FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Assistente para Coaching Vocal

André Presteux Santos

Mestrado Integrado em Engenharia Electrotécnica e Computadores

Orientador: Aníbal Ferreira

21 de Julho de 2017

Resumo

A voz é uma das ferramentas mais importantes na interação humana. A forma como é utilizada tem uma grande influência na mensagem que se pretende transmitir. Isto é especialmente verdade em cenários em que a comunicação é direcionada para um grupo relativamente largo de ouvintes, tais como apresentações e discursos. O aperfeiçoamento da capacidade de comunicação de um orador pode ser feito através da análise de características da sua voz. Estes resultados podem ser obtidos por intermédio de uma solução tecnológica.

O principal objetivo deste projeto de dissertação é o desenvolvimento de uma aplicação para plataformas móveis iOS que permita a apresentação de *feedback* visual em tempo real acerca de certas características do sinal de voz. Esta ferramenta visa servir de suporte para treino vocal, ajudando o utilizador a comunicar de forma mais eficaz. É necessário que seja intuitiva e de fácil utilização.

O desenvolvimento desta aplicação resultou numa interface interativa, constituída por um gráfico temporal e um fonetograma. Este último consiste numa forma de representação de um sinal em função da sua frequência fundamental, em Hertz, e intensidade, em deciBel. Inclui um teclado de piano interativo, que pode ser utilizado para estabelecer uma correspondência entre uma nota musical e o *pitch* da voz.

Posteriormente, a ferramenta desenvolvida foi sujeita a testes. Os parâmetros de frequência fundamental e intensidade foram medidos e comparados com valores reais, de forma a determinar a sua precisão. Foi também realizado um teste de usabilidade, que permitiu avaliar a perceção de utilizadores relativamente às funcionalidades desenvolvidas.

Tendo sido testados valores de frequência para toda a gama representada no fonetograma, verificou-se uma excelente precisão, exceto em valores correspondentes a baixas frequências, em que se verifica a ocorrência de erros. Quanto às medidas de intensidade, verificou-se que a sua variação seguiu os valores reais, apesar de existir um desvio significativo de calibração que, apesar ser simples de compensar, necessita de ser corrigido.

Os resultados dos testes de usabilidade foram globalmente satisfatórios, permitindo identificar pontos fortes e aspetos a melhorar do ponto de vista do utilizador.

Abstract

Voice is one of the most important tools allowing humans to communicate and interact. The way it is used has a great deal of influence on the intended message. This is especially true in scenarios where the communication is directed towards a large group of people, as it is the case of speeches and presentations.

The main goal of this dissertation is the development of an iOS app that provides real-time visual feedback of the voice dynamics. This tool aims to serve as a support for voice training, helping the user to communicate more efficiently. In this regard, it has to be intuitive and easy to use.

The app consists of an interactive interface in which the signal is represented in a time plot and, most importantly, a phonetogram. The latter is a form of graphical representation that maps a signal as a function of its fundamental frequency, in Hertz, as well as its intensity, in deciBel. It incorporates an interactive piano keyboard that can be used to check the musical note that best matches a specific frequency region of the voice pitch.

Finally, the app was subject to testing. The pitch and intensity parameters were measured and compared to real values, in order to determine their precision. A usability test was also conducted, in order to evaluate user perception concerning the implemented features.

The precision of the *pitch* determination was deemed excellent for the majority of the frequency range, except for low frequency values, that gave rise to significant errors. The sound intensity measurements showed that, while they vary quite consistently relative to the real values, there is a significant calibration deviation that although it is simple to compensate, needs to be corrected.

The usability test results were globally satisfactory, and lead to the identification of major strengths and possible improvements from an user's point of view.

Agradecimentos

Não posso deixar de agradecer ao meu orientador, o Professor Aníbal Ferreira, pelo acompanhamento efetuado e ajuda providenciada ao longo deste trabalho. Dirijo também um profundo agradecimento ao Professor Sérgio Ivan Lopes, por ter facultado as bases que me permitiram realizar esta aplicação em tempo útil.

Um forte abraço a todo o pessoal da sala I224 do DEEC, pelo convívio proporcionado ao longo de todos estes meses de trabalho.

Obrigado a todos os bons amigos que fiz ao longo destes 5 anos de curso, e a todos os meus amigos vilacondenses de longa data.

Aos meus companheiros do Discord, um pedido de desculpas por não ter podido estar tão presente durante a realização desta etapa.

Por fim, quero agradecer a toda a minha família, e especialmente aos meus pais, pelo apoio demonstrado ao longo desta longa jornada.

André Presteux Santos

"Ever tried. Ever failed. No matter.
Try Again. Fail again. Fail better."

Samuel Beckett

Conteúdo

Lista de Figuras

Abreviaturas e Símbolos

VRP	*Vocal Range Profile*
SPL	*Sound Pressure Level*
CPU	*Central Processing Unit*
FFT	*Fast Fourier Transform*
VBO	*Vertex Buffer Object*

Capítulo 1

Introdução

1.1 Contexto

A voz falada é um dos mais importantes meios de comunicação naturais utilizados pelo ser humano. É uma ferramenta de enorme versatilidade, cuja utilidade se estende muito para além da simples transmissão de informação. Sendo um dos meios mais importantes para transportar emoções, que constituem uma parte fundamental da interação humana, a maneira como se comunica uma mensagem possui, em várias situações, tanta importância como o seu conteúdo.

Deste modo, a utilização adequada da fala, dependendo do contexto em que se enquadra, constitui a marca de um bom orador. A capacidade de comunicar eficazmente, apesar de ser em parte inata, pode ser ensinada e treinada, o que pode beneficiar todo o tipo de pessoas, mais especificamente aquelas que praticam uma atividade condicionada pela comunicação.

Os problemas associados à capacidade de uma pessoa para comunicar eficazmente podem ser de vários tipos, que se encontram, de um modo geral, catalogados em duas secções: dificuldade em apresentar um discurso cativante, isto é, falta de capacidade para interessar o ouvinte, ou dificuldade em produzir fala por diversos motivos de causa médica.

Em ambos estes casos, uma melhora da capacidade de comunicação pode ser atingido com a ajuda de profissionais especializados, mas também com aplicações desenvolvidas para esse efeito, com a ajuda ao desenvolvimento dos profissionais anteriormente referidos. Estas aplicações constituem ferramentas que se enquadram na realidade atual, em que as tecnologias continuam a ocupar um lugar cada vez mais importante, tanto pela sua facilidade de acesso como pelo seu lado prático e económico.

1.2 Motivação e objetivos

Num mundo em que floresce a indústria do *smartphone* e dos seus produtos associados, da qual fazem parte as numerosas aplicações disponibilizadas de forma gratuita ou com custos associados, é cada vez maior a procura de ferramentas que possam ajudar os seus utilizadores a desenvolver novas competências aplicadas às mais variadas áreas do conhecimento.

O desenvolvimento de tais aplicações constitui um interessante desafio de ordem tecnológica, aplicado a várias áreas da engenharia.

Além do aspeto tecnológico, existem também motivações da ordem social, uma vez que o trabalho que se propõe realizar traz consequências positivas e pode ser utilizado por um grande número de pessoas de diferentes realidades, constituindo um fator de enriquecimento pessoal e que incide positivamente na sociedade.

O objetivo do trabalho a realizar é desenvolver uma aplicação para plataformas móveis que permita a um utilizador observar, em tempo real, informações relativamente à sua voz. Pretende-se que exista um *feedback* visual apelativo, simples e intuitivo, que permita uma leitura rápida por parte do utilizador, sem que este tenha necessidade de possuir conhecimentos prévios muito aprofundados.

A aplicação servirá como um assistente para treino vocal, apresentando informação útil que poderá ser interpretada de modo a perceber características do discurso de um orador. Não se pretende entrar no domínio da subjetividade, mas sim apresentar pistas objetivas que permitam a interpretação e o diagnóstico. Por esta razão, a aplicação a desenvolver não apresentará qualquer tipo de *feedback* que não seja baseado na obtenção de parâmetros objetivos do sinal de voz.

É uma prioridade que a aplicação seja facilmente utilizável e aberta a todo o tipo de utilizadores, bastando para isso possuir um *smartphone* ou *tablet* equipado com o *software*, apenas com o requisito de estar situado numa posição relativamente próxima da boca do orador, de forma a obter bons resultados. A existência de uma interface gráfica intuitiva e de fácil compreensão é, por isso, de grande importância.

A dificuldade em encontrar aplicações orientadas a treino vocal que cumpram com estes objetivos constitui uma motivação suplementar neste projeto, por permitir contribuir para uma área relativamente inexplorada.

1.3 Estrutura da dissertação

A presente dissertação possui uma organização específica, pensada de forma a facilitar a compreensão do assunto desenvolvido, em termos de etapas fundamentais para a sua realização.

O capítulo 2 expõe o estudo prévio que foi realizado de forma a ganhar um conhecimento mais aprofundado sobre o tema em discussão. As informações obtidas permitiram ter uma perceção mais avançada dos objetivos realizáveis, assim como das funcionalidades mais prioritárias a desenvolver.

Seguidamente, o capítulo 3 estabelece o ponto de partida do trabalho. São apresentadas duas aplicações desenvolvidas previamente que serviram de base para a realização deste projeto, bem como dos módulos constituintes que foram aproveitados para a realização deste trabalho.

O desenvolvimento do código da aplicação é exposto no capítulo 4. Pretende-se mostrar as principais partes constituintes da *app* e explicar, sem demasiado detalhe, como estas foram implementadas. Tenta-se dar uma perspetiva das alternativas consideradas e de alguns melhoramentos trazidos.

O capítulo 5 corresponde à exposição de resultados associados ao teste das funcionalidades e desempenho da aplicação.

Finalmente, o capítulo 6 serve de conclusão para a dissertação, em que é feita uma breve análise sobre os objetivos conseguidos.

Capítulo 2

Revisão Bibliográfica

2.1 Introdução à Voz

Dado que o tema desta dissertação é fortemente ligado ao fenómeno de produção de voz, é importante descrever os mecanismos de produção desta, bem como as formas de análise temporal e espectral do sinal de fala.

2.1.1 Representação da voz no tempo e nas frequências

De uma forma muito simplificada, o mecanismo de produção da fala no ser humano pode ser descrito através dos seguintes passos: primeiro, é expelido ar pelos pulmões, que vai ser encaminhado para a zona da laringe; em segundo lugar, o ar passa através das cordas vocais, que produzem som ao vibrar; em terceiro lugar, o som primário gerado pelas cordas vocais passa pelo trato vocal, constituído, entre outros, pela faringe, cavidade oral e lábios, que modificam o espectro do sinal conferindo-lhe as características fonéticas próprias [1].

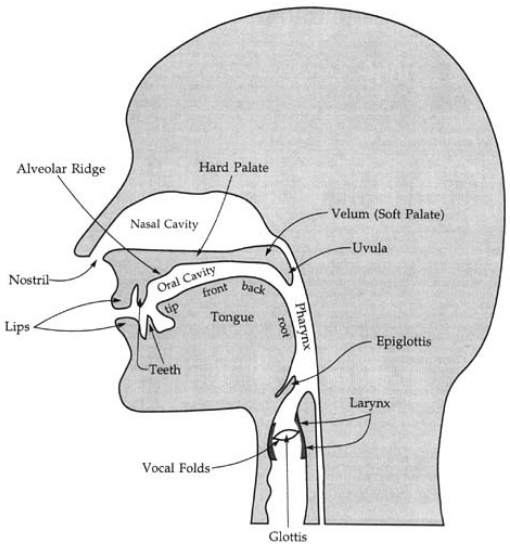
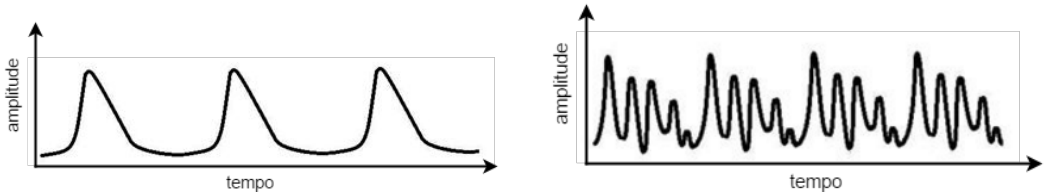


Figura 2.1: *Sistema fonatório humano*[1]



(a) *À saída da glote* ***(b)*** *À saída da boca*

Figura 2.2: *Formas de onda do sinal vocal em diferentes pontos do sistema fonatório*

Este processo pode ser descrito através do modelo fonte-filtro da produção da voz. A primeira parte da expressão advém do facto de que os constituintes do sistema fonatório até às cordas vocais constituem uma fonte de sinal, uma vez que a exalação de ar pelos pulmões constitui a excitação do meio, e as cordas vocais conferem a componente periódica de vibração. A segunda parte vem de que as estruturas supra-laríngeas constituem um filtro, pois são responsáveis especialmente pela modificação da forma da envolvente espectral do sinal de fonte. A Figura 2.2 mostra a transformação aplicada ao ar que vem dos pulmões pelas cordas vocais (impulsos glotais) e, à direita, o sinal de fala radiado, depois da passagem pelo filtro.

O modelo fonte-filtro abordado anteriormente pode ser aprofundado de forma a ser menos simplista. É importante, por exemplo, reparar que o sinal produzido pela fonte pode ser decomposto em duas componentes: uma componente ruidosa (portanto aperiódica), produzida pela turbulência característica do ar expelido pelos pulmões, e uma componente periódica gerada pelos movimentos cíclicos das cordas vocais. O sinal de fonte é, por conseguinte, a sobreposição destas duas componentes periódica e aperiódica, e isto permite-nos tirar rapidamente conclusões apenas pela observação da forma do sinal no domínio dos tempos [2]. Com efeito, diferentes vozes e diferentes sons caracterizam-se por diferentes pesos de cada uma das duas componentes, isto é, há certos sons ou vozes que apresentam uma representação temporal que se assemelha mais a ruído, ou pelo contrário cujo período é muito facilmente discernível.

Para melhor perceber esta ideia, pode olhar-se para dois casos distintos, por um lado a forma de onda de um sinal resultante de um exercício de vogal prolongada, por outro o resultante de um sinal de voz sussurrada, que se encontram na Figura 2.3.

Como facilmente se verifica após análise dos gráficos nos tempos dos sinais de voz, o sinal de vogal modal (em "voz alta") apresenta um comportamento claramente periódico, devido à vibração das pregas vocais [3], enquanto que no caso da vogal sussurrada, não se consegue verificar nenhum comportamento periódico, uma vez que é o resultado de ar a ser expelido pelos pulmões, sem vibração das cordas vocais.

[1]Imagem extraída de https://www2.leeward.hawaii.edu/hurley/Ling102web/mod3_speaking/mod3docs/3_images/midsagittal_bw.jpg

[2]Imagem extraída de http://www.ling.cam.ac.uk/li9/lab3_m08_speechandspectralanalysis.pdf

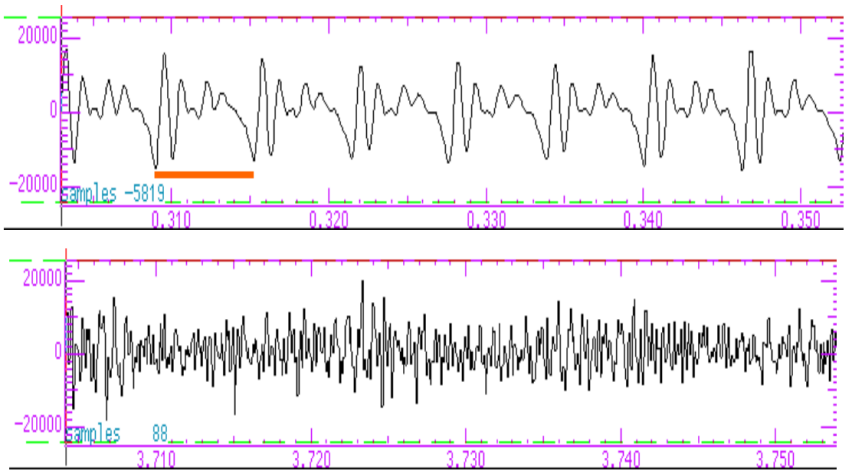


Figura 2.3: *Em cima, representação de vogal sustentada (correspondente ao som "A") no domínio dos tempos; em baixo, mesma vogal mas sussurrada*[2]*. Nesta figura, o eixo horizontal representa tempo (em segundos) e o eixo vertical representa amplitude.*

Tendo sido observada a representação gráfica de sinais nos tempos, é importante falar numa perspetiva de análise de frequências mas, para isso, é necessário compreender certas noções, que são abordadas a seguir.

