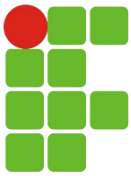


Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Curso Superior de Tecnologia em Automação Industrial

ARTHUR MEDEIROS GUIMARÃES

DAVIDSON MARCEL SEABRA DANTAS

Processamento Digital de Sinais – PDS

Relatório 02 – K-means

João Pessoa – PB,
Outubro de 2018

Sumário

1. Introdução

Nas últimas décadas tornou-se evidente a evolução do ambiente computacional, dado ao desenvolvimento de computadores cada vez mais robustos, ou seja, computadores com maior capacidade de processamento e armazenamento de informações. Apesar do grande avanço na área computacional os computadores ainda possuem limitações quanto a capacidade de processamento e armazenamento de informações, por exemplo, não é possível descrever computacionalmente um sinal do mundo real (analógico) com infinitos valores no tempo e na amplitude. Para que o computador possa representar um sinal do mundo real (analógico) é necessário converter esse sinal para um sinal amostrado no tempo e amplitude (digital), ou seja, um sinal com finitos valores no tempo e na amplitude. Apesar do sinal digital possuir números finitos de valores, qualquer operação realizada demanda de um custo computacional, e quanto maior esse custo mais demorado será o processamento e armazenamento dos dados, por conseguinte, procura-se técnicas de reduzir esse custo sem perdas significativas dos dados do sinal. A área de processamento digital de sinais (PDS) visa a otimização dos processos computacionais, utilizando técnicas de manipulação do sinal (Ref1).

Este trabalho tem o objetivo de utilizar o processamento digital de sinais (PDS) para classificar as características de um sinal de voz, classificando a composição do sinal em duas categorias, sendo a primeira: som sonoro, som surdo e silêncio; e a segunda em: vogal, consoante e silêncio/ruído. Para tal, utiliza-se do processamento digital de sinais de voz (PDSV), cujo é um ramo do PDS. O PDSV estuda as características da voz com o intuito de detectar patologias, emoções, reconhecimento de fala e entre outras áreas. As técnicas utilizadas foram as: divisão do sinal em Frames, Energia Segmental, Magnitude Segmental, Taxa de Cruzamento por Zeros e Clustering. O Clustering é um recurso computacional de alta aplicabilidade comercial no ramo do processamento de dados, classificador de dados, quantização vetorial e entre outros (Ref 2). Há diversas formas de realizar o processo de clustering, nesse trabalho será utilizado o algoritmo do K-means uma técnica de baixo custo computacional.

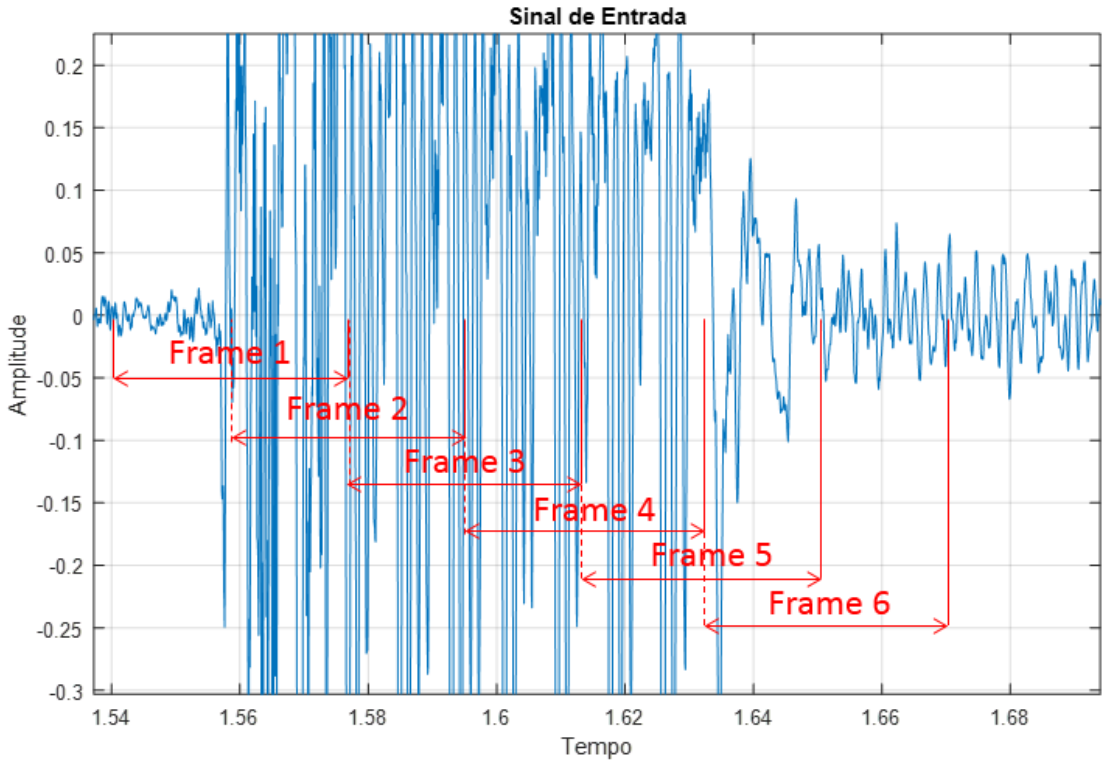
2. Fundamentação Teórica

Para realizar um processamento de sinais de áudios e identificar uma voz ou selecionar elementos da voz, é necessário extrair os elementos do sinal de áudio. As características do sinal de áudio, geralmente se baseiam em amplitude do sinal, frequência do sinal, numero de picos do sinal, e entre outras propriedades. Neste trabalho os elementos básicos avaliados para um sinal de áudio de voz é os sons surdos, os sons sonoros e o silencio. Os sons surdos são letras não vibram a glote para produzir o som. Os sons sonoros, a glote vibra em diferentes intensidades dependendo da letra a ser pronunciada. O silencio é o período em que não há nenhum som no sinal, ou seja, o sinal reproduz apenas o valor zero.

2.1. Frames

Segmentar o sinal de áudio em frames é uma tecnica que permite analisar curtos intervalos de tempo do sinal individualmente. Os frames utilizados comumente na indústria possuem um intervalo de tempo de 16 ms até 30 ms. Quando se refere a segmentar um sinal em frames, é necessário pensar também na sobreposição dos frames. A sobreposição é um artificio no qual um frame possui determinada porcentagem de dados iguais ao frame anterior. A porcentagem de sobreposição mais utilizadas sãos de 50% ou 75% do frame. A Figura 1 exemplifica a divisão dos frames em um sinal de áudio.

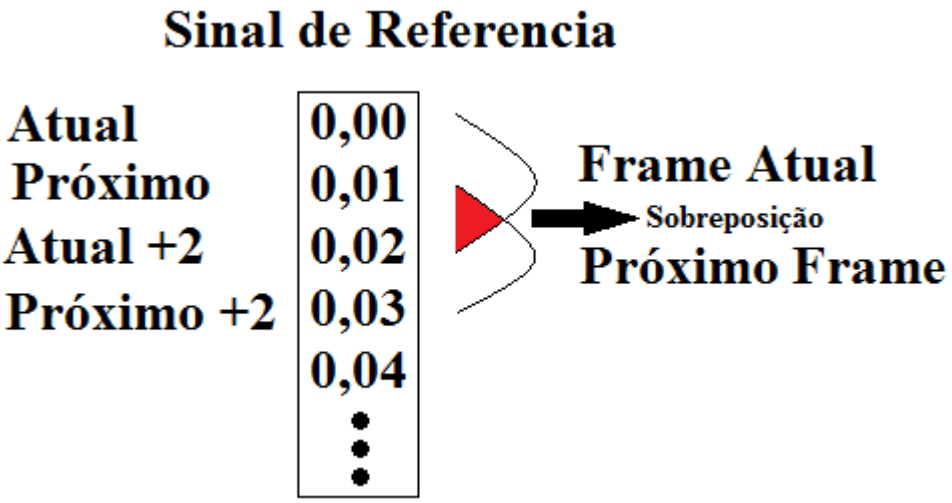
Figura 1: representação dos frames em um sinal de áudio.



Fonte: autoria própria.

