Universidade Federal de Pernambuco
Centro de Informática

Graduação em Engenharia da Computação

Aplicação de Técnicas de Recuperação da Informação de Música para Análise da Voz Cantada

Matheus Soares Monteiro

Trabalho de Graduação

Recife
21 de dezembro de 2016

Universidade Federal de Pernambuco
Centro de Informática

Matheus Soares Monteiro

Aplicação de Técnicas de Recuperação da Informação de Música para Análise da Voz Cantada

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientador: *Prof. Dr. Giordano Cabral*

Recife
21 de dezembro de 2016

Universidade Federal de Pernambuco
Centro de Informática

Matheus Soares Monteiro

Aplicação de Técnicas de Recuperação da Informação de Música para Análise da Voz Cantada

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Trabalho aprovado pela banca examinadora:

Prof. Dr. Giordano Cabral - CIn UFPE

Prof. Dr. Geber Ramalho - CIn UFPE

Recife
21 de dezembro de 2016

Dedico este trabalho a Deus, à minha família e a todos os amigos e mestres da música que me fazem crer nesse caminho.

Agradecimentos

Em primeiro lugar, agradeço a Deus pela vida, conquistas e pelo dom da música, que inspirou e motivou este trabalho.

Agradeço à minha família por todo o suporte e incentivo. Nos momentos mais difíceis e também nos melhores, eles sempre estiveram comigo.

Agradeço aos amigos que tornaram cada momento dessa caminhada um pouco mais leve e que também me ensinam e ajudam sempre que preciso.

Agraço a todos os professores que compartilharam tão generosamente os seus conhecimentos e lições durante todo o decorrer da graduação. Muito especialmente, agradeço ao professor Flávio Medeiros do departamento de música da UFPE que, através do coro universitário, me fez descobrir o músico que eu queria ser e me dá oportunidades de mostrar isso ao mundo; Aos professores Adriano Pinheiro e Luiz Kleber Queiroz também do departamento de música da UFPE, por me lapidarem como cantor e me fazerem acreditar que eu posso ir além; Aos professores Geber e Giordano do Centro de Informática, por me mostrarem que eu poderia juntar duas das coisas que mais amo na vida: música e tecnologia.

Também destaco todos os amigos cantores que se dispuseram a participar do meu experimento cedendo suas vozes tão especiais e também os amigos do CIn que sempre me apoiaram e estiveram perto em toda a caminhada.

À música, à tecnologia, à ciência, à vida!

A música é, ao mesmo tempo uma arte e uma ciência. Portanto, ela deve ser, ao mesmo tempo, emocionalmente apreciada e intelectualmente compreendida.

—OTTO KÁROLYI (Introdução à Música)

Resumo

A voz cantada é o mais antigo instrumento musical de que se tem notícia e possui diversas características empregadas pelo cantor, que estão diretamente relacionadas à sua fisiologia, técnica, saúde vocal, estilo e outros fatores. O monitoramento dessas características é essencial durante o processo contínuo de estudo desta arte.

Desde muito tempo, técnicas de processamento de sinais são utilizadas para análise e síntese da voz, com motivação principal no âmbito médico, para detectar doenças e anomalias de maneira não invasiva e, mais recentemente, no mercado de entretenimento, como karaokês, robôs e mesmo alguns jogos eletrônicos.

Recuperação de Informação de Música (do inglês, *Music Information Retrieval* – MIR) é uma área de pesquisa multidisciplinar que, como o próprio nome sugere, busca recuperar informações a partir sinais de áudio, utilizando diversas técnicas com diferentes propósitos.

Neste contexto, este trabalho propõe seguir uma abordagem *top-down* para o caso dos sinais de voz gravados por cantores líricos. Com um banco de dados exclusivo, esta pesquisa utilizou softwares de visualização como o PRAAT e implementou técnicas de procesamento de sinais, procurando destacar a associação de parâmetros acústicos à características interpretativas e técnicas da voz cantada.

Palavras-chave: MIR, Voz Cantada, Processamento Digital de Sinais

Abstract

Singing voice is the oldest musical instrument known and it has a lot of specific characteristics inherited from the singer. These characteristics are related to the vocal health, physiology, singing technique, among others important aspects in the continuous learning process of the singing art.

Since many years, digital signal processing techniques are used to develop speech synthesis systems and to analyse the singing voice, motivated by medical reasons and, most recently, by the entertainment industry.

MIR (Musical Information Retrieval) is an interdisciplinary area that aims to retrieve important information from an audio signal, using different ways to satisfy different purposes.

So, this works aims to use a *top-down* methodology in the singing voice signals recorded by classical singers. With a dedicated database, this research used audio visualisation softwares such as PRAAT and implemented digital signal processing techniques, seeking for the association between acoustic parameters and interpretative characteristics in the singing voice.

Keywords: MIR, Singing Voice, Digital Signal Processing

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

A voz cantada é o mais antigo instrumento de que se tem notícia: unindo música, letra e expressão, a voz tem a capacidade de impressionar as pessoas de uma maneira singular. Ao longo de anos, as características da voz cantada têm sido alvo de estudos por diversos autores, destacando-se a análise através de técnica lírica de canto que possui estética particular e rígida [dos Santos Ventura, 2011]. Existem diversas características da voz cantada como afinação, vibrato, amplitude, projeção, passagem de registro etc., que guardam informações importantes sobre o desempenho de um cantor e até mesmo sua saúde vocal. Apesar dos esforços empregados para o estudo da voz cantada desde o inicio dos anos 60 com a sintetização do canto, a complexidade deste fenômeno que utiliza os sistemas respiratório, fonatório, articulatório, ressonante e auditivo do corpo humano, é um grande desafio que tanto impulsiona pesquisas quanto limita resultados [Murphy, 2008].

Recuperação de Informação de Musica (do inglês, Music Information Retrieval – MIR), é definida por [Downie, 2004] como uma área de pesquisa multidisciplinar que desenvolve esquemas de busca baseados em conteúdo, interfaces inovadoras e evolui mecanismos de entrega interligados como um esforço para tornar o vasto mundo musical acessível a todos. Diversos sistemas de MIR já foram desenvolvidos para analise da voz cantada, como por exemplo, MIRACLE [Jang et al., 2001], SoundCompass [Kosugi et al., 2000], dentre outros, que possuem como objetivo a recuperação de informações como transcrição melódica, identificação de cantor, transcrição de letra, separação da voz [Murphy, 2008]. No entanto, a identificação de características na voz cantada esbarra nos desafios citados anteriormente e deixa este quesito fora do foco da grande maioria das aplicações.

Técnicas de Processamento Digital de Sinais proporcionam o estudo de um sinal de voz de forma cuidadosa, tornando possível descontruir as diversas formas de onda criadas durante a fonação, que, por conseguinte serve como base para analises minuciosas quanto a fisiologia e saúde do aparelho fonatório [Li and Wang, 2007] . A busca por características específicas do canto dentro do processo de analise da voz cantada pode ser justificada pela necessidade de entender melhor o funcionamento do corpo humano durante o processo do canto, para ajudar a prevenir danos físicos em cantores e também obter conclusões sobre o funcionamento da voz quando utilizada de forma considerada ótima e auxiliar no processo de ensino-aprendizagem de técnica vocal [Murphy, 2008]. Além disto, proporciona a oportunidade de estudo pratico de harmônicos e ondas de som musical e serve como base para áreas promissoras como a impressão digital vocal, que vem sento bastante explorada devido ao foco em segurança mundial e terrorismo [Downie, 2004].

1.1 Objetivos

1.1.1 Objetivo Geral

Relacionar características perceptíveis da voz cantada com parâmetros acústicos que podem ser extraídos a partir de um sinal de voz cantada analisado computacionalmente.

1.1.2 Objetivos Específicos

1. Montar um banco de dados específicos com sinais de voz cantada
2. Estudar a literatura
3. Formalizar conceitos relativos a caracteristicas vocais
4. Inspeccionar propriedades acusticas e fisicas associadas a caracteristicas vocais
5. Desenvolvimento e discussão de métodos de extração de parâmetros
6. Analisar os resultados a fim de obter a relação de características e parâmetros

1.2 Descrição do Documento

O primeiro capítulo deste trabalho trouxe uma introdução, contendo a justificativa do trabalho assim como seus objetivos. Também mostrou um histórico de recuperação de informações musicais através da voz cantada, assim como um perfil cronológico de avanços realizados nesta área e, por fim, o estado da arte, comparando métodos e plataformas existentes no mercado.

O segundo capítulo busca explicar o estado da arte e um breve histórico sobre a análise da voz cantada. O próximo capítulo, explicar o fenômeno da voz do ponto de vista fisiológico e físico, mostrando como o corpo humano produz som pelas cordas vocais e, mais especificamente, quais são as peculiaridades da voz cantada, assim como como esta é estudada e analisada. Ainda neste capítulo, são apresentadas e definidas características vocais buscadas e realizadas durante o canto, dentre as quais estão as que foram alvo de estudo neste trabalho.

O capítulo 4 aprofunda o conceito de frequência fundamental que é o princípio de extração de todos os parâmetros acústicos e a principal definição da emissão vocal no âmbito musical e também apresenta vários parâmetros acústicos que são obtidos por meio desta frequência e que servem para inferir estados e conceitos sobre o emissor (no caso, cantor).

O capítulo 5 tem como objetivo descrever em detalhes a metodologia empregada no trabalho, passando por cada etapa e, ainda, mostrando como os algoritmos foram desenvolvidos e quais adaptações foram feitas para que se alcançasse o objetivo final.

O sexto capítulo traz os resultados que são a aplicação dos algoritmos desenvolvidos no banco de dados criado especificamente para este fim, mostrando gráficos e tabelas e também realizando comparações entre os resultados obtidos e metodos e tecnicas ja sólidos e disponíveis no mercado.

O sétimo e último capítulo discute as conclusões e as contribuições alcançadas com este experimento, assim como aponta possibilidades de trabalhos futuros e desafios que ainda precisam ser vencidos na área de análise da voz cantada com o intuito de recuperar informações principalmente voltadas a feedback sobre características vocais. E, finalmente, referências bibliográficas e anexos, com trechos de código e algumas outras imagens, são postos no final deste documento.

CAPÍTULO 2

Estado da Arte

2.1 Histórico

O desejo e a necessidade de se estudar mais a fundo o processo de fonação e mais especificamente de aplicar técnicas de engenharia e ciência para tal objetivo se potencializaram bastante desde o início da indústria de telecomunicação. Para se ter noção, a sintetização da voz, por exemplo, começou no ano de 1773 quando Kratzenstein conseguiu reproduzir o som de vogais utilizando foles em cavidades ressonantes que, quando vibravam, produziam os sons. Mais tarde, em 1835, o alemão Josesh Faber criou uma máquina que literalmente imitava sons humanos e causou espanto na época: a Euphonia. Na estreia da máquina, o inventor fez com que ela "cantasse" um trecho do hino "God Save the Queen", em Londres – Inglaterra [Murphy, 2008].

Uma das primeiras tentativas de analisar a voz humana no século passado pode ser encontrada nos experimentos com um Vocoder, que é um instrumento capaz de analisar e sintetizar a voz, funcionando como um codificador vocal, criado primordialmente para o ramo da telefonia. O Vocoder separa o espectro das frequências geradas por uma voz (que serve como entrada) e grava esses espectros em bandas menores. Então, cada frequência dessas bandas menores é analisada e os parâmetros são salvos para serem reutilizados no processo de sintetização. Também idealizado por Bell, um pouco depois em 1989, o Voder foi criado: uma máquina capaz de produzir sinais de voz utilizando sinais elétricos [Fung, 2009].

Nos últimos anos, é possível notar um grande avanço na análise da voz cantada: O uso da transformada de Fourier que possibilita levar um sinal de áudio do domínio do tempo para o domínio das frequências, por exemplo, trouxe grandes contribuições para esse contexto, especialmente quando este método passou a ser executado em computadores, na década de 60. A descoberta da Transformada Rápida de Fourier que possui resultados aceitáveis, também impulsionou o processamento da voz [Cooley and Tukey, 1965].

O surgimento dos algoritmos de codificação preditiva linear (do inglês LPC) nos anos 60 e 70, trouxe ainda mais avanços para a área das telecomunicações e análise de voz: essa técnica extrai os formantes da voz, subtraindo-os do sinal de voz original (técnica conhecida como filtragem inversa) e então analisa o sinal resultante desta subtração. Existem versões adaptadas desta técnica como, por exemplo , quando se aplica um filtro digital de variação temporal que tenta prever a próxima amostra em um sinal de voz a partir de uma combinação linear das amostras anteriores. As correlações lineares correspondem a características espectrais, como os formantes. Vale salientar que essa filtragem gera um ruído que, se aplicado ao filtro inicial, gera o sinal de voz original [Alku and Backstrom, 2004]. O sucesso dos algoritmos LPC dentro da área da análise vocal dá-se pela semelhança entre a maneira como o sinal de voz é desconstruído e o modelo vocal fonte-filtro, que será discutido na seção 3 deste trabalho

[Alku and Backstrom, 2004]. É importante ressaltar que, apesar dos diversos modelos matemáticos desenvolvidos ao longo dos anos para a modelagem do processo de emissão vocal, o modelo fonte-filtro é o mais aceito e utilizado [Murphy, 2008].

Também na área médica é possível observar esforços durante a história para que o aparelho fonador fosse estudado mais detalhadamente: A criação da laringologia, que compreende todo o aparelho fonador e também a fonoaudiologia são exemplos de investimentos da ciência médica para melhor compreender a fonação. Um importante fato histórico dentro desse aspecto foi a criação do laringoscópio: aparelho utilizado para observar a laringe através de espelhos, criado por um professor de ópera chamado Manuel García. Este aparelho foi aceito na medicina em 1895 por esforços de Alfred Kirstein [Zeitels et al., 2002]. Hoje, obviamente, com o avanço tecnológico, o laringoscópio tornou-se em um tubo flexível que contém uma câmera em sua extremidade e é inserido através da garganta ou nariz do paciente a fim de analisar o processo de fonação. Neste exame, o paciente é solicitado a emitir algum som (seja fala ou canto) e o movimento das cordas vocais é gravado pela câmera. Em alguns casos, uma luz estroboscópica é emitida sobre as cordas vocais para que seja possível a visualização da vibração das pregas em baixa velocidade [Mota et al., 2009]. Além deste método, existe também a eletroglotografia (EGG), exame no qual dois eletrodos são posicionados no pescoço do paciente a fim de captar informações como frequência fundamental e outros parâmetros acústicos [Blowes,]. Mesmo com todos esses avanços, vale salientar que a maneira mais barata, menos invasiva e, segundo [Murphy, 2008], eficaz de se analisar o funcionamento da voz é através de uma gravação. Essa afirmação motivou e ainda impulsiona estudos para que o processo de filtragem e análise de características vocais (acústicas ou perceptíveis) por métodos computacionais seja cada vez mais preciso.

Desde que modelos de representação e técnicas de filtragem da voz foram criados e implementados, diversos parâmetros acústicos como o Jitter e o Shimmer que serão discutidos posteriormente, são utilizados para a inferência e o diagnóstico de patologias vocais [Titze, 1995]. Mais recentemente, esses parâmetros foram associados a características vocais, especialmente no processo de canto [dos Santos Ventura, 2011].

Hoje em dia, a análise vocal tem sido usada primordialmente pelas áreas médicas e pela indústria de jogos. Em [Bonada et al., 2001] é mostrado um modelo de excitação e ressonância que é utilizado no karaoke craze, encontrado especialmente na Ásia [Fung, 2009], que funciona de maneira similar ao modelo fonte-filtro. O trabalho desenvolvido em [Alemán and Carlosena, 2004] mostra outra aplicação moderna da análise vocal, mais especificamente para cantores através do estudo do vibrato, combinando trabalhos anteriores na área de análise de patologias da voz e processamento do sinal de voz. Assim, este trabalho se justifica como inserido no contexto da modernização da análise da voz, mais especificamente a voz cantada e traz benefícios para estudiosos e praticantes desta área.

2.2 Estado da Arte

Muitos esforços foram feitos no estudo da voz cantada para que fosse possível a sua sintetização e aplicação em diversas áreas, especialmente entretenimento e robótica. Paralelo a esta tendência e, ainda, com menos enfoque, existe o seguimento que estuda a voz cantada com o

intuito de prover feedback para cantores e ainda servir como aliado a área médica e fonoaudiológica no cuidado com a saúde vocal dos profissionais da voz. Com o intuito de discutir trabalhos realizados MIR com o foco na voz cantada, esta seção discutirá três importantes trabalhos que serviram como base para este trabalho e também mostrará um resumo de aplicações e softwares que tem como objetivo obter características da voz cantada, utilizados em diversas esferas do mercado musical.

O trabalho feito em [dos Santos Ventura, 2011], estudou diversas características da voz do ponto de vista teórico e escolheu uma, que foi o vibrato, para ser desenvolvida e possivelmente incorporada a plataforma Sing Studio, proporcionando um mais detalhado nível de feedback ao cantor que utilizasse esta ferramenta como material de estudo. O vibrato, muito comum entre os ca ntores que utilizam principalmente a técnica lírica ainda é considerado uma lacuna entre os softwares disponíveis no mercado. O algoritmo desenvolvido neste trabalho tem como entrada a frequência fundamental de um sinal de voz (portanto, é dependente de um método de extração desta frequência para o seu funcionamento) e tenta detectar trechos nos quais existe a presença do vibrato baseado na modulação da frequência fundamental em uma faixa de frequências periódicas e, também, calcula a duração, extensão e frequências envolvidas neste trecho. Utilizando diversos passos e técnicas, foi desenvolvido um estudo importante desta característica e ainda, o autor propôs uma maneira de transmitir ao emissor que estivesse utilizando o Sing Studio as informações calculadas. Foram desenvolvidos testes utilizando vozes sintetizadas e também naturais a fim de avaliar a robustez e precisão deste algoritmo. Informações preciosas sobre técnicas de extração de características e ainda a formalização de conceitos estão entre as principais contribuições adicionais do trabalho citado.

Em [Murphy, 2008], um excelente trabalho de explicação do processo de fonação foi desenvolvido, com detalhes sobre funções de transferência e modelagens específicas do som produzido em cada parte do processo de fonação. O objetivo principal foi utilizar técnicas de processamento de sinais para a obtenção de parâmetros acústicos da voz e então o desenvolvimento de métodos estatísticos para a comparação entre indivíduos treinados e não treinados (cantores e amadores) e vozes saudáveis e doentes. Diversos métodos estatísticos foram estudados e avaliados com o objetivo de viabilizar inferências mais precisas em distinguir uma população em relação aos parâmetros vocais destacados.

Com o objetivo de estudar e entender melhor a associação entre parâmetros acústicos e características artísticas, [de Sá Ferreira, 2012] realizou um experimento de, também por um método estatístico, tentar encontrar uma associação entre alguns parâmetros e questões artísticas e de estilo como portamento (condução de uma nota para outra de forma lenta e destacada) e ainda, questões de saúde (como excesso de ar na voz, que pode ser patológico). O método empregado pelo autor serviu como base para este trabalho, que adaptou o cenário e buscou características técnicas ao invés de interpretativas.

2.2.1 Plataformas Existentes

Comercialmente, embora muitas das aplicações que utilizam MIR para obter informações da voz cantada sejam com outro foco (conforme já discutido neste trabalho), existem diversas plataformas principalmente que tem como público alvo cantores e/ou estudantes de canto que buscam feedbacks técnicos sobre o seu desempenho. É preciso salientar que nenhuma destas

plataformas e também este trabalho tem como objetivo substituir o papel de um professor de canto. A grande busca e vontade é servir como aliado ao trabalho de ensino do canto (e até mesmo do estudo pessoal e individual) por fornecer informações rapidamente (as vezes até em tempo real). A seguir, serão mostradas algumas das principais plataformas de análise da voz cantada utilizadas para recuperar características vocais.

2.2.1.1 Sing and See

Desenvolvido pela Cantovation Technology, este software foi inicialmente criado por uma equipe multidisciplinar de pesquisadores da área da voz na Austrália. O Sing and See mostra em tempo real as notas cantadas pelo usuário e também o espectrograma do sinal de voz, com o intuito de mostrar os harmônicos desenvolvidos e suas mudanças durante o canto. Ainda, as notas são mostradas em um piano, dando feedbacks sobre afinação. Também compara as notas cantadas com uma pauta musical e representa a amplitude da voz em semitons.