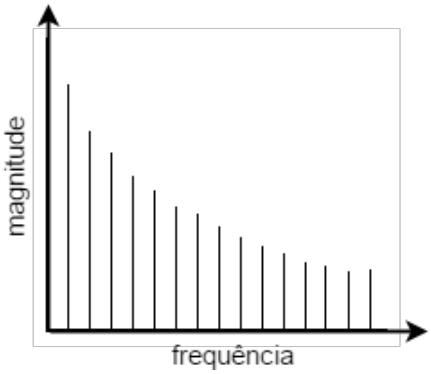
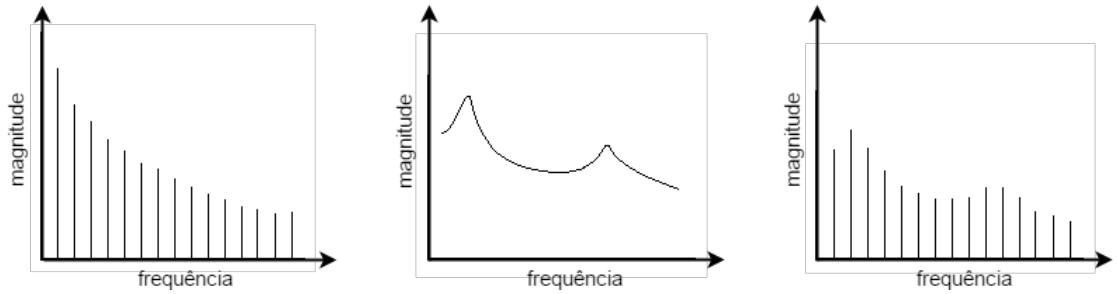


Figura 2.4: *Representação do sinal do impulso glótico no domínio das frequências*

Para se estudar um sinal do ponto de visto da frequência, é necessário passar do domínio dos tempos para o domínio das frequências, o que é obtido aplicando a transformada de Fourier. O sinal mais simples que se possa imaginar, uma sinusóide pura, uma vez que possui apenas uma frequência, tem uma transformada de Fourier que consiste apenas numa risca correspondente à frequência da sinusóide, em casos ideais [2].

Sinais periódicos mais complexos, que possuem um número variável de componentes em frequência, apresentarão espetros com mais riscas. Com efeito, sinais periódicos ideais podem ser sempre decompostos em sinusóides [4].



(a) *(b)* *(c)*

Figura 2.5: *Transformação do sinal glótico pelo filtro do tracto vocal*

Observando a Figura 2.4, verifica-se que o impulso glótico, excluindo a componente ruidosa, possui uma transformada de Fourier que consiste em várias riscas decrescentes monotonamente à medida que a frequência aumenta. A primeira risca é denominada **frequência fundamental**, por ser aquela que corresponde a uma sinusóide de menor frequência e independentemente de possuir maior magnitude. As restantes riscas são designadas de **harmónicos** encontrando-se a frequências múltiplas da frequência fundamental.

A designação de filtro corresponde, como já anteriormente mencionado, a todos os elementos do sistema fonatório que modificam a estrutura espectral (ou seja, no domínio das frequências) do sinal de voz. No entanto, pode-se subdividir este filtro em dois mais específicos: um que representa as modificações aplicadas ao sinal com base na forma do trato vocal, e outro que representa os efeitos de radiação. O filtro de trato vocal é o responsável por criar ressonâncias em certas frequências, chamadas formantes, e que conferem as características próprias aos diferentes sons que usamos para comunicar [5]. A variabilidade deste filtro ao longo do tempo também acrescenta identidade à voz humana, pois existem pequenas flutuações que fazem com que o timbre de uma voz não seja robótico.

A Figura 2.5 representa a alteração que o filtro de trato vocal aplica ao sinal gerado pelas cordas vocais. As Figuras 2.5a e 2.5c correspondem, respetivamente, às representações no domínio das frequências, dos sinais temporais representados nas Figuras 2.2a e 2.2b. Na Figura 2.5b pode observar-se a resposta em frequência do filtro. Os dois picos representam **formantes**, associadas a ressonâncias no trato vocal que conferem um ganho mais elevado nas frequências correspondentes. A Figura 2.5c representa o sinal à saída da boca, e que é resultado da aplicação do filtro ao sinal gerado na glote. Observa-se que os harmónicos situados nas regiões dos formantes têm maior magnitude, o que confere um timbre característico à voz. No entanto, estas transformações em nada alteram a frequência fundamental do sinal.

2.1.2 Espectrograma

Previamente, foram observados mecanismos de produção de voz, bem como duas formas de representar o sinal de voz de uma forma bidimensional:

- nos **tempos**, em que se representa a amplitude do sinal nas ordenadas e o tempo nas abcissas
- nas **frequências**, em que se representa a densidade espectral ou magnitude nas ordenadas e a frequência nas abcissas

Estes dois tipos diferentes de representação permitem verificar diferentes características quanto ao sinal em análise e devem, portanto, ser usados de forma complementar. No entanto, gráficos bidimensionais são uma forma muitas vezes demasiado simplista de representar sinais e muitas vezes são insuficientes numa análise mais aprofundada. É por isso muito importante, em muitos casos, poder analisar um sinal em frequência ao longo do tempo. Isto obtém-se através de uma representação tridimensional designada por **espectrograma**. Com efeito, quando observamos a densidade espectral de um sinal ao longo das suas frequências, falta-nos o elemento temporal, isto é, conseguimos apenas observá-lo num determinado instante, não tendo por isso a noção da sua variação ao longo do tempo. O espectrograma é portanto obtido juntando as diferentes representações espectrais ao longo do tempo e "colando-as", obtendo assim uma única representação tridimensional para todas elas.

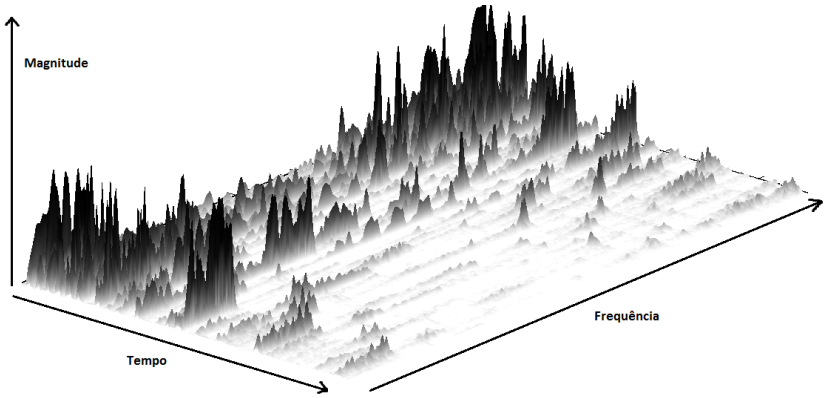


Figura 2.6: *Espectrograma tridimensional*

A Figura 2.6 representa esta ideia para representar um sinal a três dimensões. Como se observa, o conteúdo espectral do sinal vai variando ao longo do tempo, tendo algumas frequências maior densidade em alguns instantes de tempo do que noutros, como pode ser mais facilmente verificado com a ajuda das cores. Neste exemplo, foram utilizadas cores da escala dos cinzentos, com as cores mais escuras a representar as zonas em que a densidade espectral é mais elevada.

Este espectrograma, apesar de interessante, é relativamente difícil de ler, devido ao facto de a representação incluir as ideias de perspetiva e profundidade. Foi necessário então representar os espectrogramas de uma forma mais facilmente decifrável. Isto pode ser feito aproveitando uma ideia já anteriormente mencionada, que é a de incluir cor na representação. Se a cor representar a magnitude, podem ser representados apenas os eixos do tempo e da frequência, dando as diferentes cores as sensações de relevo. Usando as cores da Figura 2.6, poder-se-iam representar as magnitudes com cores da escala dos cinzentos. No caso de espectrogramas a cores, convencionou-se então que as magnitudes mais elevadas seriam representadas por cores mais quentes (mais próximas do

vermelho), e as mais baixas seriam representadas por cores mais frias (mais próximas do azul).

O espectrograma da Figura 2.7 representa um sinal de voz que corresponde à dicção de vogais com e sem pausas entre elas. As primeiras considerações que devem ser imediatamente retiradas são que o eixo das abcissas representa o tempo, e que a gravação tem portanto uma duração que ultrapassa ligeiramente os dez segundos, e que o eixo das ordenadas representa as frequências, sendo que estas estão representadas até aos 6000 Hz, sendo que o conteúdo espectral para além desse valor é quase nulo, tratando-se de harmónicos de muito baixa magnitude.

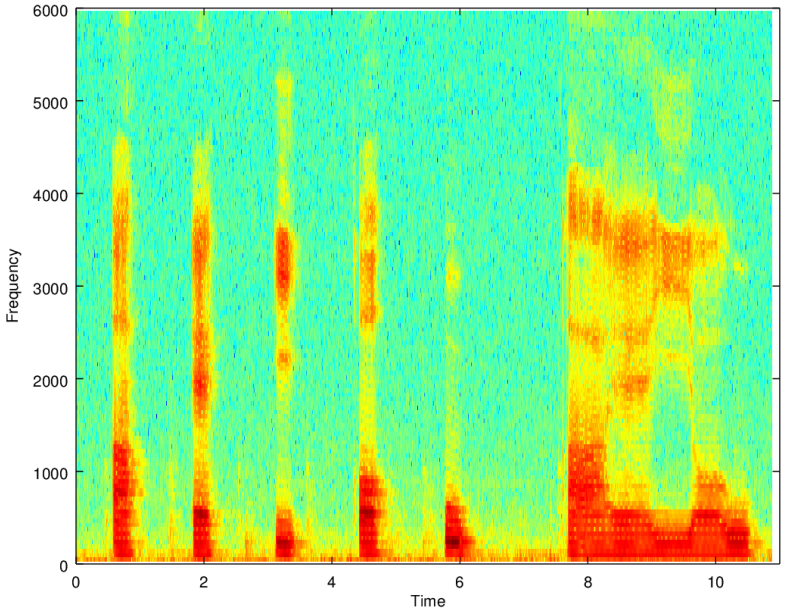


Figura 2.7: *Espectrograma do sinal de fala que corresponde à produção das vogais «a e i o u», primeiro com pausas entre cada uma, e seguidamente de forma contínua, para um indivíduo do sexo masculino.*

São facilmente reconhecíveis 5 "riscas" verticais em tons que vão do amarelo ao vermelho, durante os primeiros 7 segundos, e que correspondem a cada uma das vogais "a, e, i, o, u". Um indivíduo do sexo masculino tem normalmente uma frequência fundamental da voz falada situada entre os 80 e os 150 Hz, o que é coerente com o conteúdo do espectrograma, apesar de este ter demasiado pouca precisão para se determinar o *pitch*. É interessante verificar que cada vogal é diferente na medida em que cada uma tem um diferente peso dos harmónicos ao longo das frequências. Por exemplo, se for comparado o som "a" com o som "i", verifica-se que o primeiro tem um densidade espectral relativamente plana desde a frequência fundamental até cerca dos 1200 Hz, sendo que os harmónicos seguintes decrescem de magnitude até que voltam a aumentar entre os 2500 e os 4000 Hz, enquanto que a vogal "i" possui harmónicos fortes (formantes, para usar um termo já mencionado anteriormente), até cerca dos 500 Hz, à volta dos 2100 Hz e novamente dos 2800 aos 3600 Hz. Estas diferenças da envolvente espectral do sinal são causadas pelas variações

dos filtro do trato vocal e de radiação, mencionados anteriormente.

Seguidamente, quando analisamos a parte do espectrograma relativa às vogais sem interrupções, já não existem zonas maioritariamente azuis correspondentes às pausas que permitam muito facilmente delimitá-las. No entanto, podemos usar as diferenças de envolvente espectral para identificar as transições entre as vogais e identificá-las. A Figura 8 mostra mais claramente, com a ajuda de linhas pretas, as zonas dos espectrogramas que correspondem a cada um dos sons.

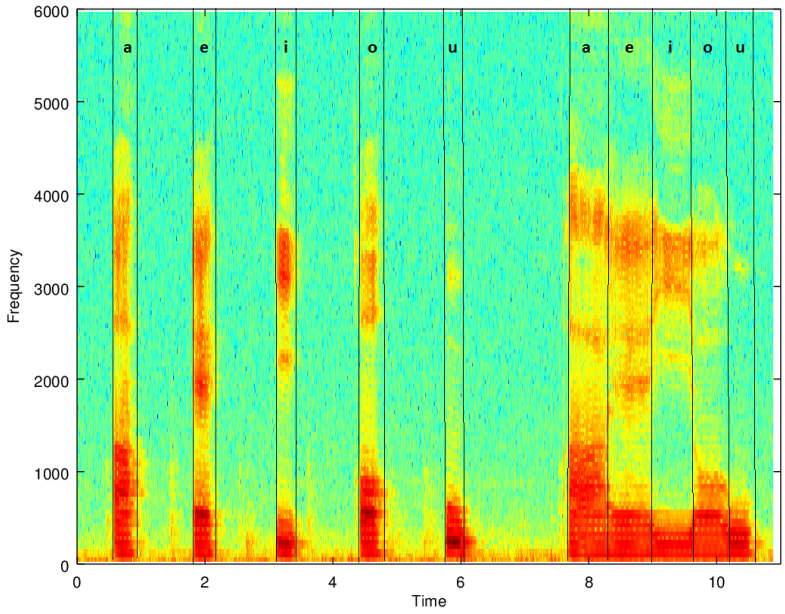


Figura 2.8: *Mesmo espectrograma presente na Figura 2.6, desta vez com as vogais identificadas.*

Tendo já sido analisado um caso real com a ajuda do espectrograma, já estão colocadas as bases para a compreensão de certos tipos de características que devem existir para a constituição de um bom discurso e que constituem o objetivo deste trabalho.

2.2 Qualidade de discurso

A utilização eficaz da voz para efeitos de comunicação é uma capacidade extremamente valorizante para as pessoas, especialmente as que necessitem de utilizar a fala para transmitir mensagens a ouvintes. A maneira como se comunica é tão importante como aquilo que se comunica, pois uma mensagem, por muito boa que seja, não interessará ou convencerá uma plateia se for divulgada de uma forma pouco interessante.

Em primeiro lugar, é necessário apurar quais os elementos fundamentais que um discurso tem de ter para ser considerado de qualidade. Após esse passo, será fundamental identificar quais

desses elementos podem ser traduzidos por critérios objetivos de forma a que possam ser detetados sem intervenção humana direta.

2.2.1 Características de um bom discurso

Existem algumas características que podem ser tidas como fundamentais no que toca a avaliar a qualidade de um orador, e que são enumeradas seguidamente [6] [7]:

- **Variação de *pitch***: Este atributo é relacionado com a monotonia do discurso. Uma vez que *pitch* significa frequência fundamental, um fraca variação deste parâmetro fará com que o orador fale sempre com a mesma frequência, o que torna o discurso pouco cativante e leva o ouvinte a desinteressar-se rapidamente. Uma rica variação de *pitch* corresponde a um discurso com maior entoação, que permite transparecer emoções e sentimentos diferentes.

- **Variação de volume sonoro**: É muito importante que exista variação de volume no discurso, ou seja, que algumas palavras sejam pronunciadas com maior intensidade. Este fator acompanha muitas vezes a variação de *pitch*, sendo que permite realçar certas partes de um discurso, conferindo-lhe maior dinamismo.

- **Pausas**: É importante que um discurso seja marcado por pausas em certos momentos. Estas permitem a mais fácil delimitação e assimilação pelo ouvinte das ideias do orador, assim como a criação de pontos de *suspense* no discurso.

- **Cadência silábica**: Pode ser mencionada de uma forma mais familiar como rapidez de discurso. Se for demasiado elevada, perde-se inteligibilidade, e há menos tempo de assimilação para o ouvinte, que faz com que as ideias não sejam consolidadas.

- **Articulação**: Este conceito relaciona-se com a clareza na pronúncia de cada sílaba. Sem uma boa articulação perde-se compreensão e inteligibilidade, o que naturalmente leva à degradação da capacidade de comunicação.

Além destes atributos, existem também certos critérios menos facilmente avaliáveis a ter em conta, tal como a soprosidade de uma voz, que pode também provocar mais difícil compreensão.

Os parâmetros fonatórios de maior relevo para esta dissertação são a variação de *pitch* e de volume, sendo que constituem valores passíveis de serem calculados com alguma precisão, além de permitirem avaliar a dinâmica de um discurso.

2.2.2 Extensão vocal

Uma das formas de representação mais úteis na avaliação de certas características fonatórias é o chamado "fonetograma"[8]. Este gráfico bidimensional permite visualizar o *Vocal Range Profile* (termo que, em português, poderia ser traduzido por "Perfil de extensão vocal") de um indivíduo, sendo que no eixo das abcissas se encontra representada a frequência, medida em Hertz (Hz), e no eixo das ordenadas a intensidade sonora ou *Sound Pressure Level* (SPL), medida em deciBel (dB).

Este tipo de representação permite, portanto, representar a extensão vocal de um indivíduo tanto a nível das frequências como das intensidades. A partir da forma adotada pelo VRP, podem extrair-se características de elevada importância para um estudo mais aprofundado das características vocais de um indivíduo.

Além de um eixo horizontal de frequências, um fonetograma pode possuir também um segundo referencial que toma o aspeto de um teclado de piano, o que fornece mais um indicador de medição da frequência fundamental da voz, permitindo associar esta a uma nota musical. Passa a ser facilmente possível, portanto, verificar a extensão de frequência da voz em semitons da escala musical, permitindo uma leitura e interpretação mais intuitiva dos resultados. Uma das principais vantagens de usar semitons para descrever uma diferença em frequência prende-se com o facto de esta escala ser exponencial, tal como a escala musical. Esta mapeia de forma mais aproximada a forma como o ouvido humano distingue diferentes tons, sendo por isso mais intuitiva[9].

Este tipo de representação possui várias aplicações, sendo mais usada em contexto musical, nomeadamente para treino de cantores, e medicinal, em que são muitas as suas aplicações na deteção de anomalias e patologias associadas ao sistema fonatório, bem como no acompanhamento de doentes.