Para exemplificar os processos de sobreposição, considere um frame com 20 ms e 50% de sobreposição. O primeiro frame irá começar em 0 e irá até o tempo de 20 ms, porém o segundo frame irá começar em 10 ms e irá até 30 ms. Ou seja, o primeiro e segundo frame compartilham entre sim o intervalo de tempo de 10 ms até 20 ms, este intervalo compartilhado é a sobreposição. Tendo estes valores é alocado em cada frame, o valor atual do vetor, até os próximos dois valores do vetor. A Figura 2 ilustra o processo de criação dos frames.

Figura 2: Processo de criação dos frames de 20 ms com uma sobreposição de 50%.



Fonte: autoria própria.

2.2. Energia Segmental

O cálculo da energia semental de acordo com Rabiner e Shafer (1978), como descrito em Passos, Silva, Costa, Neto e Fechine (2006), é igual ao somatório de todas as amostras de cada frame elevado ao quadrado. A Equação 1 descreve a formula para calcular a energia segmental.

$$E_{segmental}(i) = \sum_{n=0}^{N-1} [S(n)]^2 \quad (1)$$

Após calcular a energia segmental de cada frame, é possível plotar o gráfico dele no Matlab utilizando a variável de referência para o tempo de cada frame no eixo x, e utilizando o “frame.energ” no eixo y.

2.3. Magnitude Segmental

Para calcular a magnitude segmental, é necessário utilizar um somatório para o modulo de cada amostra de um frame. A Equação 2 descreve a forma da magnitude segmental.

$$M_{segmental}(i) = \sum_{n=0}^{N-1} |S(n)| \tag{2}$$

Porém, foi acrescentado a técnica de janelamento na magnitude segmental para servir de testes de aprendizado para o aluno. A técnica de janelamento utilizado foi a janela de Hamming. A janela de Hamming pode ser descrita pela Equação 3.

$$M_{segmental}(i) = \sum_{n=0}^{N-1} 0{,}54 \cdot |S(n)| - 0{,}46 \cdot cos(\frac{2\pi n}{L-1}) \tag{3}$$

2.4. Taxa de Cruzamento por Zeros

A Taxa de Cruzamento por Zeros (TCZ), é uma variável que ela identifica quantas vezes o sinal cruzou o valor zero, simplificando, é uma variável que determina quantas vezes o segmento variou o sinal de positivo para negativo, e quantas vezes ele variou do negativo para o positivo. A Equação 4 e 5 apresenta a formula da TCZ, de acordo com Passos, Silva, Costa, Neto e Fechine (2006).

$$TCZ(i) = \sum_{n=1}^{N-1} |sgn[s(n)] - sgn[s(n-1]| \tag{4}$$

No qual:

$$sgn[s(n)] = \begin{cases} 1 & , \ se\ s(n) \geq 0 \\ -1 & , \ se\ s(n) < 0 \end{cases} \tag{5}$$

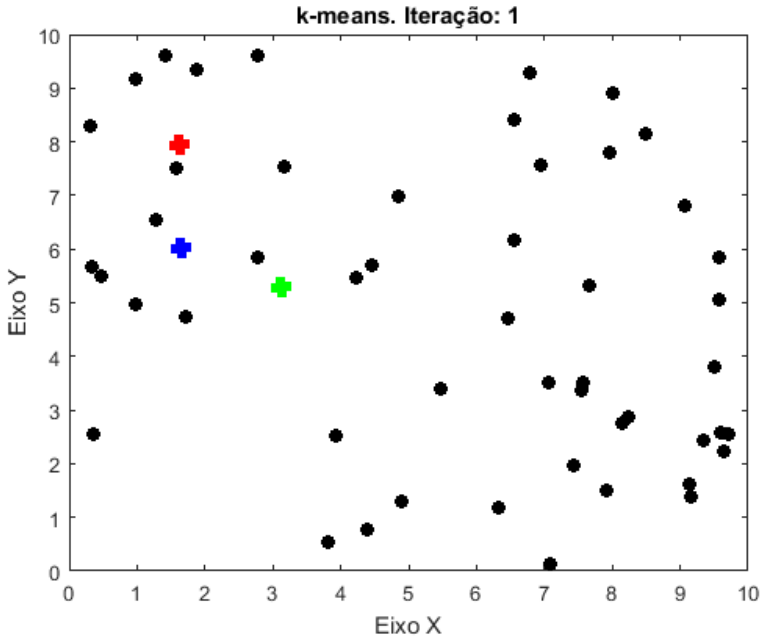
A Taxa de Cruzamento por Zeros é comumente utilizada para a detecção de sons surdos.

2.5. K-means

O K-means é um algoritmo classificador de baixo custo de processamento. O k-means agrupa determinadas características apresentadas ao programa em cluster utilizando ferramentas denominadas de centroides. Os centroides são pontos centrais inicializados em conjunto com os valores das características apresentadas ao programa.

Os clusters irão dividir entre si as características inseridas no algoritmo. A Figura 3 ilustra as características de um K-means utilizando características de valores randômicos.

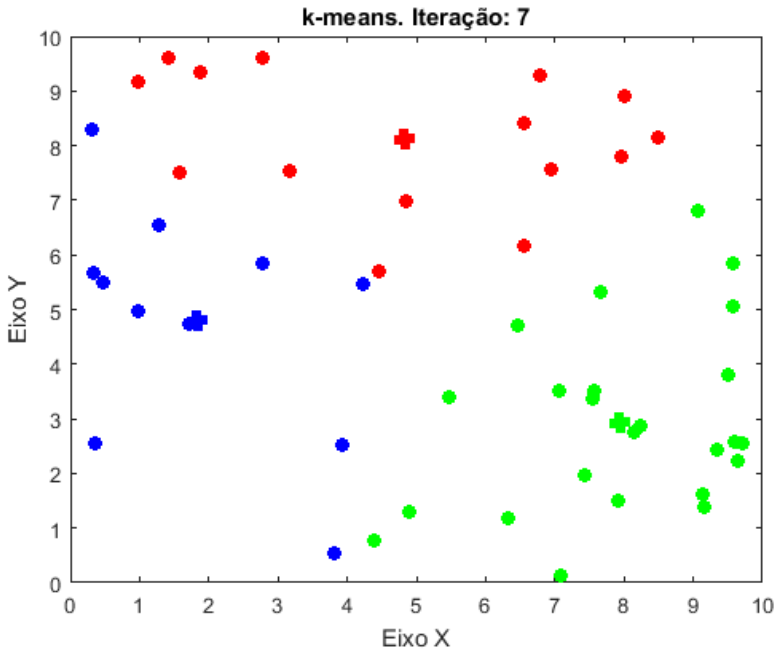
Figura 3: Representação dos centroides e do das características em um algoritmo K-means.



Fonte: autoria própria.

Os itens representados por bolinhas pretas são as características do sistema. Os itens coloridos em forma de cruz, são os centroides. A divisão das características entre os centroides ocorre por intermédio da distância Euclidiana entre os centroides e as características apresentadas. A Figura 4 apresenta a classificação das características apresentadas na Figura 3 após 7 iterações.

Figura 4: Classificação do K-means em características de valores aleatórios.



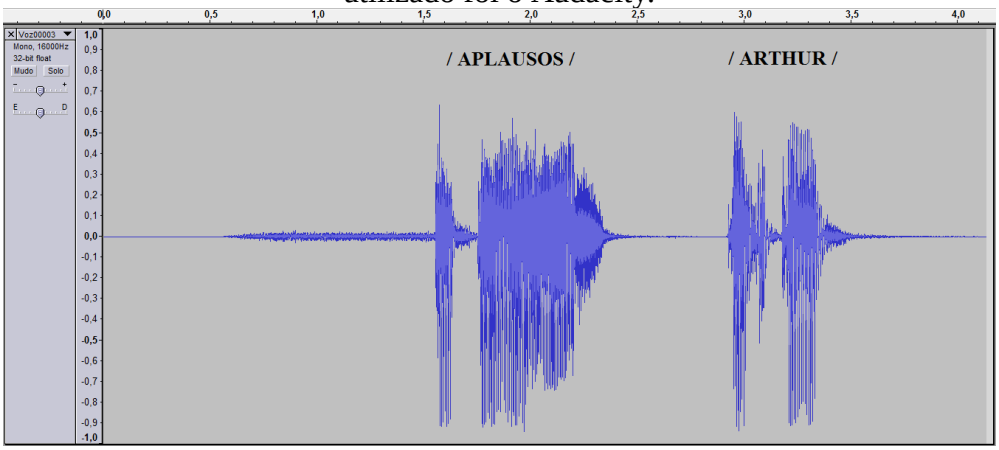
Fonte: autoria própria.