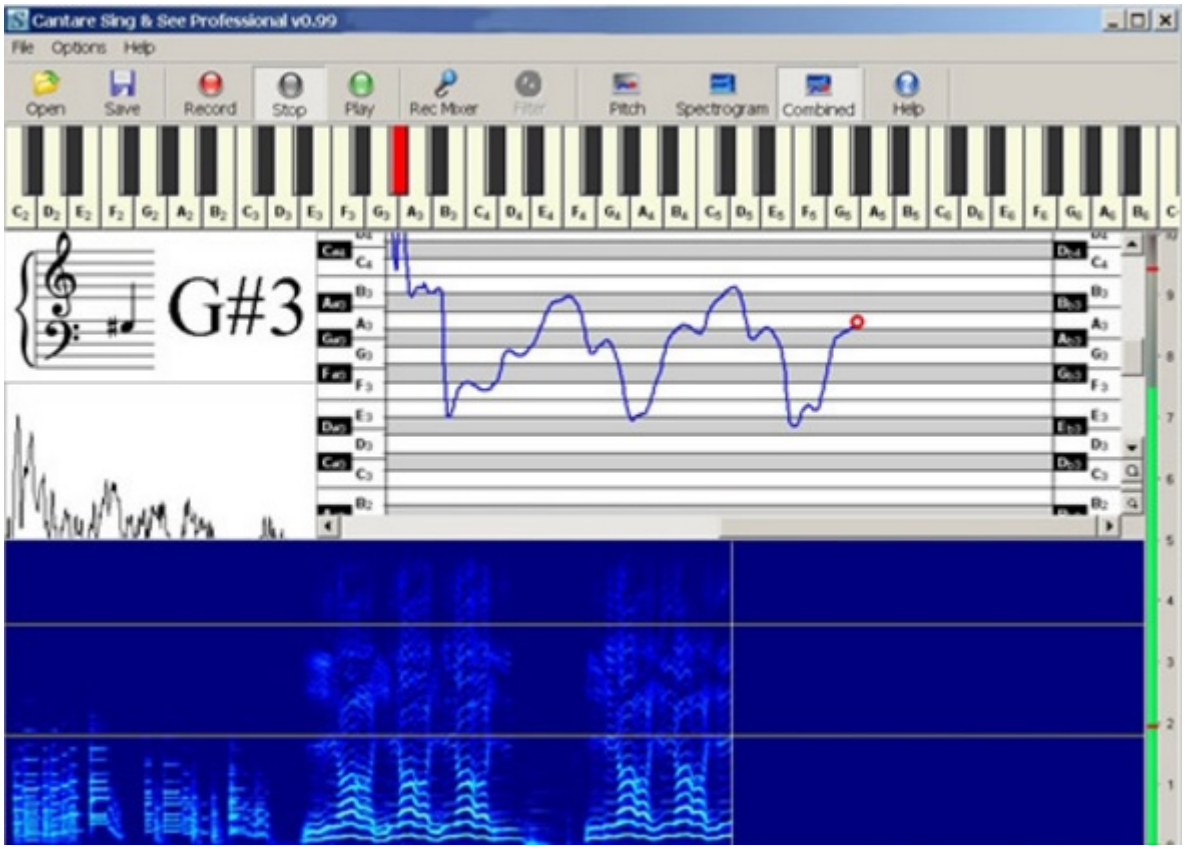

Figura 2.1 Screenshot do Sing and See.

2.2.1.2 Music Master Works

Produto da Aspire Software, este produto serve também como um editor de partituras e notações musicais. No que diz respeito a voz cantada, ele permite a importação de um arquivo e

comparação do mesmo com outro arquivo MIDI e também com uma pauta musical. Representa o pitch e a amplitude da voz em tempo real.

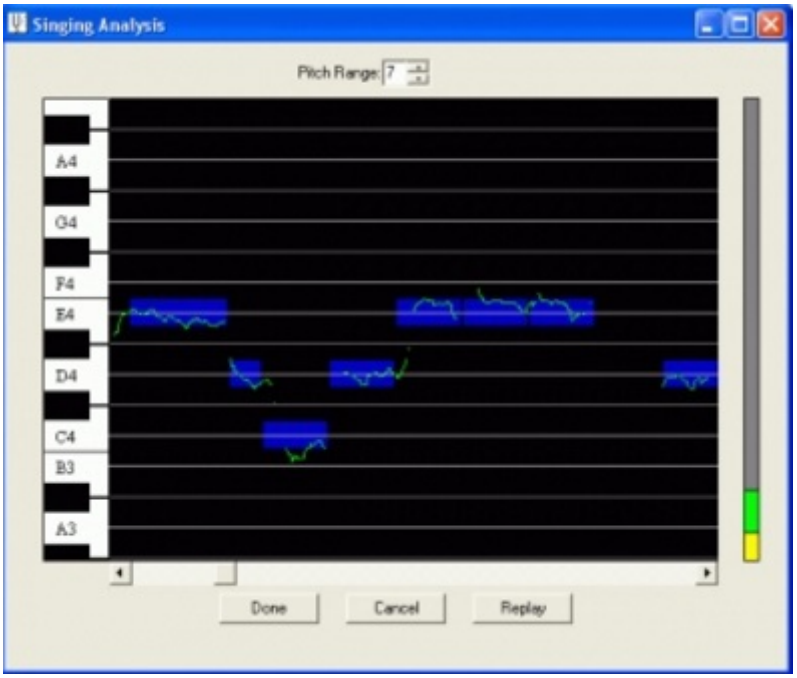

Figura 2.2 Screenshot do Music Master Works.

2.2.1.3 Singing Studio

O singing Studio é um ambiente interativo que proporciona também feedback em tempo real da voz cantada. Ele usa a voz captada de um microfone e exibe a nota cantada junto a uma representação das teclas de um piano. Basicamente, a afinação é o único conceito explorado nesta aplicação e permite ainda um tipo de pontuação para o caso de comparação com uma pauta musical ou exercício vocal pré-estabelecido no programa. É desenvolvido e comercializado pela empresa portuguesa Seenegal.

Figura 2.3 Screenshot do Singing Studio.

2.2.1.4 Sing Star

Voltado um pouco mais para o mercado de entretenimento, o Sing Star eh um jogo no formato de aplicativo e tambem disponivel para consoles, que funciona como um karaokê que pontua por afinacao. Possui um algoritmo para calcular o pitch através da voz captada e compara com uma nota esperada.

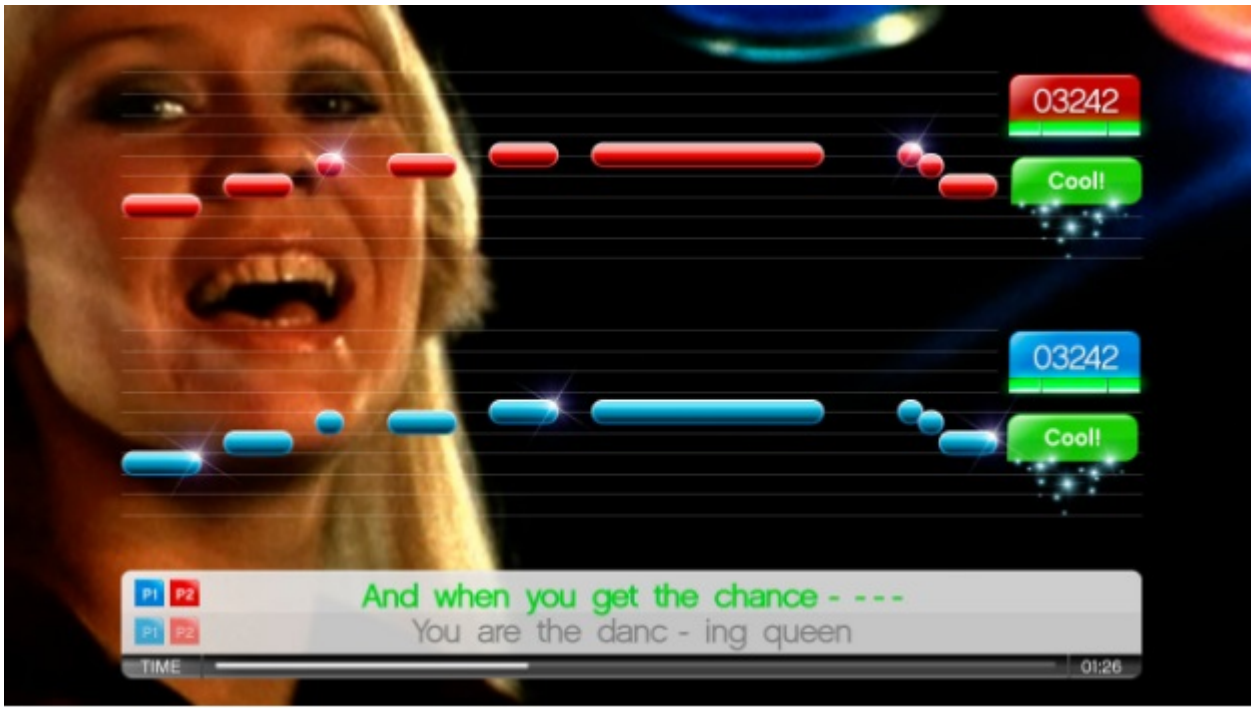

Figura 2.4 Screenshot do Sing Star.

2.2.1.5 Music Tutor

Software descontinuado e extinguido do mercado, foi um dos pioneiros no mercado de software para voz cantada e também, traz feedback da afinação a partir da captação da voz e comparação em uma pauta musical da nota cantada. Foi desenvolvido pela Sestek.

Uma importante consideração a ser feita é que absolutamente todas as plataformas pesquisadas tratam do quesito afinação, a partir do cálculo do pitch e comparação deste quesito com notas esperadas.

CAPÍTULO 3

O Fenômeno da Voz

A voz humana é produzida pelo movimento de vibração das pregas vocais por consequência do ar que vem dos pulmões devido a ação do diafragma e esse ar sofre então modificações (principalmente no que diz respeito a espectros) feitas pelo trato vocal, incluindo língua, lábios e dentes [Guimarães, 2007]. As pregas vocais (também conhecidas como cordas vocais) são duas pregas musculares encontradas na região da laringe e, de forma bem resumida, são o elemento que vibra no processo de fonação, por conta do movimento de adução que emprega resistência a saída do ar e, então, uma modulação dos fluxos de ar. A velocidade com que essas pregas abrem e fecham (chamado de frequência típica) é de, em média, 210 vezes/segundo entre as mulheres e 110 vezes/segundo entre os homens. Entretanto, essas taxas podem variar bruscamente dependendo da fisiologia de cada indivíduo e sobretudo com o ato de cantar, que naturalmente varia a emissão da frequência fundamental [Högset, 2001].

Levando em consideração o fato das pregas vocais se situarem na laringe, diz-se que o som produzido pelos fenômenos citados acima é chamado de som laríngeo. Este, é composto pela frequência fundamental (a frequência mais baixa da onda produzida correspondente a vibração das pregas vocais) e pelos seus harmônicos parciais. A frequência fundamental tem ligação direta com as características fisiológicas e morfológicas das cordas vocais como por exemplo o tamanho, a grossura, a elasticidade, entre outros. Logo, há uma enorme variabilidade nos valores desta frequência fundamental e é esse um dos principais fatores que faz com que cada pessoa tenha uma voz ou um timbre diferenciado [Högset, 2001].

A figura 2.1 abaixo mostra as cordas vocais em funcionamento. Cada fotografia presente na figura foi tirada em um intervalo de segundo. Nas primeiras seis fotografias, a fala é interrompida, gerando um afastamento entre as pregas vocais. Nas outras, a fala é retomada e observa-se como as pregas vocais se juntam [de Sá Ferreira, 2012].

O som laríngeo produzido a nível de pregas vocais, ainda precisa ser amplificado para que seja ouvido. Então, a próxima etapa deste processo de fonação consiste na passagem do som pelas cavidades supraglóticas que são a laringe, a faringe, boca e cavidade nasal que constituem o trato vocal. O trato vocal funciona como uma espécie de caixa ressonadora para as frequências emitidas e amplifica não apenas a frequência fundamental mas também os harmônicos parciais. Os sons produzidos são agregados em duas grandes classes: sons vozeados e sons não vozeados, que são determinados pela vibração ou não das cordas vocais. Sons vozeados estão ligados a predominância de vogais, enquanto os não-vozeados, a predominância de consoantes. Como exemplo, temos como sons vozeados os produzidos pela pronúncia das vogais [a], [e], [i], [o], [u]. Para os não vozeados, a pronúncia das consoantes [f] e [s] são um exemplo. Essa definição é de extrema importância pois é largamente utilizada no processo de análise e síntese da voz.

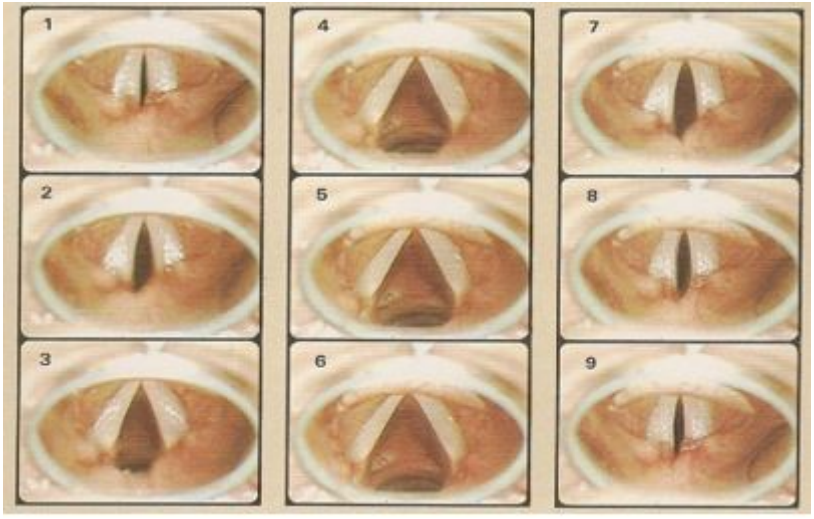

Figura 3.1 Cordas vocais em funcionamento.

3.1 Voz Falada X Voz Cantada

Se analisarmos o contexto histórico, podemos ver que a voz cantada e a voz falada andam juntas no que diz respeito a pesquisas científicas, mas, obviamente, existem diferenças muito importantes nesses dois métodos de fonação. Aproximadamente 90% dos sons produzidos durante o canto (voz cantada) são vozeados, enquanto que na fala, por exemplo, da língua inglesa, esse percentual atinge no máximo 60% [Cook, 1991]. Em uma das técnicas clássicas de canto mais comum, o bel canto, um dos princípios ensinados aos cantores eh a sustentação das vocais pelo maior tempo possível entre os fonemas justo porque são mais audíveis. Um dos benefícios disso é que os cantores pronunciam as vogais mais consistentemente e isso facilita, por exemplo, a determinação de vogais por meio de análise de um sinal de voz cantada. Cantores que utilizam a técnica clássica, por princípio da própria técnica, abaixam a laringe durante o canto, criando uma ressonância de alta frequência adicional (em torno de 5 kHz), que não é encontrada em outros tipos de fonação. Essa ressonância, conhecida como formante do cantor, é o que faz com que um cantor seja ouvido sem utilizar um microfone mesmo na presença de outros instrumentos de uma orquestra durante uma ópera ou um concerto, por exemplo [Sundberg and Rossing, 1990].

Por conta de sua natureza, sons vozeados são mais fáceis de analisar e também sintetizar utilizando a teoria linear de processamento de sinais [Sundberg and Rossing, 1990]. No Ocidente, por exemplo, a variedade de frequências fundamentais utilizadas na voz cantada é bem maior do que na voz falada. Ainda, a voz cantada tem nitidamente uma variação de dinâmicas maior em termos de amplitude do que na voz falada.

Tomando mais uma vez a técnica clássica de canto como referencial, uma semelhança entre esta e a voz falada é a atuação (teatro, por exemplo): Da mesma forma que cantores lêem uma partitura ou cantam uma melodia com características previamente definidas pelo compositor, um ator interpreta um texto de acordo com um script. Assim como cantores, atores precisam

projetar suas vozes para que sejam ouvidos em um auditório cheio e não somente se fazerem ouvir mas transmitir emoções, sentimentos e intenções, assim como na música.

Como o foco deste trabalho é a voz cantada e a técnica clássica de canto imprime uma maior rigidez e padronização na formação de harmônicos e utilização do corpo, além de outros benefícios discutidos anteriormente, as próximas seções serão focadas na voz cantada, mais especificamente na técnica clássica de canto.

3.2 Voz Cantada

No canto, diversos sistemas do corpo são envolvidos para que o som produzido contenha e transmita sentimentos e mensagens e isso determina pontos específicos, por exemplo, da respiração, que é orientada pela frase musical.

Como matéria prima deste fenômeno, o ar está presente durante todo o processo de emissão de voz. Com o intuito de diminuir o ruído originado na inspiração do ar e também como parte de requisitos técnicos, durante a inspiração, os cantores normalmente expandem as costelas inferiores aumentando a caixa torácica, tendo uma maior quantidade de ar (que é dada pela soma do ar residual com o ar da inspiração) [Sundberg and Rossing, 1990]. No processo de expiração, utiliza-se a musculatura abdominal para que o cantor tenha um maior controle da saída de ar. Vale ressaltar que o suporte respiratório é uma das principais condições para uma boa emissão vocal [Sundberg and Rossing, 1990]. O processo de canto começa com a pressão de ar produzida pelos pulmões. Para o caso de sons vozeados, os músculos cricoaritenóideos aduzem as pregas vocais e justamente a pressão de ar faz que as pregas se abram.

Conforme discutido anteriormente, na fonação, o som laríngeo sofre modificações por parte do trato vocal, sendo amplificado. No canto, esta ressonância é muito mais presente e se concentra na parte superior deste trato vocal, com o objetivo de aliviar a sobrecarga muscular na laringe. Quando cantores não treinados tentam dissipar essa energia sonora na laringe, muito provavelmente irão canalizar isso para a região nasal, gerando uma voz nasalada [Yan et al., 2005].

Uma outra característica muito importante presente na voz cantada é a sua estabilidade. O emprego de uma boa técnica, boa audição e controle emocional contribuem diretamente para este fator.

3.2.1 Análise da Voz Cantada

O modelo de análise mais comum da voz cantada mais comum foi proposto por Fant, em 1970. Ele divide a fonação em três partes independentes sendo estas a fonte sonora, o filtro acústico e a radiação acústica. O som laríngeo, cujo processo de produção foi discutido anteriormente neste capítulo é a matéria prima da voz cantada [Henrique, 2002] e representa a parte da fonte. O trato vocal que modula e divide o som por suas cavidades representa a parte do filtro. Por fim, a projeção do som que se dá pela ação dessas cavidades como amplificadoras junto com a radiação dos lábios, que é o mais externo componente do trato vocal, representa a radiação acústica. Este modelo caracteriza os fenômenos acústicos do domínio das frequências [Henrique, 2002].

Partindo para a extração das características através de um sinal de voz, podemos usar como exemplo um Eletrocardiograma (ECG), onde eletrodos captam a ação elétrica do coração por

meio de um aparelho que funciona como um galvanômetro. Existe um exame chamado eletroglotografia (EGG) que funciona como uma espécie de ECG para as cordas vocais: colocar dois eletrodos junto a laringe para extrair parâmetros desejados. Entretanto, devido à pouca praticidade e da necessidade de uma estrutura material e pessoal (médicos) para a realização deste exame, a forma mais comum é efetuar uma gravação e utilizar métodos computacionais para extrair os parâmetros. Esses métodos computacionais têm evoluído ao longo dos anos devido a avanços importantes dos algoritmos como por exemplo a FFT (do inglês, Fast Fourier Transformation) e outros algoritmos que, falando de maneira sucinta, visam extrair os formantes da voz, subtraí-los ao sinal original e depois analisar as informações existentes. Nas próximas subseções, serão discutidas as características da voz e também os parâmetros que se utilizam para a detecção dessas características.

3.2.1.1 Parâmetros Perceptivos da Voz Cantada

Durante o canto existem diversos parâmetros que são considerados perceptivos e que se baseiam na técnica e no objetivo de cada cantor. Alguns desses parâmetros estão ligados a padrões estilísticos de determinados tipos de música. As definições desses parâmetros muitas vezes são subjetivas e acompanham os cantores desde as aulas de canto. Isso faz com que, muitas vezes, esses parâmetros sejam difíceis de distinguir pelo público leigo.