Existem várias medidas que podem ser retiradas da leitura de um fonetograma, sendo que as mais comuns dizem respeito a valores máximos e mínimos de intensidade e frequência fundamental, bem como a diferença entre essas medidas.

A Figura 2.9 representa o *Vocal Range Profile* (VRP) de um indivíduo do sexo masculino, não sofrendo de qualquer anomalia no sistema fonatório. Verifica-se que este sujeito consegue atingir uma gama de frequências que se estende, aproximadamente, dos 80 Hz aos 880 Hz, e uma gama de intensidades dos 60 dB aos 110 dB.

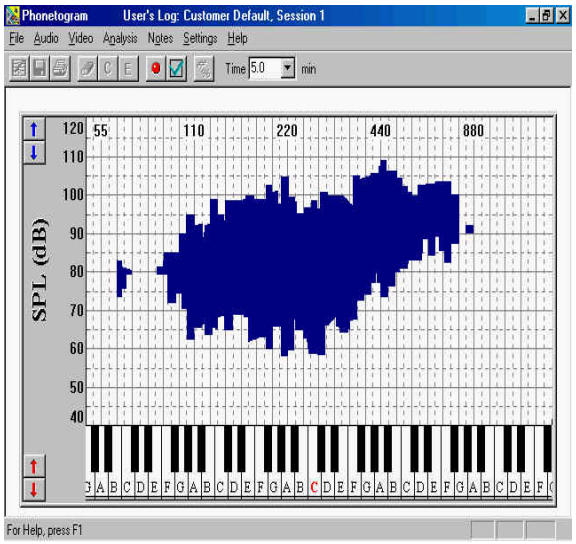


Figura 2.9: *Fonetograma correspondente a um indivíduo do sexo masculino*

Mais do que isto, verifica-se que o ser humano não consegue produzir sons à mesma intensidade para toda a sua gama de frequências. Com efeito, um aumento de frequência na voz é acompanhada por um aumento de intensidade, conferindo um declive característico. Este fenómeno explica-se pelo facto de que a tensão nas cordas vocais e no peito aumentam, o que conduz a um aumento da pressão sub-glotal, levando a uma maior intensidade [10].

Uma característica imediatamente reconhecível num fonetograma deste tipo é a quebra de intensidades existente no limite superior do gráfico (neste caso específico, na zona dos 300 Hz). Este fenómeno é explicado pela passagem do registo de voz "de peito" para falsete[11], de forma a permitir produzir sinal de voz mais agudo.

O *Vocal Range Profile* representa, portanto, a extensão total da voz de um indivíduo. No entanto, no caso particular da fala, apenas uma relativamente pequena fração desta região é utilizada, região essa que depende inteiramente do *pitch* médio do orador, bem como da sua variação em frequência e intensidade. Tipicamente, a frequência fundamental média de um orador situar-se-á entre os 80 e os 150 Hz no caso de um homem, e entre os 150 e os 260 Hz no caso de uma mulher [12]. No que diz respeito à intensidade sonora, por exemplo, no caso de um microfone colocado a 30 centímetros da boca, o SPL médio de fala de um ser humano estará situado entre os 60 dB e os 80 dB [11].

É possível, portanto, mapear uma determinada amostra de discurso num fonetograma e analisar o *Vocal Range Profile* correspondente, de modo a tirar conclusões importantes quanto à variedade utilizada pelo orador, relacionando os resultados com características importantes de um discurso tal como a variação de *pitch* e a variação de intensidade (ver secção anterior).

A visualização destas duas características através da representação do VRP pode ser feita, tendo em conta que uma elevada variação de *pitch* poderá ser visualizada como uma maior extensão da representação relativamente ao eixo das abcissas. O mesmo se verifica quanto à variação de intensidade, mas neste caso relativamente ao eixo das ordenadas.

O tipo de representação do VRP presente na Figura 2.9 tem, no entanto, uma limitação óbvia. Apesar de se poderem verificar as potencialidades de um sinal de voz quanto aos limites máximos e mínimos que podem ser atingidos, este tipo de representação não fornece qualquer informação visual sobre a intensidade e *pitch* médios de um discurso, bem como sobre a sua variância. Ora, no contexto de uma análise de discurso, não é apenas importante representar as notas e pressões sonoras que foram atingidas, é necessário também ter uma ideia de qual foi a distribuição e frequência de ocorrência daquelas.

Consideremos 3 tipos de oradores:

1. Mantém um tom monótono ao longo de todo o discurso;

2. Comunica em registo maioritariamente monótono, tendo apenas um breve período de entusiasmo;

3. Comunica eloquentemente, de uma forma variada e adequada à situação.

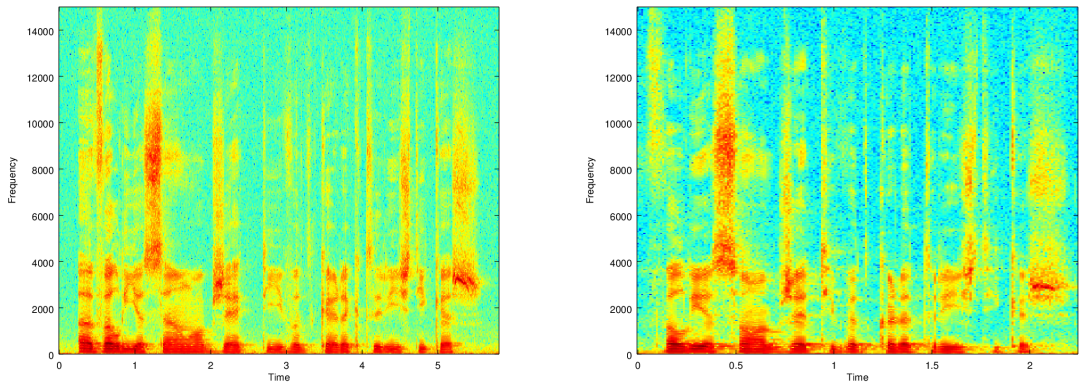
Caso fossem representados num fonetograma os sinais correspondentes a cada um destes oradores, poder-se-iam tirar certas conclusões. Verificar-se-ia que o orador 1 possuiria um VRP cuja região seria muito limitada, denotando uma pequena utilização da sua gama de frequências e intensidades. No caso dos oradores 2 e 3, com uma representação semelhante à da Figura 2.9, não seriam observadas grandes diferenças, apesar de o discurso do orador 3 ser de melhor qualidade.

De forma a diferenciar estes dois casos, é necessária a inserção de algum tipo de indicação quanto às frequências e intensidades mais utilizadas pelo orador, ou seja, uma medida de média e variância. Isto pode ser feito através da introdução de cores ou transparência, dando maior realce às regiões mais frequentemente utilizadas e tornando menos percetíveis as regiões pouco utilizadas.

2.2.3 Avaliação objetiva de características

Tendo já sido identificadas algumas das características mais importantes que contribuem para um discurso com qualidade, é necessário encontrar formas de proceder à sua avaliação objetiva, de forma a que estas possam ser medidas e representadas de forma compreensível, o que constitui o objetivo principal do trabalho a realizar.

A **cadência silábica** é um fator relativamente simples de calcular tendo como base a análise de espectrogramas. Na Figura 2.10 apresenta-se a comparação de espectrogramas que correspondem à produção da mesma expressão pelo mesmo indivíduo, dita primeiro de forma lenta, e seguidamente de forma rápida.



(a) *Baixa cadência silábica* ***(b)*** *Elevada cadência silábica*

Figura 2.10: *Espectrograma do sinal de fala que corresponde à produção da frase «avaliação objetiva de características», para um indivíduo do sexo masculino, com cadências silábicas diferentes*

Em primeiro lugar, comparativamente com a situação analisada na secção anterior, e que consistia na representação de vogais, este caso inclui representação de consoantes, dado tratar-se de uma frase. Estas possuem uma lógica diferente da das vogais. Com efeito, enquanto que as vogais correspondem a sons vozeados, ou seja, existe vibração das cordas vocais, as consoantes são, em

grande parte, produzidas sem vibração das cordas vocais (sons não vozeados), portanto correspondem apenas a ar vindo dos pulmões a sofrer alterações com a ajuda de elementos que constituem o trato vocal, não fazendo por isso sentido falar em *pitch* para estes sons. Estas consoantes, conhecidas como desvozeadas[3], podem ser divididas em vários grupos, que não serão aqui descritos. Importa apenas saber que certas consoantes correspondem apenas a bloqueio de ar com a língua ou lábios, correspondendo por isso apenas a instantes de transição entre vogais (tais como por exemplo sons como "p" ou "t"), enquanto que outras podem ser "sustentadas", ou seja, consistem num fluxo de ar constante moldado pelo trato vocal, como é o caso das sibilantes ("ç" ou "z", por exemplo).

Se tomarmos como exemplo a região da Figura 2.10a que começa pouco depois do 1 segundo e que possui uma forte densidade espectral na região dos 4 aos 11 kHz, podemos deduzir que esta corresponde à produção da letra "ç" em "avaliação". Isto pode ser deduzido uma vez que apresenta um comportamento completamente diferente do das vogais, que possuem elevada densidade espectral apenas até cerca dos 4 kHz, e não apresenta um conteúdo harmónico marcado (ausência de "riscas" na horizontal), assemelhando-se mais a ruído. Este tipo de densidade espectral aleatória é esperada nas sibilantes, uma vez que estas são produzidas por ar turbulento a ser continuamente expelido. Volta a verificar-se este comportamento no último som apresentado no espectrograma, e que corresponde à produção da última letra da palavra "características", que também é uma sibilante correspondente ao som "ch".

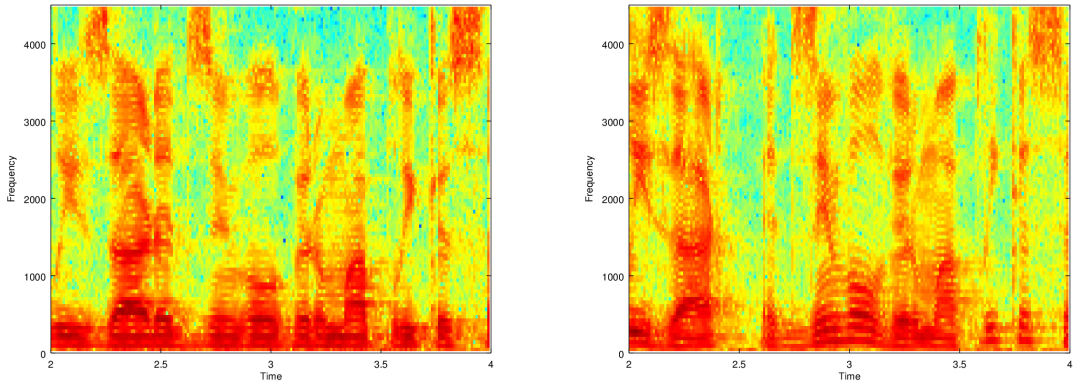
Comparando os dois espectrogramas da Figura 2.10, verificamos que existe uma forte semelhança entre ambos, como seria de esperar, e que a única diferença notável que se pode verificar é na escala temporal, que indica que um som foi produzido em cerca de 5.5 segundos, enquanto que o outro o foi em pouco mais de 2 segundos.

A duração das vogais é facilmente calculada a partir desta amostra. A partir de valores nominais que delimitem a rapidez de um discurso, pode ser calculada a duração média das vogais num determinado intervalo de tempo e verificar se esta se encontra acima ou abaixo dos limiares recomendados.

Outro dos aspetos importantes a considerar para um orador é a sua **variação de *pitch***, ou seja, o uso de inflexões e variações de frequência fundamental com o objetivo de obter um discurso mais dinâmico.

Com vista a comparar as diferenças do ponto de vista espectral entre um discurso monocórdico e um discurso dinâmico, obtiveram-se os espectrogramas correspondentes à gravação da mesma frase, primeiro de uma forma completamente monótona, e seguidamente com variações de *pitch*. A Figura 10 corresponde à ampliação dos espectrogramas de forma a serem evidenciadas as diferenças. Foram selecionadas frequências dos 0 aos 4500 Hz, dado que é neste intervalo que estão as frequências fundamentais e formantes das vogais, que por serem sons vozeados são aqueles em que facilmente se denotam inflexões e variações de frequência.

[3]Há também consoantes vozeadas ou sonoras como "m" ou "n".



(a) *Tom monocórdico* ***(b)*** *Tom dinâmico*

Figura 2.11: *Espectrogramas de sinais de fala correspondentes à produção da mesma frase, com diferenças de entoação*

Como se pode observar pela comparação das vogais nestas duas representações, as que correspondem à fala monocórdica possuem harmónicos quase perfeitamente horizontais, o que significa que a frequência fundamental e os seus harmónicos se mantiveram sempre constantes, denotando monotonia. Na representação da direita, observa-se pelo contrário que as linhas laranjas e vermelhas que constituem os harmónicos são curvadas, o que mostra que houve inflexões nessas vogais, o que lhes faz transparecer maior emoção.

No que toca às outras características que conferem carisma a um discurso, algumas delas podem ser facilmente verificadas. A frequência das pausas pode ser facilmente verificada na representação temporal do sinal de voz, verificando a frequência com que a amplitude do sinal fica no nível do ruído. A variação de volume também pode ser verificada no domínio dos tempos, pois a variação dos picos de amplitude do sinal de voz dão uma medida das variações de intensidade.

Outra característica bastante importante que pode ser extraída de uma análise nas frequências é o grau de soprosidade de uma voz, que influencia a distinção da voz e a sua projeção de modo a ser entendida sem dificuldade por uma larga audiência. Com o apoio de um espectrograma, uma voz soprosa aparece com maior "ruído" associado, isto é, os harmónicos não se encontram tão bem marcados, existindo naturalmente um maior "espalhamento" nas frequências do que numa voz brilhante.

2.3 Análise de aplicações existentes

Dentro destas aplicações não foi encontrada nenhuma que cumprisse exatamente todos os objetivos propostos neste trabalho, mas algumas das suas funcionalidades coincidem com aquilo que se

pretende desenvolver, além do interesse que existe em observar as estratégias que foram utilizadas a nível de usabilidade e implementação.

Seguidamente, são apresentadas três das aplicações encontradas que potencialmente serão mais interessantes tanto do ponto de vista das suas funcionalidades, como do ponto de vista da sua organização e cotação.

2.3.1 OperaVox (sistema iOS)

Esta aplicação enquadra-se mais no contexto de análise médica, na medida em que permite medir a qualidade de voz graças a parâmetros como *shimmer*, *jitter*, *Maximum Phonation Time* (o tempo máximo em que se consegue prolongar uma nota), *pitch* e *pitch range*.

Estes parâmetros são maioritariamente aplicados à área da saúde, ajudando no diagnóstico de perturbações fonatórias. É incluída uma funcionalidade de histórico, de forma a poder acompanhar uma possível evolução na qualidade dos parâmetros avaliados. A capacidade de detetar a frequência fundamental e a sua extensão ao longo de um intervalo de tempo constitui um indicador interessante e de elevado interesse para o trabalho que se pretende desenvolver.

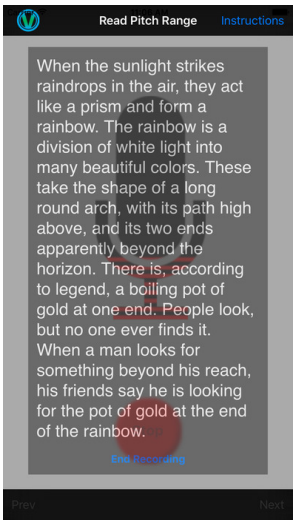


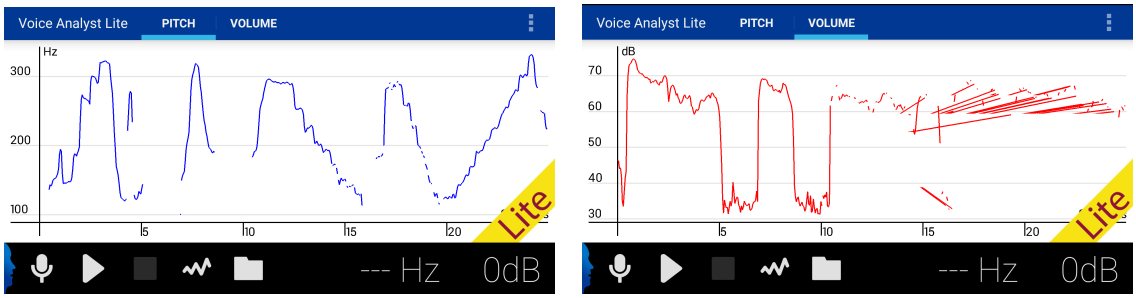
Figura 2.12: *Funcionalidade de medir o pitch máximo, mínimo e médio na leitura de um texto*

A avaliação dos características fonatórias mencionadas é feita através da leitura de uma amostra de texto, sem qualquer acompanhamento de gráficos que descrevam a sua variação em tempo real, como pode ser observado na Figura 2.12. Isto torna a interface gráfica muito enfadonha, sem qualquer acompanhamento em tempo real da análise que é efetuada.

Por estas razões, apesar da inclusão de estudos interessantes de certas características fonatórias, esta aplicação possui grandes lacunas no apelo ao utilizador.

2.3.2 Voice Analyst (sistema Android)

Esta aplicação é principalmente direcionada para cantores, bem como para diagnóstico de perturbações na voz. É possível verificar, em tempo real, a variação do *pitch* e da pressão sonora do sinal vocal captado pelo microfone, que são registados ao longo do tempo em dois gráficos distintos, que podem ser observados na Figura 2.13.