3. Metodologia

Para realizar este trabalho, foi gravado um sinal de áudio “aplausos”, seguido de uma breve pause de em seguida o primeiro nome do autor deste trabalho “Arthur”. O áudio teve duração de 4,1332 segundos. Para realizar gravação da voz, foi utilizado o microfone do celular LG 90 410hn. O formato gravado foi 3GP. O software Matlab foi utilizado para realizar o processamento do sinal de voz, porém o Matlab aceita apenas áudios no formato WAV. Diante disto, foi utilizado o site “https://audio.online-convert.com/convert-to-wav” para realizar a conversão do áudio para WAV. Neste site também é possível modificar a taxa de amostragem, a quantidade de bits do sinal, quantos canais o sinal vai ter e mais algumas características que não foram necessárias para este trabalho.

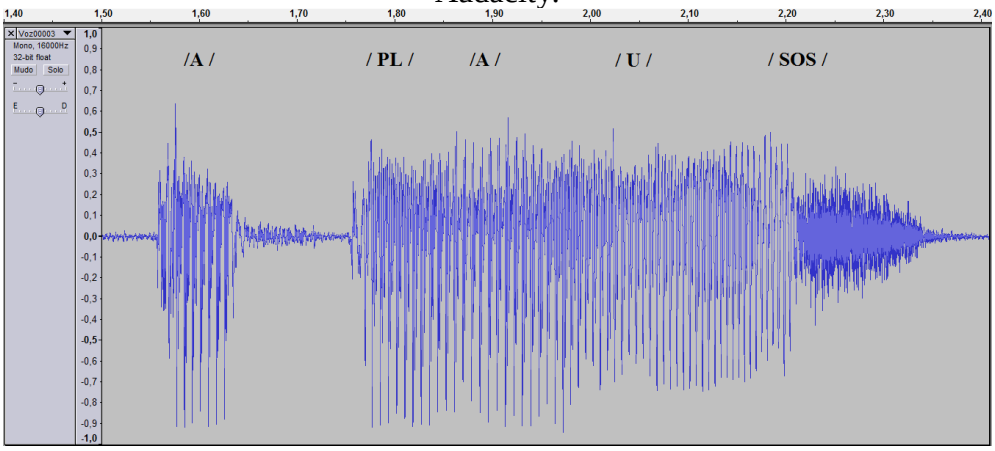
A taxa de amostragem utilizada foi de 16000 que é igual a $6{,}2500\,e^{-5}$ a/s (amostras por segundo). O sinal foi convertido para 16 bits e o canal de gravação do sinal foi mono. O sinal foi representado no software Audacity para realizar o entendido dos locais dos fonemas. A Figura 5, 6 e 7 representam os sinais de áudio utilizando o Audacity e os locais do fonema.

Figura 5: Representação de quais palavras foram faladas no sinal de áudio. O software utilizado foi o Audacity.



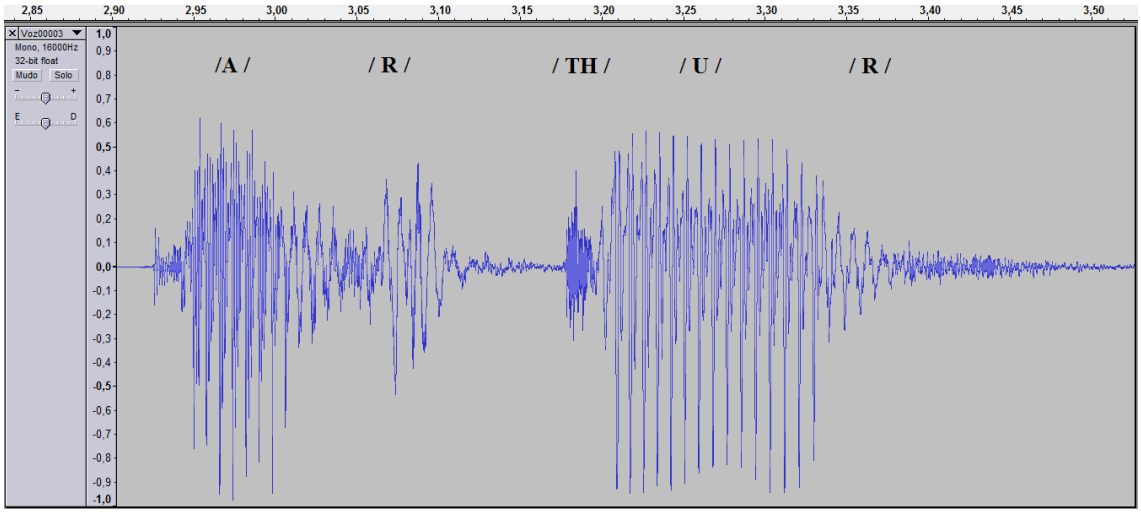
Fonte: autoria própria.

Figura 6: Representação os fonemas da palavra “aplausos”. O software utilizado foi o Audacity.



Fonte: autoria própria.

Figura 7: Representação os fonemas da palavra “aplausos”. O software utilizado foi o Audacity.



Fonte: autoria própria.

Apesar de ter configurado o site para ajudar converter o sinal para 16 bits, ao reproduzir o sinal no Audacity, a informação apresentada era que o sinal estava em 32 bits. Entretanto, ao utilizar a função “audioinfo” do Matlab, ele confirma que o sinal de áudio está 16 bits. Para realizar a extração de características, foram adotados frames de 20 ms e uma sobreposição de 50%. As características foram alocadas em uma variável struct do Matlab para facilitar o processamento de dados e a procura manual por valores dentro do Matlab. A variável struct foi denominada de frames. Nos frames foram agrupadas 5 variáveis: frame.time, frame.amp, frame.energ, frame.magnet e frame.TCZ .

- frame.time: nesta variável foi armazenado os valores do tempo de cada frame;
- frame.amp: nesta variável foi armazenado a amplitudade de cada frame;
- frame.energ: nesta variável foram armazenados os valores de energia de cada frame;
- frame.magnet: nesta variável foi armazenado os valores da magnitude de cada frame;
- frame.TCZ: nesta variável é armazenado os valores da taxa de cruzamento por zeros de cada frame.

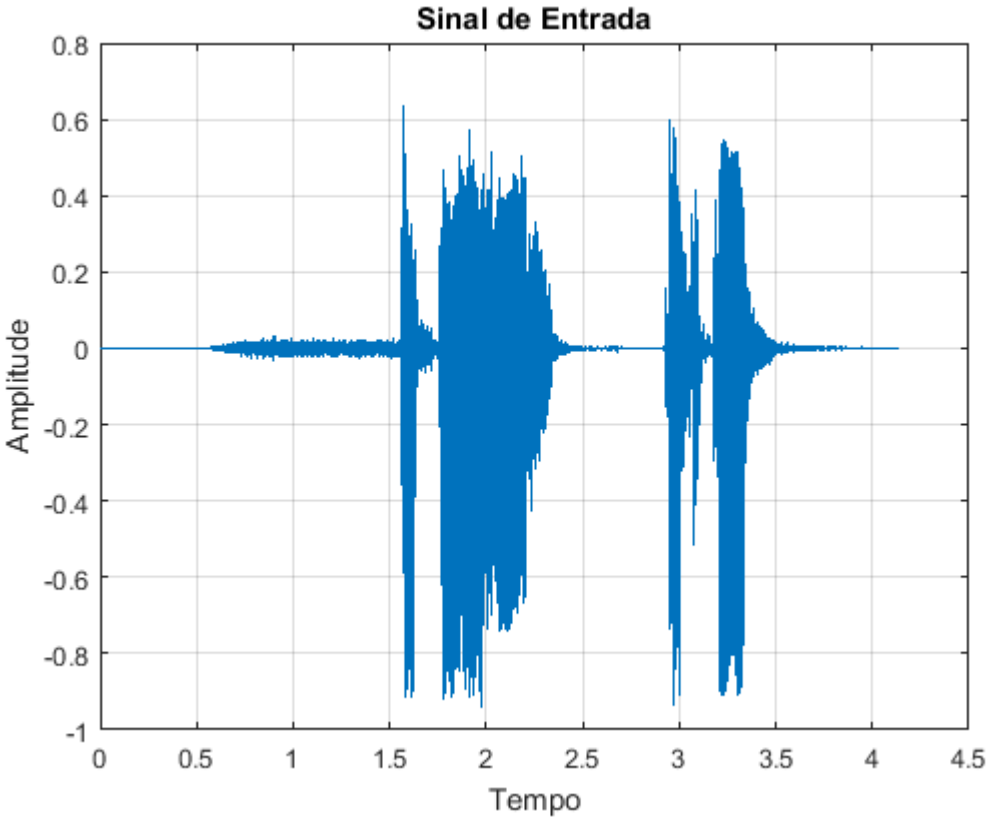
O código do k-means foi recriado do início com base em seu funcionamento teórico. Esse processo foi realizado para ter um maior domínio acerca da manipulação da função do k-means dentro do sinal de voz. Para estudar a influencia dos centroides no código, o programa teve seus centroides variados de 2 até 5 centroides. Este estudo serve para identificar a influencia da escolha da quantidade de centroides na resultante final classificada pelo k-means.

