

Universidade do Minho

Escola de Engenharia

**deteção e Remoção de ruído em sinais vozeados**

Processamento Digital de Sinal

*Docente: Professor Carlos Lima*

*Trabalho realizado por José Henrique Machado Pires, A65326*

**Introdução**

O processamento da fala é o estudo de sinais vozeados e dos seus métodos de processamento.

São normalmente processados digitalmente, o que constitui um elemento de estudo de processamento digital de sinal.

O processamento de sinal vozeado inclui o a aquisição, armazenamento, manipulação, transferência e produção de saída do sinal processado.

Utilizando algoritmos de deteção de voz “VAD” (*voice activity detection*), o objetivo deste trabalho é detetar e eliminar regiões de silêncio em sinais vozeados corrompidos com ruído, permitindo alterar o valor de *threshold* conforme o rácio sinal-ruído (SNR). Este algoritmo está também preparado para adição de ruído branco ao sinal vozeado.

***Suporte teórico***

**Modelo gaussiano**

Uma distribuição normal ou gaussiana é um conjunto de amostras de eventos aleatórios.

A partir da análise de conjunto tiram-se conclusões que os eventos tendem a aproximar-se de uma determinada posição, que representa a sua esperança matemática, ou média μ.

O modelo gaussiano é inteiramente descrito pelos parâmetros média (μ), valor para o qual se concentram os dados de uma distribuição e desvio-padrão (σ), que representa o valor de dispersão em relação à média.

Um desvio-padrão baixo indica que os dados se encontram próximos da média.

******

***Outlier***

Em estatística, temos um *outlier* (valor atípico) quando um determinado valor se encontra muito afastado em relação às restantes amostras.

Um dos métodos de identificação de *outliers* mais utilizado é o desvio-padrão. Pelo desvio padrão, será considerado *outlier* um determinado valor que se encontre a um certo valor de desvios-padrões da média.

Essa quantidade varia consoante o tamanho da amostra.

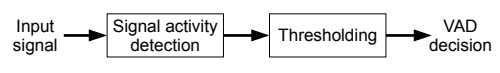


**VAD – Voice Activity Detection**

Classe de métodos que detetam se um sinal sonoro contém ou não fala. Uma tarefa estreitamente conectada a este método é a estimação da probabilidade de presença de fala. Em vez de a decisão do algoritmo VAD ser “voz presente” ou “voz não presente”, este estimador quantifica a probabilidade de o sinal conter fala.

Um VAD pode ser derivado de um estimador deste género, definindo um *threshold* probabilístico acima do qual o sinal é considerado que contém voz relevante.

É usado como algoritmo de pré-processamento para a maioria dos restantes métodos de processamento de fala. O seu objetivo é reduzir a quantidade de dados inúteis transmitidos ou processados, e remover ruído dos excertos vozeados.



Encontrando áreas de silêncio (não-vozeadas), podem-se estimar características de ruído, para que seja possível remover tudo o que seja parecido com ruído.

Em VAD, tentam-se medir algumas propriedades do sinal que dão indicação se o sinal é ou não vozeado. Estas propriedades denominam-se *features*.

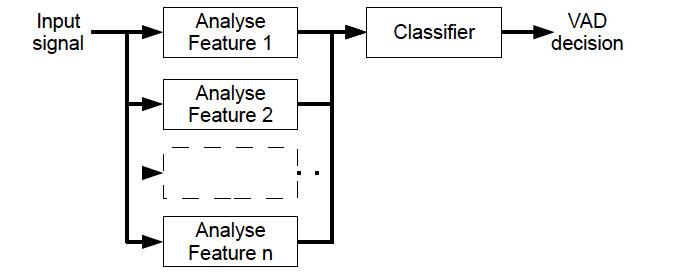
A energia do sinal é uma *feature* importante, uma vez que a energia de um sinal vozeado varia bastante. Normalmente, retêm a energia nas baixas frequências, de tal modo que estimadores de *tilt* espetral são úteis.