(a) *Representação da variação de pitch* ***(b)*** *Representando da variação de volume*

Figura 2.13: *Capturas de ecrã ilustrativas de gráficos representativos do sinal de voz, em tempo real.*

A aplicação apresenta também estatísticas sobre a precisão da frequência fundamental, que apenas estão disponíveis na versão paga e portanto não puderam ser testadas.

A medição e representação gráfica em tempo real das variações de frequência fundamental e volume sonoro, em dB, constitui um aspeto de elevado interesse no contexto da realização desta dissertação. Conferem uma dinâmica interessante à aplicação, do ponto de vista do utilizador.

A possibilidade de gravar e reproduzir o sinal gerado constitui também uma adição interessante, que pode constituir um objetivo secundário de desenvolvimento.

2.3.3 Speech Master (sistema Android)

A aplicação *Speech Master* possui um *design* rudimentar, mas particularidades interessantes e originais.

Após a sua execução, é necessário o utilizador indicar o tipo de contexto de fala, como por exemplo entrevista, conversa de escritório ou debate. De seguida, inicia-se a gravação de voz, sendo que o discurso proferido é submetido a *speech recognition* e impresso na interface da aplicação, sendo que no fim o utilizador tem que dar fim à gravação através de um botão.

Após a gravação, ocorre o processamento e avaliação de parâmetros tais como a variação de volume, cadência, variação de cadência, bem como as emoções associadas à pronúncia de certas palavras, fazendo uma avaliação da qualidade do discurso em função do contexto escolhido, uma vez que em função do tipo de discurso, os parâmetros de referência para as características referidas são modificados.

Apesar de possuir características singulares que a distinguem das outras aplicações pesquisadas e testadas, esta possui uma interface gráfica muito pouco apelativa, e que possui muito pouca dinâmica, não se podendo verificar a variação dos parâmetros testados em tempo real. Além disto,

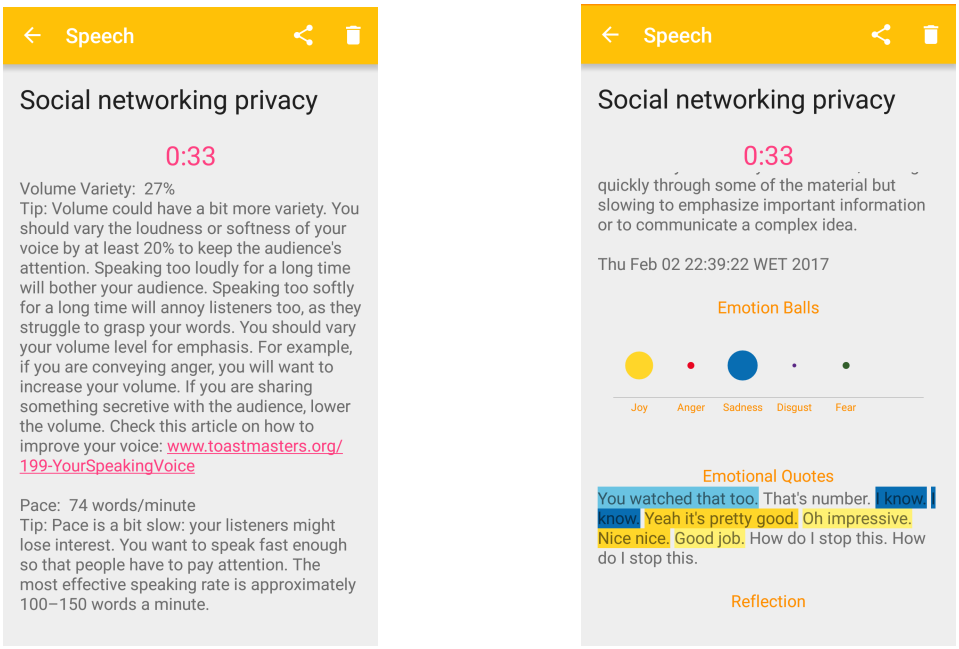


Figura 2.14: *Capturas de ecrã mostrando avaliação de resultados do discurso*

os algoritmos desenvolvidos para a análise subjetiva da qualidade de discurso não são conhecidos, o que constitui uma falta de transparência que não inspira confiança.

2.4 Conclusão

Ao longo deste capítulo, foram vistos conceitos de análise e representação do sinal de voz estudados, de modo a ter uma perspetiva sobre o tipo de informação a representar na interface da aplicação, e a forma de o fazer, de modo a satisfazer os objetivos propostos nesta dissertação e produzir resultados satisfatórios do ponto de vista do utilizador.

Além disto, foram analisadas aplicações existentes, de forma a ter uma ideia de soluções existentes atualmente relativamente à análise de parâmetros vocais e comparar diferentes tipos de interface gráfica, que permitam a recolha de ideias para o desenvolvimento deste trabalho de dissertação.

Capítulo 3

Ambiente *Baseline*

A aplicação desenvolvida no contexto desta dissertação apoiou-se em bases pré-existentes, constituídas por duas aplicações desenvolvidas pelo investigador Sérgio Ivan Lopes no contexto do projeto de investigação ARTTS[1]. Estes alicerces foram fundamentais na realização deste trabalho, uma vez que permitiram a sua realização dentro da margem temporal existente, devido facto de terem permitido implementar, de forma rápida, a parte de baixo nível da aplicação. Além disto, puderam ser aproveitadas algumas das suas funcionalidades, de forma a que não fosse perdido muito tempo a implementar algumas das partes mais secundárias da aplicação.

Além disto, foi utilizado um conjunto de ficheiros que implementam funções de processamento de sinal, facultado pelo Professor Aníbal Ferreira, orientador desta dissertação.

3.1 Ambientes pré-existentes

3.1.1 MasterPitch

A aplicação *MasterPitch* foi motivo de estudo e inspiração aquando do início do desenvolvimento prático deste projeto. Trata-se de uma ferramenta utilizada para combater a gaguez, através de métodos simples mas de eficácia comprovada. O princípio básico é a realimentação da voz do orador para os seus próprios ouvidos, utilizando auscultadores, após algumas modificações. Estas permitem alterar a frequência fundamental da voz, assim como regular o *delay* e volume do som produzido. Estas alterações ao *feedback* natural permitem reduzir ou eliminar certos tipos de gaguez.

A aplicação *MasterPitch* permitiu a familiarização com os procedimentos a adotar de forma a criar uma ferramenta deste tipo, nomeadamente em termos de estrutura do código e divisão em módulos independentes mas articulados entre si.

Esta aplicação serviu de base para a ideia de construir uma interface gráfica simples, com uma vista única e sem sobrecarregamento de informação.

[1] www.fe.up.pt/~voicestudies

3.1.2 SingingStudio

A aplicação *SingingStudio* consiste numa plataforma interativa de treino para canto, enriquecida por diversas funcionalidades, tal como a possibilidade de gravar em formato MIDI e WAV, observar a evolução do pitch em tempo real ou construção de um *piano roll* correspondente à gravação.

Foram adaptados alguns blocos constituintes do *SingingStudio* de forma a realizar a aplicação proposta nesta dissertação, nomeadamente grande parte dos módulos de baixo nível, assim como várias ideias, tais como a inclusão de um teclado interativo de piano e o aspeto geral da interface gráfica.

O *design* da aplicação a desenvolver é inspirado do *SingingStudio*, apesar de existir uma diferença muito substancial no tipo de informação apresentada, bem como no propósito do seu desenvolvimento. Com efeito, *SingingStudio* é orientado para canto, sendo, portanto, de cariz musical, enquanto que a aplicação a desenvolver é orientado para a voz falada, utilizada num contexto de comunicação, apesar de também possuir aplicações musicais.

3.2 Módulos adaptados

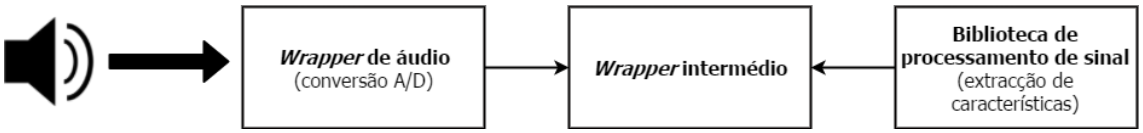


Figura 3.1: *Representação da articulação entre os três principais blocos que constituem a parte de de mais baixo nível da aplicação.*

3.2.1 *Wrapper* de áudio

Tendo em conta os objetivos que se pretendem atingir com o desenvolvimento desta aplicação, a captação de som através do microfone e a sua conversão em formato digital é absolutamente essencial.

A *Apple* possui um grupo de *frameworks*, agrupadas sob o nome de *Core Audio*, que permitem gerar e processar *streams* de áudio, através da criação de *Audio Units*.

O objetivo deste *wrapper* é, utilizando funções contidas na *framework* de áudio da Apple, criar um *stream* de áudio que segue características bem definidas, do qual se vão retirando amostras de sinal. Estas podem ser seguidamente processadas e guardadas num *buffer*, inicializado dentro do *wrapper*. Este *buffer* pode depois ser acedido pela aplicação, podendo o programador da aplicação abstrair-se de tudo o resto, uma vez que a única coisa que importa é saber que o *buffer* vai sendo preenchido com as amostras de sinal em tempo real.

Algumas especificações adicionais quanto ao *stream* de áudio inicializado, e quanto à função de *callback* podem ser dadas, uma vez que conferem certas características ao sinal digital.

O *stream* de áudio gravado é codificado em PCM linear (com quantização uniforme) de 16 bits, com uma frequência de amostragem de 44.1 kHz, ou seja, com qualidade de CD, mas com

a diferença de que, ao invés de usar dois canais, ou seja, modo estéreo, usa apenas um canal, portanto modo monofónico. Neste tipo de aplicação não haveria vantagens em usar modo de gravação estéreo, fazendo sentido usar mono para evitar complexidade desnecessária.

A função de *callback* vai atualizando os *buffers* com novas amostras, convertendo-as previamente para o formato de vírgula flutuante (*float*), de forma a facilitar as operações de processamento de sinal.

São definidas ainda duas funções relativamente simples de inicialização e fim da *Audio Unit*, que são utilizadas para alternar entre os dois modos em que a aplicação funciona, um de captura de som em tempo real, em que a *Audio Unit* precisa de estar cativa, e o outro em que não ocorre gravação de som.

3.2.2 Biblioteca de processamento de sinal

Esta biblioteca contém um ficheiro principal, designado por *PitchMeter*, que contém diversas funções de extração de características de um sinal. Estas funções foram construídas a partir de outras, que implementam operações mais básicas, tal como transformadas, cálculos com números complexos ou leitura de ficheiros temporários, que também se encontram incluídas na biblioteca.

Entre os algoritmos incluídos neste conjunto, encontram-se as duas operações que retornam os resultados pretendidos para atingir os objetivos da aplicação, a saber, o cálculo da frequência fundamental e a obtenção da potência em cada *frame*, convertida em dB.

3.2.3 *Wrapper* intermédio

Este *wrapper* tem como objetivo implementar a articulação entre o *wrapper* de áudio e os algoritmos de processamento de sinal. O *buffer* que contém as amostras de sinal, e que é obtido no *wrapper* de áudio, necessita de ser processado pelas funções de processamento de sinal apropriadas, de forma a obter os resultados esperados, ou seja, neste caso, os valores de intensidade sonora e de frequência fundamental. O *wrapper* intermédio tem conhecimento das funções que permitem a obtenção destes resultados, e invoca-as de forma apropriada.

Após o processamento das amostras, os valores obtidos são retornados ao *wrapper* intermédio, que os guarda em dois vetores diferentes, um de intensidades e outro de frequências fundamentais. Estes vetores podem ser vistos como uma abstração para o resto da aplicação, pois é deles que são retirados os dados utilizados na interface gráfica, sem necessitar de conhecimento algum quanto ao que se passa nas camadas inferiores.

3.2.4 Teclas de piano e MIDI wrapper

Uma das funcionalidades aproveitadas e adaptadas da aplicação SingingStudio é a de um teclado de piano que pode ser tocado.

Tal como referido no Capítulo 2, secção 2, um fonetograma pode ser enriquecido com a representação de um teclado de piano, permitindo comparar valores de *pitch* com a nota musical correspondente e medir a extensão vocal em semitons.

À partida para o desenvolvimento desta aplicação, pretendia-se integrar a representação do piano como uma adição puramente visual. No entanto, dado o facto de já ter sido desenvolvido um teclado funcional, que utiliza o MIDI, foram aproveitados esses desenvolvimentos para enriquecer a aplicação.

De forma a poder acionar as teclas de piano e estas produzirem a nota correspondente, foi criado um wrapper de MIDI. Tal como no caso do *wrapper* de áudio, este necessita de criar uma *Audio Unit* baseada na *framework Core Audio* da *Apple*, que possui também uma biblioteca de MIDI. Mais uma vez, todas estas operações são feitas no *wrapper*, sendo que as únicas funções que são invocadas ao carregar na tecla respetiva são aquelas que estão associadas aos eventos **onPress** e **onRelease**, que invocam os métodos **noteON** e **noteOFF** contidos no MIDI *wrapper*.

Os métodos correspondentes ao desenho das teclas na interface gráfica precisaram de ser modificados e adaptados, na parte de desenvolvimento, à aplicação que constitui o propósito deste projeto, e serão, portanto, mencionados no capítulo seguinte.

3.3 Conclusão

Neste capítulo, foram brevemente abordadas funcionalidades que foram inicialmente desenvolvidas para outra aplicação, que serviu de baseline para este projeto, e que foram integradas no desenvolvimento da nova aplicação. Estas foram de extrema importância para a realização deste trabalho, uma vez que, principalmente o wrapper de áudio e a biblioteca de processamento digital de sinal, se tratam de classes que necessitaram um conhecimento aprofundado para serem desenvolvidas, e que não poderiam ter sido desenvolvidas no âmbito deste projeto pela relativamente escassa margem temporal.

No capítulo seguinte são explicadas as principais etapas do desenvolvimento da aplicação, já realizadas no contexto deste trabalho de dissertação, apoiando-se nas funcionalidades referidas neste capítulo.

Capítulo 4

Desenvolvimento da aplicação

4.1 Abordagem conceptual

Dados os objetivos para esta aplicação em termos de simplicidade de utilização, optou-se por recorrer a um tipo de interface gráfica relativamente simples, constituída apenas por uma vista. Esta é mostrada aquando da execução, e pode ser dividida em três partes fundamentais:

- **Representação temporal:** Consiste num simples gráfico de amplitude em função do tempo, que se vai atualizando à medida que mais amostras de sinal vão chegando, pretendendo ilustrar a ideia de tempo real e de sinal a variar.
- **Fonetograma:** Tal como explicado no capítulo 2, o fonetograma permite visualizar a informação em termos de frequência e pressão sonora. Faz também parte desta representação um teclado de piano.
- **Barra de controlos:** Constitui a interface com o utilizador, permitindo iniciar e parar a captura sonora, alternando assim entre dois modos de representação.

Dada a natureza da aplicação, que possui gráficos a serem alterados em tempo real, é necessário a janela ser atualizada a cada *frame*. Isto implica redesenhar toda a parte da interface gráfica correspondente ao fonetograma 30 vezes por segundo, dado este ser o *frame rate* do *display* da aplicação.

Como se pode verificar na Figura 4.1, os blocos de mais baixo nível, mencionados no capítulo 3, são responsáveis pela obtenção da informação que será apresentada nos dois tipos de representação gráfica. As amostras de sinal obtidas no *wrapper* de áudio são diretamente representadas no gráfico temporal, dado não necessitarem de ser sujeitas a qualquer tipo de processamento adicional. A informação de frequência e pressão sonora, apresentada no fonetograma, é recebida do *wrapper* intermédio, que guarda estes valores em vectores que vão sendo atualizados em tempo real com informação que passou pelos algoritmos de processamento de sinal.

A Figura 4.2 representa a organização funcional da aplicação, dividida em diferentes classes, que representam os diferentes tipos de objetos que fazem parte da aplicação.

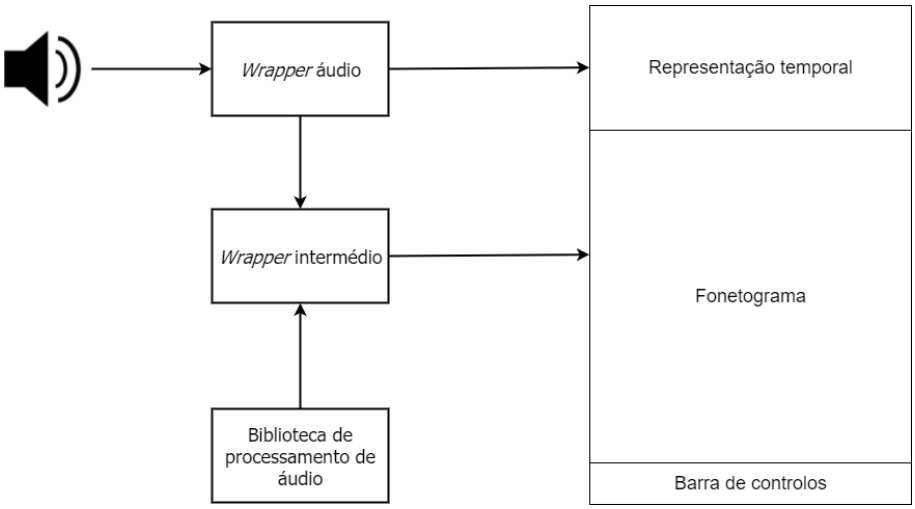


Figura 4.1: *Do lado esquerdo, encontram-se representados os blocos responsáveis pela obtenção das amostras de sinal e seu processamento. Estes valores são representados na interface gráfica da aplicação, que pode ser vista, de forma simplificada, do lado direito dividida nos seus três principais componentes.*

A classe **App** e o **View Controller** apresentam uma importância central nesta organização, uma vez que constituem os primeiro objetos inicializados, imediatamente após a aplicação ser lançada.