Com o intuído de estudar a importância das características para o resultado final do k-means, o algoritmo foi aplicando utilizando dois diferentes grupos de características energia e magnitude, e energia e TCZ. Este estudo serve para comprovar que a escolha das características a serem utilizadas no programa, são de fundamental importância para determinar o agrupamento dos centroides.

4. Resultados

A função “audioread” do Matlab foi utilizada para que fosse realizada a leitura do sinal de áudio no Matlab. A Figura 8 apresenta o sinal de áudio lido pelo Matlab.

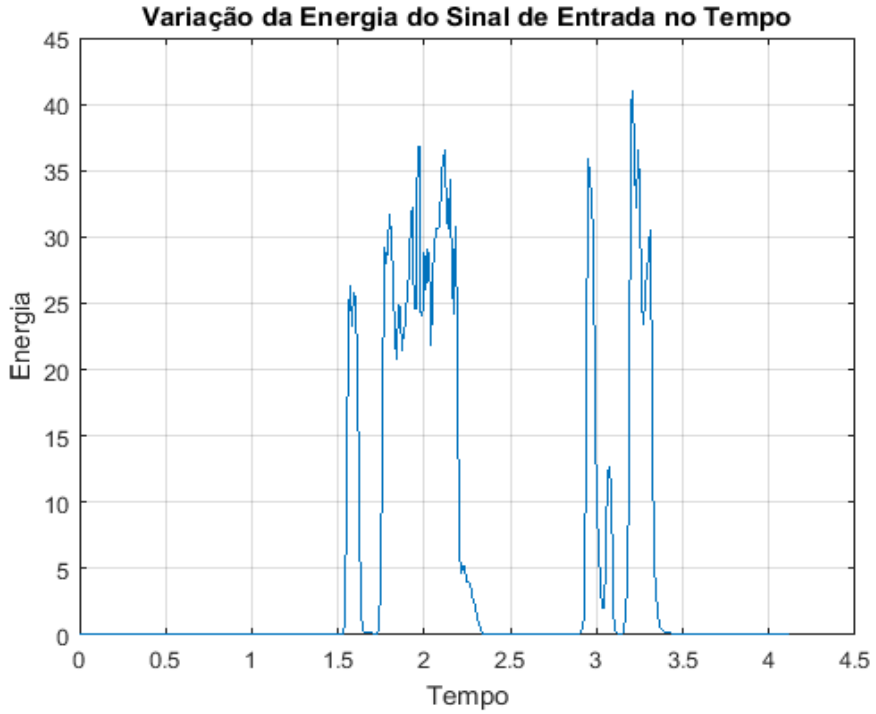
Figura 8: Sinal de áudio da fala "Aplausos" e "Arthur" representado no software Matlab.



Fonte: autoria própria.

A energia segmental foi calculada utilizando a formula da Equação 1 do tópico “2.3 Energia Segmental” e os valores foram alocados na variável frame.energ, como os valores da energia segmental é apenas um valor por frame, para plotar o gráfico da energia segmental é necessário criar um vetor de referência contendo os instantes no tempo em que cada frame inicia. A Figura 9 apresenta o gráfico da energia segmental do sinal de áudio deste trabalho.

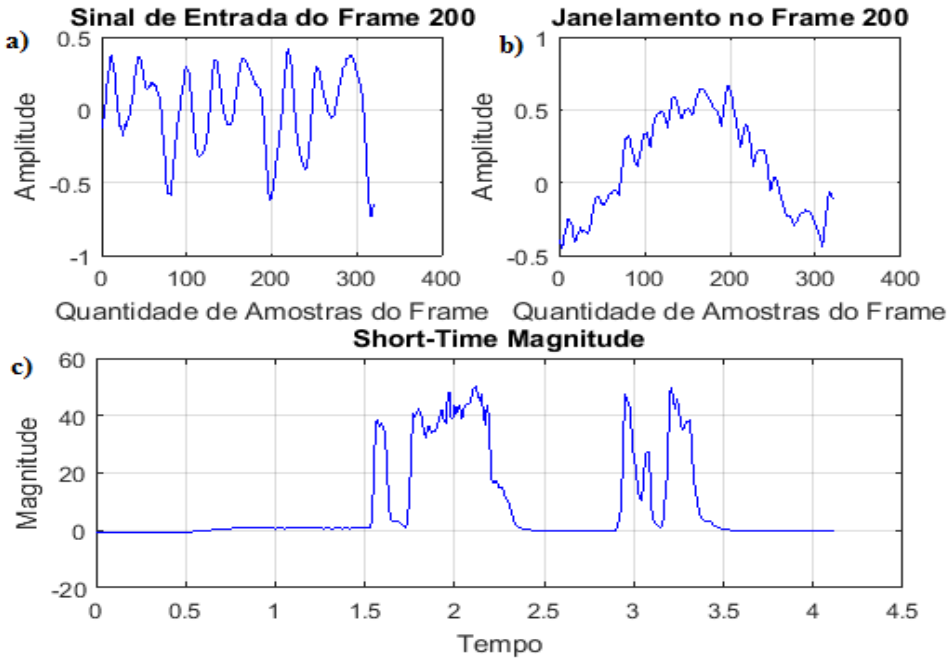
Figura 9: Gráfico da energia segmental do sinal de áudio "aplausos" e "Arthur". Foi utilizado o software Matlab para criar o gráfico.



Fonte: autoria própria.

Em seguida ocorreu extração da característica da magnitude do sinal de áudio utilizado neste trabalho. A Figura 10 apresenta o gráfico da magnitude do sinal.

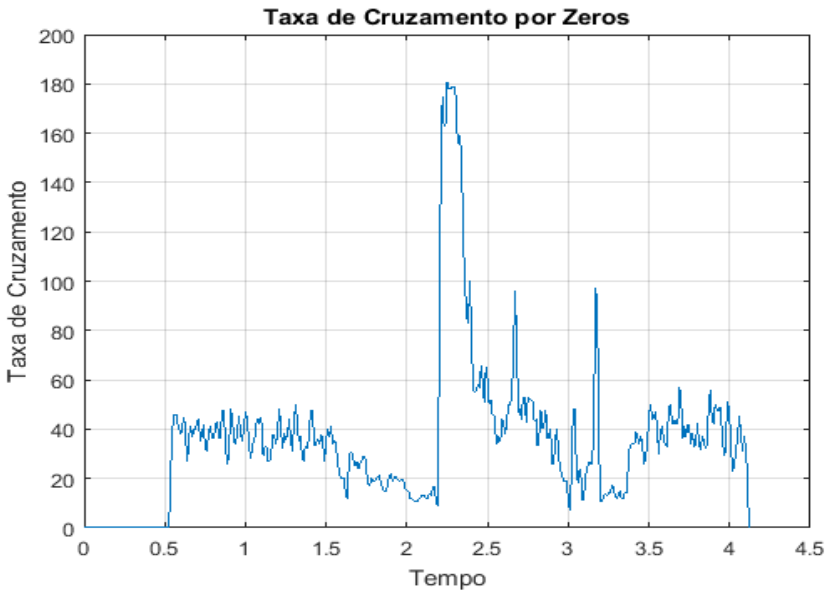
Figura 10: Construção da magnitude do sinal utilizando a janela de Hamming. a) Imagem dos valores das amostras do frame 200. b) Aplicando a técnica das janelas de Hamming no frame 200. c) Magnitude segmental de todo o sinal.