Nesta seção, serão definidos e caracterizados alguns dos parâmetros mais comuns encontrados na prática do canto e que serão objetos de pesquisa deste trabalho que tem como objetivo, através de uma abordagem top-down, identificar quais dessas características podem ser alcançadas utilizando técnicas de MIR (mais especificamente aquelas ligadas ao processamento de sinais) em um sinal de áudio.

3.2.1.1.1 Afinação Pode-se definir como afinação a capacidade de produzir um que tenha a mesma frequência que outro, partindo sempre de um mesmo referencial (por exemplo, a nota Lá 440 Hz) [Cook, 1991]. A afinação pode variar com questões de natureza cultural: na europa, por exemplo, tem-se a necessidade de referência ao padrão utilizado (seja um modo ou uma escala). A escala mais utilizada chama-se escala igualmente temperada e sua característica principal é a igualdade entre todos os meio-tons [Donoso, 2012].

A relação R entre notas separadas por um meio tom é descrita:

$$R = \frac{1}{2^{\frac{1}{12}}} \tag{3.1}$$

3.2.1.1.2 Tessitura A tessitura é a região de emissão de notas ou o conjunto de notas que são emitidas com conforto por um cantor. Nesta região, a voz é emitida sem esforço e com qualidade. A tessitura está ligada diretamente com a fisiologia das cordas vocais: o tamanho a grossura e a elasticidade das pregas influencia diretamente no timbre do indivíduo. Ainda, esta característica é responsável por classificar os cantores: vozes femininas são classificadas como contraltos, mezzo-sopranos ou sopranos e as vozes masculinas como baixos, barítonos, tenores ou contratenores. Existem ainda outras subclassificações que estão mais ligadas a outros fatores de timbre: por exemplo, sopranos podem ser subclassificadas como spinto, dramático, ligeiro

Tabela 3.1 Tessitura Vocal Masculina

Classificação Vocal	Nota Inicial	Nota Final
Baixo	Dó 2	Ré 4
Barítono	Fá 2	Sol 4
Tenor Dramático	Lá 2	Si 4
Tenor Spinto	Si 2	Ré 5
Tenor Lírico	Si 2	Ré 5
Tenor Ligeiro	Dó 3	Mi 5
Contratenor	Fá 3	Si 6

etc. As tabelas 2.1 e 2.2 abaixo mostram as classificações vocais e a região de tessitura de acordo com o sistema Fach [Mangini et al., 2013]:

Tabela 3.2 Tessitura Vocal Feminina

Classificação Vocal	Nota Inicial	Nota Final
Contralto	Mi 3	MI 5
Mezzo-soprano	Sol 3	Lá 5
Soprano Dramático	Sol 3	Si 5
Soprano Spinto	Si 3	Dó 5
Soprano Lírico	Si 3	Dó 5
Soprano Ligeiro	Dó 4	Ré 5
Soprano Coloratura	Ré 4	Fá 6

3.2.1.1.3 Loudness Loudness é a sensação auditiva causada pelo nível de intensidade à medida que ocorre uma variação de frequências: é justamente a relação da intensidade de um som com a sua frequência. Vale salientar que, apesar de estarem intrinsecamente relacionados, loudness não é apenas a intensidade da voz.

3.2.1.1.4 Vibrato O vibrato tem como característica uma variação da frequência fundamental de forma regular durante a emissão de uma nota musical. De acordo com [Sundberg and Rossing, 1990], a frequência f aceitável do vibrato varia de f = 5.5 Hz a f = 7.5Hz e ainda, a extensão pode ser considerada aceitável entre 1 e 2 semitons.

Em termos técnicos, o vibrato é resultante do relaxamento da musculatura da laringe ou ainda, em alguns casos, da modulação de tensão na região laríngea ou variação na pressão subglótica, fazendo com que as pregas vocais tenham sua tensão média elevada (vale salientar de que este último método citado não é utilizado por cantores com técnica sólida).

A figura 2.2 mostra a detecção do vibrato pelo software Singing Studio. Vê-se claramente uma variação da frequência fundamental porém em torno da nota que está sendo emitida (no caso do exemplo, um La).

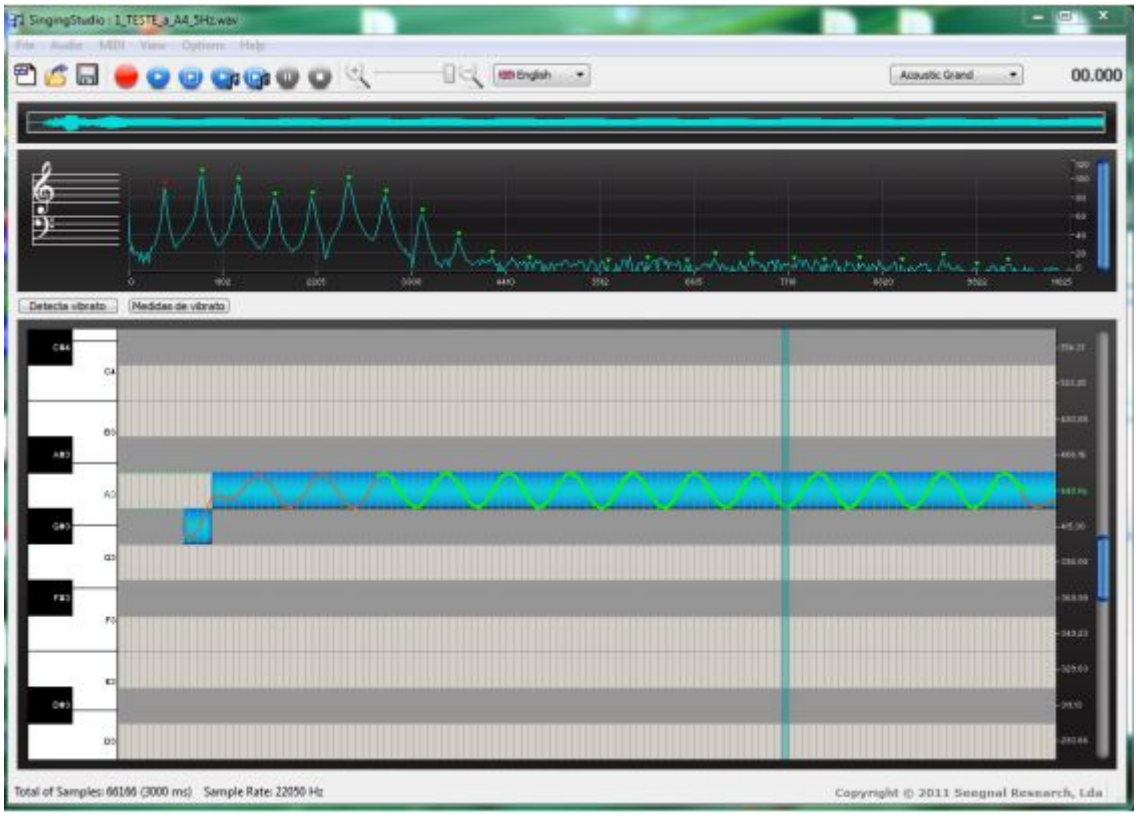

Figura 3.2 Representação do Vibrato no Singing Studio.

3.2.1.1.5 Timbre De acordo com [Henrique, 2002], o timbre eh uma característica sonora que nos possibilita diferenciar sons de mesma frequência e intensidade emitidos por diferentes fontes sonoras. Ele eh justamente o conjunto de características que se somam a frequência fundamental, podendo incluir distribuição de energia espectral, envolvente temporal, grau de inarmonicidade dos parciais e frequência. Na voz cantada, o timbre está completamente relacionado à técnica e a característica fisiológica de cada indivíduo, ou seja, o trato vocal. Por vezes, é comum utilizar o termo "voz timbrada" que, na verdade, corresponde a presença (ou ausência) de algumas características específicas. Essas características serão listadas com seus respectivos termos antagônicos:

- **Claro e Escuro**: Na voz, esse termo se refere a presença de harmônicos reforçando baixa ou alta frequência. Uma voz clara possui predominância de harmônicos agudos e uma vez escura possui predominância de harmônicos graves.

- **Voz na máscara e voz recuada:** Estes aspectos dizem respeito a projeção. A voz na máscara explora o som a partir das cavidades dos seios da face e gera mais nitidez e frontalidade. A voz recuada é mais difícil de se ouvir e entender, pois explora demais espaços internos que não projetam tanto a voz. Vale salientar que a voz na máscara não se refere a nasalização, que eh um uso errado e, infelizmente, comum entre cantores que

não estudam.

- **Limpeza e soprosidade:** Estes aspectos estão muito relacionados a saúde do cantor e são a representação da quantidade de ar na emissão da voz. A soprosidade (quando há excesso de ar) resulta de uma fenda glotal (quanto as pregas vocais não entram em contato uma com a outra). Além do fato patológico, a soprosidade pode ser resultado também de falta de técnica, sendo esta característica raríssima em cantores líricos (se existe, é porque há um erro técnico).

3.2.1.1.6 Falsete O termo falsete vem do italiano falsetto, que significa falso. Este termo é associado a voz cantada produzida pela vibração de apenas uma fração das pregas vocais. Consiste justamente em mudar registros utilizados pelo cantor para a emissão do som para o registro da cabeça. O falsete gera níveis de mais altos na frequência fundamental do que normalmente o cantor produziria (por isso o nome). Durante o processo de canto utilizando o falsete, as pregas vocais estão mais esticadas e, portanto, a região de contato é menor entre elas e, em compensação, demanda mais energia. Isso faz que com a amplitude da fonação seja menor. Apesar de presentes, harmônicos são mais escassos na voz de falsete e isso traz uma enorme diferença desta em relação a voz normal (ou modal) [Sundberg and Rossing, 1990].

3.2.1.1.7 Portamento Portamento nada mais é do que uma ligação entre duas notas. Característica muito comum principalmente no estilo erudito, pode ser designada até mesmo numa partitura. Em termos mais práticos, é quando o cantor muda de uma nota para outra passando por semitons entre essas. Existe uma infinidade mais de características associadas a voz cantada mas essas listadas acima são de extrema importância para a prática do canto e, ainda, englobam as análises realizadas neste trabalho e, sendo assim, outras características não serão definidas ou exploradas.

CAPÍTULO 4

Frequência Fundamental e Parâmetros Acústicos

Com o intuito de analisar e identificar os parâmetros perceptíveis da voz cantada, alguns parâmetros acústicos foram estudados e analisados e tiveram a sua a sua relação com os parâmetros listados na seção de parâmetros perceptíveis da voz, mapeada. Esses parâmetros representam significativamente a limitação que as técnicas utilizadas de para recuperação de informação da voz cantada encontram. Mesmo com tal diversidade, a frequência fundamental é o ponto de partida para a análise de todos os outros parâmetros. Então, esta seção irá discutir a definição da frequência fundamental e sua relação com os demais parâmetros acústicos utilizados para mapear algumas características da voz cantada alcançáveis por meio de técnicas computacionais.

4.1 Frequência Fundamental

De acordo com [dos Santos Ventura, 2011], podemos definir a frequência fundamental ($f0$) como sendo o valor de frequência mais baixa numa estrutura harmônica ao qual se relacionam os harmônicos de uma onda periódica. Com o passar dos anos, diversos avanços sobre a maneira de estimar a F0 foram alcançados entretanto, a pluralidade de contextos nos quais se faz necessária esta estimativa faz com que a precisão dos algoritmos seja um desafio contínuo. Comumente, são utilizados dois principais métodos para análise da frequência fundamental: análise temporal e espectral.

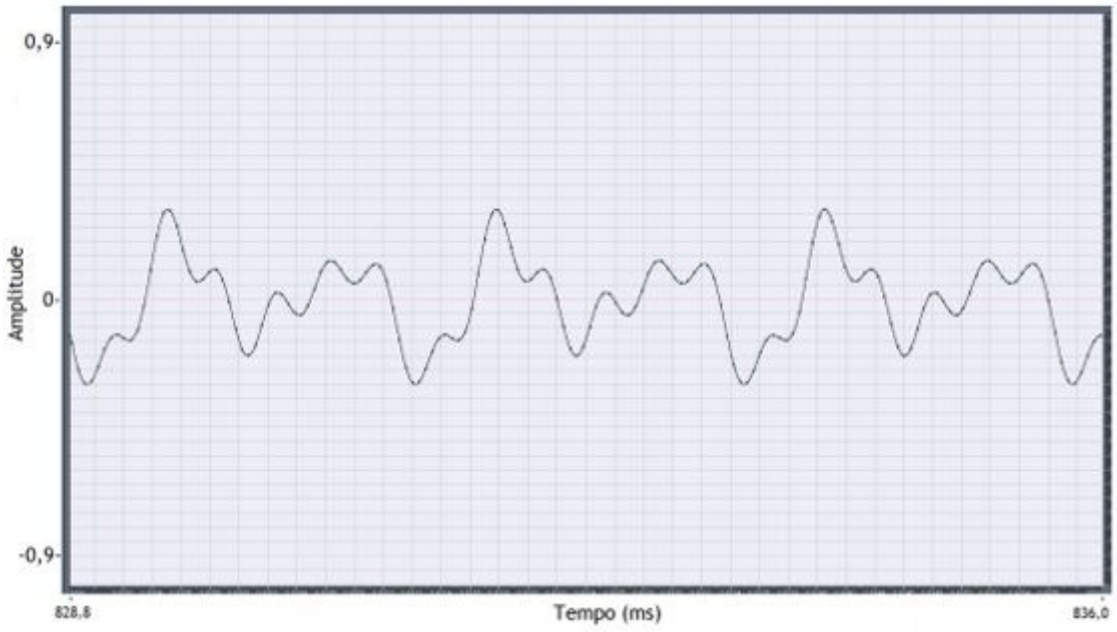

Figura 4.1 Frequência de voz sintetizada.

A figura 3.1 acima mostra um trecho de um sinal de voz cantada sintetizada com a frequência fundamental 440Hz. Pode-se observar claramente a periodicidade desse sinal e, portanto, a extração da frequência por meios temporais é factível.

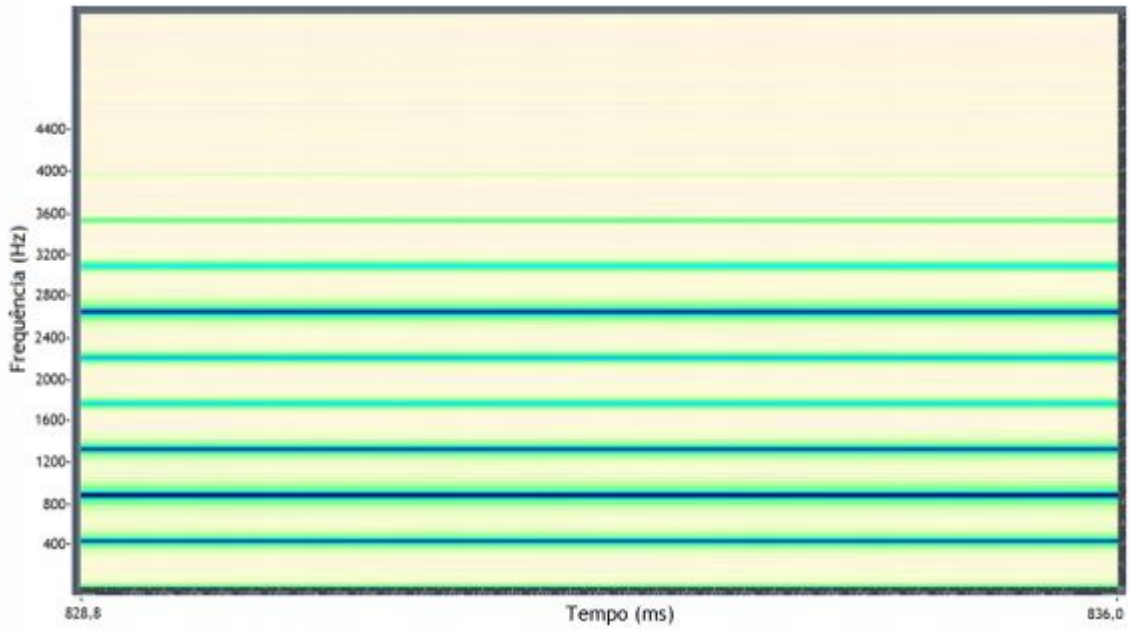

Figura 4.2 Espectrograma de voz sintetizada.

A figura 3.2 acima mostra o mesmo sinal de voz sintetizado (440 Hz) porém por meio de um espectrograma. Vê-se nitidamente que a frequência mais baixa (referente a 440 Hz, que é a $f0$) está destacada por ser uma linha azul. Se analisássemos esse espectrograma de forma que evidenciássemos a densidade espectral, como por exemplo, na figura 3.3 abaixo, o valor da frequência poderia ser evidenciado pelo primeiro máximo local [dos Santos Ventura, 2011]. Mais abaixo, serão descritos alguns algoritmos e técnicas dentro de cada abordagem (espectral e temporal) para estimar a $f0$.

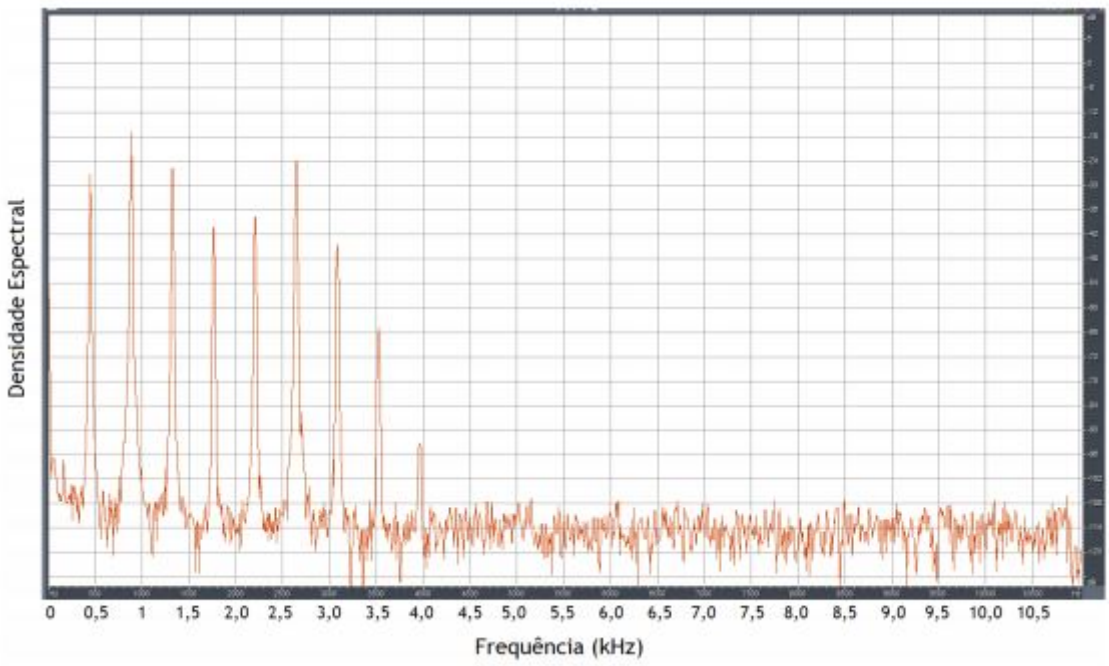

Figura 4.3 Densidade Espectral de voz sintetizada.

4.1.1 Análise Temporal

Como sugerido pelo próprio nome, esta abordagem tenta estimar a $f0$ através do sinal original ao longo do tempo. Existem diversas técnicas para a estimativa da frequência fundamental. Alguns métodos que são mais simples (como por exemplo Zero-crossing rate, peak detection e slope event rate) são de baixo consumo computacional mas também se revelaram imprecisos no que diz respeito a voz cantada, uma vez que essa é rica em harmônicos [dos Santos Ventura, 2011].

Abaixo, serão discutidos alguns métodos para a estimação da frequência fundamental na perspectiva de análise temporal.

4.1.2 Zero-cross rate

Como o próprio nome sugere, o princípio deste método consiste em estimar a quantidade de vezes em que a onda passa por zero por unidade de tempo. De acordo com [Brandão et al., 2007], quando a potência espectral está em torno da frequência fundamental, a onda passará por zero duas vezes no mesmo período. Essa é considerada uma técnica simplista para estimar a frequência fundamental porque fica presa em relação ao período da onda. Outro problema é a presença de harmônicos: Quando, na onda, existem componentes de mais alta frequência (que são muito presentes na voz cantada por conta dos harmônicos obtidos no processo de emissão vocal), a onda pode passar mais vezes por zero num mesmo período e, consequentemente, o método falharia ao estimar a frequência fundamental. As duas imagens a seguir (3.4 e 3.5) mostram

dois exemplos de funções. A primeira, representada pela função seno, onde o método funcionaria. A segunda, com uma equação que representa a inserção de componentes de mais alta frequência na função seno, pode evidenciar como este método falharia.