O objeto **App** é responsável pelo *setup* da aplicação, tanto pela definição da frequência de amostragem e do tamanho do *buffer* de amostras a ser usado, quer quanto à inicialização dos objetos que dependem dele, que podem ser observados na Figura 4.2. Pode ser visto como a ponte entre a parte de obtenção de dados e a interface gráfica.

Numa aplicação, podem existir vários *view controllers*, que definem os comportamentos da aplicação em função das ações do utilizador. Neste caso, existe apenas um, que está associado à barra de controlos. É inicializado aquando do lançamento da aplicação e corresponde ao único módulo do programa independente do objeto **App**. Comunica com a máquina de estados, que por sua vez informa a aplicação das alterações a fazer em função das ações do utilizador na barra de controlos.

O bloco **GUI** define os métodos responsáveis pela criação do gráfico temporal, além de criar o objeto **PianoKeyboard**. Este último é responsável pelo desenho do fonetograma e das funções associadas a cada uma das representações gráficas, além da inicialização e desenho do teclado de piano, em que cada uma das teclas é definida através de um objeto **PianoKey**.

4.2 Barra de controlos

Todo o controlo realizado pelo utilizador de modo a selecionar a informação apresentada é realizado através da barra de controlos da aplicação, situada na parte inferior da janela. Os módulos que permitem a implementação da barra de controlos são o **ViewController** e a **StateMachine**.

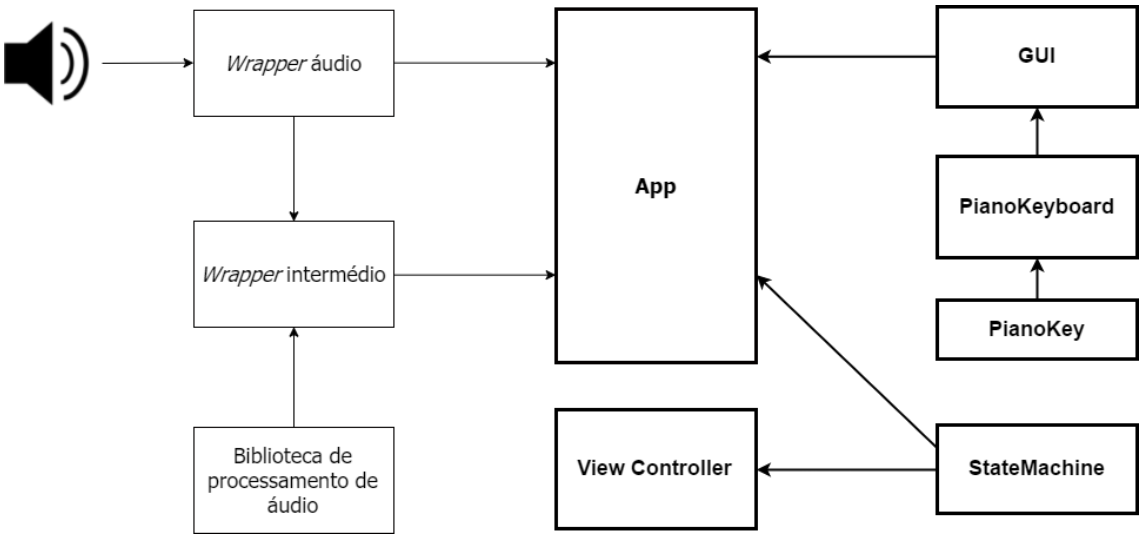


Figura 4.2: *Esquema simplificado de funcionamento da aplicação.*

Um *View Controller* é uma parte da aplicação responsável pelo controlo de objetos de interface com o utilizador. Contém métodos que definem o aspeto visual da barra de controlos, e eventos associados ao clique de cada botão. Cada um destes eventos provoca uma mudança de estado e/ou modo de funcionamento da aplicação. De cada vez que o utilizador aciona um botão da aplicação, é chamado o respetivo evento do **ViewController**, que por sua vez notifica a máquina de estados da mudança de estado. A **StateMachine** (nome do objeto que implementa a máquina de estados), por sua vez, comunica com a classe principal **App**, dando instruções correspondentes à alteração do estado.

Como já foi explicado anteriormente, foram previstos dois modos de funcionamento para a aplicação:

- Modo **inativo**,
- Modo de **captura** (em tempo real).

De forma a alternar entre estes dois modos, existe o botão **start/stop**. Ao ser pressionado, é dada a ordem de começar/parar a captura de som, através das funções do *wrapper* de áudio que são invocadas pelo objeto **App**.

Em função do modo de funcionamento da aplicação, o tipo de representação gráfica apresentada no fonetograma muda, de acordo com o que será explicado na secção 4.4.

Além da opção de iniciar e parar, foi acrescentada a funcionalidade de pausar e retomar a captura. Neste caso, o *input* de sinal áudio é igualmente interrompido, mas não é terminada a sessão de áudio nem reinicializados os vetores de dados, de tal forma que a captura pode ser retomada normalmente. Desta forma o modo de captura tem dois estados associados, **capturing** e **capturing pause**. Esta nova funcionalidade leva à necessidade de criar mais um botão, designado de **pause/resume**.

Estes requisitos permitem chegar ao diagrama de transição de estados apresentado na Figura 4.3.

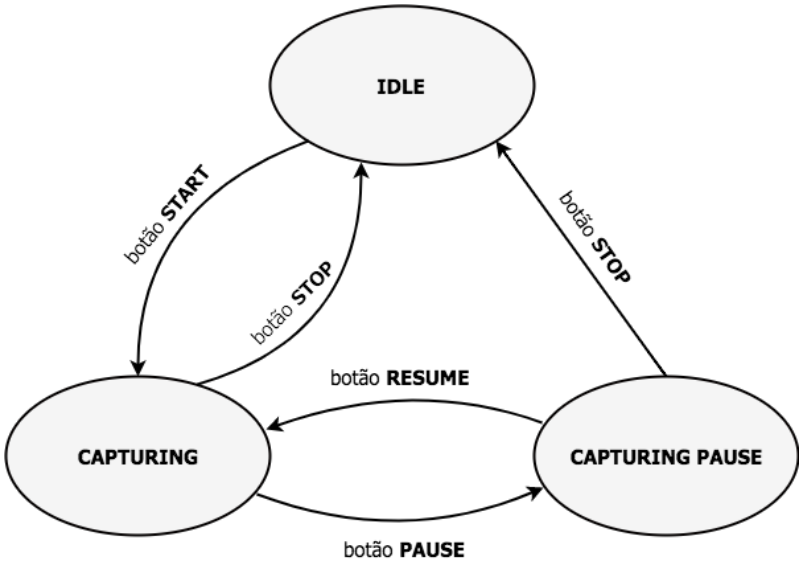


Figura 4.3: *Diagrama de transição de estados associado à barra de controlos. O estado* ***idle*** *corresponde ao modo* ***inativo****, sendo que os estados* ***capturing*** *e* ***capturing pause*** *estão associados ao modo de* ***captura****.*

4.3 Gráfico temporal

As funções de implementação do gráfico temporal encontram-se definidas na classe **GUI**. Numa primeira fase, é necessário reservar espaço para a sua localização na interface gráfica. O espaço reservado corresponde a 100% da largura da janela e 25% da sua altura, situado no topo.

O segundo passo corresponde à inicialização do objecto **ofxUIWaveform**, que implementa o gráfico temporal, e está definido na *framework* utilizada. Os argumentos de inicialização são os valores de largura e altura do gráfico, um apontador para o *buffer* com os valores a representar, comprimento do dito *buffer*, bem como os valores máximo e mínimo presentes na representação, em termos de amplitude.

Dado as amostras do sinal de entrada terem sido normalizadas, os seus valores encontram-se mapeados entre -1 e 1. Estes valores são, por isso, atribuídos aos valores mínimo e máximo na escala das amplitudes.

Devido ao facto de a frequência de amostragem definida no *wrapper* de áudio ser de 44100 amostras por segundo, são geradas muitas mais amostras do que aquelas que podem ser representadas no gráfico.

De forma a obter os valores que são colocados no *buffer* temporal, é necessário proceder a uma sub-amostragem do sinal digital recebido, o que significa que serão apenas conservadas certas amostras do sinal recebido.

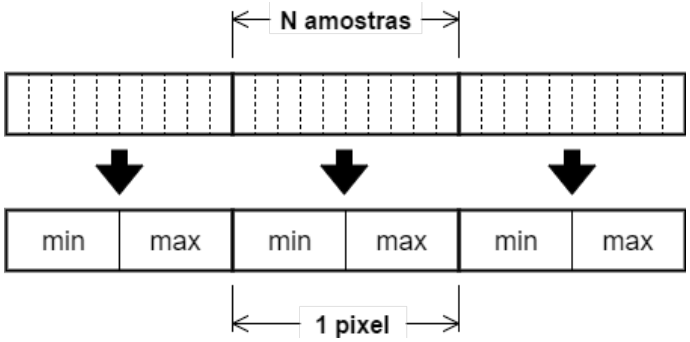


Figura 4.4: *O buffer de amostras de sinal, situado na parte de cima, é dividido em partes. Cada parte é percorrida, são retirados os valores do máximo e mínimo, que são adicionados ao buffer temporal. Cada par máximo-mínimo corresponde à informação a representar em 1 pixel do gráfico temporal.*

O tamanho do *buffer* temporal é definido como sendo o dobro do número de pixéis de largura do gráfico. Isto deve-se ao facto de se pretender guardar o valor máximo e o valor mínimo para cada intervalo de amostras, o que significa ter uma relação de 2 valores do *buffer* por cada pixel de resolução temporal.

O valor da sub-amostragem pode ser calculado como sendo

$$\frac{t \times F_S}{\frac{tamanho\_buffer}{2}}$$

em que t é o tempo, em segundos, representado no gráfico e F_S é a frequência de amostragem. Visto que, para cada intervalo, são guardados os valores máximo e mínimo, o tamanho do *buffer* temporal é dividido por dois. O valor resultante corresponde ao número de amostras do sinal digital original para cada pixel do gráfico temporal, e corresponde ao valor de **N** visto na Figura 4.4.

O gráfico temporal possui dois regimes distintos de funcionamento. O primeiro ocorre enquanto o *buffer* do sinal temporal não está cheio. O gráfico entra no segundo regime de funcionamento quando o seu *buffer* é totalmente preenchido. Neste caso, é necessário retirar amostras das posições iniciais e acrescentar as novas nas posições mais avançadas, funcionando o *buffer* como uma janela deslizante que se vai atualizando em tempo real.

4.4 Fonetograma

A representação de valores de frequência fundamental em função da pressão sonora num fonetograma constitui o ponto mais enfático do desenvolvimento desta aplicação. Deste gráfico faz igualmente parte um teclado de piano, que acompanha a escala de frequências e que pode ser utilizado como referência, de forma a associar o *pitch* obtido a uma nota musical.

O fonetograma desenvolvido tem como objetivo representar dois tipos de informação distintos, dependendo do modo em que a aplicação se encontra, definido pelo utilizador através da barra de

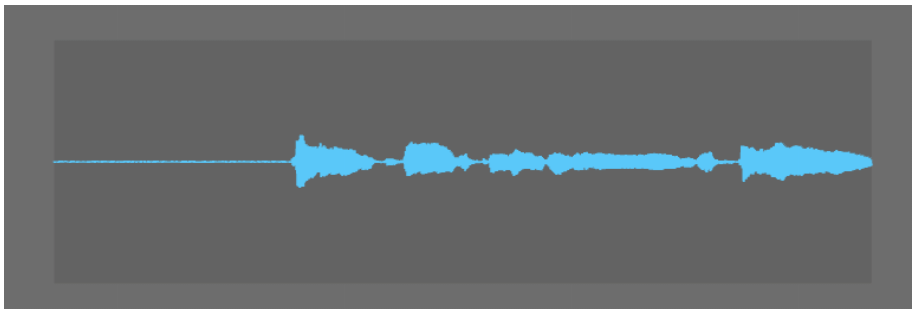


Figura 4.5: *Aspecto final da parte da interface gráfica respeitante à representação temporal.*

controlos. Estes dois modos de funcionamento são descritos nas subsecções 4.4.2 e 4.4.3.

4.4.1 Representação do fonetograma

A interface correspondente ao fonetograma pode ser dividida em duas partes fundamentais: o teclado de piano e o gráfico de *pitch* e SPL, em que são representados os valores obtidos para estes parâmetros.

4.4.1.1 Considerações gerais de implementação

A interface gráfica deve ser desenhada de forma a representar uma extensão de frequências fundamentais que englobe as diferenças de género e faixa etária. Todavia, não deve ser escolhida uma gama de frequências demasiado elevada, devido a vários fatores, entre os quais a falta de precisão de leitura do fonetograma.

A gama de frequências fundamentais num discurso vozeado estende-se dos 80 aos 150 Hz nos homens e dos 150 aos 260 Hz nas mulheres.

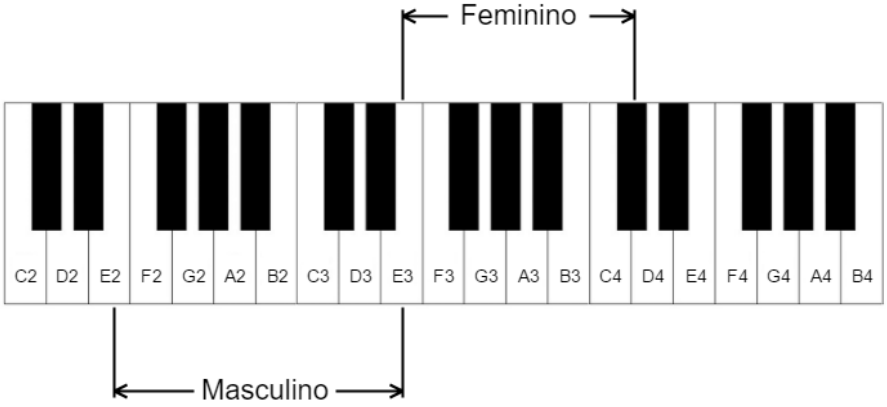


Figura 4.6: *Extensões vocais relativas a indivíduos adultos do sexo masculino e feminino, representadas no teclado que se pretende implementar.*

O desenho do teclado de piano, que deve acompanhar o eixo das frequências, sugere a delimitação em oitavas. Segundo a notação musical, uma escala ou oitava tem início na nota C (dó) e

fim na nota B (si). Deste modo, a primeira tecla representada no fonetograma deve ser a primeira nota C cuja frequência fundamental é inferior ao limite inferior da gama vocal mais grave, correspondente ao sexo masculino. Sendo este valor 85 Hz, a primeira nota a ser representada deve ser o C2 (C da segunda oitava de um piano completo), cujo *pitch* é, aproximadamente, 65.4 Hz. Pela mesma lógica, a última tecla de piano a ser representada deve corresponder a uma nota B cuja frequência fundamental seja superior ao máximo da extensão vocal feminina. Dado este valor ser de 265 Hz, a nota B imediatamente acima (B4) possui um *pitch* de 493.9 Hz.

4.4.1.2 Representação do teclado de piano

O teclado de piano consiste num vetor de objetos da classe **PianoKey**, que contém, entre outras, as seguintes variáveis:

- código MIDI,
- oitava,
- altura da tecla em percentagem da altura do piano,
- largura da tecla em percentagem da largura do piano,
- posição da tecla no eixo das abcissas, em percentagem da largura do piano,
- cor da tecla.

O código MIDI corresponde a um valor numérico de 0 a 127 que identifica notas da escala musical. As notas que se estendem do C2 ao B4 podem, então, ser representadas pelos seus respetivos códigos, que correspondem aos números inteiros entre 36 e 71 (inclusive).

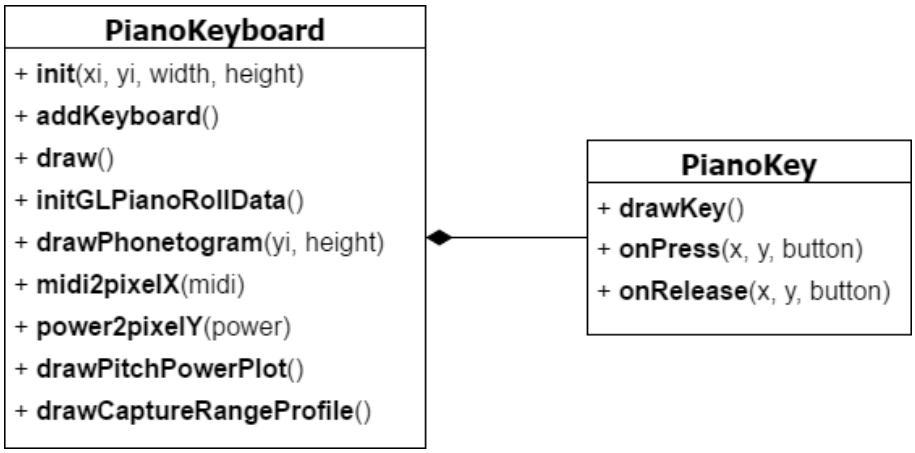


Figura 4.7: *Estrutura básica das classes **PianoKeyboard** e **PianoKey**, e seus métodos e argumentos.*

O bloco **GUI** é responsável pela criação do objeto **PianoKeyboard** e pela sua inicialização, através do método **init**. Os argumentos recebidos permitem especificar o posicionamento, em

pixéis, do teclado do piano na janela da interface. Dado que o teclado deve ocupar a largura da janela e deve ficar situado imediatamente abaixo do gráfico temporal, x_i deve valer 0, de forma a estar encostado à esquerda, e y_i deve corresponder à altura do referido gráfico. O parâmetro *width* deve tomar o valor da largura do ecrã, e ao parâmetro *height* é conferido o valor da altura das teclas (pré-definido como macro).

