

PARAMETRIZAÇÃO DE SINAIS DE FALA UTILIZANDO O MFCC

Wellington Freitas¹

Wellington Lopes de Freitas²

Fabiano Tondello Castoldi³

Resumo:

Atualmente existe um crescimento de aplicações que realizam tarefas por comandos de fala utilizando sistemas de reconhecimento automático de fala (AutomaticSpeech Recognition -ASR). Exemplos de uso desse tipo de aplicações são os aplicativos conhecidos como Google now, Apple Siri e Microsoft Cortana. Um sistema ASR é composto por diversos subsistemas como: sistema de processamento de sinais, sistema de decodificação, sistema de modelagem acústica etc. Em especial o bloco de processamento de sinais é responsável pela parametrização e extração de características linguísticas utilizadas para classificar um sinal de fala. Existem hoje diversas técnicas para a realização desta tarefa, as quais aproveitam-se de mecanismos de produção e compreensão humana da fala para a implementação do bloco de processamento de sinais. Atualmente a técnica que possui o maior sucesso para essa finalidade são os coeficientes cepstrais de frequência Mel (Mel-Frequency Cepstral Coefficients -MFCC). Baseado nessas considerações, este trabalho tem por objetivo apresentar o projeto de todos os submódulos que compõem um subsistema de processamento de sinais em um sistema ASR utilizando a técnica MFCC para extração de parâmetros de sinais de fala relevantes para sua futura classificação.

Palavras-chave: Processamento de Fala, Extração de Parâmetros, MFCC

Modalidade de Participação: Iniciação Científica

PARAMETRIZAÇÃO DE SINAIS DE FALA UTILIZANDO O MFCC

¹ Aluno de graduação. wellington.frts@gmail.com. Autor principal

² Aluno de graduação. wellington.frts@gmail.com. Apresentador

³ Docente. fabianocastoldi@unipampa.edu.br. Orientador

PARAMETRIZAÇÃO DE SINAIS DE FALA UTILIZANDO O MFCC.

1 INTRODUÇÃO

A comunicação por fala tem sido a maneira dominante com a qual os humanos interagem socialmente e trocam informações. Esse tipo de comunicação, hoje, estendeu-se por meio da tecnologia através do telefone, rádio, cinema, televisão e internet (HUANG et al., 2001). Atualmente, com emprego de técnicas de processamento de fala em computadores, dispositivos embarcados e smartphones, já existe a possibilidade de se realizar tarefas por comandos de fala utilizando sistemas de reconhecimento automático de fala (*Automatic Speech Recognition -ASR*). Exemplos de uso desse tipo de aplicações são os aplicativos conhecidos como *Google now*, *Apple Siri* e *Microsoft Cortana*.

Um sistema ASR é composto por diversos subsistemas como: sistema de processamento de sinais, sistema de decodificação, sistema de modelagem acústica etc. Em especial o bloco de processamento de sinais é responsável pela parametrização e extração de características linguísticas utilizadas para classificar um sinal de fala. Este procedimento é extremamente importante para o sucesso de uma aplicação ASR, visto que, a fala é um sinal de natureza complexa, influenciada por duas características distintas, que se não tratadas corretamente podem afetar o desempenho geral do sistema (HUANG et al., 2001): as dependentes de locutor e as independentes de locutor.

As características dependentes de locutor podem ser biológicas de cada indivíduo como: idade, sexo, composição biológica do aparelho vocal (BENESTY et al., 2008), ou de cunho pessoal e cultural como sotaque e pronúncia das palavras (SANTOS, 2013). Já as componentes independentes de locutor são denominadas formantes e possuem características ligadas ao processo de formação de sons vozeados, que por sua vez estão relacionados a pronúncia de vogais em palavras. Dependendo da pronúncia de cada palavra pode-se ter diversos padrões de formantes em um idioma (SANTOS, 2013). Entretanto, a formação da fala também conta com processos de formação de sons não vozeados, geralmente relacionados a consoantes fricativas, que são interpretados como ruído (SANTOS, 2013).

Existem hoje diversas técnicas de parametrização e extração de características de classificação linguística, as quais aproveitam-se das características acústicas dos formantes para a implementação do bloco de processamento de sinais.