Vale salientar que a opção de filtrar os harmônicos não pode ser fundamentalmente levada em conta pois, para o caso da voz cantada, existem diversos parâmetros acústicos que se baseiam nestes.

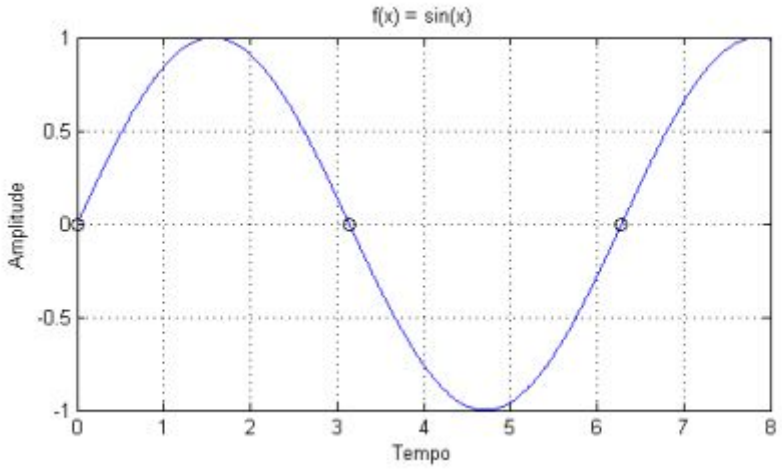

Figura 4.4 Função seno.

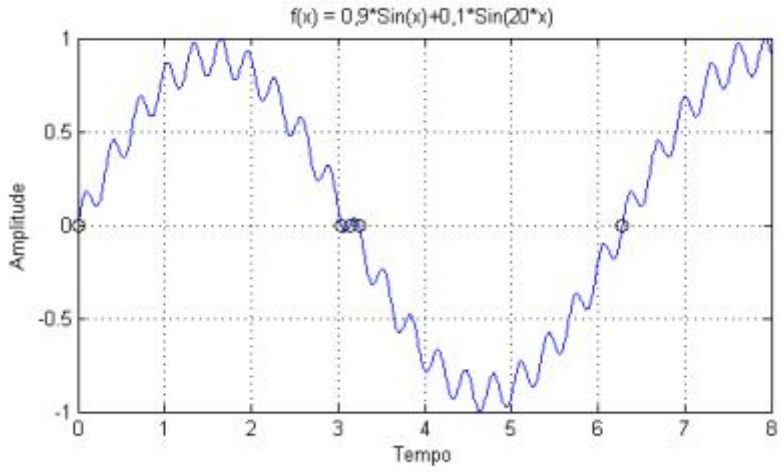

Figura 4.5 Função seno mais harmônicos.

4.1.2.1 Slope Event Ratio

Da mesma maneira que o método citado anteriormente, este não tem uma grande robustez. Considerando que a forma de onda tem um período, é possível concluir que a mudança de sinal relacionada ao declive também terá um período. Então, partindo desse pressuposto, a frequência fundamental pode ser estimada em uma predição e contagem desses eventos, de forma semelhante a explica no método ZCR, na seção 3.1.1.1.

4.1.2.2 Autocorrelação

O conceito de autocorrelação de um sinal de onda é, como o próprio nome sugere, a relação entre o sinal e ele mesmo, afetado de um deslocamento, com o objetivo de obter uma medida da semelhança da forma de onda.

A função de autocorrelação de um sinal estacionário é dada pela equação:

$$R_x(\tau) = \lim_{t0\to\infty} \frac{1}{t0} \int_{\frac{-t0}{2}}^{\frac{t0}{2}} x(t+\tau)x(t)dt \tag{4.1}$$

A partir desta equação, pode-se inferir, de acordo com [Brandão et al., 2007], que se as duas partes estão correlacionadas, as somas das autocorrelações das partes de um sinal representa a sua autocorrelação geral. Se o sinal for periódico, a autocorrelação vai também ser periódica. Segundo [dos Santos Ventura, 2011], uma das particularidades desse método é a existência de um máximo global quando $\tau = 0$. Então, a partir do período fundamental $t0$, pode-se obter o valor da frequência fundamental, uma vez que esta é o inverso de $t0$ e, para determinar $t0$, leva-se em consideração os máximos locais da autocorrelação de um sinal.

A figura 3.6 abaixo, mostra a determinação da $f0$ para o caso de um sinal de voz.

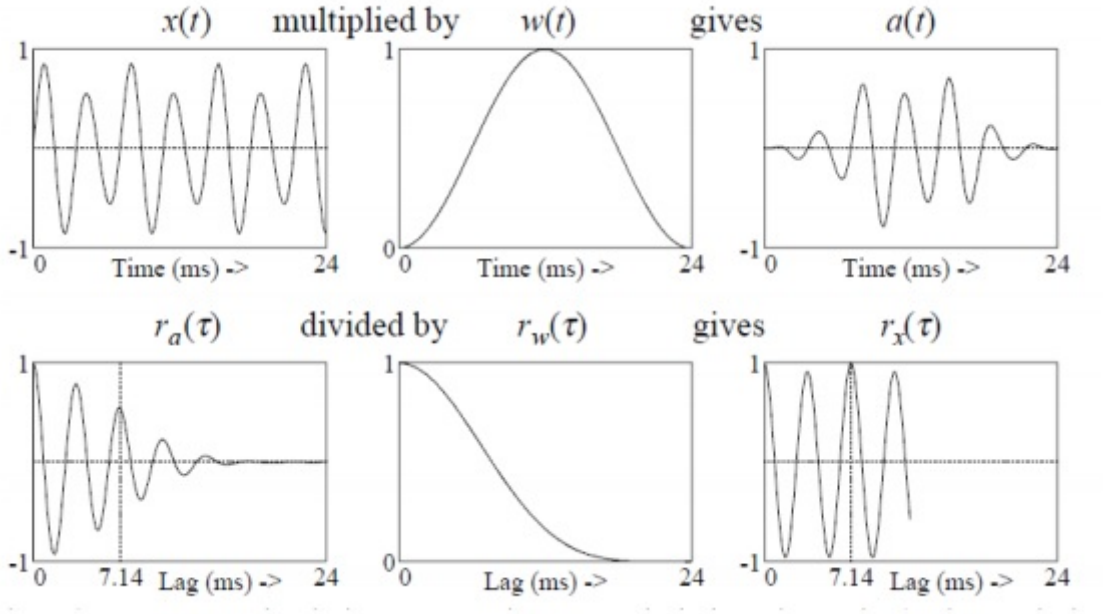

Figura 4.6 Definição da $f0$ na voz cantada [Murphy, 2008]

Este método será melhor discutido na seção posterior, uma vez que foi implementado neste trabalho.

4.1.2.3 Algoritmo de Yin

Este método que é utilizado em diversas aplicações também se baseia no método de Autocorrelação porém, tenta melhorar principalmente o desempenho através da utilização de outros passos [De Cheveigné and Kawahara, 2002]. Desses novos passos, dois são destacados pelo autor como sendo diferenciais diante dos outros para potencializar o desempenho: Uso de uma função cumulative mean normalized difference em vez de uma simples função de diferenças; Execução de uma interpolação parabólica de modo a aumentar a precisão.

Com os dados extraídos de [Kedem, 1986], a tabela 3.1 abaixo mostra os erros de cada passo desenvolvido no algoritmo Yin.

Tabela 4.1 Percentual de erro por passo de implementação do algoritmo Yin

Passo do Algoritmo	**Erro (%)**
Autocorrelação	10,0
Função das diferenças	1,95
Cumulative mean normalized difference	1,69
Absolute threshold	0,78
Interpolação parabólica	0,77
Estimação do melhor valor	0,5

4.1.3 Análise Espectral

Especialmente em se tratando da voz cantada, o domínio das frequências carrega muitas informações importantes no que diz respeito a determinação da frequência fundamental. Composto por uma série de harmônicos e parciais, um sinal de voz cantado pode ter a sua $f0$ estimada a partir da análise dessas partes.

Abaixo, estão relacionados três métodos para tal abordagem.

4.1.3.1 Cepstrum

A análise Cepstral, ou Cepstrum (trocando a ordem das letras de "Spectrum"), tem origem em 1963 [Bogert et al., 1963] e tem sua definição como sendo a transformada inversa de Fourier do logaritmo do espectro. As propriedades matemáticas conferidas pela aplicação da transformada de Fourier e do logaritmo, permitem aos especialistas trabalharem com o sinal do trato vocal e da glote separadamente, o que facilita a identificação de anomalias nas pregas vocais [Murphy, 2008].

Sabemos que, pelo modelo fonte-filtro adotado, a fonação pode ser dividida em três principais partes, mostradas na figura 3.7:

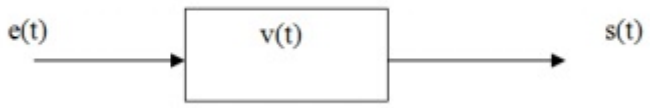

Figura 4.7 Modelo fonte-filtro.

O sinal de voz resultante $s(t)$ é fruto da aplicação da excitação $e(t)$ no trato vocal, com função resposta $v(t)$. Ou seja,

$$s(t) = e(t) * v(t) \tag{4.2}$$

A equação acima corresponde a uma convolução.

Passando isso para o domínio da frequência, temos a representação das transformadas de Fourier dadas por:

$$S(w) = E(w) * V(w) \tag{4.3}$$

Como $E(w)$ e $V(w)$ são combinados multiplicativamente, é possível separá-los utilizando funções logarítmica:

$$\log[S(w)] = \log[E(w)] + \log[V(w)] \tag{4.4}$$

Isso mostra um caminho que é utilizado para a separação do espectro final resultante.

Depois de separados, se utilizarmos uma nova DFT (Discrete Fourier Transformation) e aplicarmos um filtro passa-baixas, o espectro resultante será apenas com características harmônicas devido ao filtro do sinal original que, no caso da voz, vai ser o trato vocal. Neste espectro logarítmico, a componente de período levando em conta um som emitido (vocal) num intervalo de frequências inverso ao período fundamental, aparece no Cepstro na forma de pico.

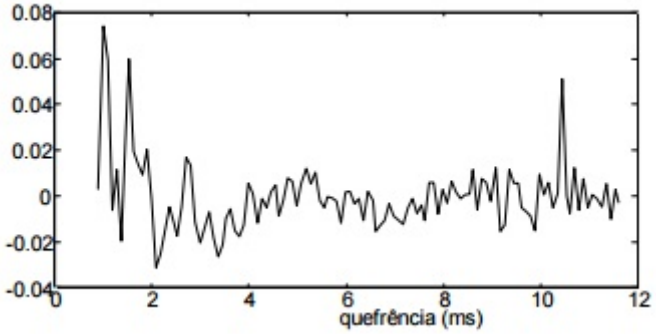

Figura 4.8 Cepstro de um segmento de fala [Teixeira, 1995]

A figura 3.8 acima mostra o Cepstro de um segmento de voz. Na função cepstral, o eixo das abcissas é chamado de quefrequências. As componentes do período fundamental aparecem com valores de quefrequência mais altos.

O conjunto de valores de saída da transformada inversa de Fourier, que são os valores Cepstrais discretos, formam o Cepstro. Assim, se aplicarmos um "lifter" e uma função janela retangular mais gradual, como indicado por [Teixeira, 1995] e tomarmos a transformada inversa discreta de Fourier do sinal resultante, teremos uma versão "alisada" do espectro logarítmico do filtro do trato vocal, formando o envelope espectral, mostrado na figura 3.9 abaixo.

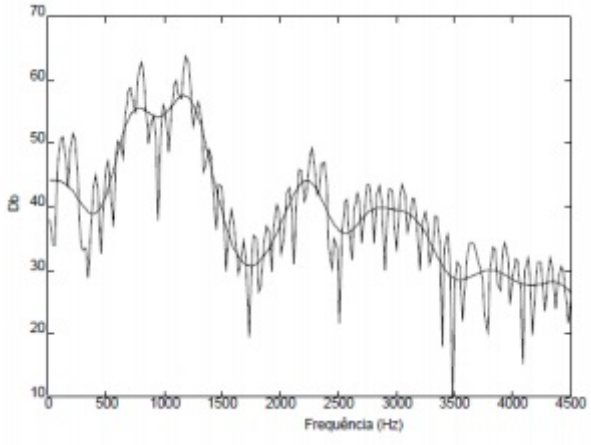

Figura 4.9 .
]Espectro e envelope cepstral de um segmento de fala da vogal [a] [Teixeira, 1995]

Finalmente, a partir do envelope cepstral, é possível obter a as frequências (inclusive a fundamenta), largura de banda e amplitude das formantes.

4.1.3.2 Search Tonal

Este método está entre os mais recentes e também entre os mais robustos por ter sido construído com este objetivo. A figura 3.10 abaixo mostra os passos do algoritmo Search Tonal.

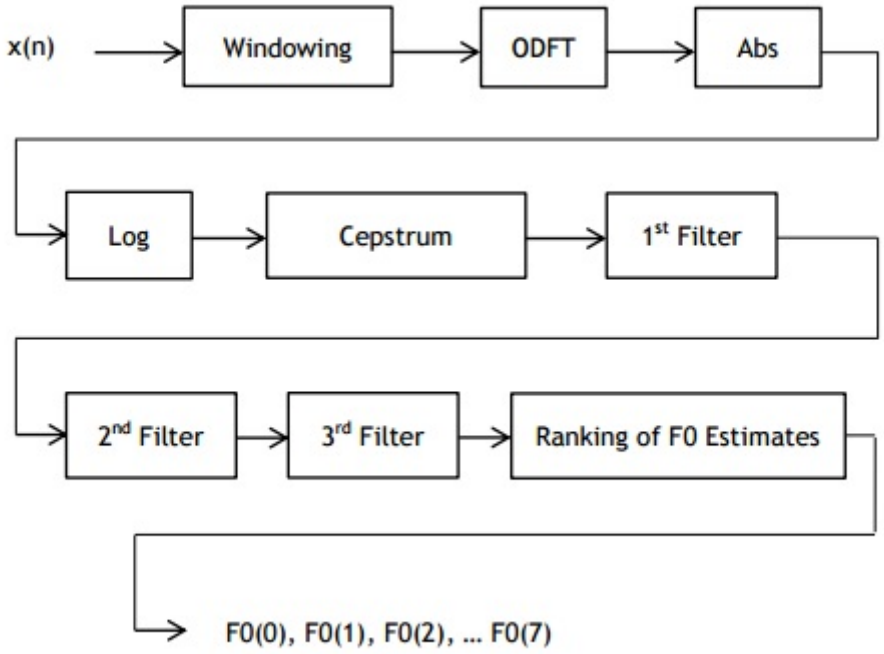

Figura 4.10 Passos do algoritmo Search Tonal [dos Santos Ventura, 2011]

Além destes passos descritos no diagrama, após estimar os 8 valores prováveis da frequência fundamental, o algoritmo implementa um método de seleção de valor, baseado na análise Cepstral, uma vez que leva em conta não apenas o valor de F0 mas também as parciais harmônicas [dos Santos Ventura, 2011].

4.1.3.3 Component Frequency Ratios

Desenvolvido por Martin Piszczalski em 1979, este método é considerado um dos pioneiros para análise vocal no domínio das frequências [Piszczalski and Galler, 1979]. Consiste, basicamente, em aplicar uma transformada ao sinal original de maneira que este seja visualizado no domínio das frequências. Feito isso, os máximos locais são estimados por meio de um método de detecção de picos. Para cada um destes parciais encontrados, o algoritmo estima o menor número harmônico possível que teria relação a uma série harmónica que contivesse esses par-

ciais. Após isso, os resultados seriam pesados e avaliados para que a f0 seja encontrada. O peso de cada parcial é diretamente proporcional a amplitude destes.

Uma das vantagens mais evidentes deste método é que, mesmo que a frequência fundamental não estivesse contida no sinal, ela seria detectada, bastando para isso a existência no sinal de pares de harmónicos suficientes. Já existem diversas versões mais aprimoradas desse método mas a ideia principal permanece a mesma.

4.2 Parâmetros Acústicos

4.2.1 Pitch

Podemos definir o pitch como a sensação que se ouve em relação à voz cantada. Exatamente a nota que o ouvido humano identifica e que traz a sensação de grave, médio ou agudo. Este parâmetro está completamente ligado a frequência fundamental, uma vez que se caracteriza como a sensação audível desta frequencia mas, é importante salientar que pitch e frequência fundamental são coisas diferentes.

A equação 3.5 abaixo relaciona a frequência fundamental com uma nota musical (levando em conta uma escala onde a nota LÁ é considerada 440 Hz).

$$f = 2^{\frac{n}{12}} * 440 \tag{4.5}$$

Na equação, n equivale ao numero de intervalos entre notas. A constante 440Hz, como explicado anteriormente, representa a nota LA, numa escala mais largamente utilizada.

4.2.2 Jitter e Shimmer

O jitter e o shimmer são parâmetros acústicos associados a variação da frequência fundamental ciclo-a-ciclo: O jitter é propriamente a variação da $f0$ e o shimmer é a variação da amplitude associada, ambos considerados ciclo-a-ciclo e, portanto, chamadas de medidas de curto termo [de Krom, 1993]. Em outras palavras, o jitter é uma medida percentual de irregularidade na nota vocal (perturbação da $f0$) e o shimmer é uma medida percentual de irregularidade na amplitude da nota vocal. Esses parâmetros são largamente utilizados para a detecção de patologias associadas a fonação e, segundo [Guimarães, 2007], a variação entre 0,5% e 1% é considerada normal para um jovem adulto, em uma nota sustentada.

Segundo Boersma e Kattharne, o Jitter pode ser dividido em quatro principais subcategorias:

4.2.2.1 Jitter absoluto

Representa a medida absoluta da diferença da frequência fundamental entre dois períodos seguidos. É chamado de *jitta* entre os profissionais de saúde no campo da voz.

$$jitta = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i-1}| \tag{4.6}$$

4.2.2.2 Jitter local

Calculado pela diferença média absoluta entre a frequência de dois períodos consecutivos, dividida pelo período médio. Esse parâmetro é chamado de *jitt*.

$$jitt = \frac{jitta}{\frac{1}{N}\sum_{i=1}^{N} T_i} * 100 \tag{4.7}$$

4.2.2.3 Jitter rap

Tem o objetivo de representar a média relativa de perturbação. A sigla rap vem do inglês *Relative Average Perturbation*. A média da diferença absoluta em um período e a média desse período com o período anterior e o posterior, dividido pelo período médio.

$$rap = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - (\frac{1}{3}\sum_{n=i-1}^{i+1} T_n)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} * 100 \tag{4.8}$$

4.2.2.4 Jitter ppq5

É o quociente de perturbação num período de cinco pontos, ou seja, a média da diferença absoluta entre um período e a média desse período com os dois períodos anteriores e os dois posteriores.

$$ppq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2} |T_i - (\frac{1}{5}\sum_{n=i-2}^{i+2} T_n)|}{\frac{1}{N}\sum_{i=1}^{N} T_i} * 100 \tag{4.9}$$

Já o shimmer, também de acordo com [Boersma, 2009] e [Murphy, 2008], pode ser dividido em:

4.2.2.5 Shimmer (dB)

é uma variação da amplitude do sinal, pico-a-pico. Pode ser calculado através da diferença do logaritmo na base 10 das amplitudes em dois períodos consecutivos. Tem como unidade de medida decibéis (dB).

$$ShdB = \frac{1}{N=1}\sum_{i=1}^{N-1} |20 * \log\frac{A_{i+1}}{A_i}| \tag{4.10}$$

4.2.2.6 Shimmer local

É a diferença absoluta da amplitude de dois períodos consecutivos, dividida pela amplitude média, expressa em porcentagem.