Por exemplo, o valor de *Zero-Crossings* (por unidade de tempo) é alto para sinais de alta frequência (ruído), e baixo para sinais de baixa frequência (voz), o que faz dele uma *feature* importante.

A autocorrelação *lag-1* (correlação entre valores que estão separados por um período) é alta (perto do valor 1) para sinais de alta frequência e baixa (perto de -1) para sinais de alta frequência.

Sinais vozeados podem ser facilmente modelados através de predição linear. Se o erro do preditor é grande, provavelmente estamos perante fala. Caso contrário, é um sinal não vozeado.

Até agora, recolhemos indicadores e *features* de sinais vozeados ou não vozeados. O próximo passo é recolher informação destas *features* e tomar uma decisão entre sinal vozeado ou não.



O classificador utilizado é um *Gaussian Mixture Model (GMM)*. O espaço de *features* é modelado por uma soma de Gaussianas.

Resumindo, a ideia básica do algoritmo é a seguinte:

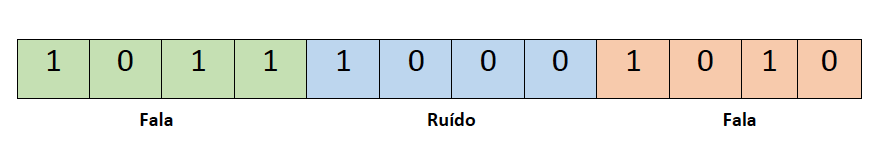
1. Calcular o conjunto de *features* do sinal.
2. Reunir a informação destas *features* num classificador, que retorna a probabilidade de o sinal ser ou não vozeado.
3. Aplicar um *threshold* ao classificador.

Utilizando o classificador de distribuição gaussiana, calculamos o desvio padrão de cada amostra.

Se a condição **|X - µ| > σ** se verificar, esta amostra trata-se de um *outlier* (valores que se encontram muito afastados dos restantes, e que podem ser identificados através do desvio padrão).

Efetua-se esta análise a cada amostra e armazena-se num vetor o resultado da condição **|X - µ| > σ**, correspondendo o valor 0 a ruído e 1 a fala.

Este vetor é analisado por janelas de tamanho variável (window\_size). Cada janela é considerada fala ou ruído através da contagem de 1’s. Se esta contagem for igual ou ultrapassar o parâmetro desejado para o número de 1’s (fator de número de 1’s), então consideramos a janela analisada como fala. Caso contrário, a janela ou segmento de amostras é considerado ruído. O outro parâmetro deste método de janelas é o valor do deslize da janela a cada iteração de classificação.



Exemplificação do método de classificação com um tamanho de janela de 4, e uma decisão de número de 1’s com fator ½, ou seja, se o número de 1’s for igual ou maior que o dobro do número de 0’s, é considerado fala. Caso contrário, é considerado ruído.

**Ruído Branco:**

Em processamento de sinal, ruído branco é um sinal aleatório de intensidade igual em todas as frequências, conferindo-lhe PSD (densidade espetral de potência) constante.

Em tempo discreto, é uma sequência cujas amostras são consideradas variáveis aleatórias não-correladas, com média nula e variância finita.

Caso as amostras sejam variáveis aleatórias independentes e identicamente distribuídas, com distribuição normal e média nula, diz-se que o sinal é ruído branco aditivo gaussiano.

A largura de banda de um sinal ruído branco é, na prática, limitada pelo mecanismo que o gera, pelo meio de transmissão e pelas óbvias capacidades finitas de observação.

Resumindo, sinais aleatórios são ditos ruído branco se observamos um espetro constante sobre a gama de frequências que são relevantes para cada contexto.

Para um sinal de áudio, a gama relevante de frequências audíveis está entre 20 e 20.000 Hz.

**SNR: Relação Sinal-Ruído**

Signal-to-noise ratio ou SNR envolvem medidas de um sinal num meio ruidoso, e é definido como a razão entre a potência de um sinal e a potência do ruído sobreposto ao sinal, geralmente expresso em Decibel.