Atualmente a técnica que possui o maior sucesso em aplicações ASR são os coeficientes cepstrais de frequência Mel (*Mel-Frequency Cepstral Coefficients – MFCC*) (HUANG et al., 2001). Esta técnica baseia-se em experimentos que visam compreender os mecanismos humanos de produção e percepção dos formantes para uma correta classificação linguística do sinal.

Baseado nessas considerações, este trabalho tem por objetivo apresentar o projeto de todos os submódulos que compõem um subsistema de processamento de sinais utilizando a técnica MFCC para extração de parâmetros de sinais de fala. O resto do artigo segue a seguinte disposição: a seção 2 apresenta o funcionamento do subsistema de processamento de sinais, descrevendo cada submódulo que o compõe; na seção 3 o resultado e considerações sobre o desempenho são realizadas; por fim, considerações finais sobre o trabalho desenvolvido são apresentadas na seção 4.

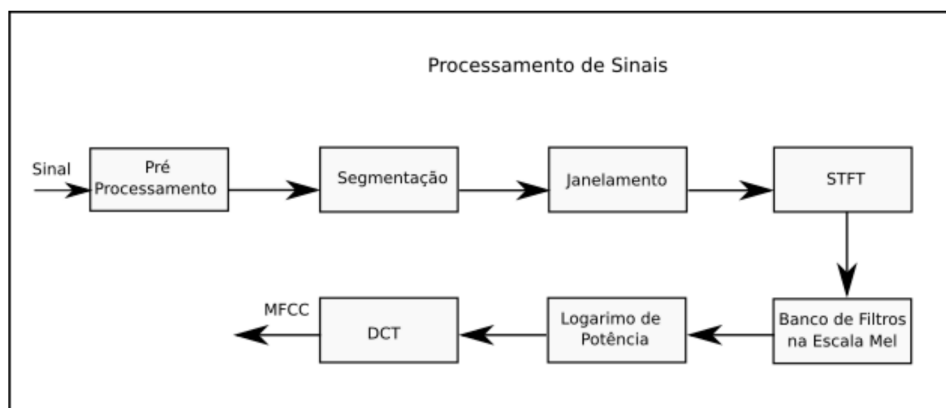
2 METODOLOGIA

O processamento de sinais utilizando neste trabalho é descrito pelo diagrama de blocos da Figura 1. O trabalho foi desenvolvido utilizando a linguagem de programação Python® devido a sua agilidade e robustez. No diagrama de blocos ilustrado na Figura 1, o bloco de pré processamento representa a primeira etapa do processo e consiste na aquisição, conversão analógico-digital (A/D) do sinal de entrada e a realização de um processo de pré filtragem para redução de ruídos. O *software* Audacity® foi utilizado para a implementação desta etapa.

A conversão (A/D) foi executada a partir da gravação de um sinal de fala no formato .WAV, utilizando um microfone do tipo *headset* direcional, com apenas um canal de

som (mono), com frequência de amostragem de 16 kHz e 16 bits de quantização, como recomenda a literatura (HUANG *et al.*, 2001). A faixa de áudio utilizada consiste na gravação da palavra “Unipampa”.

Figura 1- Diagrama de blocos do módulo de processamento de sinais.



Fonte: do Autor, 2018.

Os dois próximos blocos realizam a segmentação do sinal e o janelamento, utilizando uma janela de Hamming, visando segmentar o sinal em quadros de duração de 25 milissegundos (ms) com sobreposição de quadro a cada 15 ms. Executa-se esse procedimento para evitar a perda de informações vitais para o reconhecimento das falas durante o cálculo do espectro do sinal, pois os formantes tem natureza não estacionária, ou seja, variam no tempo. O cálculo direto do espectro do sinal obteria apenas suas características estacionarias (BENESTY *et al.*, 2008).

O próximo passo é calcular o espectro do sinal por meio da transformada de Fourier em tempo curto (*Short-Time Fourier Transform – STFT*) ou seja, aplicar a transformada discreta de Fourier quadro à quadro. Após o cálculo do espectro do sinal inicia-se o procedimento de extração dos parâmetros desejados para classificação linguística.