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - Ai - 1|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{4.11}$$

4.2.2.7 Shimmer apq3

Quociente de perturbação de amplitude de três pontos, calculado como a média da diferença absoluta de amplitude de um período e a média de diferença absoluta de amplitude dos períodos anterior e posterior a esse, dividida pela amplitude média.

$$apq3 = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - (\frac{1}{3}\sum_{n=i-2}^{i+2}A_n)|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{4.12}$$

4.2.2.8 Shimmer apq5

Quociente de perturbação de amplitude de cinco pontos, calculado como a média da diferença absoluta de amplitude de um período e a média de diferença absoluta de amplitude dos períodos dois anteriores e dois posteriores a esse, dividida pela amplitude média.

$$apq5 = \frac{\frac{1}{N-1}\sum_{i=2}^{N-2}|A_i - (\frac{1}{5}\sum_{n=i-2}^{i+2}A_n)|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{4.13}$$

4.2.3 Harmonics-to-Noise Ratio (HNR)

O parâmetro HNR determina uma identificação do período total do sinal de voz através da quantificação da taxa entre os componentes periódicos e não periódicos. Os componentes periódicos são a parte harmônica do sinal de voz, enquanto que os componentes não periódicos são os ruídos no sinal. É importante salientar que o HNR é um parâmetro geral do sinal e não está apenas ligado a frequência. Ainda, a variação deste parâmetro de indivíduo para indivíduo se dá por conta da diferença entre os tratos vocais, que conferem diferentes amplitudes para os harmônicos emitidos por cada um [Teixeira and Fernandes, 2014].

O HNR é calculado pela fórmula abaixo:

$$HNR = 10 * \log \frac{AC_v(T)}{AC_v(0) - AC_v(T)} \tag{4.14}$$

4.2.4 Short Time Energy

O short time energy é a amplitude de um sinal de voz em um determinado período.

4.2.5 Centróide Espectral

Este parâmetro representa o centro de gravidade de um espectro. No ramo musical, este parâmetro está associado ao aspecto de brilho do som, conferido pelas altas frequências. Quando esse parâmetro possui valores altos, significa que há um reforço do som na região de altas frequências.

CAPÍTULO 5

Metodologia e Experimento

Diversos trabalhos na área de análise da voz cantada partem de uma abordagem bottom-up para buscar especificamente por uma característica ou conceito a partir de um sinal de voz. Entretanto, este trabalho propôs utilizar uma metodologia top-down, utilizando um banco de dados exclusivo com sinais de voz cantada e analisando diversas técnicas e parâmetros.

A metodologia empregada foi dividida em 5 principais etapas.

5.1 Etapa 1: Estado da arte e embasamento teórico

Nesta etapa, foram estudados técnicas e métodos para análise da voz com foco na voz cantada. Com o objetivo de entender melhor como essas técnicas são utilizadas na prática, fez-se uma análise de plataformas e aplicações já existentes e consolidadas que utilizam a voz cantada com princípio. A compilação das informações e resultados desta etapa estão descritos em todas as seções anteriores deste trabalho.

5.2 Etapa 2: Criação do Banco de Dados

Para a realização deste trabalho, foi necessária a criação de um banco de dados específico que contemplasse as características vocais discutidas anteriormente para que os parâmetros fossem analisados. Levando em consideração os motivos discutidos nas seções anteriores, para este projeto, foram convidados cinco cantores líricos, com pelo menos 3 anos de estudo, vinculados ao curso de canto da Universidade Federal de Pernambuco. Isso garante um maior rigor técnico na produção dos áudios e uma melhor consistência na presença das características. As gravações foram realizadas numa perspectiva de home studio e sempre supervisionadas por um outro cantor para garantir que os áudios gravados estavam contemplando as características desejadas. Para a gravação, foram utilizados os seguintes equipamentos:

- Microfone: Rode NT1A Anniversary Vocal Condenser
- Computador: Intel® Core™ i7 6500U - 3.1 GHz 4 MB L3 Cache
- Software: Audacity

O perfil dos cantores utilizados neste experimento pode ser verificado na tabela 4.1 abaixo:

Tabela 5.1 Perfil dos cantores do experimento

Cantor	Classificação Vocal	Tempo de estudo (anos)	Atuação Profissional
Cantor 1	Tenor	9	Sim
Cantor 2	Tenor	5	Sim
Cantor 3	Barítono	7	Sim
Cantor 4	Mezzo-soprano	4	Sim
Cantor 5	Soprano	7	Sim

Vale salientar que todos os cantores eram declaradamente conhecedores de todas as características que gravaram e, ainda, que a identidade desses cantores, bem como os áudios, serão mantidos em sigilo e usados especificamente para fins desta pesquisa e não serão sob nenhuma hipótese divulgados.

5.3 Etapa 3: Escolha das Características

O trabalho realizado por [Loscos, 2007], apresenta uma série de algoritmos e métodos para análise de alguns aspectos da voz mais especificamente destinados ao processo de síntese e modelagem da voz cantada. Segundo o autor, algumas limitações e complexidades ainda comprometem o avanço rápido dessa área. Diante dessa complexidade e da exequibilidade, foram selecionadas duas características extremamente importantes que estão presentes na voz cantada: afinação e mudança de registro entre voz plena e falsete. Para a afinação, dois métodos foram implementados (um temporal e um spectral) e então, a eficiência e precisão destes serão comparadas e discutidas. Para a mudança de registro, o método adotado baseia-se no trabalho de [Murphy, 2008], que apresenta métodos estatísticos para a avaliação de características apresentadas em forma de binômio (por exemplo: soprosidade da voz. Voz que apresenta soprosidade versus Voz que não apresenta soprosidade). Sendo assim, vozes plenas e vozes em falsete cantando a mesma melodia foram avaliadas para identificação dos parâmetros acústicos que mais se relacionam com essa mudança de registro.

5.4 Etapa 4 - Implementação e Extração de Parâmetros

A plataforma utilizada para a implementação de todos os algoritmos foi o Matlab [Grant et al., 2008]. Além de ser muito aceita no meio acadêmico e também industrial, essa ferramenta permitiu de forma robusta e precisa a implementação dos algoritmos, além da facilidade por dispor de uma série de bibliotecas e funções (como, por exemplo, a Transformada de Fourier). Ainda, o MATLAB permite a exportação dos resultados desejados para tabelas no Excel e gráficos que facilita o processo de análise de resultados.

Também utilizou-se a ferramenta Scilab, que, apesar de não ter diversas funções como o MATLAB, foi uma alternativa por ser gratuita e disponpivel legalmente na internet.

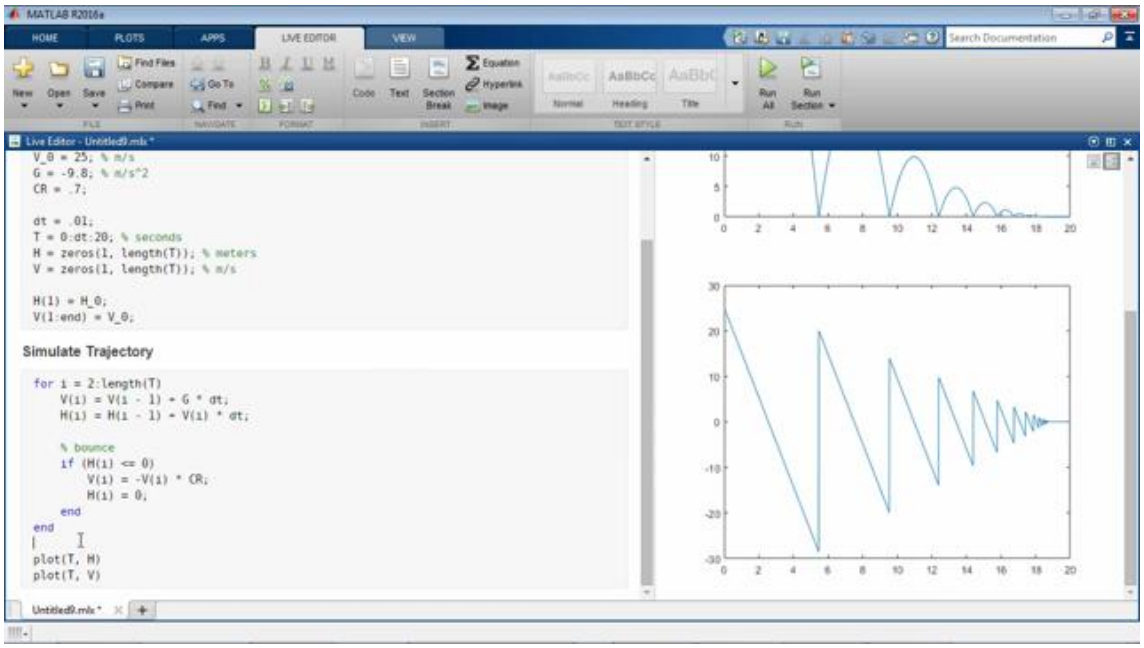

Figura 5.1 Screenshot do Matlab.

A fim de comparar alguns resultados, utilizou-se o software PRAAT, um software aberto desenvolvido na Universidade de Amsterdam, mais especificamente no Institute of Phonetic Sciences por Paul Boersma e David Weenink [Oguz et al., 2007].

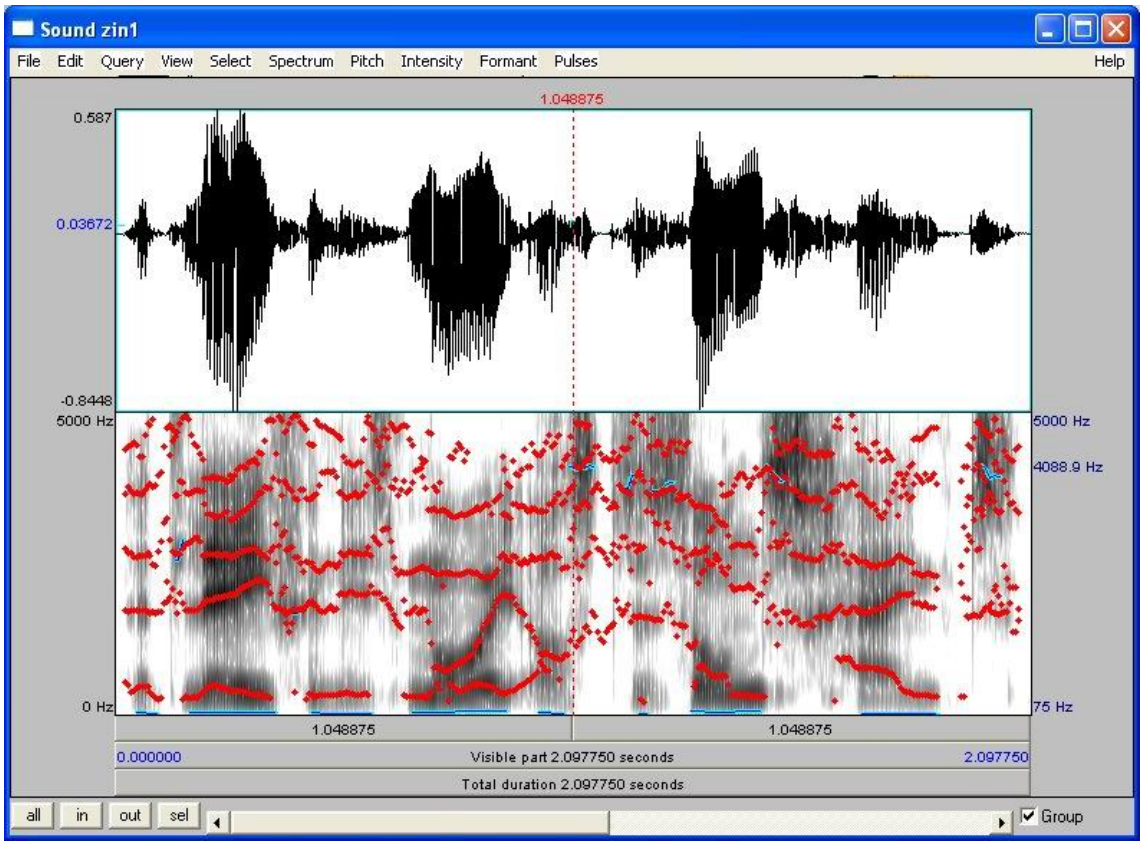

Figura 5.2 Screenshot do PRAAT.

Ainda, para cada um dos parâmetros selecionados, foram implementados métodos diferentes:

5.4.1 Afinação

Como a afinação está diretamente ligada com a nota ou pitch, uma vez que se considera afinação o acerto entre a nota cantada e a nota esperada dentro de uma melodia, para esta característica, foram desenvolvidos dois algoritmos de extração da frequência fundamental no MATLAB/Scilab. Por questões de exequibilidade e por diferenciar as abordagens (temporal e espectral), os algoritmos escolhidos foram Autocorrelação e Cepstrum.

Para ambos os métodos utilizou-se a função textitaudioread e também a função textitwaveread que recebem como parâmetro um arquivo de áudio na extensão WAV e retornam um vetor com os dados da amostra e um inteiro com a taxa de amostragem Fs.

A primeira parte do algoritmo chamada de **segmentação** divide o sinal em intervalos que contém uma nota. Essa divisão é feita graças ao envelope característico de uma nota musical na maior parte dos instrumentos, incluindo a voz, que, neste caso, e é a variação da amplitude de uma nota musical ao longo de sua duração. Este evenlope tem como característica uma subida rápida, depois um decaimento, como podemos observar na figura 4.3 abaixo:

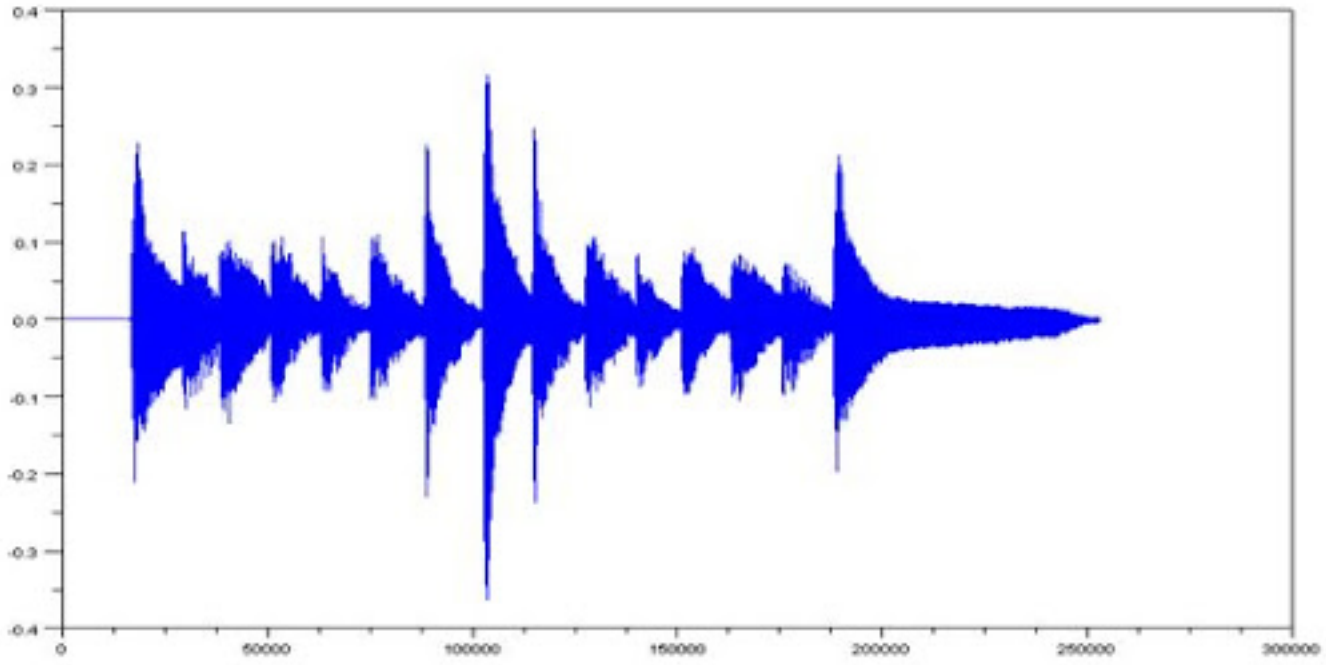

Figura 5.3 Gráfico da amplitude por amostra do sinal de aúdio com várias notas

Assim, é possível dividir esses segmentos de nota a partir desta característica, baseando-se na variação brusca da amplitude.

Este processo de detecção pode ser muito complicado se tomarmos o sinal tal como está plotado no gráfico da imagem 4.3, devido ao fato de que o envelope é modulado por ondas na frequência de sua própria nota musical então, se derivássemos esse sinal, só encontraríamos as variações de amplitude desta frequência e não os picos bruscos, como desejamos. Então, a partir do sinal original, outro sinal é gerado como se segue:

- O sinal original é processado, obtendo-se os valores absolutos da amplitude

- Este sinal de valores absolutos é processado, gerando outro sinal que contém em cada índice a integral discreta de uma janela cujo tamanho possa conter as menores frequências de moludação do envelope, a qual desloca ao longo do sinal.

O sinal resultante para o mesmo sinal mostrado na figura 4.3, está na figura 4.4 abaixo:

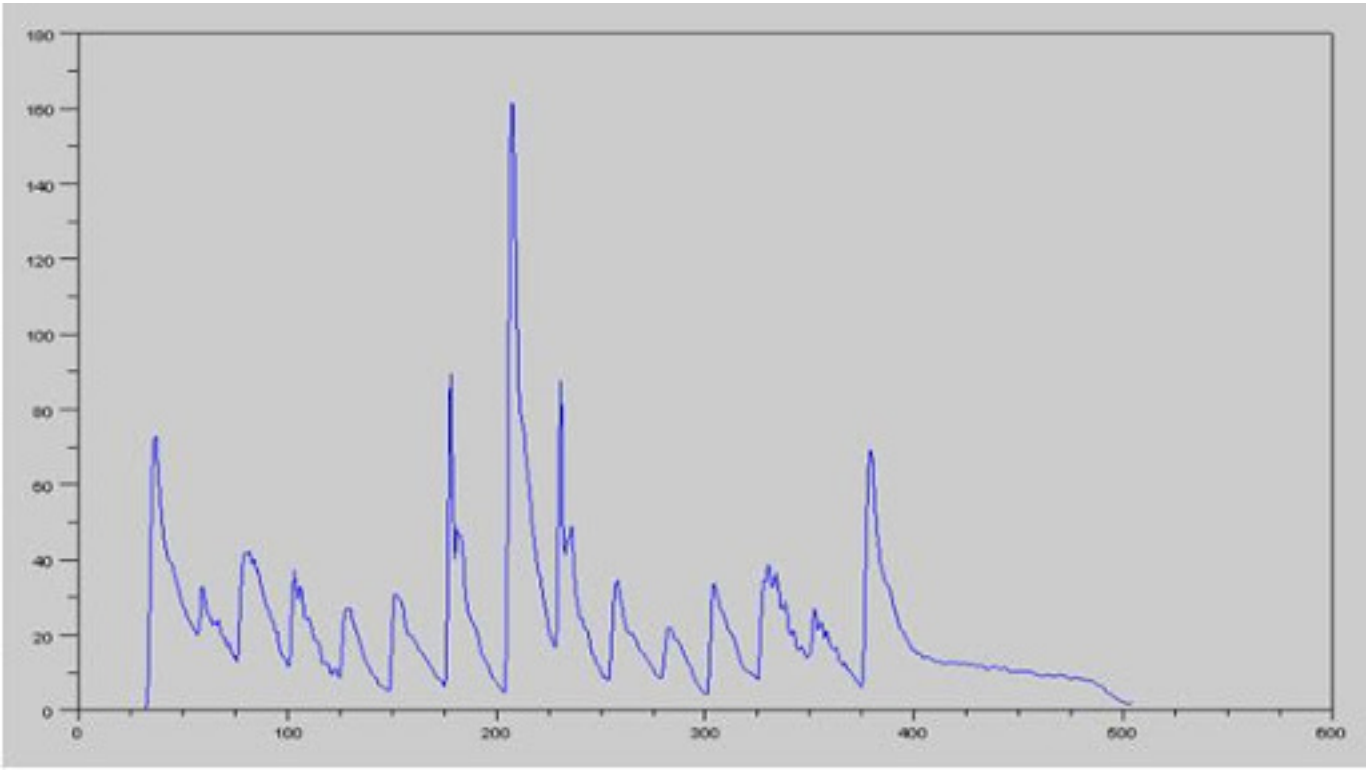

Figura 5.4 Sinal envelope do sinal da figura 4.3

Neste, as frequências de modulação são inexistentes. Uma derivação discreta deste sinal, gera picos que são exatamente os divisores de segmento do sinal, conforme [Teixeira, 1995].

Porém, os índices obtidos nesse processo para que sejam mapeados no sinal original, precisam ser multiplicados pelo intervalo de deslocamento da janela escolhido (no caso deste experimento, 500 amostras).