Fonte: autoria própria.

A última característica extraída foi a Taxa de Cruzamento por Zeros. A Figura 11, apresenta o gráfico da TCZ no sinal de entrada utilizado neste estudo.

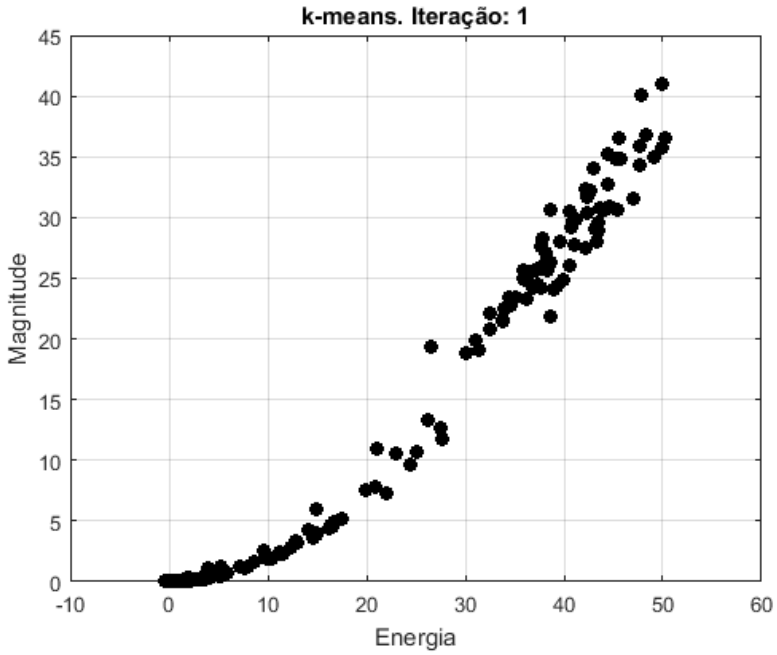
Figura 11: Taxa de Cruzamento por Zeros do sinal de áudio "Aplausos" e "Arthur". O software utilizado foi o Matlab.



Fonte: autoria própria.

Após extrair as características de energia, magnitude e taxa de cruzamento por zeros, foi aplicado seus valores no código k-means em dois tipos de permutações utilizando as três características extraídas. A primeira analise realizada é energia pela magnitude. Esta analise com 2 centroides, e depois foram aumentando a quantidade de centroides até o total de 5 centroides. A Figura 12 exemplifica a disposição das características da energia pela magnitude.

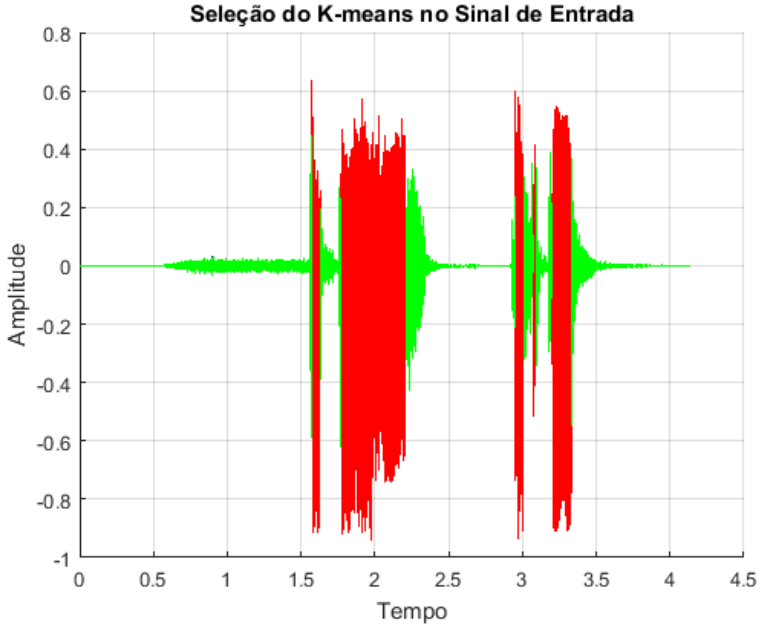
Figura 12: disposição das características da energia pela magnitude de um sinal de voz no k-means.



Fonte: autoria própria.

Estas características são separadas pelo k-means dependendo de quantos centroides foram definidos no programa. Na Figura 13 é apresentado os resultados da analise energia por magnitude utilizando 2 centroides.

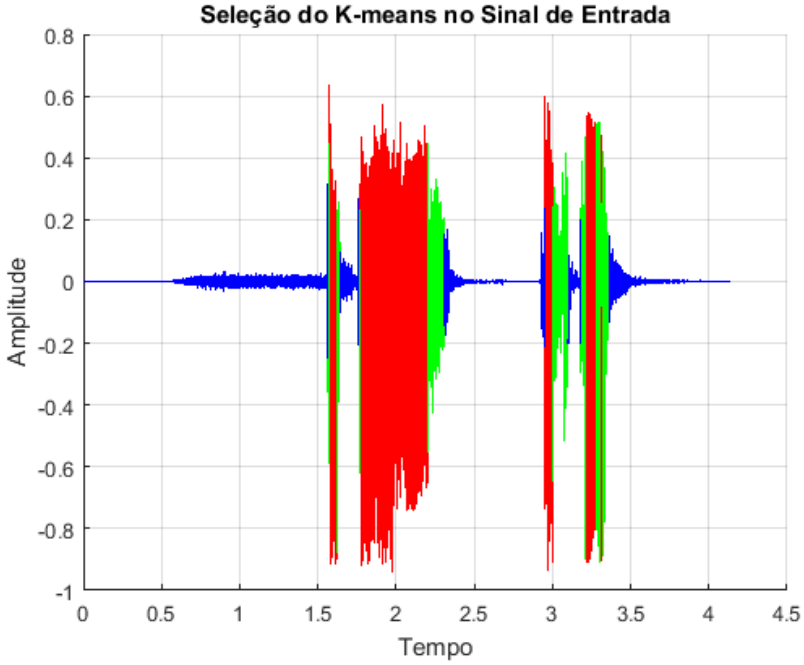
Figura 13: seleção do k-means para energia e magnitude utilizando dois centroides.



Fonte: autoria própria.

Analisando o gráfico da Figura 13 visualmente, é possível perceber que os centroides separaram o sinal em duas partes distintas. O centroide vermelho constitui a maioria dos espaços que possuem as maiores intensidades do sinal de entrada. O centroide verde, constitui a região de sons surdos e silenciosos, além dos inícios das pronuncias das regiões de maiores intensidades do sinal. Este mesmo teste da energia pela magnitude foi repetido, porém utilizando 3 centroides.