Para tal, utilizam-se duas características apresentadas por (MERMELSTEIN, 1976) em estudos que mapeiam a percepção humana a variações de frequência do som. A primeira característica mostra que percepção humana ao incremento de frequência sonora comporta-se linearmente até 1000 Hz, acima disso, torna-se logarítmica. A escala que mapeia fenômeno é denominada escala Mel (MERMELSTEIN, 1976).

A segunda característica mostra que o aparelho auditivo humano não é capaz de distinguir dois tons de frequências próximas e é perceptualmente mais discriminativo em baixas frequências, ou seja, com o incremento da frequência, a diferença de frequência entre dois tons necessita ser cada vez maior para que haja a diferenciação entre ambos.

Este comportamento pode ser reproduzido artificialmente utilizando um banco de filtros do tipo passa-faixas com suas frequências centrais e banda de passagem distribuídas linearmente na escala Mel. Dentre as diversas respostas em frequência possíveis para a implementação do banco de filtros, a função triangular é a mais popular, pois, estudos indicam que esta função é a que possui resposta em frequência que melhor imita a percepção do ouvido humano (HUANG *et al.*, 2001). Neste trabalho utilizou-se o total de 26 coeficientes, como recomendado pela literatura (HUANG *et al.*, 2001), resultando em um vetor com 26 elementos.

Após o processo de filtragem, realiza-se o cálculo do logaritmo da potência do sinal filtrado em escala Mel. Esta operação é utilizada para facilitar o processo de separação dos formantes do ruído. Devido a características da pronuncia de uma palavra, a interação de um formante com o ruído tem como característica ser a convolução dos dois sinais no domínio do tempo (BENESTY *et al.*, 2008).

Ao calcular o espectro desse sinal tem-se a transformação da convolução em uma multiplicação dos dois sinais. Aplicação do logaritmo transforma o esta operação em uma soma dos dois sinais, facilitando a operação de separação, visto que, agora o sinal de fala será composto pelos formantes e um ruído aditivo (BENESTY et al., 2008).

Por fim, o vetor de parâmetros *MFCC* são obtidos a partir da transformação o sinal novamente para o domínio do tempo. Para tal, aproveita-se da característica dos formantes serem componentes do sinal com variação mais lenta e uma maior concentração de energia nas frequências mais baixas, e o ruído gerado pelos sons não vozeados terem variação mais rápida, e uma maior concentração de energia nas frequências mais altas (HUANG et al., 2001).

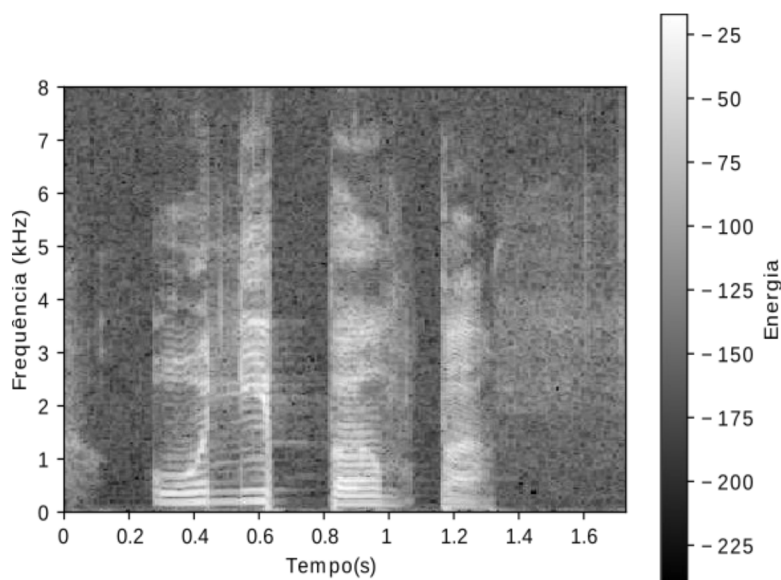
Pode-se tirar proveito desta característica para utilizando a transformada discreta do cosseno (Discrete Cossine Transform – DCT) para efetuar a transformação do sinal de fala de volta para o domínio do tempo. A DCT tem por característica armazenar os sinais de variação mais lenta nos primeiros elementos do vetor resultante, separando os formantes do ruído.