O processo de segmentação usa parte do conceito de janelas sobrepostas discutido na apresentação dos métodos de cálculo da frequência fundamental mas, o cálculo da integral discreta (que é, na realidade, um somatório) requer menos processamento de que analisar a frequência fundamental da janela. Assim, a análise de segmentos foi a melhor opção.

Agora, com os divisores de segmentos, é possível estimar a frequência fundamental. Os métodos utilizados são discutidos a seguir.

5.4.1.1 Autocorrelação

A função principal desta parte, é descrita na equação 3.1. Como o sinal de áudio está discretizado, cada valor de $R_x(\tau)$ consiste na soma de várias parcelas $x(t+\tau)x(t)$. Assim, os valores de τ dos picos de $R_x(\tau)$ são os períodos T da janela analisada. Então, extraindo-se o período fundamental, obtém-se a frequência fundamental, que é o inverso do período. A figura 4.5 abaixo mostra o sinal resultante da autocorrelação de um dos segmentos do sinal apresentado na figura 4.3.

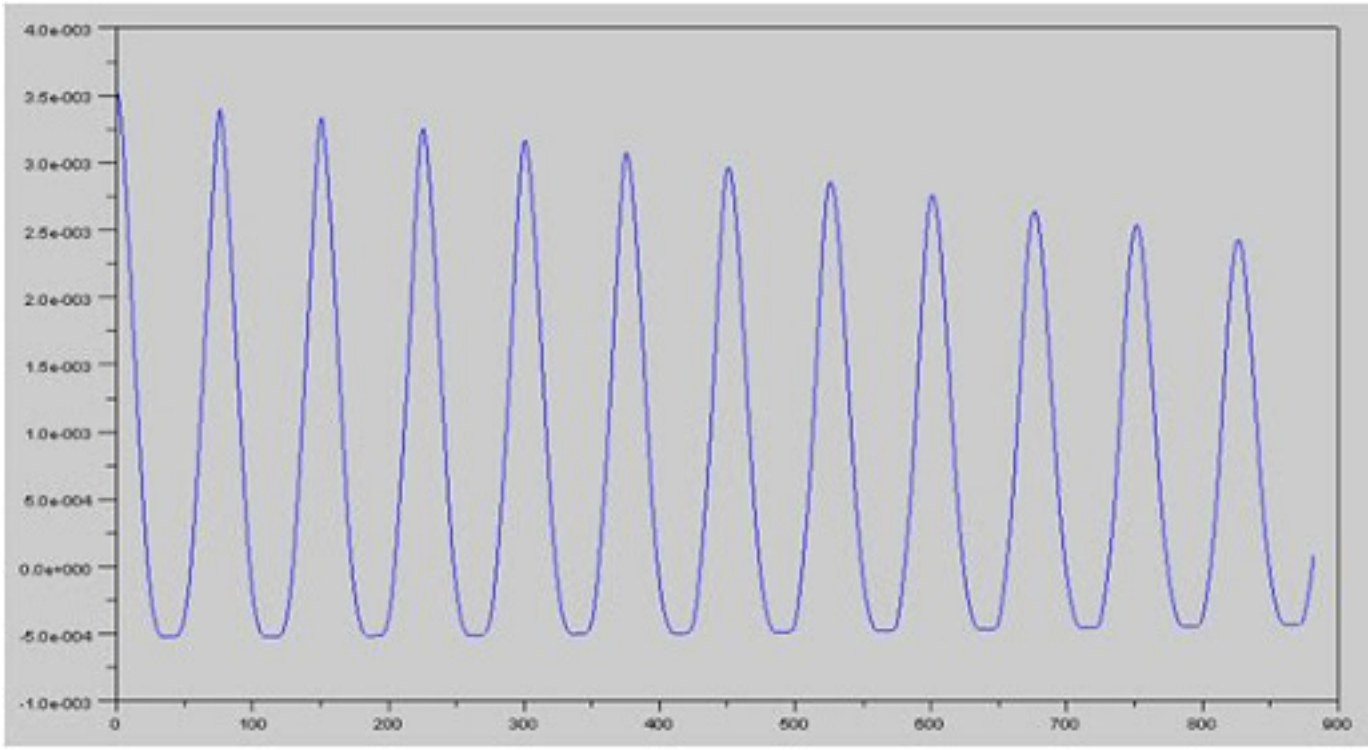

Figura 5.5 Sinal autocorrelação

O tamanho mínimo de uma janela deve ser o dobro do comprimento de onda da nota mais baixa da frequência [Teixeira, 1995]. Como utilizamos segmentação, esta preocupação foi eliminada, uma vez que o segmento pode contar todas as faixas.

5.4.1.2 Cepstrum

Esta implementação foi relativamente simples graças a quantidade de funções já programadas no MATLAB.

Uma vez que os segmentos já estavam definidos, os passos para calcular a frequência fundamental pelo Cepstrum foram os seguintes:

- Utilizou-se a função rcps do matlab, que determina o Cepstro
- Calculou-se o pico
- O periodo do mesmo foi transformado para ms
- calculou-se a freqeuência, que é o inverso do período

Uma vez estimada a frequência fundamental, a nota musical correspondente é encontrada comparando-se o valor da $f0$ com a tabela de referências da escala temperada. O valor referência foi o da nota Lá 3 (frequência 13,75 Hz).

Ainda, como um adicional, uma vez concluído o histograna das notas musicais encontradas, soma-se a quantidade de cada nota. Então, verifica-se qual a escala com a maior soma, sendo esta escala a tonalidade da melodia cantada. Verificando, apenas, a primeira nota da escala, que é a nota que nomeia a tonalidade.

5.4.2 Falsete X Voz Modal

O desejo de identificar essa característica analisando a voz cantada e obter marcadores para a transição entre voz modal e falsete veio do grande desafio que é a chamada nota de passagem para todos os cantores. A nota de passagem é a nota que marca a transição de registro de ressonância na voz cantada. Quanto mais experiente for o cantor, mais sutil será essa transição porém o processo de se conseguir essa sutileza é um dos mais difíceis principalmente para estudantes iniciantes [Echternach and Richter, 2012]. Como o registro de cabeça tem muitos princípios de falsete e esta é uma característica mais perceptível (principalmente para ouvidos não treinados), escolheu-se tentar identificar marcadores entre os parâmetros para essa transição da chamada voz modal e o falsete.

Baseando-se no trabalho desenvolvido em [Murphy, 2008] que usou métodos estatísticos para relacionar características de patologia da voz a paraêmtros acústicos, a partir da análise de vozes saudáveis e patológicas, este trabalho se propõe a utilizar os mesmos métodos porém aplicados a vozes com falsete e sem falsete.

Pela análise esperada na perturbação de frequência e harmônicos, os parâmetros desenvolvidos foram Jitter e Shimmer, bem como suas variações. Espera-se, com esses valores aplicados ao teste estatístico, criar um embasamento teórico formal que permita relacionar essa mudança de registro a esses parâmetros, servindo como base para investigações e trabalhos futuros.

O método estatístico escolhido se chama Teste Mann-Whitney e será explicado nesta subseção, juntamente com a implementação dos parâmetros acústicos extraídos do PRAAT.

5.4.2.1 Teste Mann-Whitney

Este teste é indicado para encontrar correlações entre duas populações diferentes e é indicado para amostras pequenas e/ou quando pressuposições para a análise da variância estão comprometidas [Birnbaum et al., 1956]. Logo, este teste se aplica perfeitamente a este trabalho pois será aplicado a uma amostra pequena de dados (aoenas jitter e shimmer de um trecho de audio de dois cantores - homem e mulher). No total, são 4 trechos de áudio (com falsete masculino e feminino e com voz modal masculino e feminino).

O procedimento para o teste é o seguinte:

- Formular a hipótese: Supõe-se que as duas amostras são provenientes da mesma população. Neste caso, supomos que a voz modal e falsetista são iguais em termos de jitter e shimmer.

- Colocar os dados (valores de jitter e shimmer) dos dois grupos (voz com falsete e modal) em ordem crescente. Se houverem valores repetidos, estes devem ser substituidos pela média dos postos.

- Considera-se n_1 o número de casos do grupo 1 e n_2 o número de casos do grupo 2. Entende-se por casos, o valores dos parâmetros.
- Calcula-se R_1 que é a soma dos postos do grupo 1 e R_2, que é a soma dos postos do grupo 2.
- Calcula-se as médias estatísticas com as fórmulas abaixo

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{5.1}$$

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{5.2}$$

- Escolhe-se o menor valor de U

Se o valor de U for menor ou igual aos valores contidos na tabela de Mann-Whitney, conforme [Birnbaum et al., 1956], a hipótese é descartada.

5.4.2.2 Jitter e Shimmer

Como o software PRRAT oferece um relatório do qual se pode extrair o Jitter e o Shimmer, foi configuraco para um range de frequência entre 50Hz e 600Hze então, as perturbações são extraídas.

CAPÍTULO 6

Resultados e Discussões

Os resultados apresentados neste capítulo são frutos do experimento, das referências e dos testes realizados, conforme descritos no capítulo anterior. Todas as informações relevantes, comentários e críticas serão discutidos, bem como gráficos, tabelas e outros artefatos serão apresentados.

6.1 Afinação

Os algoritmos de estimação da frequência fundamental testados neste módulo foram o Cepstrum e o de Autocorrelação.

6.1.1 Teste 1

- Arquivo: SB1MG.wav

- Cantor: Barítono

- Descrição: Sustentação da nota Lá3 com a vogal /a/.

- Objetivo: Analisar o comportamento do algoritmo em região grave, porém com harmônicos cheios. Por esse fato, escolheu-se o áudio gravado por uma voz grave.

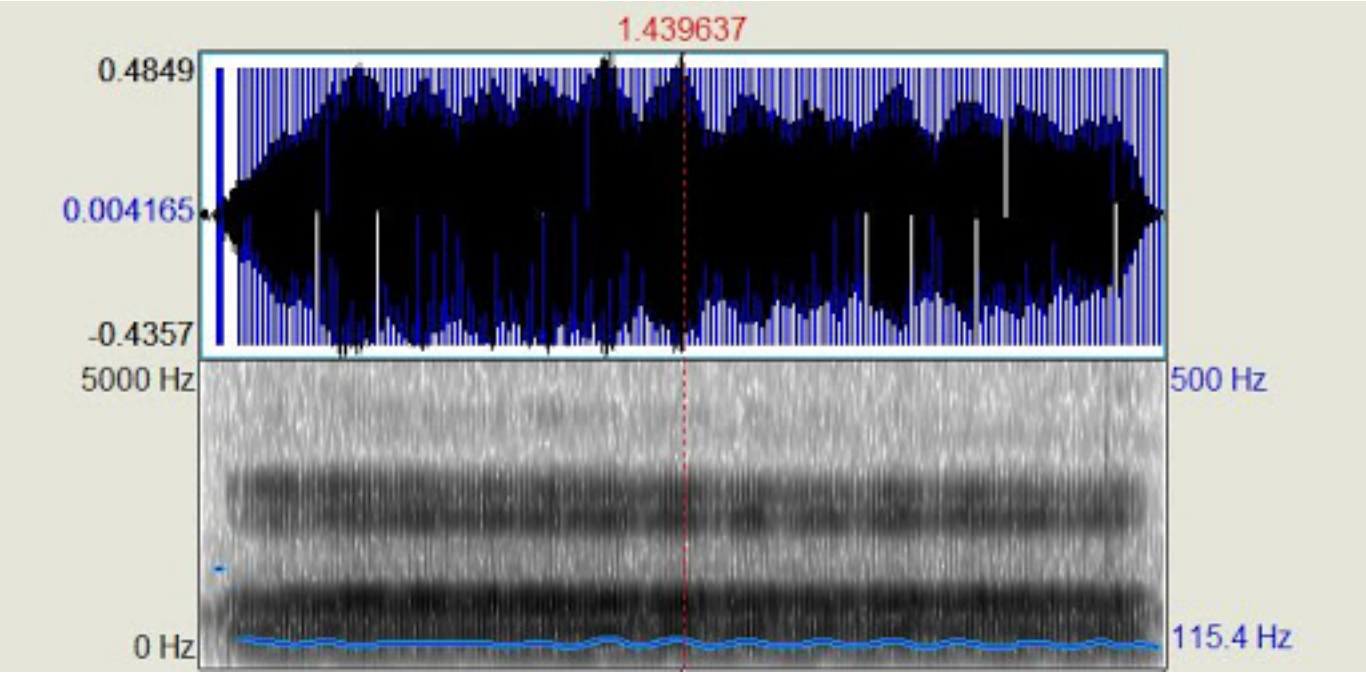

Figura 6.1 Espectrograma e frequência fundamental teste 1 pelo PRAAT

A imagem mais superior contida nesta figura, que mostra também a amplitude do sinal por amostra, pode ser comparada com a figura 5.2 gerada pelo algoritmo no MATLAB. Vê-se, claramente, uma semelhança, entre as mesmas.

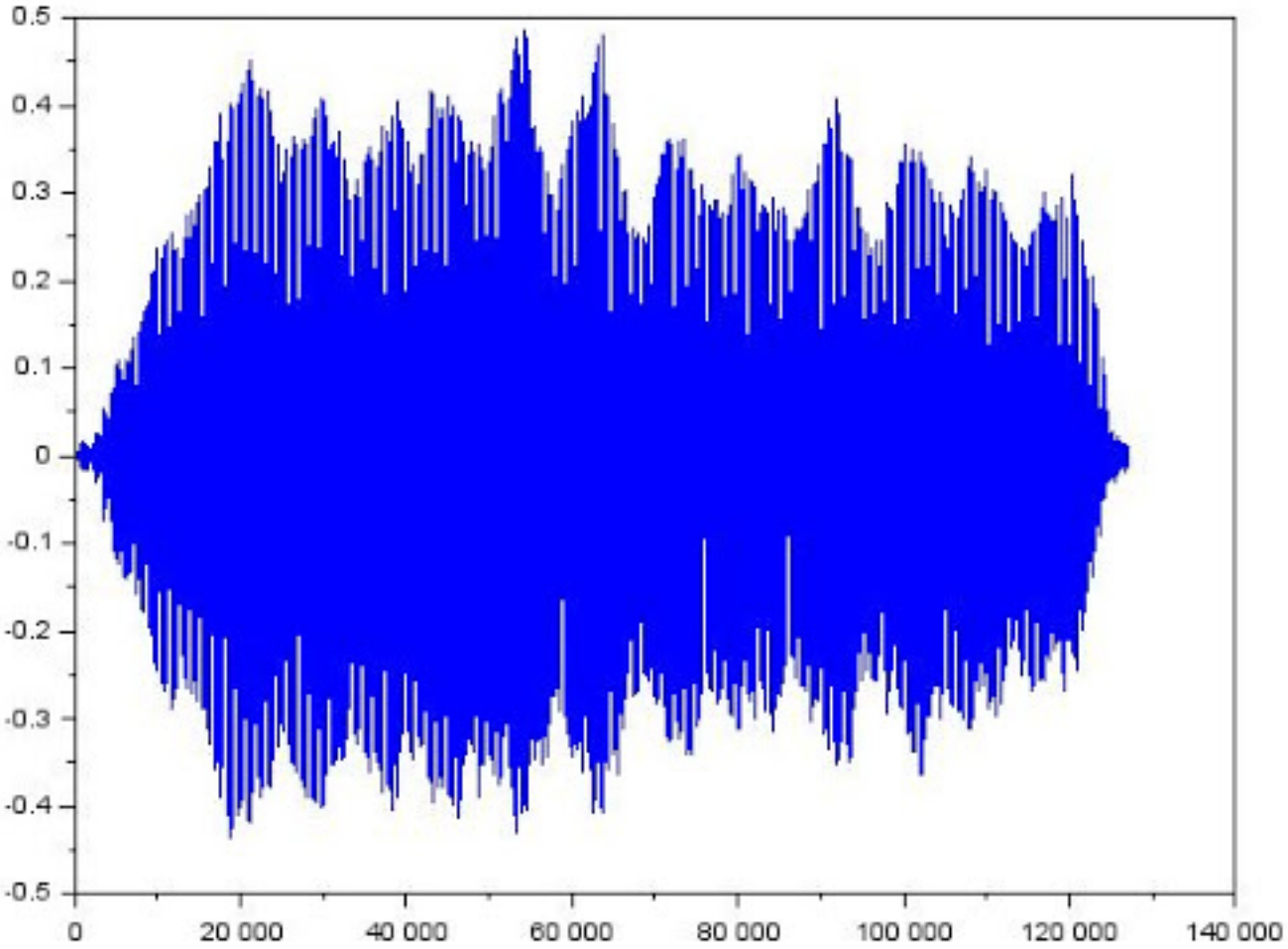

Figura 6.2 Amplitude por amostra de sinal, teste 1.

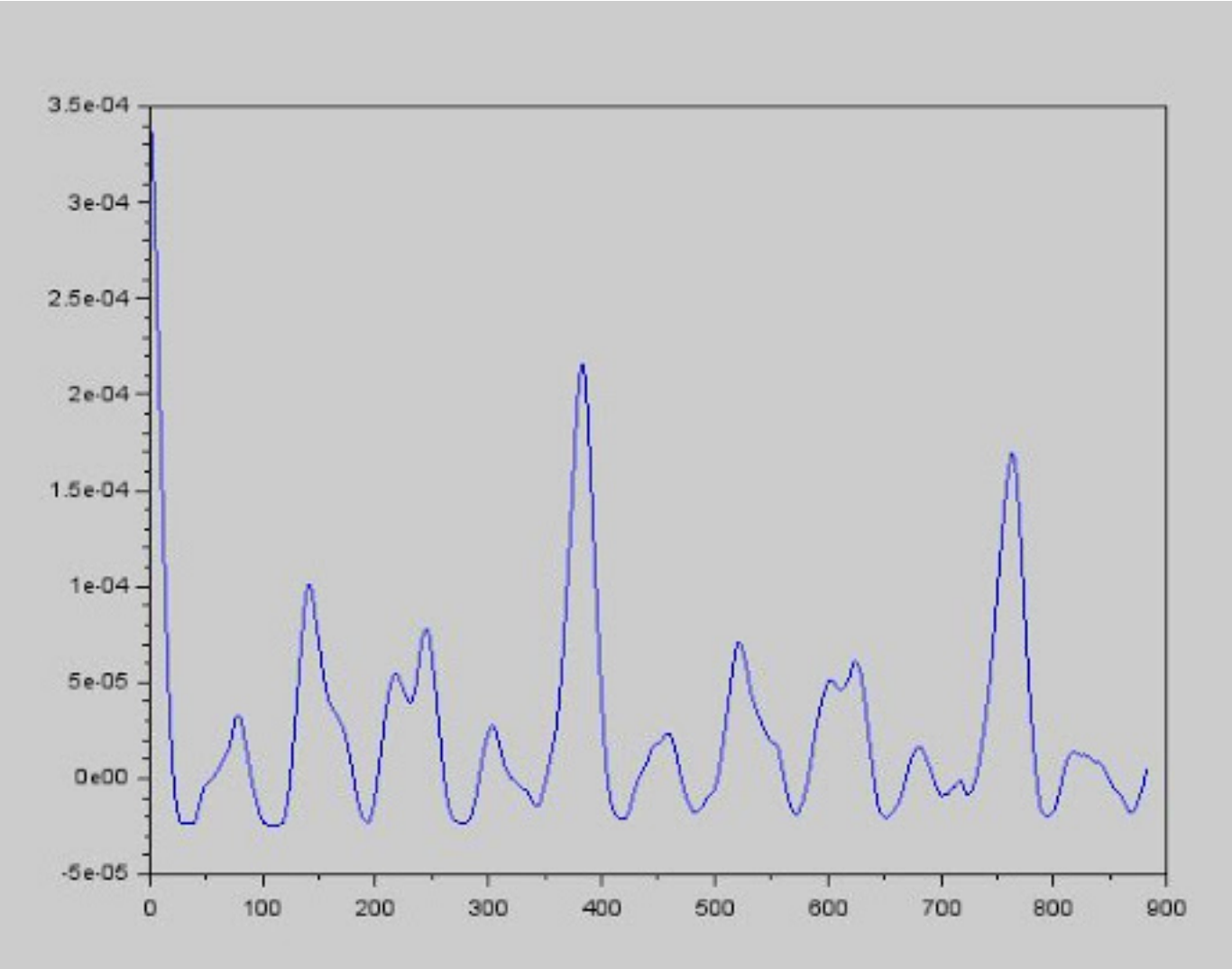

Figura 6.3 Primeira janela do teste 1 com Autocorrelação

A figura 5.3 acima mostra o sinal de autocorrelação no primeiro segmento detectado pelo algoritmo. Através deste sinal de autocorrelação, é possível obter a frequência por meio do período do pico.

