Medição de diversas características subjetivas da audição

*4º Trabalho Laboratorial de Áudio 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 trabalho propôs uma abordagem abrangente para a síntese realista das vogais do português, utilizando o software "formant.exe" e scripts MATLAB. Configurou-se os formantes para [a][e][i][o][u], incorporando elementos como controle da entoação, envolvente temporal, "jitter", "shimmer" e ruído de fundo. Os resultados demonstraram uma representação acústica da fala humana, destacando a importância do controle detalhado dos parâmetros espectrais. A adição de variações naturais e ruído contribuiu para uma síntese vocal mais natural.**

***Index Terms*—Fala Humana, Formantes, Entoação, Envolvente Temporal, Jitter, Shimmer, Ruído de Fundo**

I. Introdução

O presente trabalho propõe-se a utilizar o software "formant.exe" em conjunto com os scripts de Matlab fornecidos, nomeadamente "fgerimp.m" e "SINTESE_FORMANTES.m", para a geração de sinais de fala que representem de maneira realista as cinco vogais base do português [a][e][i][o][u]. O objetivo é não apenas sintetizar os conteúdos acústicos de cada vogal, mas também incorporar elementos que contribuam para a realidade da fala, tais como "jitter" e "shimmer". Além disso, será aplicada uma envolvente temporal e de ruído de fundo, de modo a aprimorar a autenticidade do sinal gerado.

II. Metodologia

De modo a atingir os objetivos propostos, adotamos a seguinte metodologia:

- **Utilização do software "formant.exe":** O software "formant.exe" será empregue para descobrir quais são as frequências associados a cada formante de cada vogal.
- **Seleção das vogais base:** A escolha das vogais é feita [a][e][i][o][u] e é registada a frequência de cada formante.
- **Controlo da Entoação:** A entoação será modificada através da incorporação de variações na frequência fundamental (Fo). Essas variações serão implementadas para simular mudanças naturais na entoação durante a produção da fala.
- **Controlo da Envolvente Temporal:** Nesta etapa, a atenção volta-se para o controlo da envolvente temporal do sinal de fala. Variáveis como attack e decay são ajustadas para modular a transição entre diferentes partes da fala, promovendo uma síntese mais suave e natural. A utilização de janelas de Hanning no processo contribui para atenuar possíveis artefactos indesejados, garantindo uma representação temporal coerente.
- **Incorporação de "jitter" e "shimmer":** Para adicionar realismo ao sinal de fala, serão aplicados "jitter" e "shimmer", que introduzem variações sutis na frequência e amplitude do sinal, simulando as flutuações naturais presentes na fala humana.
- **Adição de ruído de fundo:** Com o intuito de simular as condições reais de uma elocução, será introduzido ruído de fundo nos sinais gerados. Isso contribuirá para a autenticidade da fala, considerando o ambiente em que a comunicação ocorre.

Ao seguir esta metodologia, espera-se obter sinais de fala que não apenas representem as características acústicas das vogais escolhidas, mas que também reproduzam de maneira fidedigna os elementos dinâmicos e realistas presentes na fala humana. Posteriormente, serão apresentados os resultados obtidos e uma análise crítica do processo de síntese adotado.

III. Implementação

A. Utilização do Formant.exe:

Iniciamos o processo abrindo o programa "formant.exe", conhecido como Formant Synthesizer Demo. Este software é uma ferramenta valiosa para a síntese de formantes, permitindo a manipulação de diversas variáveis que influenciam as características acústicas dos sinais de fala gerados. Dedicamos um período considerável para estudar minuciosamente a interface e as funcionalidades do programa, com o objetivo de compreender como as configurações disponíveis impactam a síntese de formantes e, consequentemente, a qualidade dos sinais de fala produzidos.

Ao observarmos a interface do programa, identificamos a capacidade de escolher diferentes vogais e percebemos como as frequências F1 e F2 variam de acordo com essas escolhas. Notamos, durante esse processo, que as frequências F3 e F4 se mantêm constantes, uma característica que será abordada posteriormente no desenvolvimento do projeto. Essa observação é essencial para compreender como as características espectrais das vogais podem ser ajustadas para representar de maneira mais fiel as propriedades acústicas da fala humana.

A interface do programa também nos permitiu explorar a modificação da largura de banda (bandwidth) de cada frequência, fornecendo a oportunidade de ajustar detalhes mais refinados na síntese dos formantes. Essa capacidade de personalização contribui para a precisão na representação das características específicas de cada vogal.