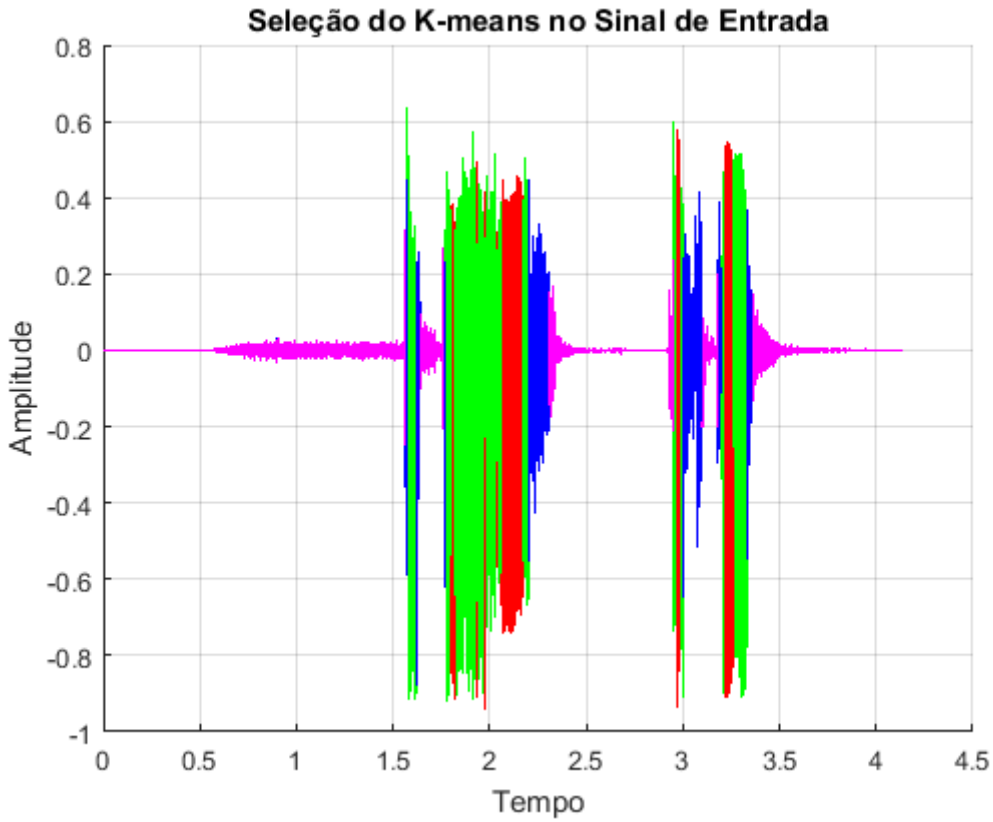
Figura 14: seleção do k-means para energia e magnitude utilizando três centroides.



Fonte: autoria própria.

Analisando o gráfico da Figura 14 visualmente, é possível perceber que os centroides separaram o sinal em duas partes distintas. O centroide de cor vermelha, selecionou as regiões mais audíveis do sinal de entrada. O centroide de cor verde, selecionou os sons surdos e sinais de baixa intensidade da voz. O centroide azul selecionou em sua maioria, as regiões que possuem silencio ou ruído do sinal. Este mesmo teste da energia pela magnitude foi repetido, porém utilizando 4 centroides.

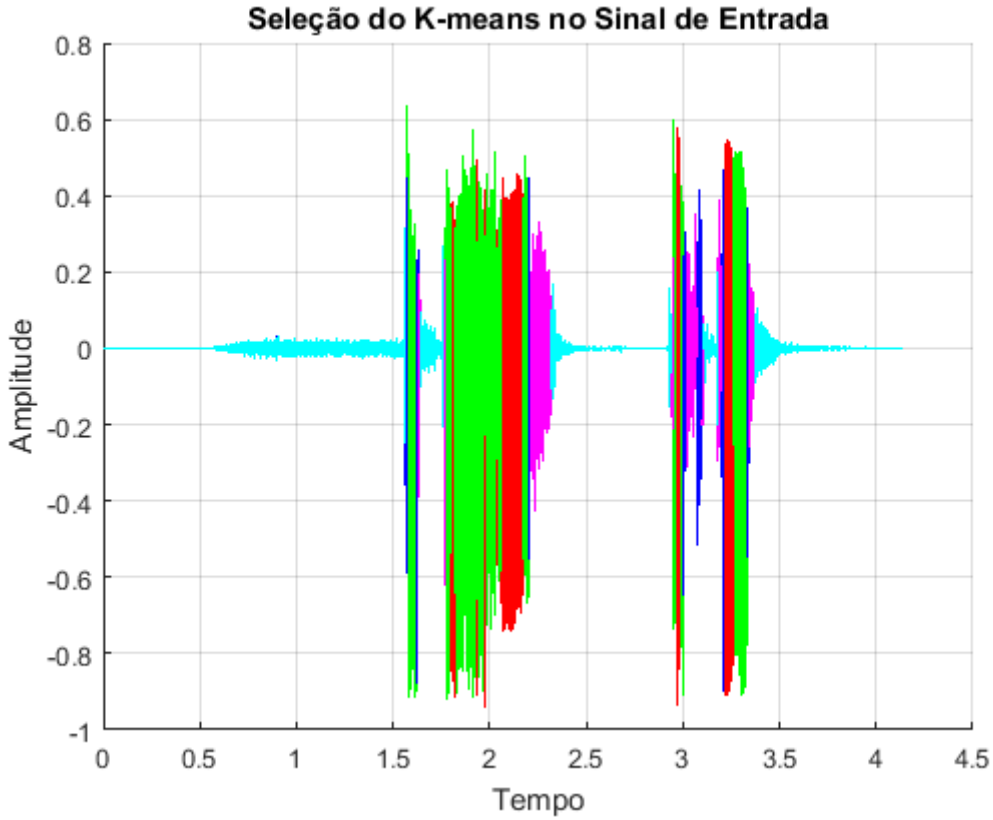
Figura 15: seleção do k-means para energia e magnitude utilizando quatro centroides.



Fonte: autoria própria.

A seleção realizada pelos centroides da Figura 15, não são tão distintas quanto os testes realizados com 2 ou 3 centroides. Neste experimento alguns centroides constituem de regiões bem definidas em sua teoria, como o centroide de cor magenta, que selecionou em sua maioria os ruídos e a região de silencio do sinal. O centroide azul selecionou em sua maior parte, as regiões contendo sons surdos. Entretanto, os centroides verde e vermelho, competiram pelas regiões de alta intensidade da voz, as regiões audíveis que não constituem sons surdos. Este mesmo teste da energia pela magnitude foi repetido, porém utilizando 5 centroides.

Figura 16: seleção do k-means para energia e magnitude utilizando cinco centroides.



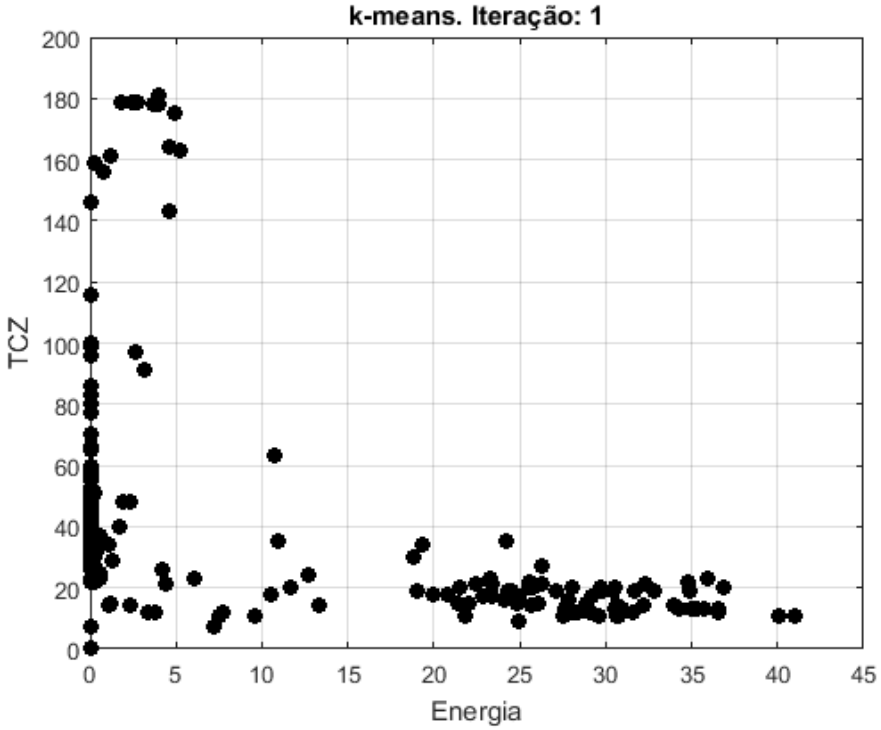
Fonte: autoria própria.

A seleção realizada pelos centroides da Figura 16, não são tão distintas quanto os testes realizados com 2 ou 3 centroides, e quase não houve mudanças em relação aos experimentos realizados com 4 centroides. Neste experimento alguns centroides constituem de regiões bem definidas em sua teoria, como o centroide de cor ciano, que selecionou em sua maioria os ruídos e a região de silencio do sinal. O centroide azul disputou com o centroide magenta pelas regiões contendo sons surdos. Entretanto, os centroides verde e vermelho, competiram pelas regiões de alta intensidade da voz, as regiões audíveis que não constituem sons surdos.