Assim, a DCT irá gerar os coeficientes MFCC armazenando os formantes nos elementos de menor valor no vetor, resultante e o ruído nos elementos de maior valor, bastando descartar os coeficientes de maior valor para obter-se apenas os coeficientes relacionados aos formantes. A literatura recomenda a utilização entre 12 e 20 coeficientes *MFCC*, dependendo do número de elementos utilizado no banco de filtros (HUANG et al., 2001).

3 RESULTADOS e DISCUSSÃO

Para demonstrar o funcionamento dos blocos de segmentação e janelamento, primeiro calculou-se o espectrograma do sinal “Unipampa” demonstrado pela Figura 2. Analisando a Figura 2, observa-se que existem pequenas faixas horizontais em diversos instantes de tempo com maior nível de energia, denotando a existência de formantes nestes intervalos.

Figura 2- Espectrograma da palavra “Unipampa”.

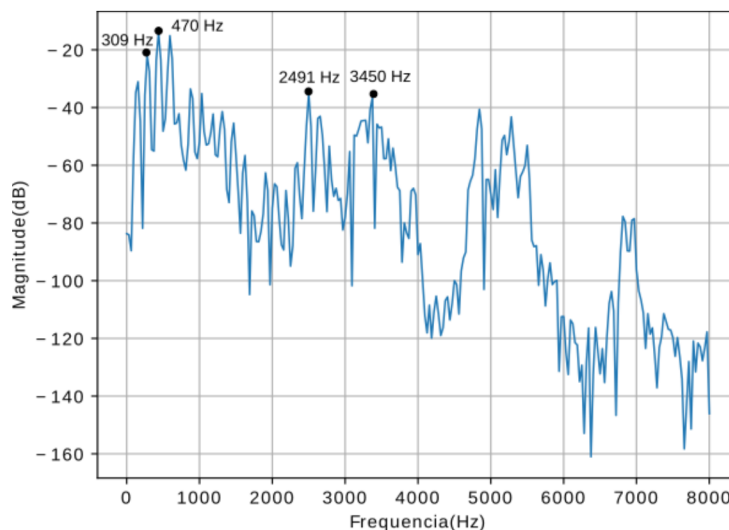


Fonte: do Autor, 2018.

Com auxílio do resultado obtido na Figura 2, foi escolhido o instante de tempo de aproximadamente 0,85 segundo para executar a análise em frequência do sinal segmentado e janelado na Figura 3. Para efeitos de comparação, apresenta-se na Figura 4 o sinal sem segmentação. Observa-se na Figura 3 a mudança de comportamento do sinal em frequência causada pela pronuncia de formantes.

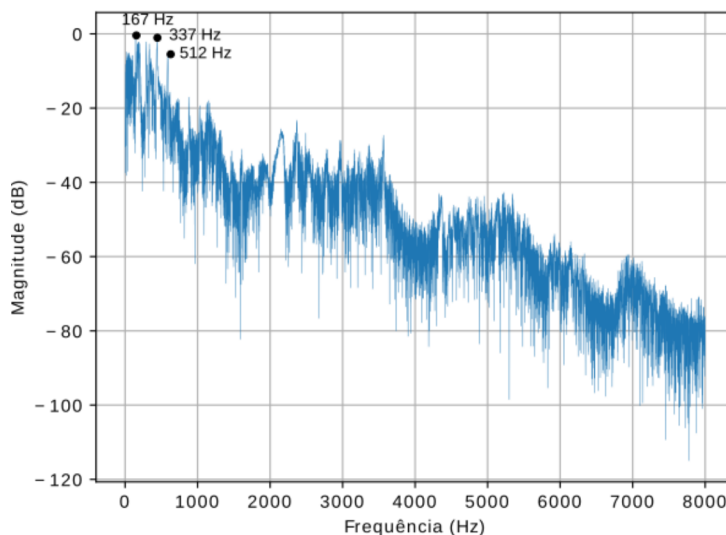
Comparando o espectro exibido pela Figura 3 com o ilustrado pela Figura 4 nota-se que o processo de segmentação e janelamento é de grande importância para a técnica, uma vez que o cálculo do espectro diretamente obtém apenas a informação da frequência fundamental da voz do autor e suas harmônicas.