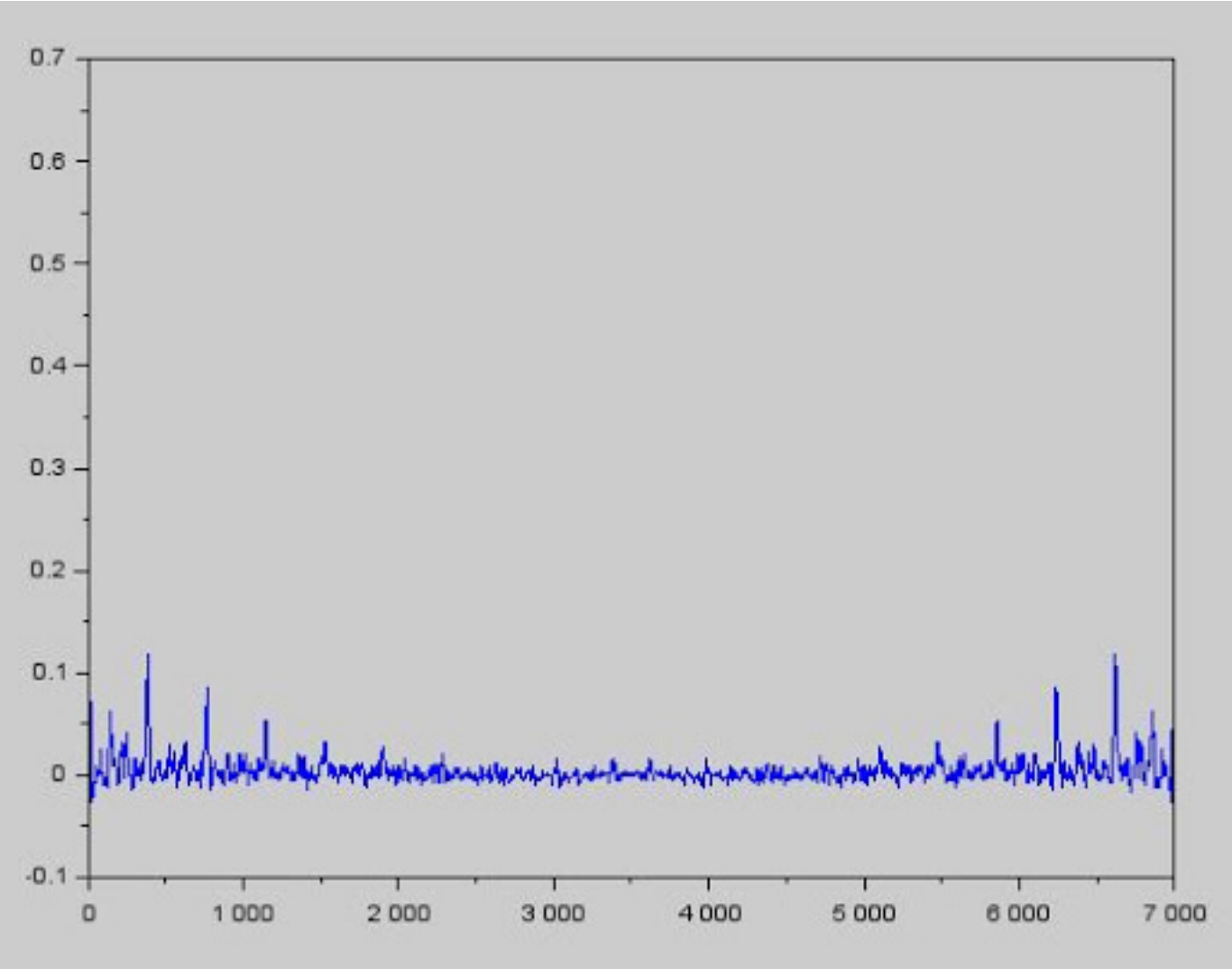

Figura 6.4 Primeira janela do teste 1 com Spectrum

Já a figura 5.4, mostra o mesmo segmento da figura 5.3, porém na abordagem Cepstrum.

A tabela 5.1 abaixo mostra as notas encontradas pelo algoritmo em cada segmento encontrado. É fácil perceber que o algoritmo se confunde por conta da presença do vibrato e do formante do cantor, características que empregam um deslocamento na frequência fundamentla e conferem mais harmônicos a voz. Com isso, em ambas as abordagens, as alterações na frequência fundamental causaram uma oscilação de meio tom.

É importante destacar ainda que, em ambos os casos, as notas encontradas foram exatamente as mesmas nesse teste e emissão de uma única nota.

Ainda, o algoritmo, apesar de acertar a nota, errou a oitava.

Tabela 6.1 Notas encontradas no Teste 1 de afinação

# Nota	Autocorrelação	Cepstrum
Nota 1	A#1	A#1
Nota 2	A1	A1
Nota 3	A1	A1
Nota 4	A1	A1
Nota 5	A#1	A#1
Nota 6	A#1	A#1
Nota 7	A1	A1
Nota 8	A1	A1
Nota 9	A1	A1
Nota 10	A1	A1

6.1.2 Teste 2

- Arquivo: SB1MG.wav

- Cantor: Tenor

- Descrição: Sustentação da nota Sol4 com a vogal lal.

- Objetivo: Analisar o comportamento do algoritmo em região aguda e cheia de harmônicos. Por isso, escolheu-se o áudio de um tenor, que é uma voz masculina aguda e, certamente, emite notas nessa região com mais brilho.

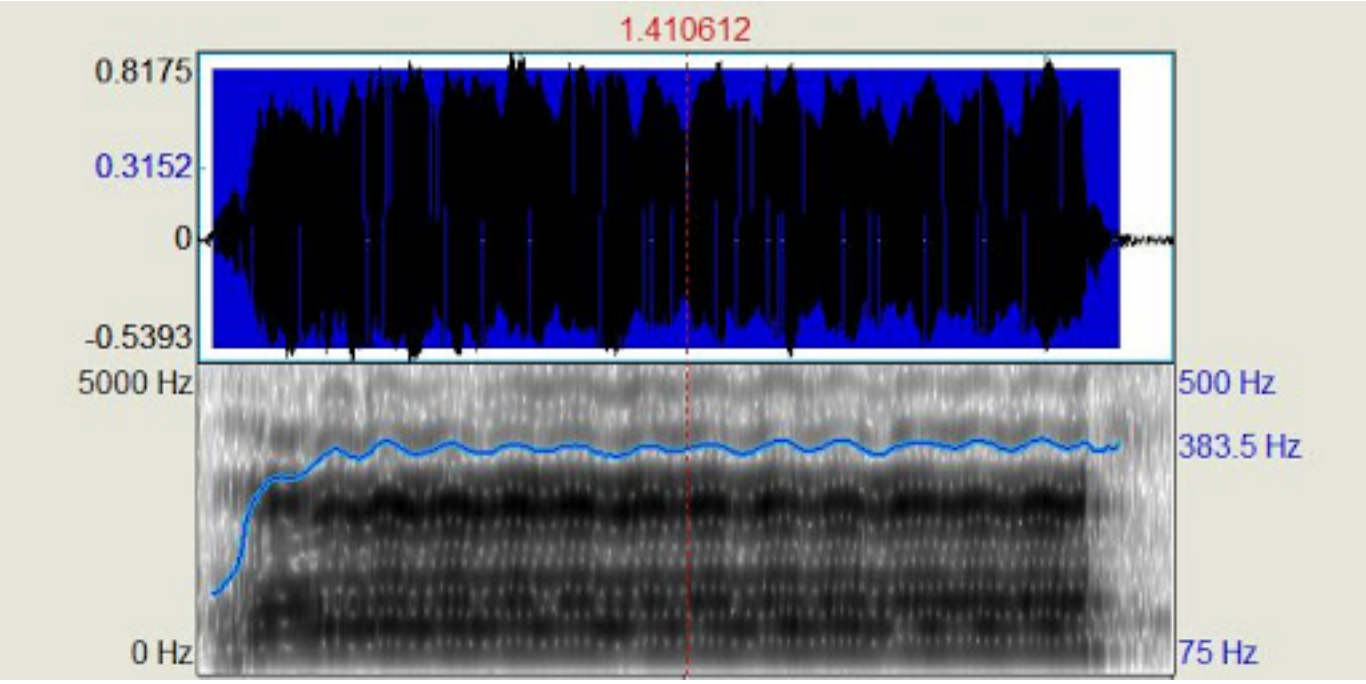

Figura 6.5 Espectrograma e frequência fundamental teste 2 pelo PRAAT

Mais uma vez, percebemos uma relação clara na demonstração da amplitude dos sinais por amostra, entre as figuras 5.5 (gerada pelo PRAAT) e 5.6 (gerada pelo MATLAB).

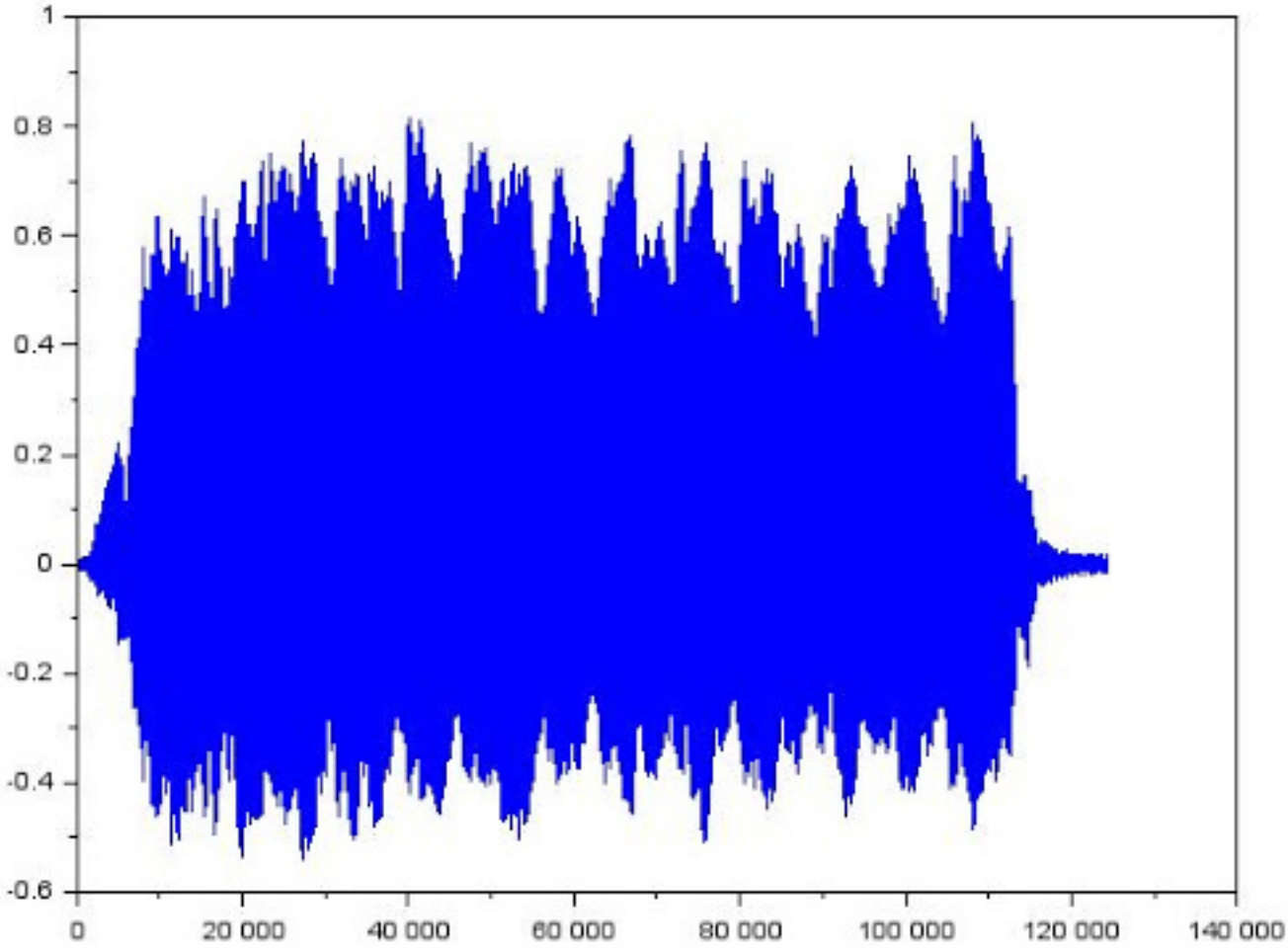

Figura 6.6 Amplitude por amostra de sinal, teste 2.

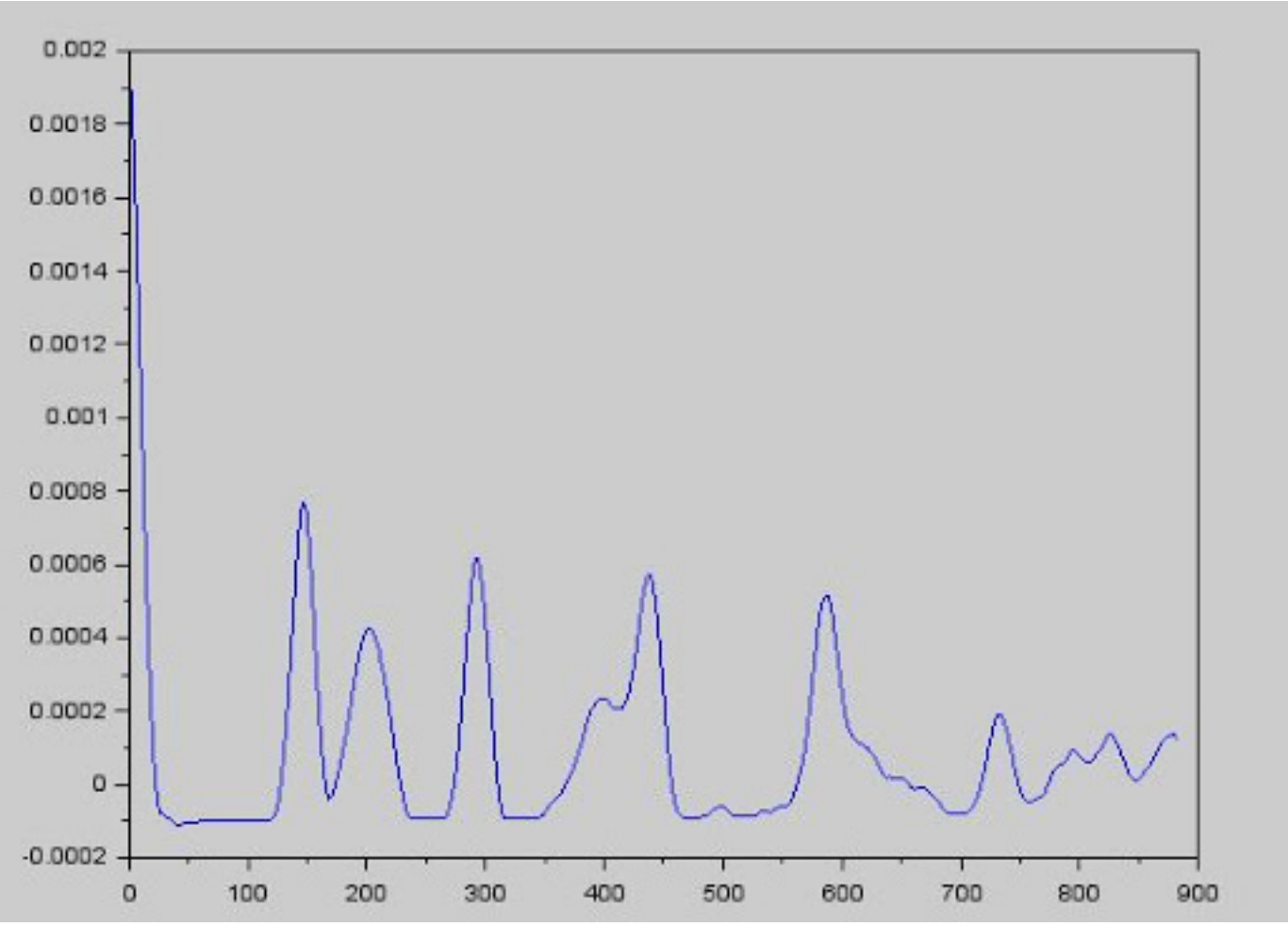

Figura 6.7 Primeira janela do teste 2 com Autocorrelação

A figura 5.7 acima mostra o sinal de autocorrelação no primeiro segmento detectado pelo algoritmo. Através deste sinal de autocorrelação, é possível obter a frequência por meio do período do pico.

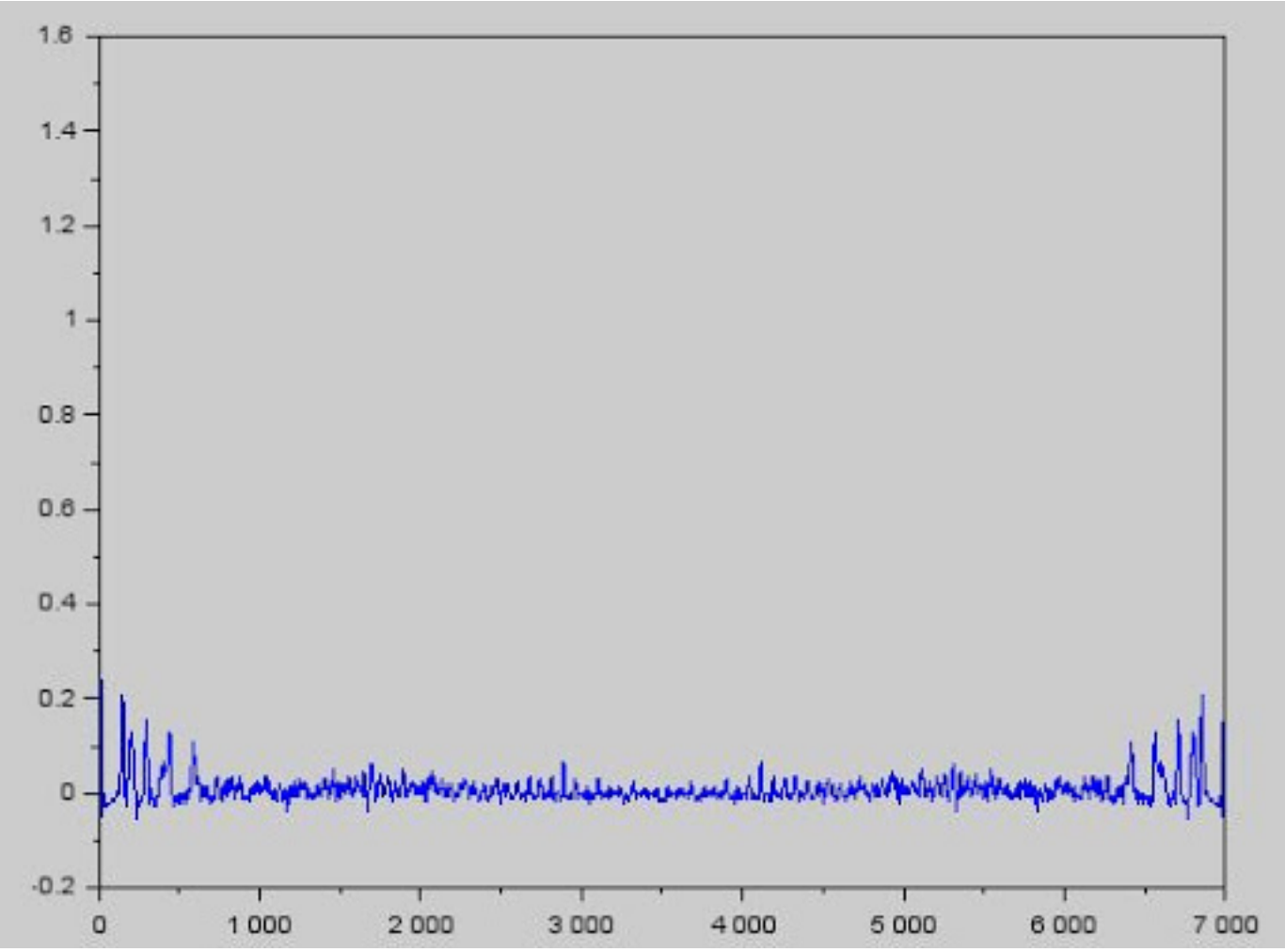

Figura 6.8 Primeira janela do teste 2 com Spectrum

Já a figura 5.8, mostra o mesmo segmento da figura 5.6, porém na abordagem Cepstrum.