Ao analisar todos os 4 experimentos, é possível concluir que nem sempre o aumento da quantidade de centroides vai lhe proporcionar uma categorização bem definida das características do sistema. Os experimentos com 2 e 3 centroides, foram os que categorizaram melhor as características do sistema, a diferença estava no fato de que o experimento com 2 centroides realizou uma características mais genérica do que o com 3 centroides, isto porque a seleção separou as regiões que são sonoras, e as que não são sonoras. Entretanto o experimento com 3 centroides além de separar as regiões sonoras, também separou as regiões de sons surdos das regiões de silencio e ruído.

Estes mesmos experimentos foram realizados com a análise da característica de energia e magnitude. A Figura 17 exemplifica a disposição das características da energia pela TCZ.

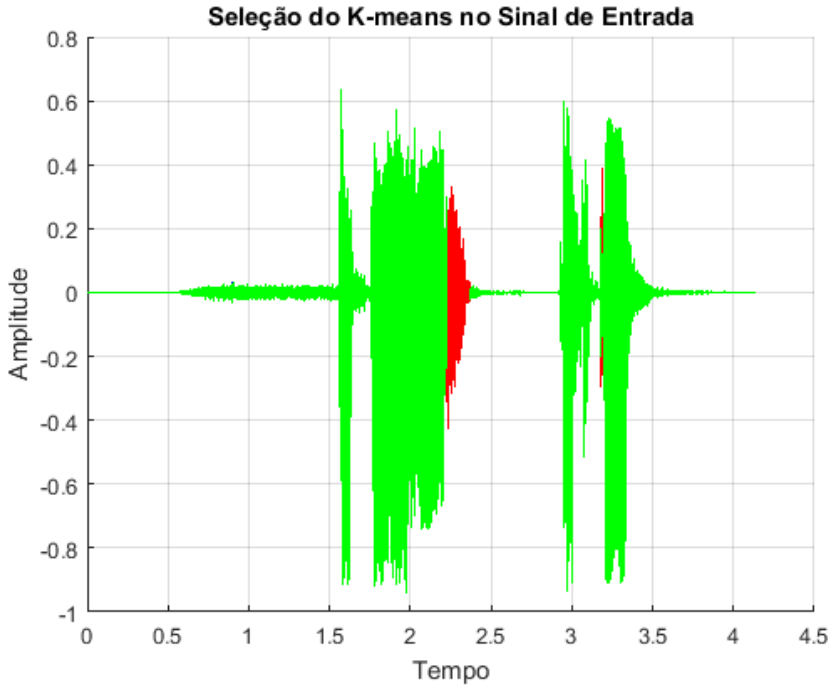
Figura 17: disposição das características da energia pela Taxa de Cruzamento por Zeros de um sinal de voz no k-means.



Fonte: autoria própria.

Estas características são separadas pelo k-means dependendo de quantos centroides foram definidos no programa. Na Figura 18 é apresentado os resultados da analise energia por magnitude utilizando 2 centroides.

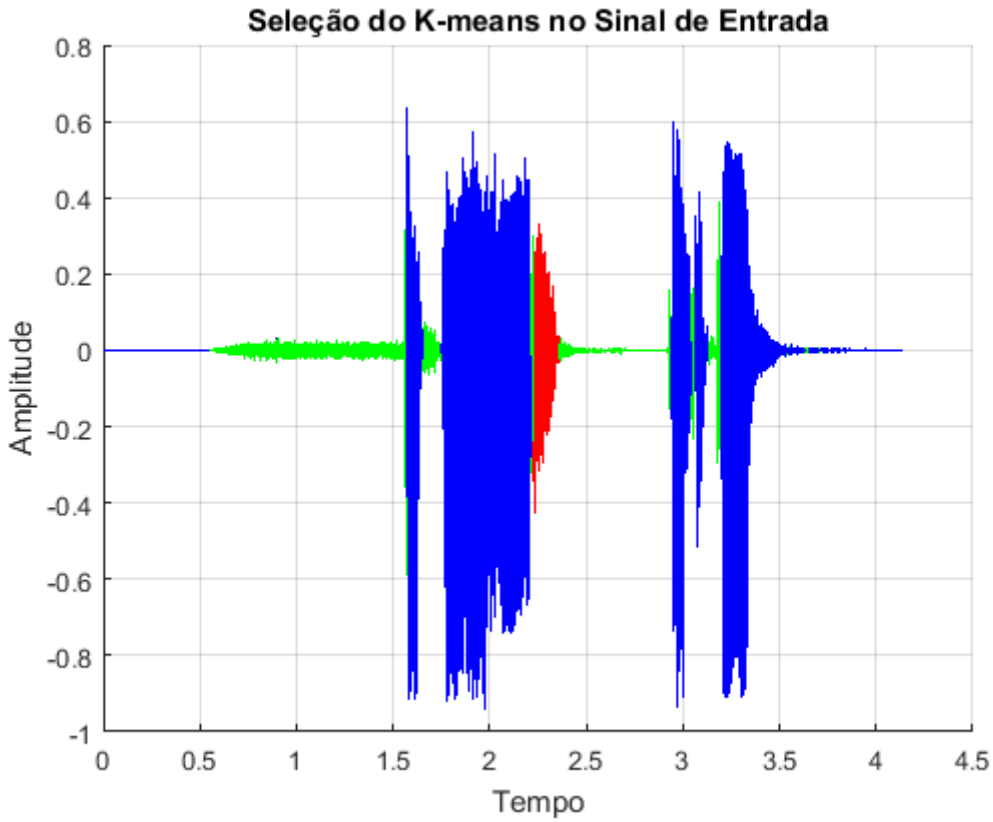
Figura 18: seleção do k-means para energia e TCZ utilizando dois centroides.



Fonte: autoria própria.

No experimento da Figura 18, os centroides conseguiram isolar quase que por completo apenas a região do som surdo “s” de “aplausos” no centroide vermelho. Esta seleção pode ser considerada boa, pois é uma classificação bem definida para isolar uma única característica de um sinal de voz. Entretanto, é perceptível que no inicio da pronuncia do “t” em Ar”t”hur, houve um pequeno erro na seleção das características. Este mesmo teste da energia pela TCZ foi repetido, porém utilizando 3 centroides.

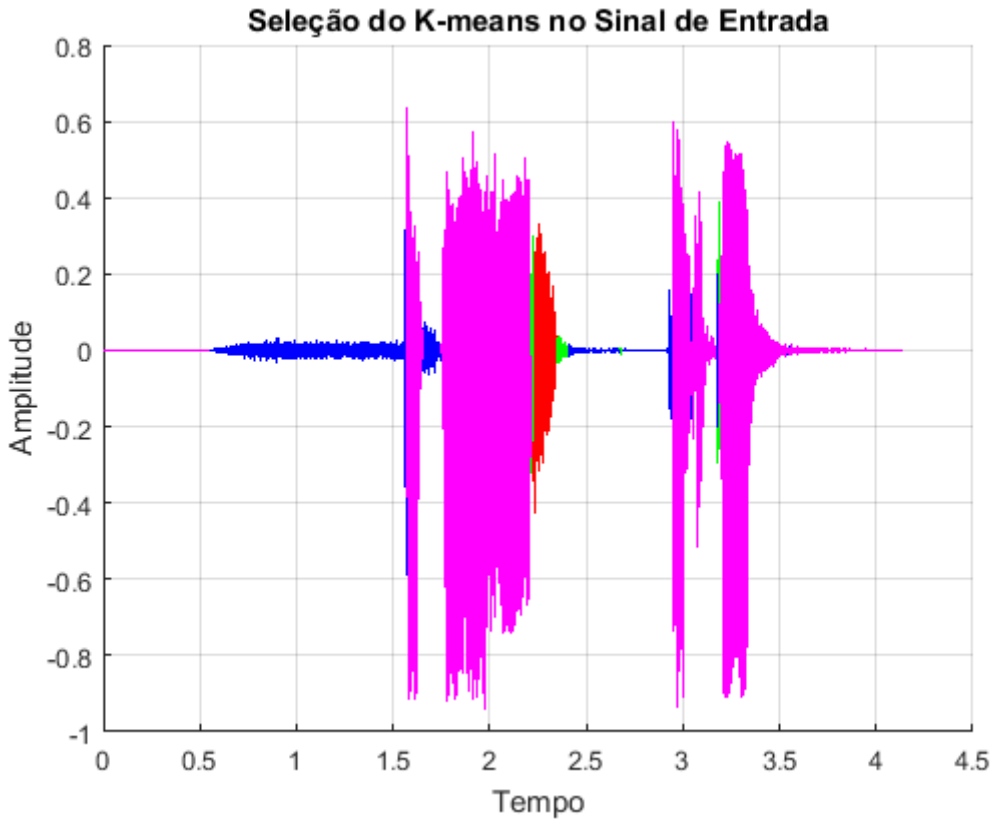
Figura 19: seleção do k-means para energia e TCZ utilizando três centroides.