O método **init** procede à inicialização de cada tecla, de acordo com os valores recebidos quanto às dimensões do teclado. Os códigos MIDI da primeira e última teclas a representar no teclado estão definidas como macros.

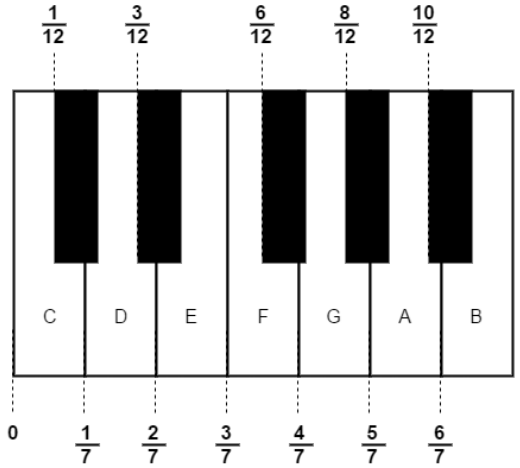


Figura 4.8: *Coeficientes de posição associados a cada uma das 12 teclas de uma oitava.*

As dimensões e posições das teclas devem ser determinadas separadamente, em função da cor. As teclas pretas, que se encontram nas posições 2, 4, 7, 9 e 11 de uma oitava, possuem uma altura relativa de 0.6 (60%) em relação à altura do teclado, e uma largura relativa de 1/12 da largura de uma oitava. As teclas brancas, correspondentes às restantes 7 teclas de uma oitava, possuem a altura do teclado, e uma largura de 1/7 da largura de uma oitava. A posição relativa de cada tecla no teclado é determinada através da seguinte expressão:

$$pos\_relativa = \frac{coef + N_{oitava} - 1}{total_{oitavas}}$$

O número da oitava é relativo às oitavas representadas na interface. Sendo que existem 3 oitavas, são-lhes atribuídos os números 1, 2 e 3, apesar de, na realidade, estarmos perante a 2ª, 3ª e 4ª oitavas em notação musical.

O método **addKeyboard** é invocado no fim da execução do método **init**, e consiste no preenchimento do vetor de teclas, de acordo com os valores obtidos anteriormente. Nesta fase, é necessário obter os valores exatos, em pixéis, da largura, altura e posição de cada tecla. A altura e largura de cada tecla são obtidas multiplicando a altura e largura relativas pela altura e largura do teclado. Para obter a posição, em pixéis, de cada tecla, basta multiplicar a posição relativa pela

largura do teclado. Após a execução deste método, o teclado de piano encontra-se pronto para ser impresso.

4.4.1.3 Representação do gráfico de *pitch* e SPL

De forma a tornar a leitura de um fonetograma mais fácil e expedita, são normalmente acrescentadas linhas de grade. No caso desta aplicação, e de forma a condizer com o teclado de piano, manteve-se, salvo vários ajustes, o aspeto do gráfico desenvolvido na aplicação *SingingStudio*, que continha um *piano roll*. Esta forma de representação está normalmente associada à transposição musical para MIDI, mas existe um forte interesse em manter esta funcionalidade. Desta forma, o espaço de representação é enriquecido com linhas verticais com dois tons de cinzento, correspondendo as mais escuras ao *pitch* das teclas pretas, e as mais claras ao das teclas brancas. De forma a representar a escala das ordenadas, são acrescentadas linhas paralelas, na horizontal, representando valores iguais de potência sonora.

A solução encontrada para representar o *piano roll* foi a utilização de **openGL**, que consiste numa interface para desenho de gráficos a duas ou três dimensões. A vantagem da sua utilização neste contexto prende-se com o facto de ser inicializado um *Vertex Buffer Object* (VBO), que armazena a informação gráfica do *piano roll*. Este VBO apenas necessita de ser inicializado uma vez, bastando depois invocar uma função para desenhar os seus conteúdos. Isto é bastante vantajoso no caso desta aplicação, dado que será necessário redesenhar o *piano roll* novamente a cada *frame*.

O método **initGLPianoRollData** consiste na inicialização do VBO que diz respeito ao *piano roll*. Para cada linha vertical de frequência fundamental, são guardados os valores associados à sua cor e à posição dos seus vértices em vetores do VBO. Este método é invocado apenas uma vez, no lançamento da aplicação, mais especificamente no método **setup** do bloco **App**.

4.4.1.4 Articulação do fonetograma

O método **drawPhonetogram** imprime na janela toda a interface gráfica do fonetograma, recebendo como argumentos a sua posição inicial e altura, em pixéis. Primeiramente, é colocado o **piano roll**, seguido de todas as teclas brancas do piano, e apenas depois as teclas pretas, dado que têm de ficar sobrepostas às brancas. Por fim, são desenhadas as linhas que representam a escala de potência sonora, com intervalos de 5 em 5 dB.

O método **draw** é responsável pelo desenho de toda a interface correspondente ao fonetograma. É invocado a cada *frame*, de tal maneira que toda a visualização é refrescada a cada nova iteração, permitindo atualizar, em tempo real, os dados representados. Em primeiro lugar, é invocado o método **drawPhonetogram**, sendo que, seguidamente, é invocado o método correspondente ao modo de funcionamento da aplicação.

Figura 4.9: *Captura de ecrã correspondente à representação do fonetograma.*

4.4.2 Gráfico de pontos (modo de captura)

O ponto fundamental de desenvolvimento deste projeto foi o desenvolvimento de uma forma de representação gráfica que permitisse caracterizar o sinal de voz do orador em tempo real, tendo como base o fonetograma. A solução implementada foi a impressão de pontos na interface, correspondentes a diferentes valores de frequência fundamental e intensidade, e que se vão desvanecendo em função do tempo. Assim, pode ser observada, em tempo real, a variação dos dois parâmetros referidos ao longo dos últimos segundos da captura.

A implementação deste tipo de representação foi realizada através da criação do método **drawPitchPowerPlot**. Este não recebe quaisquer argumentos, uma vez que acede diretamente aos vetores de dados guardados no *wrapper* intermédio, e que armazenam os valores de potência e de frequência fundamental calculados em cada *frame* de áudio. Estes vetores encontram-se sincronizados, ou seja, os valores calculados para cada *frame* estão armazenados na mesma posição.

São definidas duas fases em que ocorre o *fading* dos pontos: num primeiro momento, os pontos são colocados no fonetograma com uma cor viva, e mantêm o mesmo nível de brilho durante um certo intervalo; seguidamente, os pontos começam a desvanecer, até que, a partir de um certo número de *frames* de áudio, desaparecem mesmo.

São definidos dois valores, que representam limiares mínimos e máximos de *fading*, **min fade time** e **max fade time**. Estes definem posições, contadas a partir do fim do vetor (dado que os índices mais avançados correspondem a valores mais recentes), em que ocorrem as duas fases descritas anteriormente.

De seguida, é necessário percorrer os dois vetores simultaneamente, extraindo cada par de valores que constitui um ponto, e calcular as respetivas coordenadas.

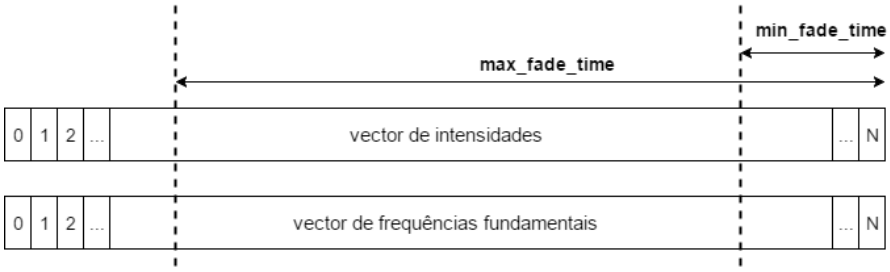


Figura 4.10: *Representação dos dois vetores que guardam os valores de intensidade e frequência fundamental. As posições mais avançadas correspondem a valores mais recentes.*

De forma a calcular as coordenadas dos pontos em relação à janela, é necessário traduzir os valores de frequência e intensidade guardados nos respetivos vetores para pixéis. De forma a obter este resultado, foram implementadas duas funções simples que chamam o método **ofMap** da *framework*. Esta função recebe o valor que tem de ser mapeado, bem como o valor máximo e mínimo da escala original e o valor máximo e mínimo da escala para o qual tem de ser convertido, sendo que é retornado o valor correspondente nessa nova escala. As funções em questão são **midi2pixelX**, que traduz valores de MIDI para pixel, correspondendo às abcissas, e **power2pixelY**, que passa de valores de dB para pixel, representados nas ordenadas. Note-se que os valores obtidos para a frequência fundamental são armazenados com o seu valor exato na escala em MIDI. Estes valores podiam ter sido guardados pela forma de representação mais comum, em Hz, mas visto ser aquela a notação usada na aplicação SingingStudio, esta não foi alterada, por não constituir nenhuma desvantagem.

A *framework* utilizada contém três métodos fundamentais, que foram usados no desenho dos pontos no gráfico:

- **ofSetColor** define a cor com a qual serão desenhados os objetos futuros. Recebe como argumentos os 3 valores R, G e B, e ainda um valor de opacidade, que varia entre 0 (transparente) e 255 (cor sólida),
- **ofCircle** desenha um círculo nas coordenadas recebidas e com o raio especificado,
- **ofFill** deve ser invocado de forma a que os círculos sejam preenchidos.

Em função da posição dos valores obtidos nos respetivos vetores, os parâmetros de cor e intensidade são variados, tal como pode ser observado na Figura 4.11.

Se estes estiverem a menos do que **min fade time** do fim do vetor (mais recentes), serão desenhados com uma cor que irá do amarelo vivo ao verde vivo. De forma a obter este gradiente, cada um dos coeficientes RGB de **ofSetColor** é mapeado entre o amarelo e o verde, através do método **ofMap**.

Se os pontos estiverem colocados entre as posições delimitadas por **min fade time** e **max fade time**, já terão assumido a sua cor verde viva obtida na condição anterior, e será feito o *fading*. Desta vez, os valores RGB mantêm-se contstantes, e utiliza-se a função **ofMap** para fazer variar os valores de opacidade entre 255 (cor sólida) e 0 (transparente).

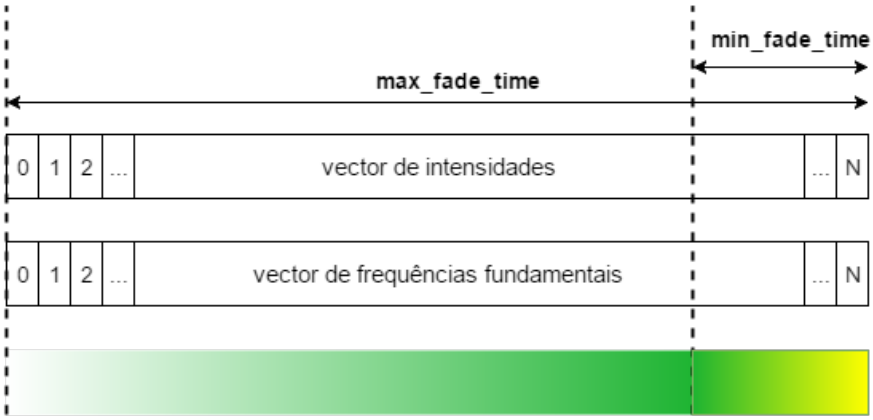


Figura 4.11: *As posições iniciais de ambos os vetores devem ser libertadas, por corresponderem a informação antiquada. É também representada a cor atribuída a cada ponto, em função da sua posição nos vetores de dados.*

A partir do momento em que os pontos desvanecem completamente, deixa de ser relevante guardar a sua informação associada, por questões de gestão de memória. Com efeito, enquanto a aplicação continua a funcionar em modo de captura, estão constantemente a ser adicionadas e processadas novas amostras de sinal, pelo que os vetores que armazenam a informação de frequência fundamental e intensidade continuam a ser preenchidos indefinidamente, levando a uma ocupação de memória cada vez mais elevada.

Por esta razão, as posições dos vetores de frequência fundamental e intensidade que ultrapassam o **max fade time** devem ser libertadas, utilizando o método **erase** da biblioteca de vetores em C++. Além de libertar o espaço reservado às primeiras posições do vetor, esta função automaticamente reorganiza-o, atualizando as posições dos seus elementos de forma a começar no zero. Após estes procedimentos, estes vetores possuirão exatamente **max fade time** posições.

4.4.3 Gráfico de regiões (modo inactivo)

Apesar de o gráfico de pontos constituir uma boa forma de visualizar, em tempo real, a evolução de características fundamentais do sinal de voz, este possui uma limitação fundamental. Esta deve-se ao facto de não subsistir informação relativa a toda a captura, dado os pontos desvanecerem-se ao fim de um certo tempo. De forma a solucionar esta lacuna, existia interesse em implementar um outro tipo de representação que indicasse, após o fim da captura de som, as regiões do fonetograma mais utilizadas (ou seja, as que continham mais pontos). Pode-se então desenhar um gráfico de aspeto semelhante a um *Vocal Range Profile*, mas que, em vez de representar a extensão total da

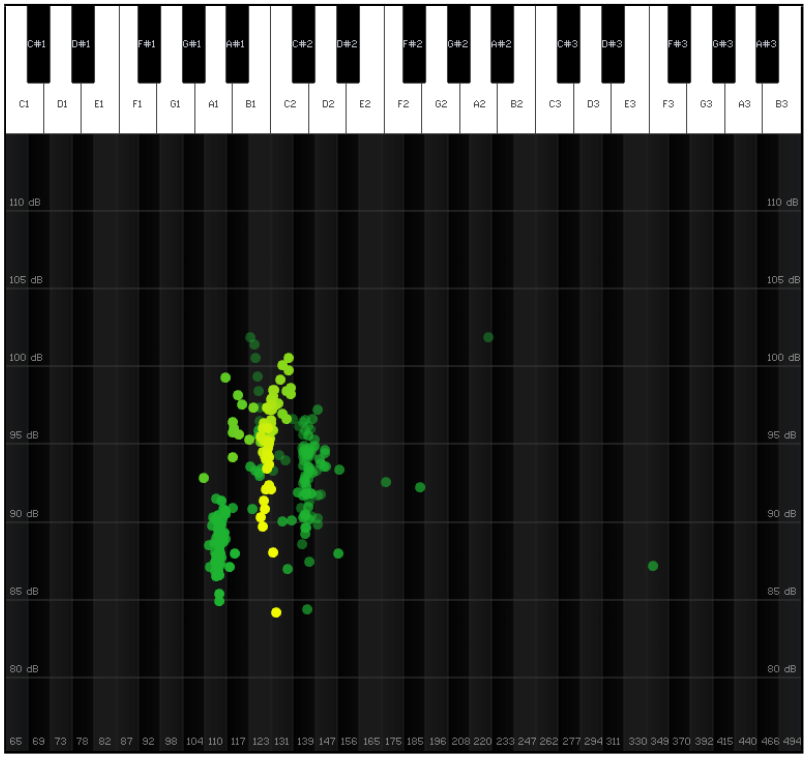


Figura 4.12: *Captura de ecrã correspondente à representação do gráfico de pontos no fonetograma.*

voz de um indivíduo, permite representar as frequências e intensidades atingidas ao longo de toda a captura.

No entanto, esta implementação obrigaria a conservar a informação relativa a toda uma captura, o que levantaria problemas adicionais de gestão de memória e de utilização do CPU por parte da aplicação.

A solução encontrada foi a divisão do gráfico em regiões de forma quadrada, em que cada região está associada a uma posição numa matriz, que vai sendo preenchida no modo de captura, durante a execução do método **drawPitchPowerPlot**. Esta matriz possui, inicialmente, todas as posições com valor zero, sendo que, à medida que se vão obtendo as coordenadas dos pontos, vão sendo incrementadas as posições relativas à região de ocorrência.

A grande vantagem desta solução é o facto de os vetores de intensidade e frequência fundamental poderem ser libertados, tal como foi explicado na secção anterior. Apesar de existir um certo aumento no tempo de execução do método **drawPitchPowerPlot**, este não é significativo.

A granularidade das regiões a serem impressas deve resultar de um compromisso: por um lado, regiões mais pequenas levam a um aumento de posições na matriz, aumentando por isso de forma quadrática o número de operações a realizar; por outro lado, regiões demasiado grandes não conferirão grande precisão à avaliação, além de serem esteticamente pouco apelativas. Neste caso, pretende-se que o lado dos quadrados correspondentes a regiões possua, no mínimo, a largura de cada tecla do piano no *piano roll*. Desta forma, é simples de fazer a ligação entre uma região do

gráfico e a nota associada.