Além disso, durante o processo de configuração, tivemos a responsabilidade de escolher a frequência fundamental, amplitude, formato do sinal (shape), e tipo de onda desejado, que poderia variar entre rectangle, triangle, sine, sampled e noise. Esses parâmetros adicionais possibilitaram uma personalização mais abrangente dos sinais de fala gerados, permitindo-nos moldar a qualidade e o caráter acústico de cada vogal de acordo com as necessidades do projeto.

B. Código MATLAB

De forma a realizar o trabalho aproveitamos o valor dos formants que obtivemos no programa "Formant Synthesizer Demo". De seguida foi tomado estas medidas:

- **Parâmetros Iniciais:**

  Estes parâmetros iniciais estabelecem as bases para a síntese de fala, determinando como o sinal será dividido em frames, a resolução temporal da análise e a taxa de amostragem utilizada para representação digital do sinal. A seleção cuidadosa desses parâmetros é crucial para assegurar a qualidade e realismo do sinal de fala gerado.

  - Fs (Frequência de Amostragem):
    A frequência de amostragem, denotada por Fs, representa a taxa na qual o sinal de áudio é amostrado. No nosso contexto, Fs é fixado em 11025 Hz, determinando quantas amostras por segundo são coletadas para representar a forma de onda do sinal.
  - jmax (Número de Frames):
    O parâmetro jmax corresponde ao número total de frames considerados durante a síntese de fala. Cada frame representa uma segmentação do sinal de fala para análise individual. Neste caso, jmax é definido como 75 frames.
  - janela (Tamanho da Janela para Análise):
    O tamanho da janela (janela) refere-se à extensão, em número de amostras, utilizada para analisar cada frame. A análise local de pequenos trechos do sinal, conhecidos como frames, é essencial para extrair características relevantes na síntese de fala. Neste contexto, janela é fixado em 256 amostras.
  - f0 (Frequência Fundamental Base):
    A frequência fundamental base, representada por f0, é a taxa de repetição dos ciclos de vibração das pregas vocais. Ela desempenha um papel crucial na percepção da altura tonal da voz. Aqui, f0 é inicializado em 100 Hz. Este valor foi optado por ser próximo da frequência tipica de um homem, cerca de 110 Hz
  - DF (Duração de uma Frame):
    A duração de uma frame (DF) indica o intervalo de tempo coberto por cada frame em termos de segundos. Calculada como o inverso da frequência de amostragem multiplicado pelo tamanho da janela, neste contexto, DF é definida como 1/11025 * 256 segundos.

- **Configuração dos Formantes para Vogais:**

  Durante a síntese de fala, a configuração dos formantes desempenha um papel fundamental na modelagem das características acústicas específicas de cada vogal ([a], [e], [i], [o], [u]). Os formantes são ressonâncias no espectro sonoro associadas às diferentes posições das articulações vocais. Os valores específicos escolhidos para os formantes são cruciais para a autenticidade e reconhecimento das vogais, e os parâmetros na matriz "vogais" têm significados específicos:

  - Frequência Fundamental (f0):

    A frequência fundamental, representada por f0, é a taxa de repetição dos ciclos de vibração das pregas vocais. Esse valor determina a altura tonal da voz e é crucial para a percepção da vogal.
  - Formantes (F1, F2, F3, F4):

    Os formantes $(F1, F2, F3, F4)$ são frequências ressonantes na resposta espectral da fala e são determinantes para a identificação das vogais. Cada formante está associado a uma posição específica das articulações vocais durante a produção da vogal. Valores mais altos de formantes indicam frequências mais altas no espectro sonoro.

    Tendo em conta o uso do programa *Formant Synthesizer Demo*:

    Para a vogal [a], os formantes são configurados para: 717, 1089, 2500 e 3500 Hz.
    Para a vogal [e], os formantes são configurados para 554, 1761, 2500 e 3500 Hz.
    Para a vogal [i], os formantes são configurados para 282, 2238, 2500 e 3500 Hz.
    Para a vogal [o], os formantes são configurados para 470, 1020, 2500, 3500 Hz.

Para a vogal [u], os formantes são configurados para 311, 875, 2500 e 3500 Hz.

As vogais geralmente apresentam 4 formantes (ressonâncias) - $F1, F2, F3, F4$. No entanto, é possível caracterizar todas as vogais com os dois primeiros formantes, uma vez que o primeiro ($F1$) corresponde à altura da língua e o segundo ($F2$), ao movimento horizontal da língua. Assim sendo, o uso de $F3$ e $F4$ mantêm-se constantes.