Fonte: autoria própria.

A seleção dos elementos do que compõe o sinal de áudio realizado pelo experimento da Figura 19, conseguiu classificar regiões bem definidas. O centroide azul separou a região sonora de maior intensidade de sinal do programa, porém ele também classificou alguns ruídos de baixa intensidade no final do áudio, como também o silencio inicial do sinal. O centroide verde, separou os ruídos de maior intensidade e algumas transições entres as palavras. O centroide vermelho conseguiu classificar a região do som surdo “s” de aplauso”s”. Este experimento também foi realizado com 4 centroides.

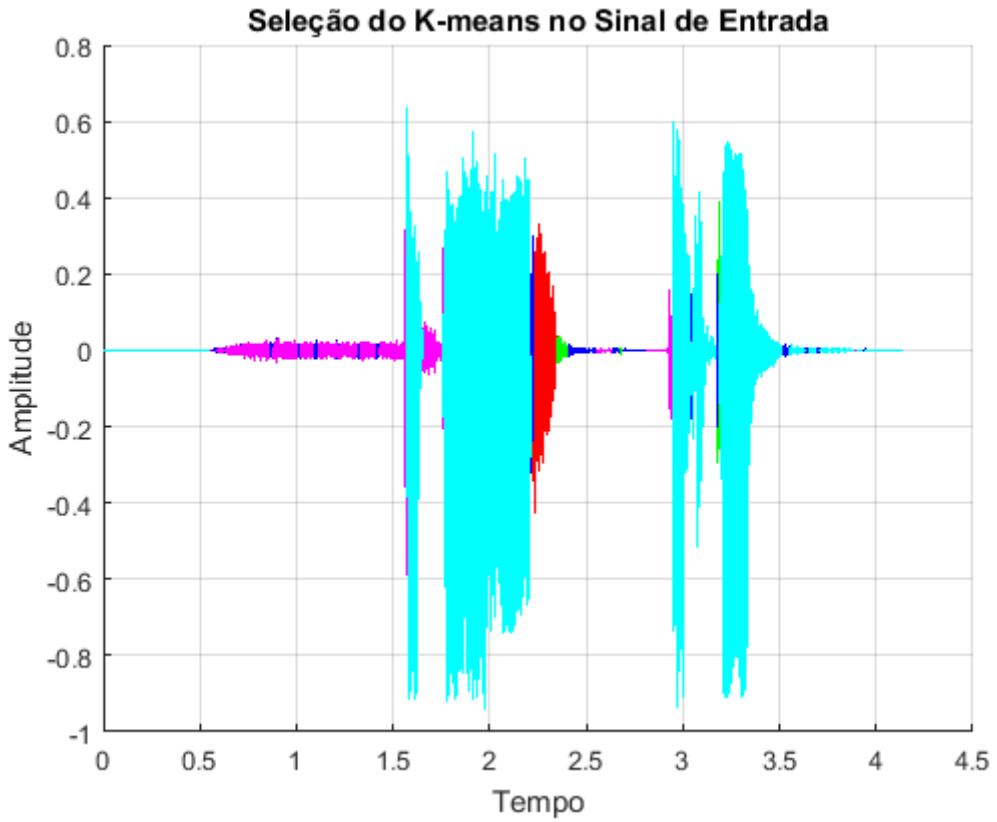
Figura 20: seleção do k-means para energia e TCZ utilizando quatro centroides.



Fonte: autoria própria.

O experimento da Figura 20, não obteve muita diferença em relação ao experimento da Figura 19, pois as regiões definidas pelo sistema são em sua maioria, idênticas, isto ocorreu pelo fato de o centroide verde selecionou poucos frames, de tal forma de que ele apresentou pouca influência na classificação total dos outros três centroides. O centroide azul classificou em sua maioria os ruídos de maior intensidade e algumas transições entres as palavras. O centroide de cor magenta selecionou a maior parte do sinal sonora do sinal, porém, incluindo o ruído de baixa intensidade após a palavra Arthur. O centroide vermelho selecionou em sua maioria o som surdo “s” de aplauso”s”. Este experimento também foi realizado com 5 centroides.

Figura 21: seleção do k-means para energia e TCZ utilizando quatro centroides.



Fonte: autoria própria.

O experimento da Figura 21, não obteve muita diferença em relação ao experimento da Figura 20, pois as regiões definidas pelo sistema são em sua maioria, idênticas, isto ocorreu pelo fato de o centroide verde e azul disputaram pela seleção de poucos frames, de tal forma de que ele apresentou pouca influência na classificação total dos outros três centroides. O centroide magenta classificou em sua maioria os ruídos de maior intensidade e algumas transições entres as palavras, com disputa entra a classificação como centroide azul em alguns momentos. O centroide ciano selecionou a maior parte do sinal sonora do sinal, porém, incluindo o ruído de baixa intensidade após a palavra Arthur em conjunto com o centroide azul. O centroide vermelho selecionou em sua maioria o som surdo “s” de aplauso”s”.

Este experimento demonstra que o aumento de centroides pode ser considerado irrelevante em algumas circunstancias. Pois o experimento com 4 e 5 centroides não apresentou mudanças significativas em relação a classificação com 3 centroides. A classificação com 2 e 3 centroides foram as mais concisas entre os experimentos. A classificação com 3 centroides classifica bem 3 regiões do sinal de áudio, similar ao experimento da energia pela magnitude com 3 centroides da Figura (procure). A classificação com 2 centroides foi a mais concisa, pois conseguiu extrair uma característica em especifico de todo o sinal de áudio, apresentando assim uma solução fácil de implementar e de baixo custo de processamento para a extração de uma característica em particular do sinal de áudio, o dos sons surdos.

5. Conclusão

Neste trabalho foi apresentado a extração de características de um sinal de áudio e a seleção de elementos do áudio pelo algoritmo classificador K-means. Por intermédio das características do sinal de áudio é capaz de detectar componentes do sinal que são correlacionados com a fala. O algoritmo K-means foi utilizado para detectar componentes não especificadas a ele. Foram apresentadas as características ao algoritmo, e ele classificaria as características em grupos. Neste projeto as características apresentadas ao programa, visam estudar a importância da seleção das características a serem apresentadas ao algoritmo k-means para que haja a extração dos elementos almejados pelo usuário. Também foi variado a quantidade de centroides do algoritmo para perceber a influencia deles para o algoritmo.

Com os testes finalizados, é possível concluir que as características apresentadas ao programa, definem os elementos que serão classificados pelo algoritmo, além de que o modo que o algoritmo classifica as características do sinal de entrada também é diferente. Em relação a variação dos centroides, é possível concluir que o aumento do número de centroides não significa que o K-means irá alcançar melhores resultados, é necessário estudar quais elementos são almejados pelo usuário, para definir quantos centroides irão lhe proporcionar os melhores resultados.

6. Bibliografia

[1] PROAKIS, J.G.; INGLE, V.K.; ***Digital Signal Processing using MATLAB***, Cengage Learning, 2012, pág: 1 e 2.

[2] KANUNGO, T; et al; ***An efficient K-means clustering algorithm: analysis and implementation***, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, 2002.

[3] PASSOS, M.G.; SILVA, P.L.; COSTA, S.L.d.N.C.; NETO, B.G.A.; FECHINE, J.M.; ***Um ambiente para Processamento Digital de Sinais Aplicado a Comunicação Vocal Homem-Máquina***, PRINCIPIA, João Pessoa, 2016.