A matriz que mapeia as regiões na interface pode ser representada da seguinte maneira:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1N} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \dots & \alpha_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{L1} & \alpha_{L2} & \alpha_{L3} & \dots & \alpha_{LN} \end{bmatrix}.$$

Cada valor α representa o número de pontos representados na região respetiva. O valor de **N** corresponde ao número de colunas da matriz. Este valor é obtido dividindo a largura do fonetograma pelo lado de cada região, que depende da granularidade. No caso desta aplicação, **N** corresponde ao número de teclas do piano, de forma a poder relacionar cada coluna da matriz com cada uma destas.

A matriz foi implementada com um *array* uni-dimensional como o representado abaixo, deforma a gastar menos poder computacional:

$$\begin{bmatrix} \alpha_{11} & \dots & \alpha_{1N} & \alpha_{21} & \dots & \alpha_{2N} & \dots & \alpha_{L1} & \dots & \alpha_{LN} \end{bmatrix}.$$

A impressão das regiões na interface é realizada através do método **drawCaptureRangeProfile**, invocada *frame* a *frame* no modo inativo. O *array* (que representa a matriz) necessita de ser percorrido duas vezes.

A primeira vez tem como objetivo determinar o máximo valor de α. Este é usado para determinar a opacidade relativa de cada região.

Na segunda vez que o *array* é percorrido, é calculada, para cada posição, a posição da região correspondente, e a sua opacidade. Esta última é determinada dividindo o valor de α pelo máximo obtido anteriormente. Para traduzir este valor relativo para a escala RGB, basta multiplicá-lo por 255.

Após a execução deste método, obtém-se o resultado visualizado na Figura 4.13, em que a região de maior ocorrência de pontos se encontra representada com uma cor branco sólido.

4.5 Conclusão

Neste capítulo, foram ilustradas as principais etapas do desenvolvimento da aplicação. Procurou-se explicar, de uma forma descritiva e com o auxílio de ilustrações, a estrutura do código desenvolvido, em termos dos seus módulos constituintes, assim como dos seus principais métodos.

Foram dadas algumas explicações quanto ao raciocínio executado de forma a chegar à solução apresentada, assim como algumas preocupações adicionais a ter, tendo em conta o desempenho da aplicação e a gestão de memória.

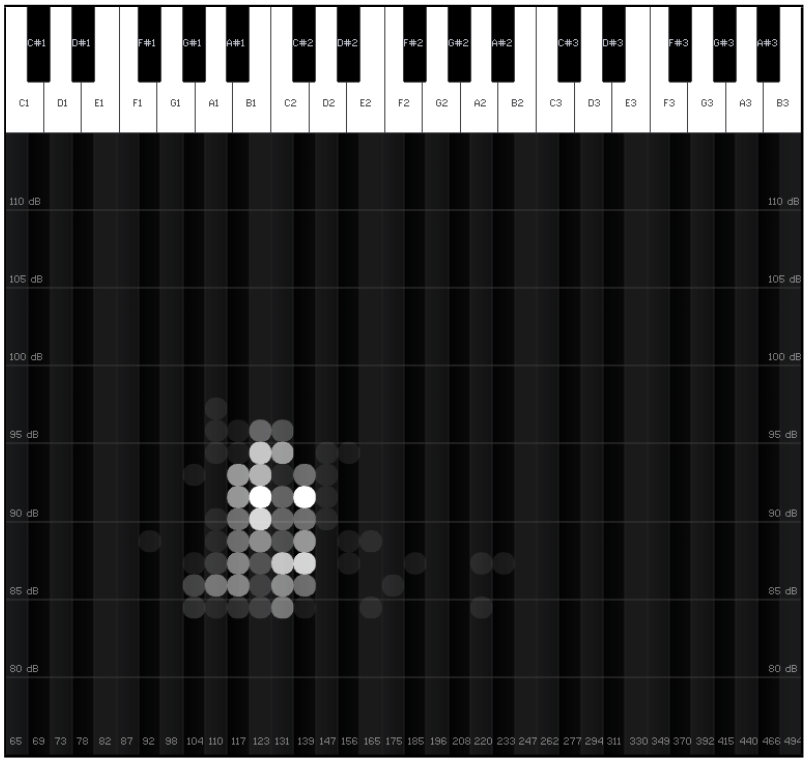


Figura 4.13: *Captura de ecrã correspondente à representação do gráfico de regiões no fonetograma.*

Capítulo 5

Resultados

Após a fase de desenvolvimento da aplicação, é necessária a obtenção de resultados que permitam tirar conclusões quanto ao seu desempenho e verificar o cumprimento de determinados objetivos. Com este intuito, foram realizados dois tipos de testes: experimentais, de modo a avaliar a precisão da informação apresentada, e testes de usabilidade, que permitam medir a satisfação do utilizador quanto à utilização da aplicação.

5.1 Medição da precisão dos algoritmos

Na primeira fase de testes, foram realizadas experiências com vista a medir a qualidade dos dados representados no fonetograma. É necessário comparar os valores de frequência fundamental e de intensidade sonora calculados com valores nominais, de forma a estabelecer a precisão daqueles. Em caso de se verificar a ocorrência de erros significativos, estes dados poderão ainda ajudar a discernir as razões pelas quais os resultados obtidos não foram os esperados, assim como o impacto que estes têm no correto funcionamento da aplicação.

5.1.1 Precisão de *Pitch*

Para a obtenção de resultados de precisão de *pitch*, foram geradas, através do *software* Octave, amostras de sinais periódicos com uma dada frequência fundamental. Foram gerados e reproduzidos sinais com frequências fundamentais correspondentes a cada nota do teclado incluído no fonetograma, tendo depois sido impressos os valores obtidos na aplicação. Comparando estes com a frequência fundamental definida manualmente, foram obtidas medidas do acerto relativo ao cálculo do *pitch* do sinal.

O sinal de teste tem forma de onda dente-de-serra, também designada de *sawtooth wave*, sintetizada com os 5 primeiros coeficientes da série de Fourier. Isto significa que a representação da sua densidade espectral de potência possui 5 picos, sendo o primeiro correspondente à frequência fundamental, e os 4 seguintes aos 4 primeiros harmónicos do sinal.

De forma a testar a precisão do cálculo da frequência fundamental ao longo de toda a gama de frequências representada no fonetograma, geraram-se amostras do sinal sintetizado para todas as

notas do piano. Foram geradas amostras correspondentes a um segundo de sinal. Seguidamente, concatenaram-se as amostras correspondentes às 36 notas do piano contido no fonetograma num único ficheiro, resultando num ficheiro áudio com duração de 36 segundos.

Os resultados do cálculo da frequência fundamental para cada um destes diferentes valores de frequência fundamental podem ser observados no gráfico de dispersão da Figura 5.1.

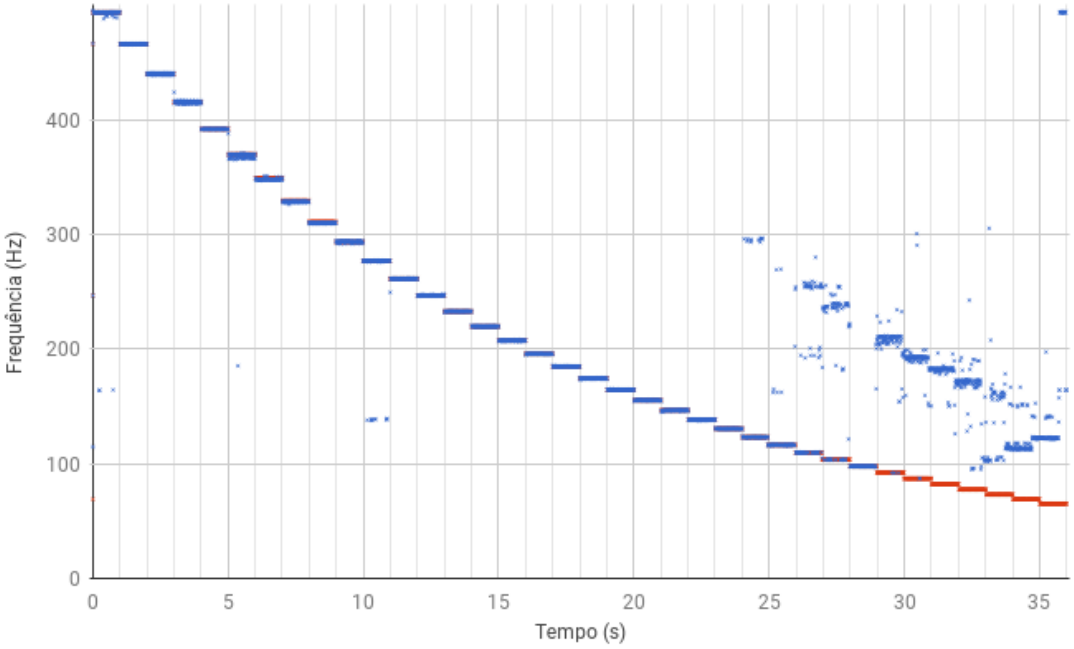


Figura 5.1: *Representação dos valores obtidos para a frequência fundamental de cada nota do teclado do piano*

Cada um dos pontos representados no gráfico corresponde a uma *frame* áudio. A vermelho encontram-se representados os valores reais da frequência fundamental do ficheiro de som ao longo do tempo, enquanto que a azul estão representados os valores calculados na aplicação.

Pelo que se pode verificar, os resultados obtidos na aplicação encontram-se muito próximos do ideal para as mais altas frequências da escala. Na gama que vai da frequência da nota mais alta (493.88 Hz) até à frequência de 130.81 Hz, verifica-se que praticamente todos os valores calculados na aplicação se encontram muito próximos do ideal. Estas conclusões são suportadas pelo gráfico da Figura 5.2. De forma a construir esta representação, foram considerados aceitáveis os valores de *pitch* que se encontram mais próximos da frequência da tecla correspondente do que das frequências das teclas adjacentes.

Porém, a partir da nota correspondente à frequência 123.47 Hz (reproduzida dos 24 aos 25 segundos), verifica-se um início de ocorrência de erros muito significativos na determinação do *pitch*. No intervalo dos 25 aos 29 segundos, ainda existe uma percentagem bastante significativa de acertos, como é evidenciado na Figura 5.2 (98-123.47 Hz). No entanto, frequências fundamentais abaixo dos 92.50 Hz praticamente nunca são detetadas corretamente.

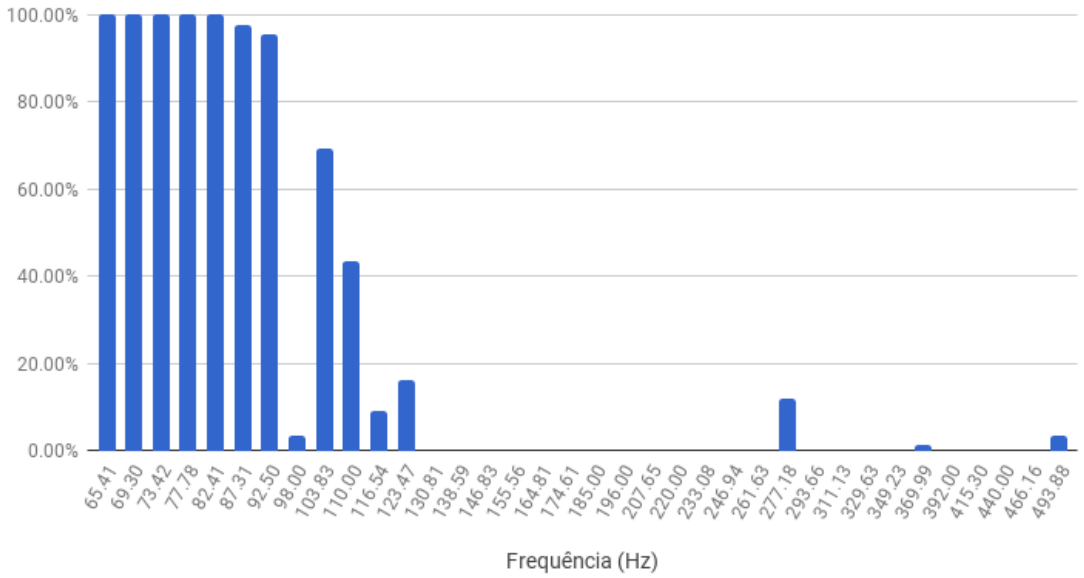


Figura 5.2: *Representação da ocorrência de valores fora do aceitável na medição da frequência fundamental das teclas de piano*

Estes erros são certamente causados pela falta de resolução da FFT (Fast Fourier Transform). Este método é uma forma mais eficiente de calcular a transformada de Fourier para sinais discretos, possuindo uma resolução que depende de certos parâmetros usados. No caso desta aplicação, a resolução de frequência da transformada é de aproximadamente 43 Hz. Regra geral, a FFT não apresenta bons resultados para frequências abaixo do triplo da sua resolução. Este ponto corresponde, neste caso, a um valor próximo de 129 Hz, o que é concordante com os resultados obtidos, em que começam a ocorrer erros significativos para as notas com frequência igual e inferior a 123.47 Hz.

No entanto, os casos em que o *pitch* nunca é determinado com correção coincidem com frequências que são atingidas por apenas alguns indivíduos do sexo masculino, pelo que, globalmente, os resultados da obtenção da frequência fundamental são muito positivos.

5.1.2 Precisão de intensidade

Com o objetivo de determinar a precisão do cálculo da intensidade do sinal de entrada, foram comparados os valores obtidos na aplicação com os valores medidos com um *sound level meter*. De forma a obter estes valores, foi colocado um dispositivo com a aplicação a correr ao lado do medidor de intensidade sonora, de tal maneira a que os seus microfones respetivos se encontrassem aproximadamente à mesma distância da fonte sonora.

A fonte sonora consistiu na reprodução de um sinal com forma de onda em dente-de-serra, num computador, através dos altifalantes. Foram realizadas várias capturas de alguns segundos, variando-se o volume de saída dos altifalantes entre capturas consecutivas, e calculou-se a média dos valores de potência obtidos na aplicação. Estes foram comparados com o valor médio medido

pelo sonómetro durante o mesmo intervalo de tempo. Os resultados, utilizando um sinal com frequência fundamental de 220 Hz, e com os dispositivos a 30 cm dos altifalantes, podem ser verificados no gráfico da Figura 5.3.

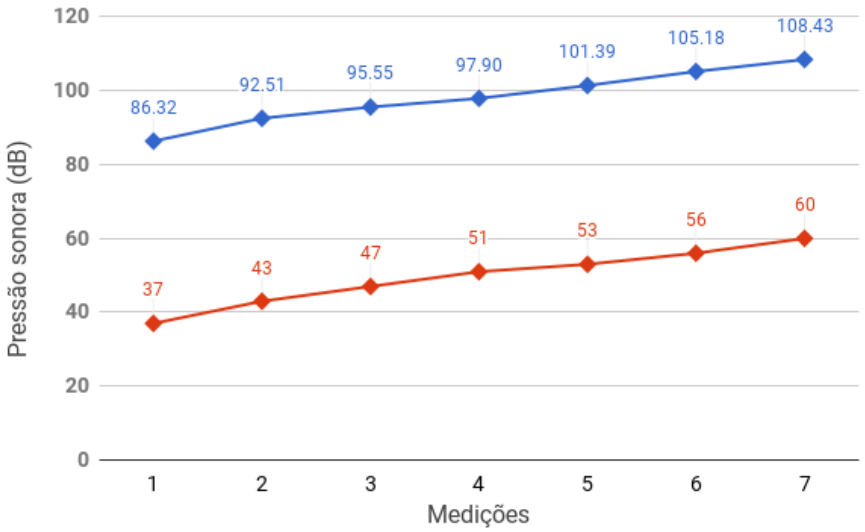


Figura 5.3: *Resultados da comparação entre valores de intensidade obtidos na aplicação e com um medidor de pressão sonora*

A partir da observação dos resultados obtidos, verifica-se que existe uma grande diferença entre os valores obtidos na aplicação e os valores reais. As intensidades obtidas encontram-se várias dezenas de deciBel acima dos valores medidos através do sonómetro.

No entanto, o fator de variação das medições é praticamente igual quando se comparam os dois casos, o que constitui um resultado muito positivo, uma vez que indica que as medições feitas pela aplicação são coerentes, a menos de um fator constante. Desta forma, valores muito próximos da realidade podem ser obtidos através de uma operação de calibração dos resultados executada na própria aplicação.

5.2 Avaliação de usabilidade

De forma a recolher a perspetiva de um utilizador sobre a aplicação desenvolvida, foi conduzido um teste de usabilidade, sob a forma de um questionário de 8 perguntas. Nas 7 primeiras, pediu-se ao utilizador que atribuísse uma nota a cada um dos diferentes componentes da interface gráfica da aplicação, assim como ao aspeto e desempenho geral da aplicação. A nota a atribuir consistiu num valor de 1 a 5, em que 1 constitui o nível "nada" e o nível 5 representa o nível "muito". Finalmente, na última pergunta, foram pedidas sugestões para melhoria da aplicação.