Assim, a configuração precisa dos formantes na matriz "vogais" é essencial para garantir que o sinal de fala sintetizado seja percebido de maneira autêntica, respeitando as características acústicas específicas de cada vogal no idioma português.

- **Controlo da Entonação**

  No âmbito deste projeto de síntese de fala, um dos elementos críticos é o *Controle de Entonação*. Este controle desempenha um papel essencial na criação de uma síntese vocal mais dinâmica e natural, permitindo a simulação de variações na entonação durante a produção da fala.
  A implementação do *Controle de Entonação* baseia-se na manipulação da frequência fundamental ($F0$), que é a frequência básica associada à vibração das pregas vocais. Para garantir uma representação realista da entonação, o código adota uma abordagem inovadora ao incorporar a variação específica da frequência fundamental para diferentes vogais e tendo em conta o que realmente acontece. Assim sendo, para chegar aos valores corretos, foi utilizado o PRAAT de forma a obter os valores especificos para cada vogal.
  Um aspecto destacado é o uso de interpolação para ajustar o número de amostras de $F0$ de acordo com o número total de quadros ($jmax$). A interpolação é uma técnica matemática crucial que preenche as lacunas entre pontos de dados discretos, assegurando uma transição suave entre valores de $F0$ e uma representação contínua ao longo do tempo. Essa abordagem é fundamental para garantir a coerência e naturalidade das variações de entonação na síntese de fala.

- **Controlo da Envolvente Temporal:**

  O controle da envolvente temporal, um aspecto crucial na síntese de fala para modelar as características dinâmicas da voz. Variáveis como attack e decay são utilizadas para ajustar a duração das transições suaves, enquanto janelas de Hanning suavizam essas transições, contribuindo para uma síntese mais natural. A multiplicação da envolvente resultante pelo vetor de ganho Av ajusta a amplitude global do sinal, garantindo uma representação fiel da variação de intensidade na fala.
  Além disso, são como já foi abordado a inicialização dos formantes (F1, F2, F3, F4) e a frequência fundamental base (f0) para as vogais A, E, I, O e U do português. Esses parâmetros são essenciais para modelar as características espectrais específicas de cada vogal, contribuindo para uma síntese acusticamente precisa. Ao controlar a envolvente temporal, o código assegura que a transição entre diferentes partes da fala seja suave, resultando em uma síntese mais natural e coerente. Essa abordagem contribui significativamente para a qualidade perceptual da fala sintetizada.

- **Incorporação de "jitter" e "shimmer":**

  A incorporação de "Jitter" e "Shimmer" no processo de síntese de fala é fundamental para simular variações naturais na produção vocal. "Jitter" refere-se à variação na periodicidade dos ciclos vocais, enquanto "Shimmer" está relacionado à variação na amplitude das ondas sonoras da fala. Ambos desempenham um papel crucial na percepção da qualidade vocal e na criação de uma síntese de fala mais realista.
  O "Jitter" é calculado como a variação percentual média nos períodos de pitch, e valores elevados podem indicar instabilidade na produção vocal, resultando em uma qualidade vocal percebida como menos suave ou mais áspera. Por outro lado, o "Shimmer" é calculado como a variação percentual média na amplitude do sinal vocal, e um aumento nesse parâmetro pode indicar variações na intensidade da voz.
  A incorporação dessas características na síntese de fala permite uma modelagem mais precisa das variações naturais na produção vocal humana. A adição controlada de "Jitter" e "Shimmer" contribui para uma síntese mais autêntica e perceptualmente rica, replicando nuances importantes da fala humana.

- **Adição de Ruído de Fundo:**

  A adição de ruído de fundo ao sinal de fala torna a síntese mais representativa de ambientes do mundo real. Essa prática simula condições mais naturais.

IV. Conclusões

Foram obtidos sinais de fala que representam não apenas as características acústicas das vogais escolhidas, mas também incorporam elementos dinâmicos e realistas. A configuração precisa dos formantes para cada vogal, aliada ao controle da entoação, envolvente temporal, e a adição de "jitter", "shimmer" e ruído de fundo, contribuíram para uma síntese vocal autêntica. A análise dos resultados destacou uma representação fiel das características espectrais específicas de cada vogal, evidenciando a importância do controle de parâmetros como F1, F2, F3, F4 e f0. Deste modo, foi possível reproduzir 5 vogais, cuja a sua identificação auditiva é clara. O facto de ainda se notar que o sinal é construído por uma máquina é em grande parte consequência da interpolação realizada na variação de f0.