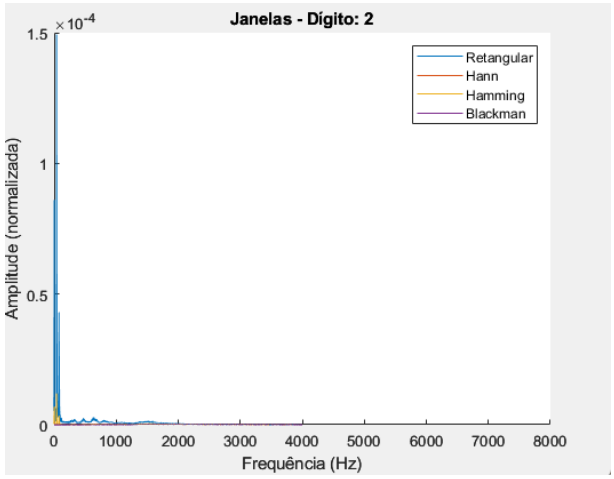


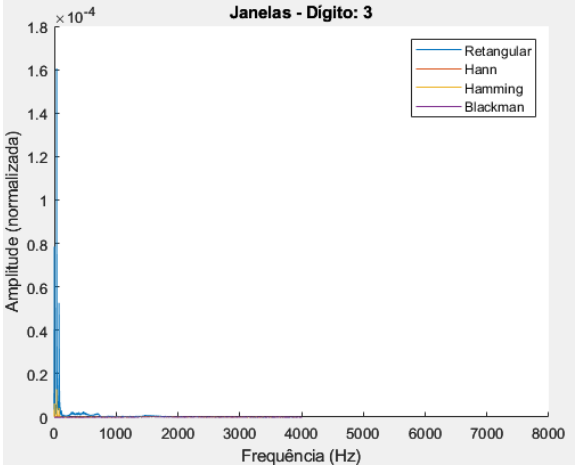


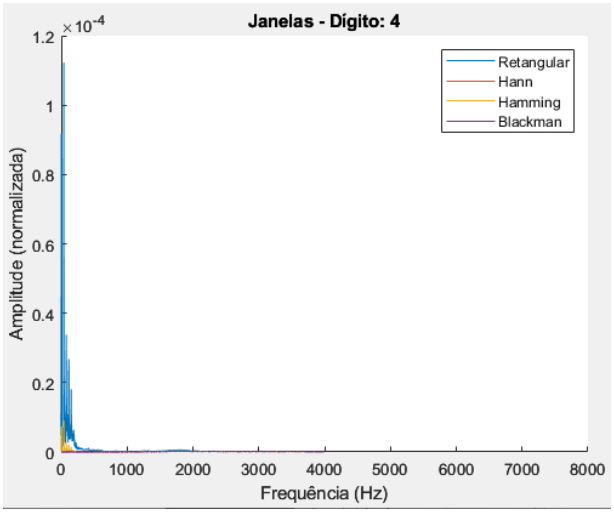
Os próximos gráficos representam os diferentes tipos de janelas utilizados para se verificar o seu efeito no sinal de cada dígito.