Compara o nível de um sinal desejado com o nível de ruído de fundo. Quanto maior for este rácio, menor é o efeito do ruído sobre o sinal de áudio.

No contexto do problema, uma adição de ruído branco com SNR = 0 dB, a fala é muito mais ruidosa do que uma adição de ruído branco com SNR = 50 dB.



Alternativamente, a Relação sinal-ruído é recíproca ao coeficiente de variação, isto é, a razão entre a Média e o Desvio Padrão de uma medida de sinal.



Nota: cálculo da potência de um sinal de áudio:

Potência = (1/Nsamples) \* Σ Si²

**Procedimentos:**

* Gravação do sinal áudio / abertura de um .wav com voz e regiões de silêncio;
* Sintetização de ruído branco de média nula e potência dependente do SNR, e adição do ruído sintetizado ao sinal áudio;
* Run do algoritmo para efetuar testes de valores de SNR [0-50 dB], em incrementos de 10 dB e sua relação com o *threshold* a escolher;
* Pelo procedimento anterior e resultados dos testes, colocar os valores de *threshold* mais indicados para os SNR aplicados.

Valores de threshold para diferentes SNRs:

if (SNR > 0 && SNR < 15) threshold = 0.45;

elseif (SNR > 15 && SNR < 25) threshold = 0.5;

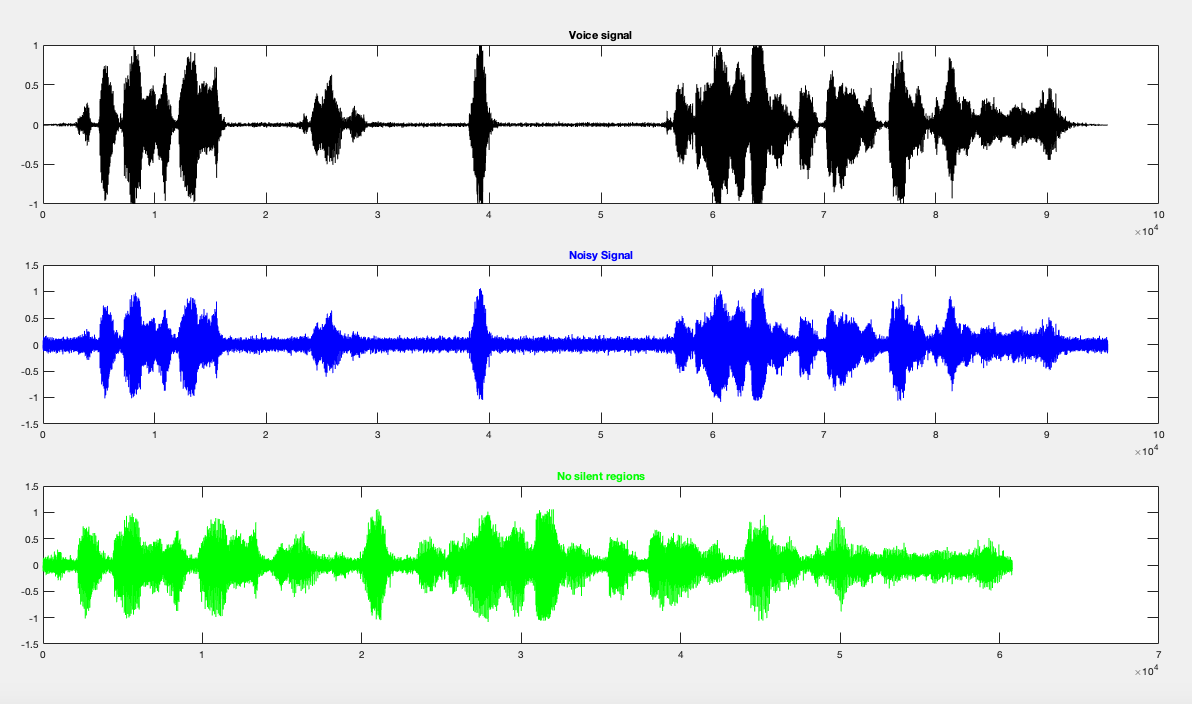
elseif (SNR > 25 && SNR < 35) threshold = 0.6;

elseif (SNR > 35 && SNR < 45) threshold = 0.75;