Figura 3- Representação espectral: segmentado em 0,85 segundos.



Fonte: do Autor, 2018

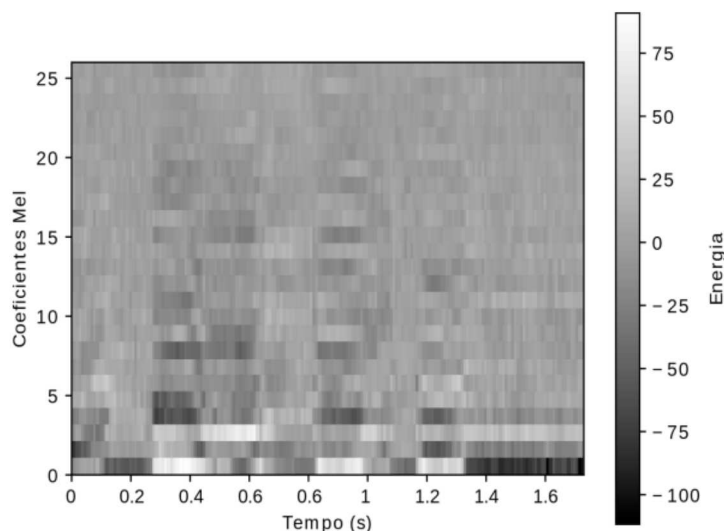
Figura 4- Representação espectral: sem segmentação.



Fonte: do Autor, 2018.

Por fim, a Figura 5 exibe os resultados obtidos após a obtenção dos MFCC. Observa-se que houve a separação dos formantes e do ruído presentes no sinal, visto que os elementos de maior energia estão alocados nos coeficientes de menor valor no vetor de parâmetros. Desse modo, pode-se descartar os coeficientes de maior valor, de maneira a obter as informações necessárias para o reconhecimento de uma palavra por um sistema ASR.

Figura 5- Espectrograma dos coeficientes MFCC.



Fonte: do Autor, 2018.

4 CONSIDERAÇÕES FINAIS

Este trabalho apresentou o projeto e os resultados da implementação de um módulo que compõe um Sistema automático de reconhecimento de fala (ASR), de processamento de sinais. Este sistema tem por característica ser o responsável pela aquisição, digitalização e parametrização das mensagens de fala que serão reconhecidas pela aplicação. Para a parametrização de sinais adotou-se neste trabalho a técnica MFCC. Esta técnica replica artificialmente alguns mecanismos do sistema auditivo humano com o objetivo de extrair os parâmetros linguísticos, os quais serão utilizados posteriormente para a classificação e reconhecimento de palavras por outros subsistemas. Os resultados obtidos neste trabalho destacam especialmente os submódulos de segmentação e janelamento, que tornam possível a identificação de características do sinal de fala no domínio espectral. Também destaca-se o submódulo de cálculo da DCT que é responsável pela separação das características de classificação linguística do ruído, permitindo a extração apenas dos parâmetros de interesse. Por fim, salienta-se que o bloco de processamento de sinais é fundamental para o bom funcionamento de um sistema ASR e o sistema aqui apresentado vem sendo utilizado pelo autor para a construção de um sistema ASR.

REFERÊNCIAS

BENESTY, J.; SONDHI, M. M.; HUANG, Y. (2008). *Springer Handbook of Speech Processing*. Berlim, Alemanha: Springer.

HUANG, X.; ACERO, A; HON, H. W. (2001). *Spoken Language Processing: A guide to Theory, Algorithm and System Development*. Nova Jersey, EUA: Prentice-Hall.

MERMELSTEIN, P. (1976). Distance Measures for Speech Recognition – Psychological and Instrumental. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v.24, 374-388. doi: 10.1109/TASSP.1976.1162849 .

R SANTOS, G. B. (2013). *Análise Fonético-Acústica das Vogais Orais e Nasais do Português: Brasil e Portugal*. (Trabalho de pós-graduação – Doutorado). Universidade Federal de Goiás. Goiânia.