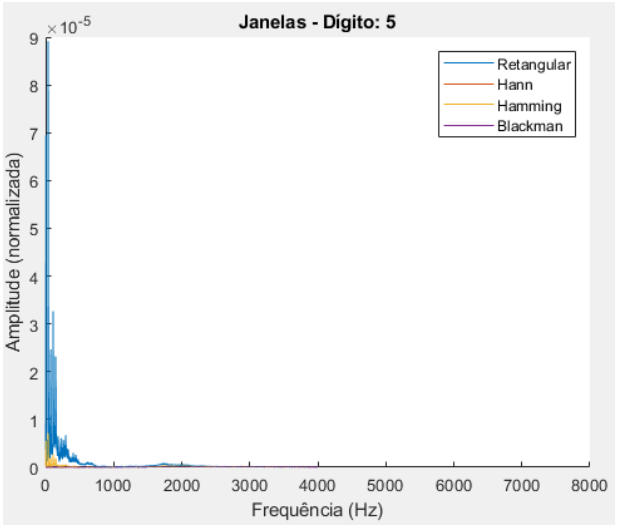


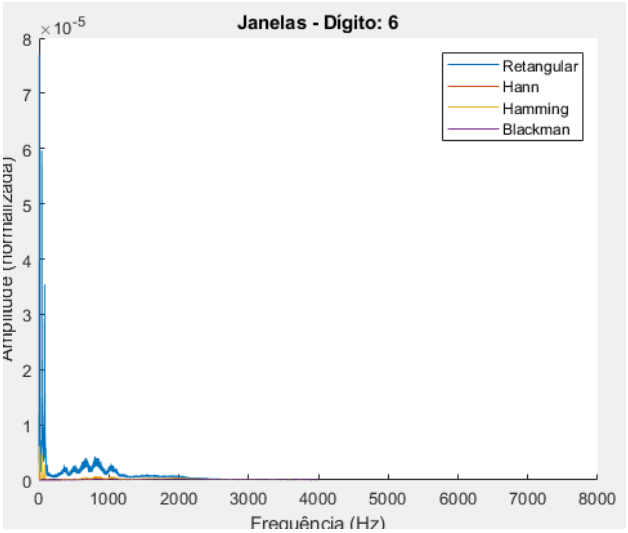


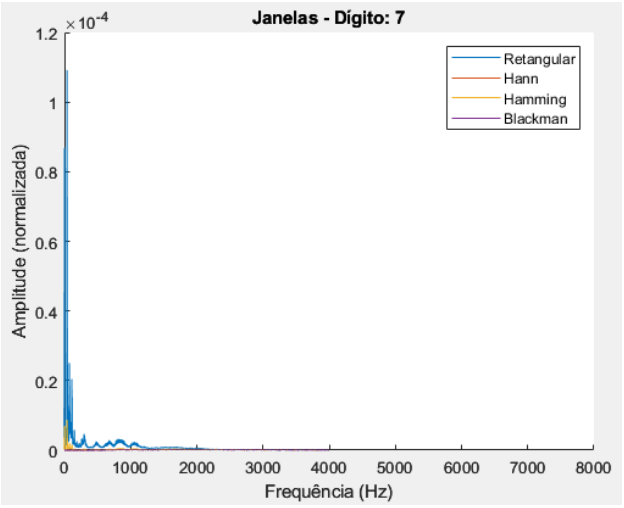


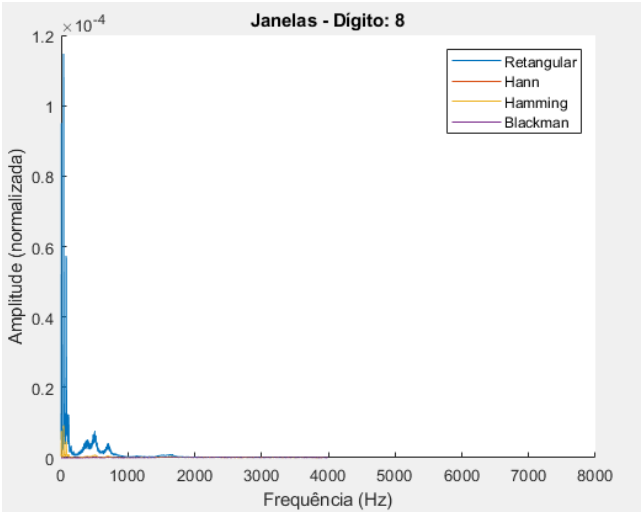


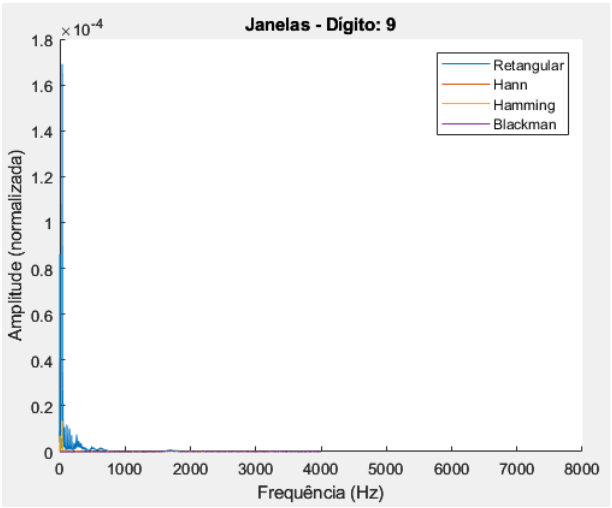












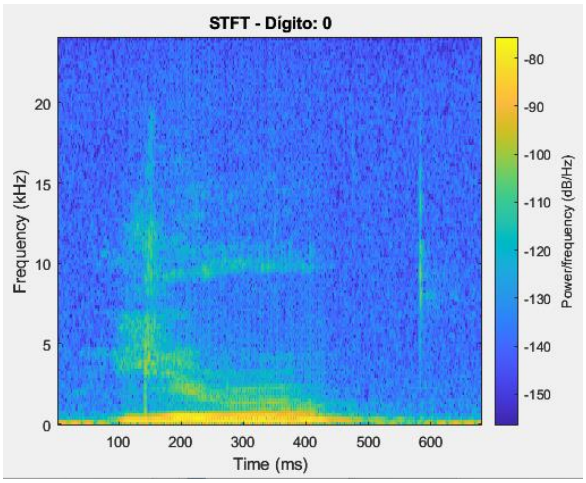
A janela retangular não aplica nenhuma ponderação ao sinal, preservando sua forma original. Isso pode resultar em vazamento espectral significativo, onde as componentes de frequência se espalham para além de seus valores reais.

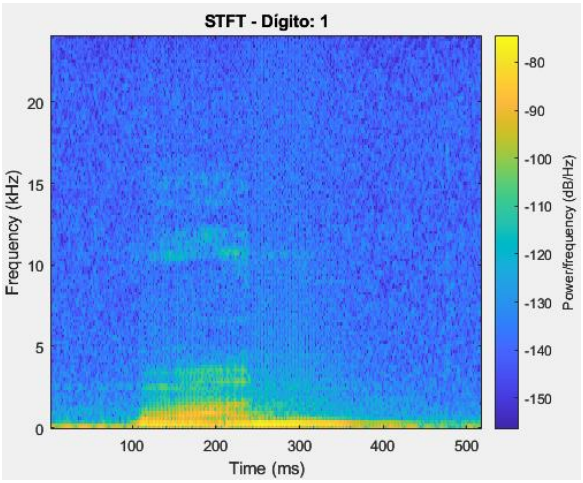
As janelas de Hamming e Hanning são janelas de suavização que ajudam a reduzir o vazamento espectral. Elas