Intuição relativa ao aspeto gráfico

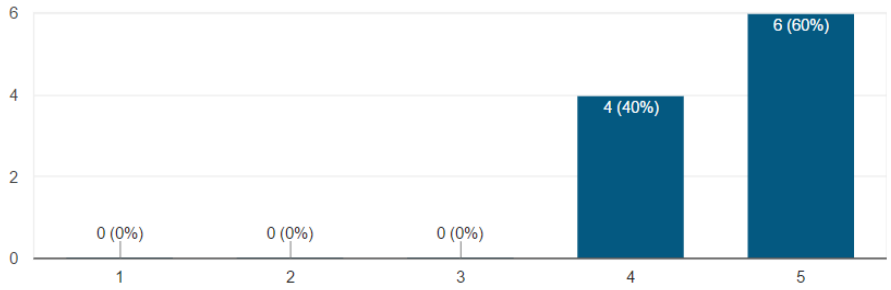


Figura 5.4: *Histograma das respostas à pergunta: "O aspeto gráfico é intuitivo?".*

Utilidade da representação temporal

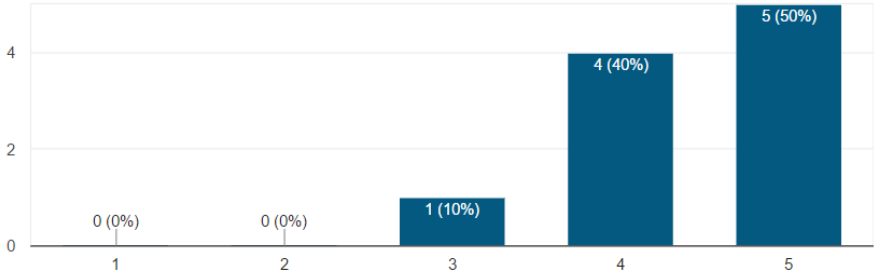


Figura 5.5: *Histograma das respostas à pergunta: "A representação a azul do sinal de voz é útil?".*

Utilidade da representação das teclas de piano

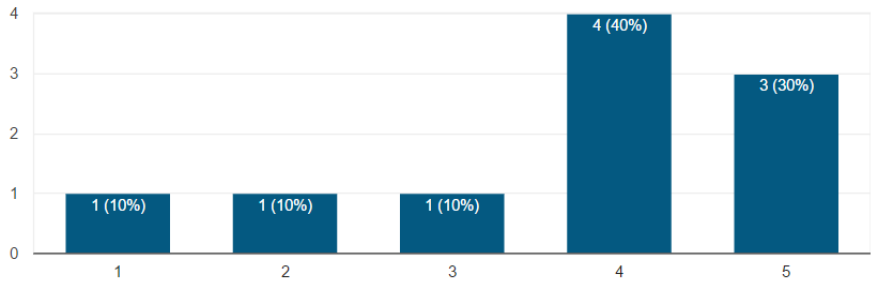


Figura 5.6: *Histograma das respostas à pergunta: "A representação das teclas de piano é útil?".*

Intuição relativa ao gráfico de pontos

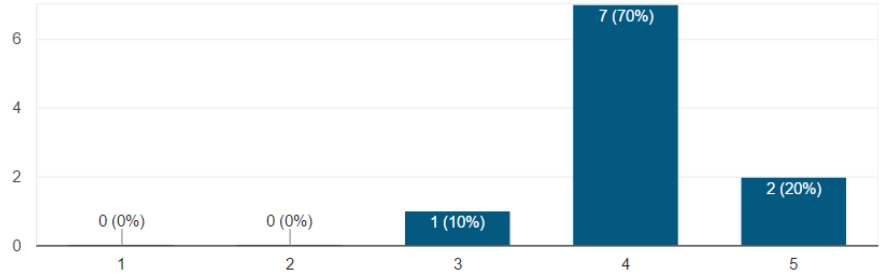


Figura 5.7: *Histograma das respostas à pergunta: "A representação do tom de voz através dos pontos amarelos/verdes é compreensível?".*

Intuição relativa ao gráfico de regiões

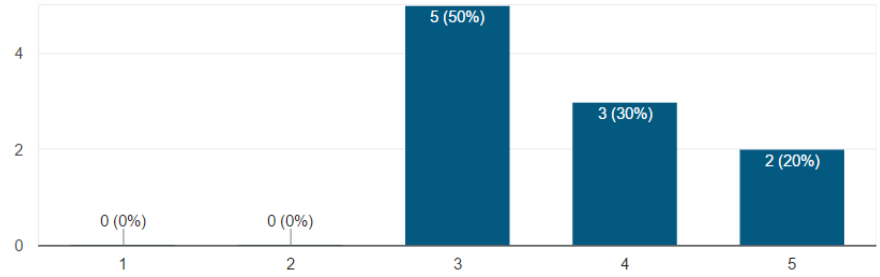


Figura 5.8: *Histograma das respostas à pergunta: "O mapa final (em tons de cinzento) da distribuição do tom de voz é intuitivo?".*

Fluidez da resposta gráfica

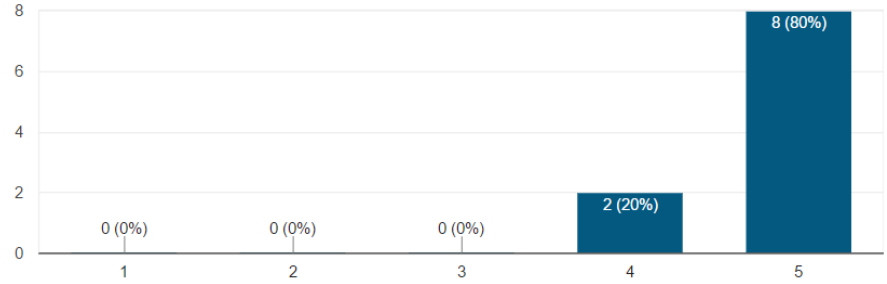


Figura 5.9: *Histograma das respostas à pergunta: "A resposta gráfica da aplicação é fluida?".*

Apreciação geral da aplicação

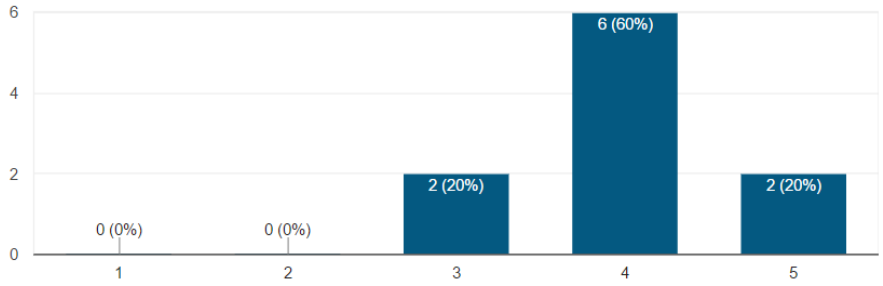


Figura 5.10: *Histograma das respostas à pergunta: "Como avalia a aplicação em termos de apreciação geral?".*

Sugestões de melhoria apresentadas

De forma a poder avaliar a aplicação de forma mais completa, foi dada a possibilidade aos utilizadores questionados de indicarem funcionalidades que gostassem de ver implementadas. Estas propostas constituem informação valiosas para o enriquecimento da aplicação, uma vez que permitem ter uma ideia das prioridades quanto a um possível desenvolvimento futuro.

As sugestões recolhidas foram as seguintes:

- extensão da gama de frequências apresentada no fonetograma, mais especificamente de modo a incluir valores de *pitch* acima da nota B4; possibilidade de representação de toda a gama vocal de um indivíduo, e não apenas o intervalo correspondente à voz falada, dado que diminui a aplicabilidade desta ferramenta,

- introdução da funcionalidade de gravação/reprodução da captura efetuada,

- aumento da largura das teclas para dispositivos de baixa resolução, compensada pela introdução de uma funcionalidade de *swipe* (possibilidade de arrastar o teclado para os lados),

- possibilidade de gravar uma imagem do gráfico de regiões obtido, para posterior análise.

Interpretação de resultados

Os resultados obtidos, em termos de média e de variância das respostas a cada pergunta, são apresentados na tabela seguinte:

Objeto de avaliação	**Média**	**Variância**
Intuição relativa ao aspeto gráfico	4.6	0.267
Utilidade da representação temporal	4.4	0.489
Utilidade da representação das teclas de piano	3.7	1.789
Intuição relativa ao gráfico de pontos	4.1	0.322
Intuição relativa ao gráfico de regiões	3.7	0.678
Fluidez da resposta gráfica	4.8	0.178
Apreciação geral da aplicação	4.0	0.444

Apesar de o número de amostras ser relativamente pequeno, é possível observar claras tendências nestes resultados. Os aspetos mais positivos e consensuais, segundo este estudo, foram o aspeto e resposta gráfica da aplicação.

Em termos de funcionalidades, a inclusão da representação temporal foi julgada útil. Apesar de não constituir o ponto focal de desenvolvimento, funciona bem em conjunção com o fonetograma, por ser uma forma de representação mais intuitiva. O facto de o gráfico de pontos ter sido considerado compreensível vai ao encontro de um dos principais objetivos desta aplicação, que era a possibilidade de ser rapidamente assimilada, sem existir a necessidade de grandes conhecimentos teóricos por parte dos utilizadores. A forma de representação caracterizada pelo gráfico de regiões não foi considerada muito intuitiva, o que constitui um resultado pouco satisfatório, dado indicar que foi difícil estabelecer a relação entre este e o gráfico de pontos. Este resultado sugere a inclusão de uma breve pista indicativa dos resultados atingidos pela representação da extensão vocal aplicada a toda a captura.

A funcionalidade julgada mais controversa foi a inclusão do teclado do piano. Na ótica de certos utilizadores, esta pouco ou nada contribuiu para o enriquecimento da aplicação. No entanto, suscitou também reações muito positivas. Isto pode ser explicado pelo facto de, uma vez que constitui uma forma de representação musical, não ser acessível a todo o tipo de utilizadores.

5.3 Conclusão

Neste capítulo, foram obtidos resultados respeitantes à precisão da aplicação na determinação do *pitch* e SPL de um sinal de entrada. Verificou-se que a determinação da frequência fundamental apresenta excelentes resultados nas frequências mais elevadas. Contudo, à medida que se medem valores de *pitch* abaixo de 129 Hz, passam a ocorrer erros. Para frequências abaixo de cerca de 83 Hz, nunca se conseguem resultados dentro do aceitável. Quanto às medições de intensidade, verificou-se que estas não são precisas, mas a sua variação está de acordo com os valores reais.

Foram também recolhidas as opiniões de utilizadores acerca da usabilidade da aplicação, que permitiram avaliar a qualidade desta em cenários reais, de uma perspetiva externa.

Capítulo 6

Conclusões e Trabalho Futuro

6.1 Satisfação dos Objetivos

O objetivo principal fixado para esta dissertação era a realização de uma aplicação que permitisse representar, dinamicamente e em tempo real, características do sinal de voz de um orador, de forma a permitir avaliar a qualidade do seu discurso.

Ao longo deste trabalho, foram desenvolvidas formas de representação do sinal de voz em função dos tempos, frequência fundamental e intensidade sonora. Foram construídos dois tipos de gráficos, um de representação temporal, e um fonetograma. Estes permitem a caracterização de um orador em função da variação de *pitch* e de intensidade da sua voz, assim como da variação da sua amplitude ao longo dos tempos, podendo assim obter uma medida da dinâmica de discurso, que contribui para uma comunicação eficaz.

A inclusão de dois modos de funcionamento reforçou a informação apresentada e constitui um enriquecimento significativo em relação aos objetivos iniciais. Assim, é possível seguir a evolução do sinal em tempo real, e, no fim da captura, ainda se pode visualizar toda a sua informação associada, permitindo ter um *feedback* completo e em retrospetiva.

Pretendia-se desenvolver uma aplicação que fosse simples e intuitiva, para poder ser utilizada por um grande número de pessoas. O objetivo da simplicidade foi atingido graças à construção de uma interface gráfica de vista única e sem sobrecarga de informação. Todavia, verificou-se a partir do questionário de usabilidade que é necessário um tempo de adaptação e uma breve explicação teórica de modo a perceber a informação apresentada no fonetograma.

Apesar de os resultados proporcionados por esta aplicação em termos de cálculo da frequência fundamental e da intensidade sonora não terem sido ideais, foi possível verificar a origem dos erros. As falhas na obtenção de valores precisos de *pitch* abaixo dos 129 Hz podem ser solucionadas com uma melhoria na resolução de frequência da FFT. Quanto às medições de intensidade sonora, os valores obtidos necessitam apenas de uma calibração, que pode ser executada de forma relativamente expedita com base nos resultados verificados.

6.2 Trabalho Futuro

Após o desenvolvimento da aplicação e da realização dos testes, surgiram diversos aspetos que podem constituir possibilidades de desenvolvimento futuro.

Existem duas principais direções que podem ser seguidas no sentido de trazer melhorias à aplicação. Por um lado, podem ser introduzidas novas funcionalidades, que venham enriquecer os resultados que podem ser obtidos a partir da sua utilização. Por outro, há a possibilidade de correção de aspetos existentes, assim como a otimização de processos e do desempenho, com vista a melhorar as funcionalidades já existentes.

6.2.1 Melhoria de funcionalidades existentes

- **Calibração das medidas de intensidade sonora** calculadas na aplicação, retirando um valor fixo aos dados obtidos, de forma a aproximá-los ao máximo dos valores apresentados no sonómetro.

- Diminuição do valor da resolução de frequência da FFT, de maneira a permitir a **obtenção de melhores resultados de medição do *pitch* para frequências mais baixas** da gama vocal humana.

- Possibilidade de **alterar a orientação do ecrã** para *landscape*.

- **Após o fim de uma captura, impressão do gráfico temporal na sua totalidade**. Em vez de mostrar apenas as últimas amostras ainda presentes no *buffer* temporal, esta funcionalidade permitiria uma melhor correspondência entre os valores de intensidade apresentados no gráfico de regiões e a amplitude das amostras da representação temporal.

- **Adaptação da interface gráfica da aplicação para ecrãs de reduzida resolução**. Atualmente, a largura das teclas do piano, assim como os *labels* que apresentam o nome das notas, dos valores de frequência e de intensidade são demasiado pequenos, principalmente em dispositivos com resolução de 4 polegadas.

6.2.2 Introdução de novas funcionalidades

- **Extensão da gama de frequências apresentada no fonetograma**, de forma a permitir a representação de dados de mais alta frequência. Isto permitiria a representação de toda a extensão vocal para qualquer indivíduo, que possibilitaria a utilização da aplicação em contexto médico e musical. Estas alterações teriam de ser acompanhadas com a introdução de métodos para fazer deslizar o fonetograma para os lados, uma vez que seria incomportável a sua representação em inteiro numa janela de resolução reduzida.

- **Possibilidade de apresentar o gráfico de regiões** (gama vocal correspondente aos dados da captura) **a qualquer momento**, ao carregar no botão de pausa, sem que isso significasse ter

de terminar a captura atual. Seria interessante poder visualizar a sua progressão, de tempo a tempo.

- **Apresentação de parâmetros estatísticos** obtidos numa captura. Poderiam ser mostrados resultados tais como as frequências máxima e mínima atingidas, extensão total e média de frequência fundamental e intensidade do orador. Além disto, podia ser apresentada uma estimativa para a cadência silábica, que acabou por constituir um fator pouco abordado no decurso deste trabalho.

- Apesar de constituir uma perspetiva mais ambiciosa, seria particularmente interessante a **introdução da capacidade de gravação e reprodução** por parte da aplicação. Isto permitiria voltar a reproduzir o sinal de entrada e observar a evolução gráfica ao longo do tempo.

Referências

[1] Ricardo Sousa. *Metodologias de Avaliação Perceptiva e Acústica do Sinal de Voz em Aplicações de Ensino do Canto e Diagnóstico/Reabilitação da Fala*. Tese de doutoramento, Faculdade de Engenharia da Universidade do Porto, 2011.

[2] Aníbal Ferreira. Análise acústica, percetiva e visual da voz. Em Ana P. Mendes, editor, *Vocologia do Fado*, chapter 4. 2016.

[3] Victor Zue. Acoustic theory of speech production, 2003.

[4] Jeffrey Fessler. Frequency analysis of signals and systems. https://web.eecs.umich.edu/~fessler/course/451/l/pdf/c4.pdf. Accessed: 20-06-2017.

[5] Yen-Liang Shue. *The Voice Source in Speech Production: Data, Analysis and Models*. Tese de doutoramento, University of California, Los Angeles, 2010.

[6] Jacqueline Whitmore. Sound advice: How to make your voice more effective. *Entrepreneur*, 2013.

[7] Toastmasters International. Your speaking voice. Relatório técnico, Toastmasters International, 2011.

[8] Estella P.-M. Ma e Edwin M.-L. Yiu. *Handbook of Voice Assessments*. Plural Publishing, First edição, 2011.

[9] Cibele; da Silva Teles Lídia Cristina; Berretin-Felix Giédre Tiemi Mituuti, Cláudia; Carméllo Santos. Características da fonetografia em indivíduos com equilíbrio dentofacial pós-muda vocal. *Revista CEFAC*, 15(5):1300–1307, setembro-outubro 2013.

[10] Anick M-J Lamarche. *Putting the Singing Voice on the Map*. Tese de doutoramento, KTH School of Computer Science and Communication, 2009.

[11] Arend Marten Sulter. *Variation of voice quality features and aspects of voice training in males and females*. 1996.

[12] Fundamental frequency and the glottal pulse. https://msu.edu/course/asc/232/study_guides/F0_and_Glottal_Pulse_Period.html. Accessed: 18-06-2017.