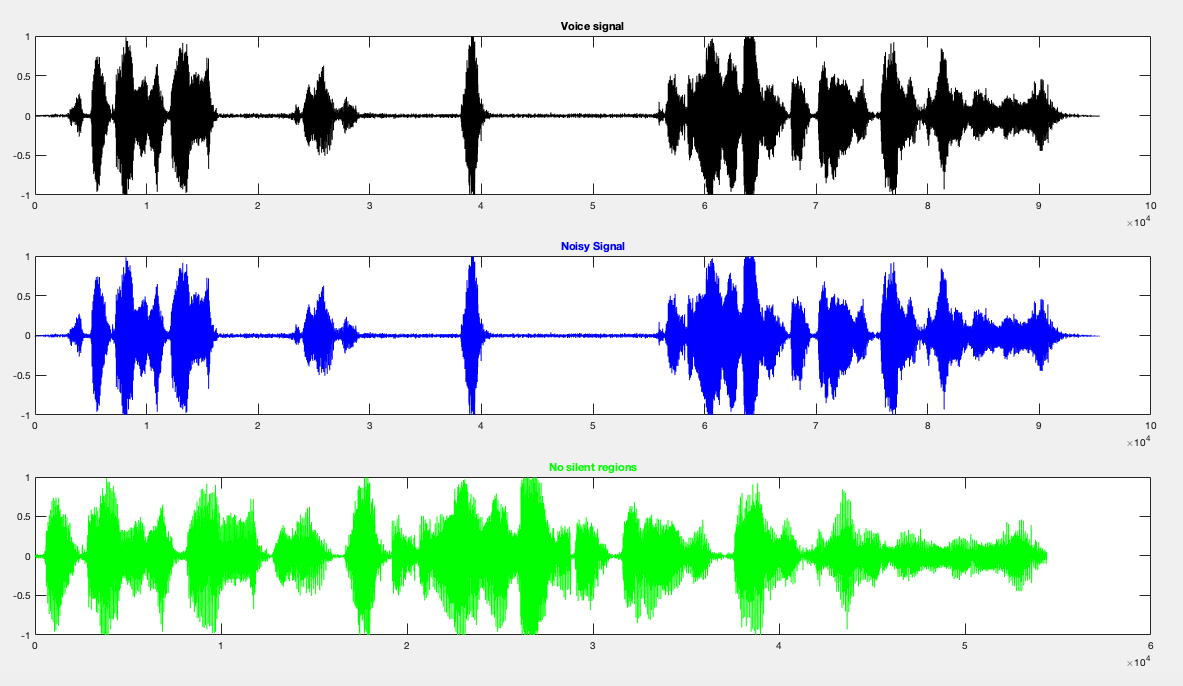
else threshold = 0.8;

**Conclusões:**

SNR = 10 dB



SNR = 50 dB:



Por comparação entre as duas imagens, o sinal vozeado aditivado com mais ruído (menor SNR em dB), mais dificilmente são detetadas pausas na fala e pior será a remoção de ruídos de maior amplitude.

Esta dificuldade é inerente às fragilidades deste algoritmo, uma vez que o ruído toma proporções elevadas e aproxima-se da amplitude da fala.

À medida que diminuímos o valor de SNR, cada vez mais segmentos de silêncio ficam armazenados no sinal detetado como fala.

Em relação ao factor de predominância de 1’s que classifica o sinal como vozeado ou ruído, conclui-se que, se “ones\_factor” for baixo, o buffer “voice” armazena não só o sinal vozeado, mas também o ruído e os silêncios.

Caso “ones\_factor” seja alto, pode perder-se informação.

O “jump” influencia a duração das formas de onda dos dois buffers, pelo que será necessário um maior número de ciclos para analisar todo o sinal.

Este “jump” tem que ser conjugado corretamente com o tamanho da janela de forma a permitir uma melhor análise do áudio.