atenuam gradualmente as componentes de frequência nas extremidades da janela, resultando em uma transição suave do espectro. A janela de Hanning tem uma resposta em frequência mais estreita em comparação com a janela de Hamming, o que significa que ela concentra mais energia nas frequências centrais.

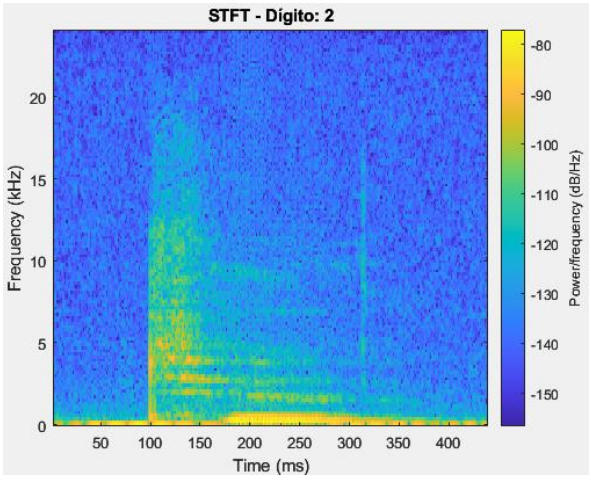
A janela de Blackman também é uma janela de suavização, mas possui uma resposta em frequência ainda mais estreita do que as janelas de Hamming e Hanning. Isso permite uma melhor resolução espectral, com uma atenuação mais acentuada nas frequências adjacentes às extremidades da janela.

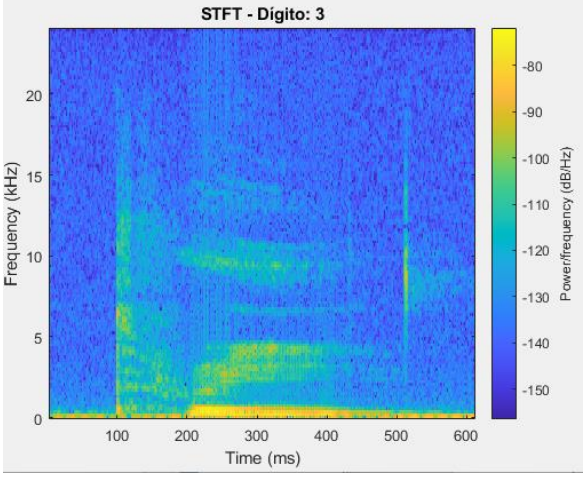
Por fim, são definidos os parâmetros da STFT, como o tamanho da FFT (tamanhoFFT) e a porcentagem de sobreposição (sobreposicao) para se obter a representação da mesma graficamente para cada dígito.

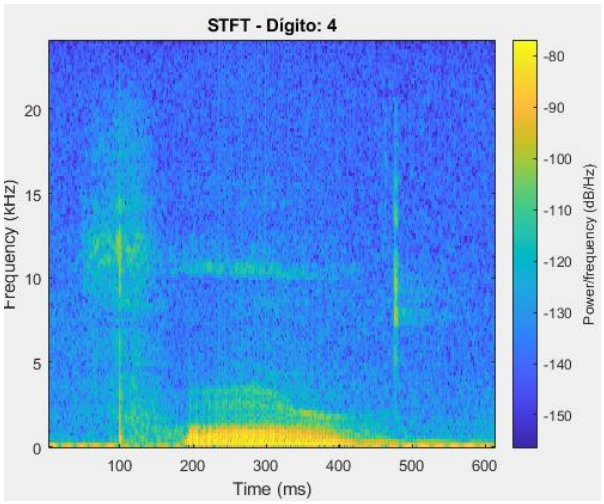
É plotado o espectrograma do sinal de áudio usando a função spectrogram. O espectrograma é gerado aplicando uma janela de Hamming ao sinal, usando o tamanho da FFT e a sobreposição definidos anteriormente calculou-se as SFTF para cada dígito e verificou-se os seguintes resultados:

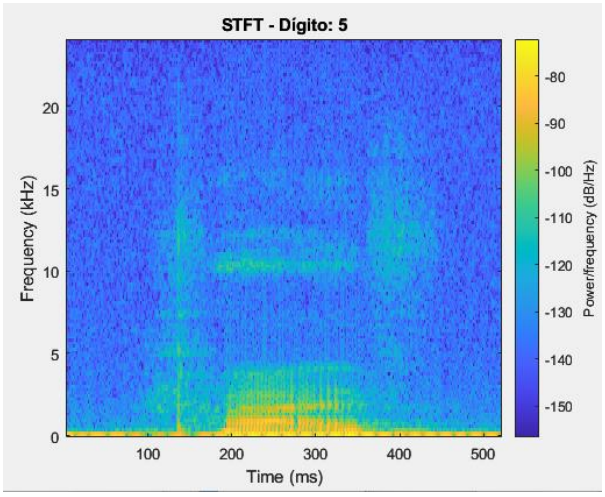


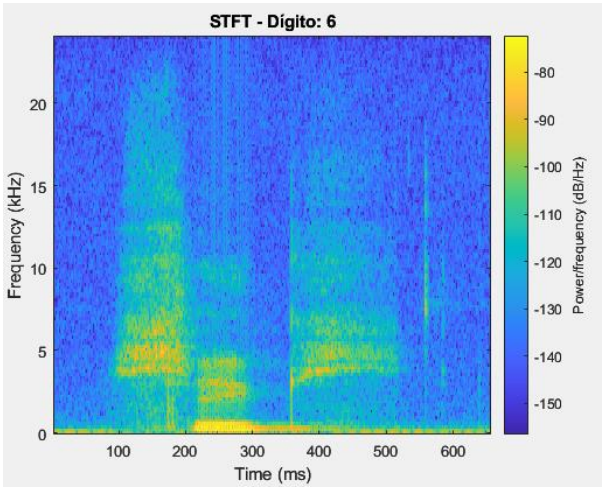


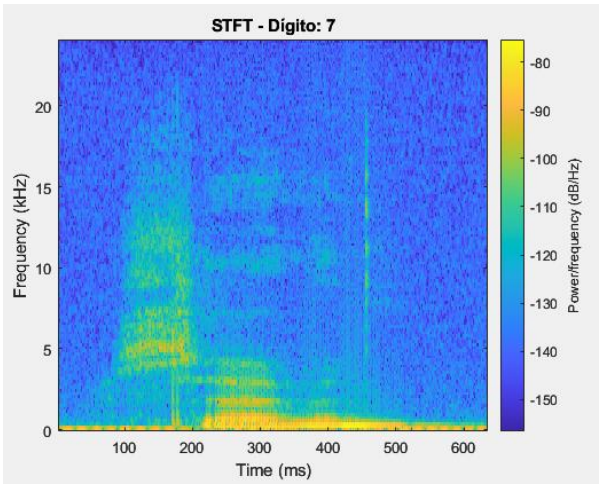


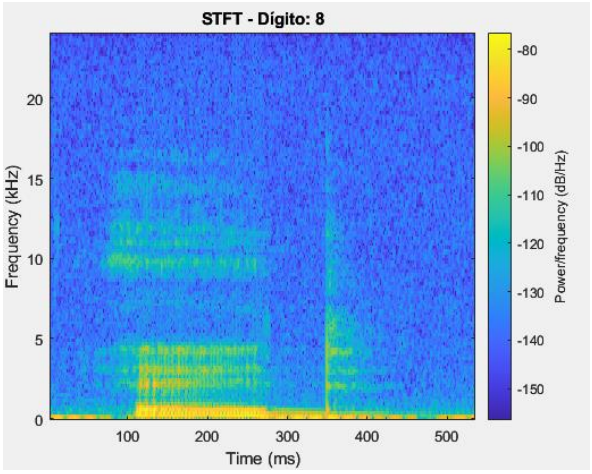


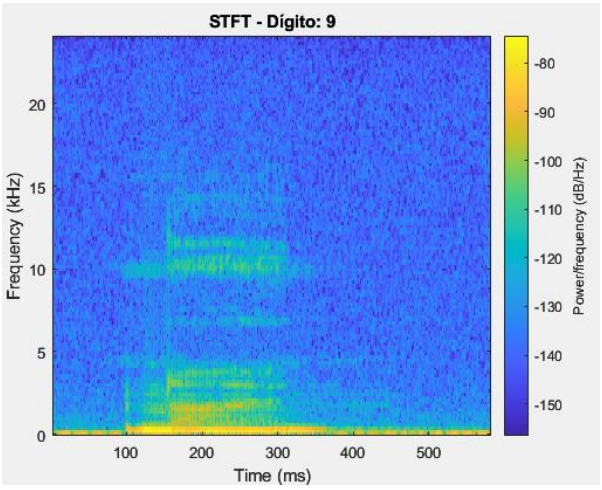












Discussão

Neste projeto foram implementadas várias metodologias para a análise dos sinais correspondentes a cada dígito para distinguir sons da fala humana, tanto a nível temporal, espectral e utilização da STFT, nos exercícios 4 5 e 6, respetivamente.

Para as suas respetivas análises, foram calculadas características adicionais tais como a energia, amplitude máxima, Zero Crossing Rate, Spectral Centroid, Spectral Slope.

Concluindo, com este trabalho foi-nos permitido identificar os dígitos em inglês entre 0 e 9, analisados em frequência de sinais de áudio.