A tabela 5.2 abaixo mostra as notas encontradas pelo algoritmo em cada segmento encontrado. Também é possível identificar a não sensibilidade do algoritmo ao vibrato, o que faz a frequência variar. Porém, neste caso de uma região aguda com harmônicos de mais alta frequência presentes de maneira abundante, percebemos divergências entre os métodos. Muito pela presença desses harmônicos mais fortes, a abordagem Ceptrum infere diferente o valor da $f0$ para algumas notas. Porém, esta diferença é muito pequena e, por exemplo, poderia ser dissipada com um módulo específico para lidar com o vibrato.

Outro ponto importante a se observar é que no início do áudio o cantor fez um portamento, detectado com sucesso por ambos Cepstrum e Autocorrelação.

Novamente, o algoritmo, apesar de acertar a nota, errou a oitava.

Tabela 6.2 Notas encontradas no Teste 2 de afinação

# Nota	Autocorrelação	Cepstrum
Nota 1	D3	D3
Nota 2	F3	E3
Nota 3	F#3	F#3
Nota 4	F#3	F#3
Nota 5	F#3	F#3
Nota 6	F#3	G3
Nota 7	G3	G3
Nota 8	G3	G3

6.1.3 Teste 3

- Arquivo: SB1FG.wav
- Cantor: Mezzo-soprano
- Descrição: Sustentação da nota E4 com a vogal lal.
- Objetivo: Analisar o comportamento do algoritmo em região grave com voz feminina. Por questões de tessitura, escolheu-se o áudio da mezzo-soprano.

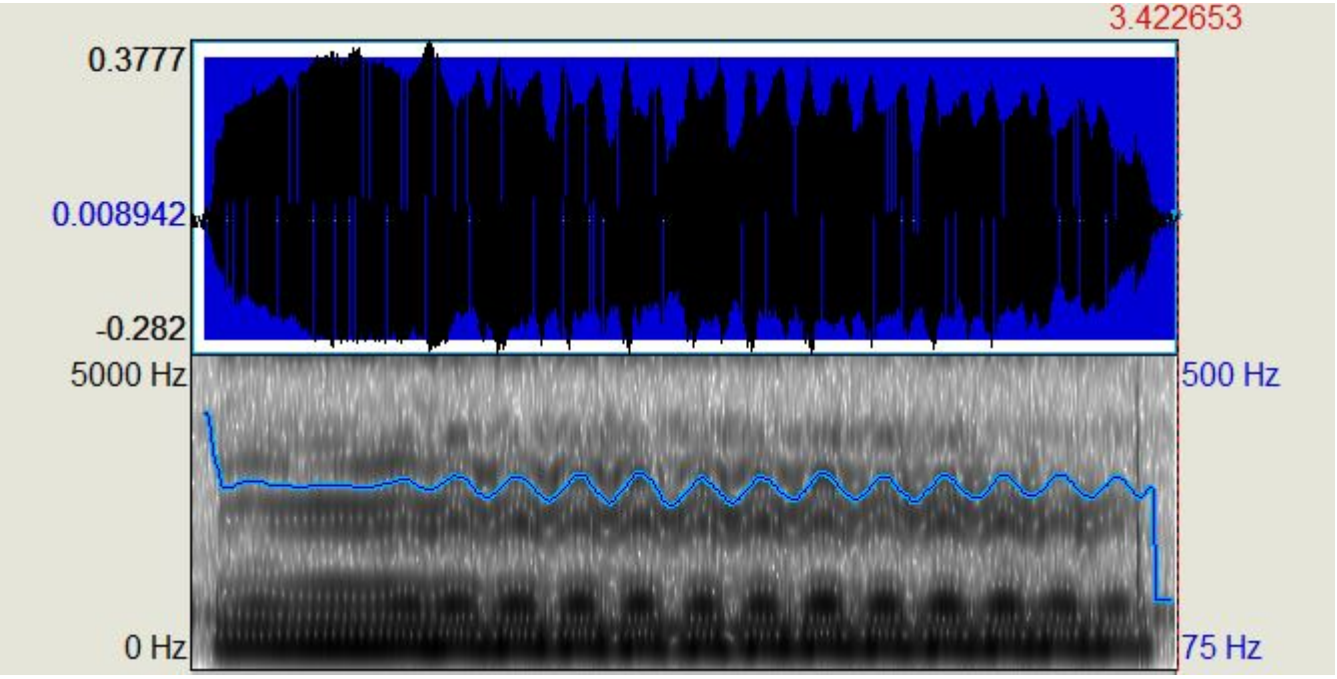

Figura 6.9 Espectrograma e frequência fundamental teste 3 pelo PRAAT

Mais uma vez, percebemos uma relação clara na demonstração da amplitude dos sinais por amostra, entre as figuras 5.9 (gerada pelo PRAAT) e 5.10 (gerada pelo MATLAB).

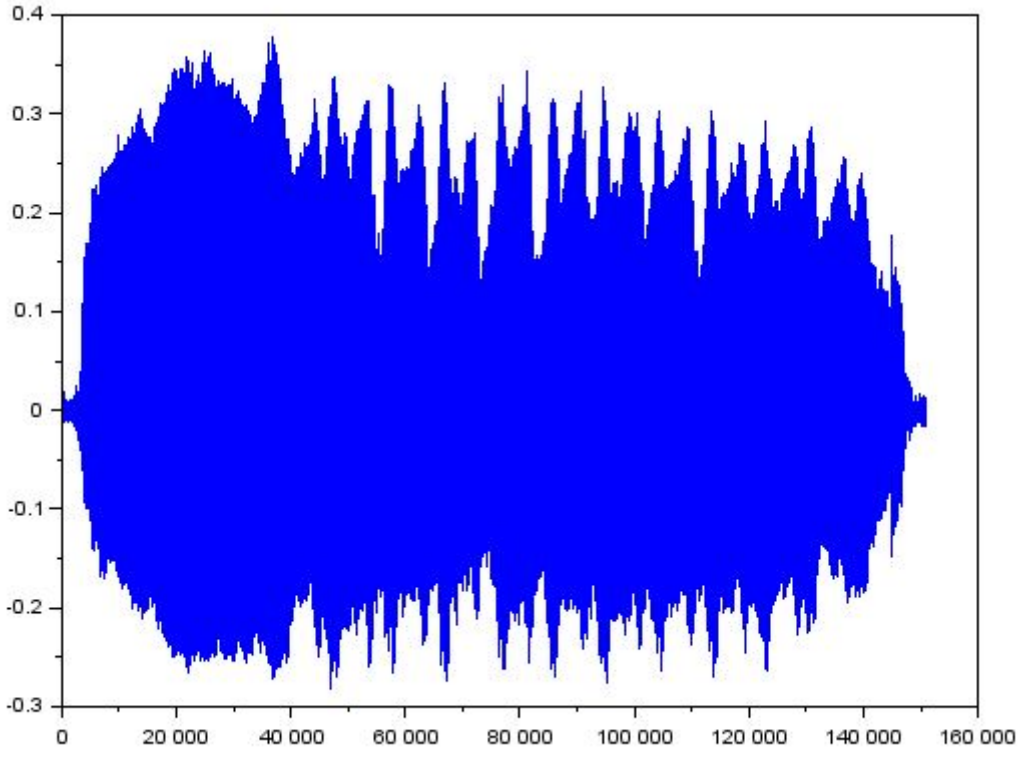

Figura 6.10 Amplitude por amostra de sinal, teste 3.

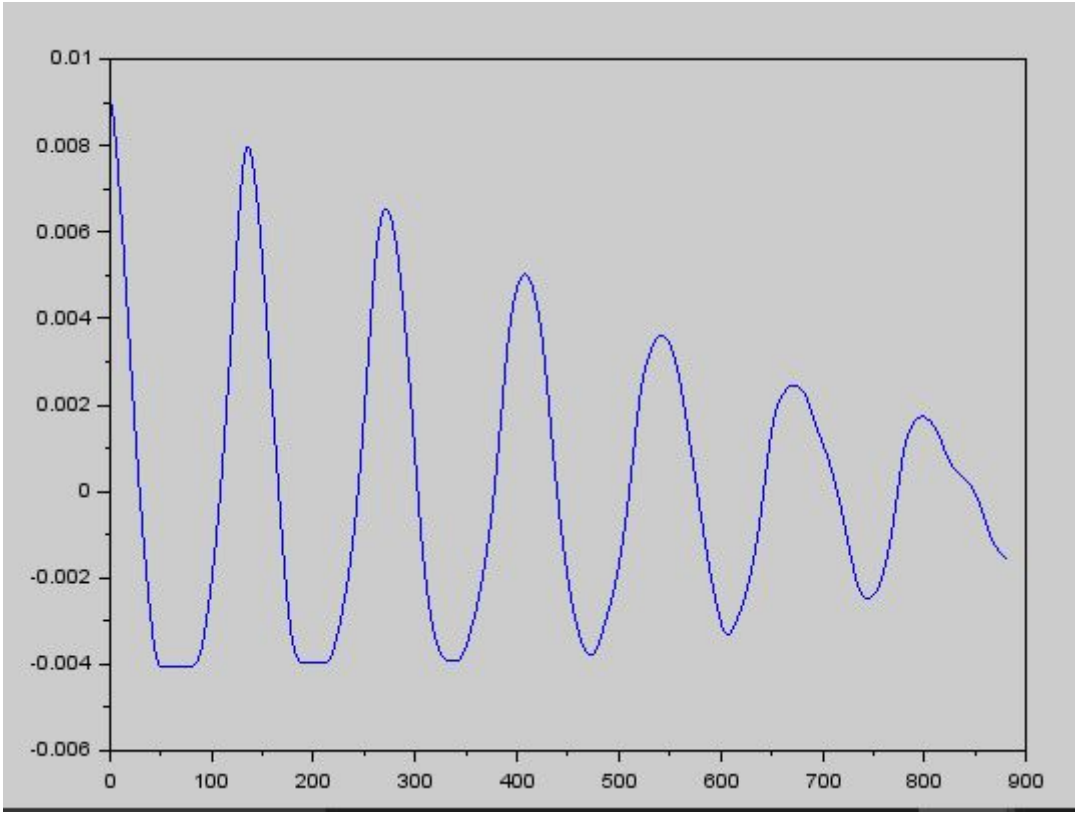

Figura 6.11 Primeira janela do teste 3 com Autocorrelação

A figura 5.11 acima mostra o sinal de autocorrelação no primeiro segmento detectado pelo algoritmo. Através deste sinal de autocorrelação, é possível obter a frequência por meio do período do pico. Já percemos diferenças notórias em relação à voz masculina neste sinal. Naturalmente, a estrutura fisiológica feminina confere mais harmônicos de alta frequência como um geral, o que deixa as regiões graves com menos amplitude.

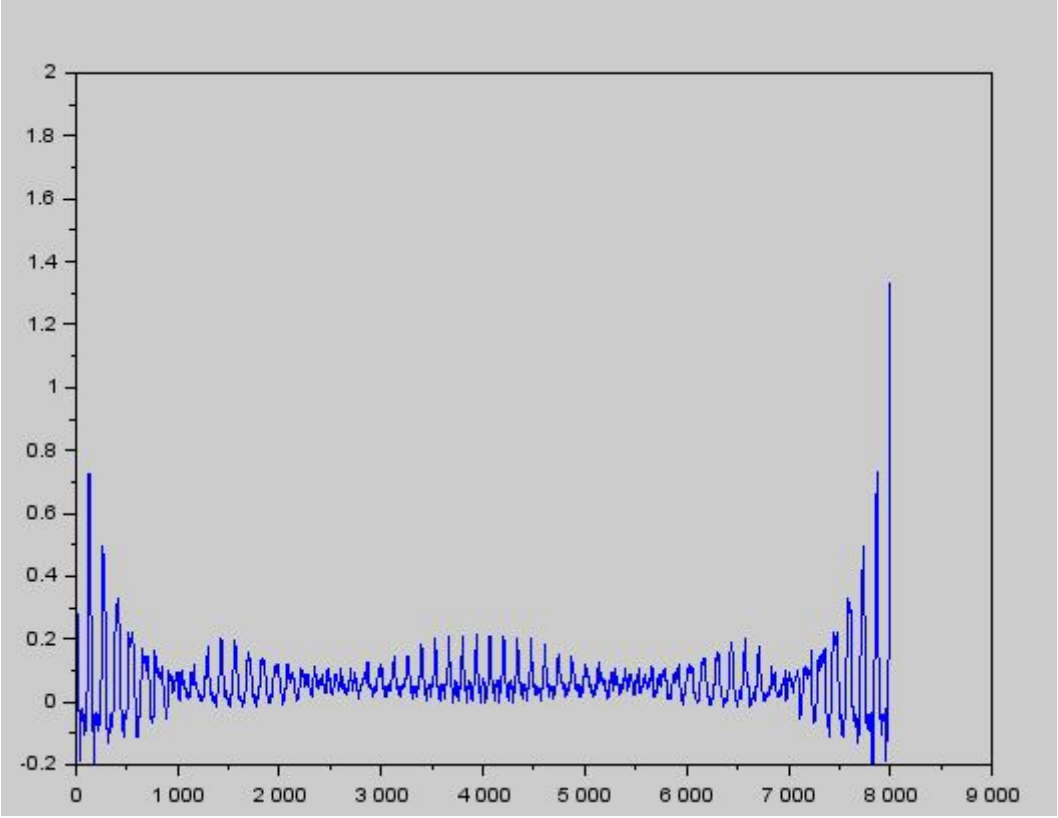

Figura 6.12 Primeira janela do teste 3 com Spectrum

Já a figura 5.12, mostra o mesmo segmento da figura 5.10, porém na abordagem Cepstrum.

A tabela 5.3 abaixo mostra as notas encontradas pelo algoritmo em cada segmento encontrado. Assim como no teste 2, a presença de do vibrato e por se tratar de uma voz feminina que têm harmônicos de mais alta frequência em geral de que na voz masculina, há uma leve diferença entre as abordagens.

Novamente, o algoritmo, apesar de acertar a nota, errou a oitava.

Tabela 6.3 Notas encontradas no Teste 3 de afinação

# Nota	Autocorrelação	Cepstrum
Nota 1	D#3	D#3
Nota 2	D#3	E3
Nota 3	D#3	D#3
Nota 4	E3	E3
Nota 5	D#3	D#3
Nota 6	D#3	D#3
Nota 7	E3	E3
Nota 8	D#3	E3
Nota 9	D#3	E3
Nota 10	D#3	E3
Nota 11	D#3	D#3

6.1.4 Teste 4

- Arquivo: SB4FA.wav

- Cantor: Soprano

- Descrição: Sustentação da nota G5 com a vogal |a|.

- Objetivo: Analisar o comportamento do algoritmo em região muito aguda, cantada por voz feminina aguda, no caso, o soprano.

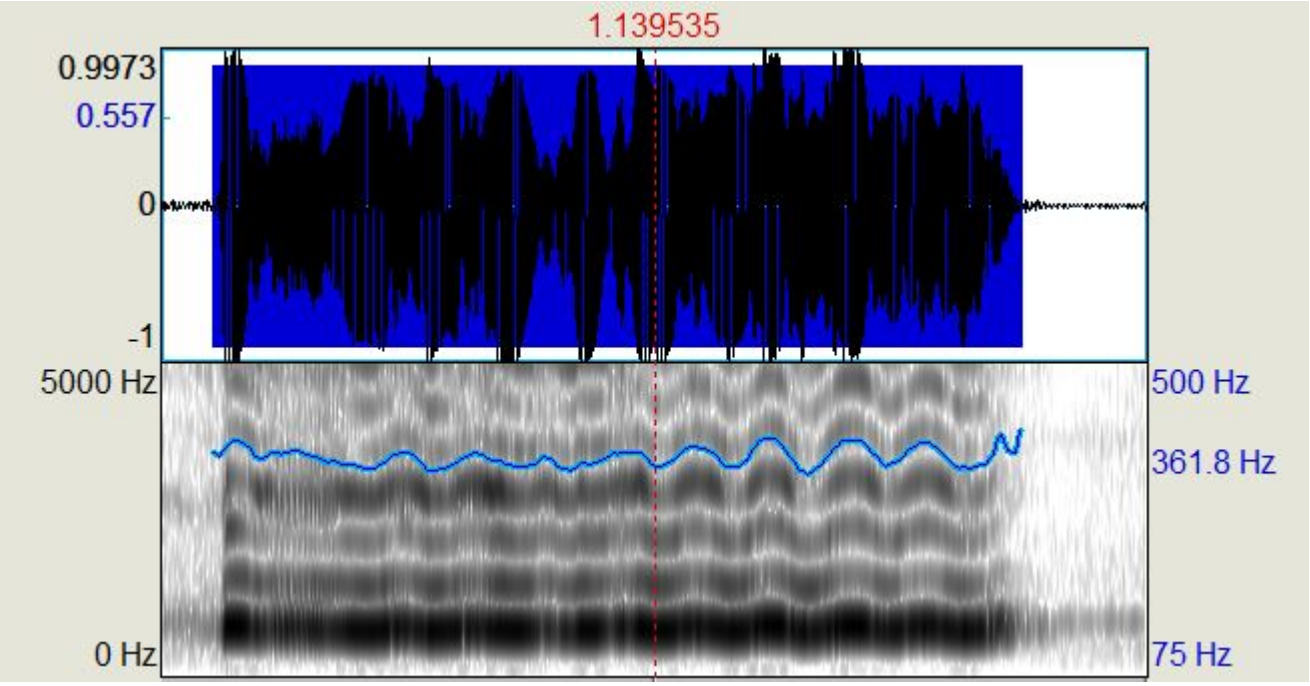

Figura 6.13 Espectrograma e frequência fundamental teste 3 pelo PRAAT

A identidade na amostra da amplitudo do sinal por amostra se mantém em todas os testes.

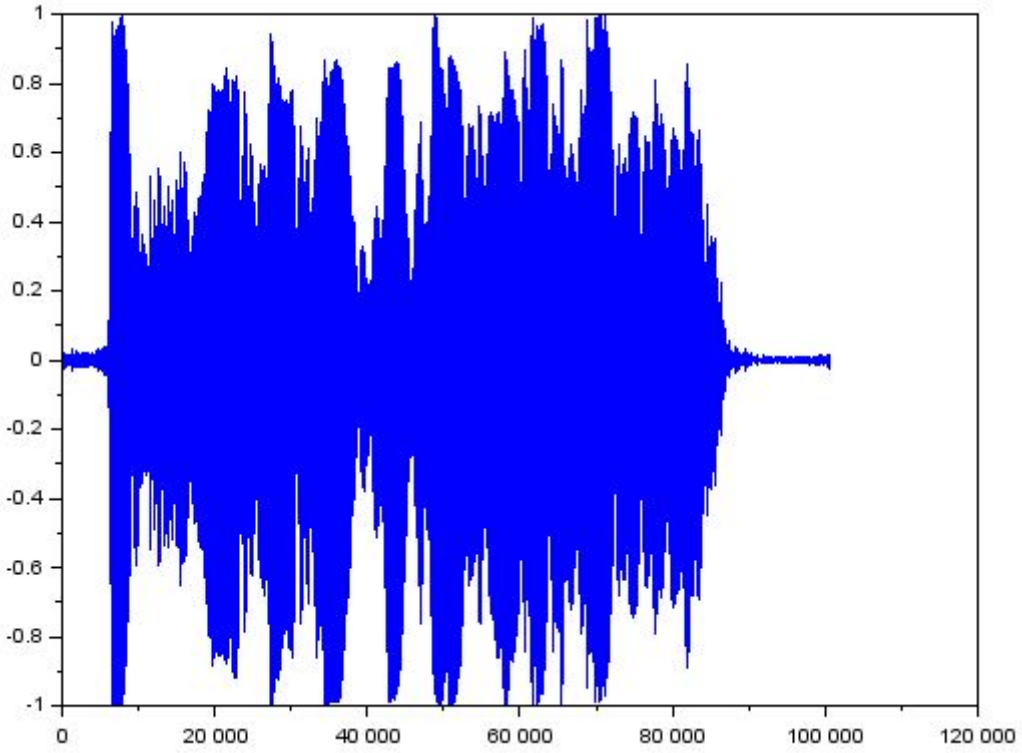

Figura 6.14 Amplitude por amostra de sinal, teste 4.

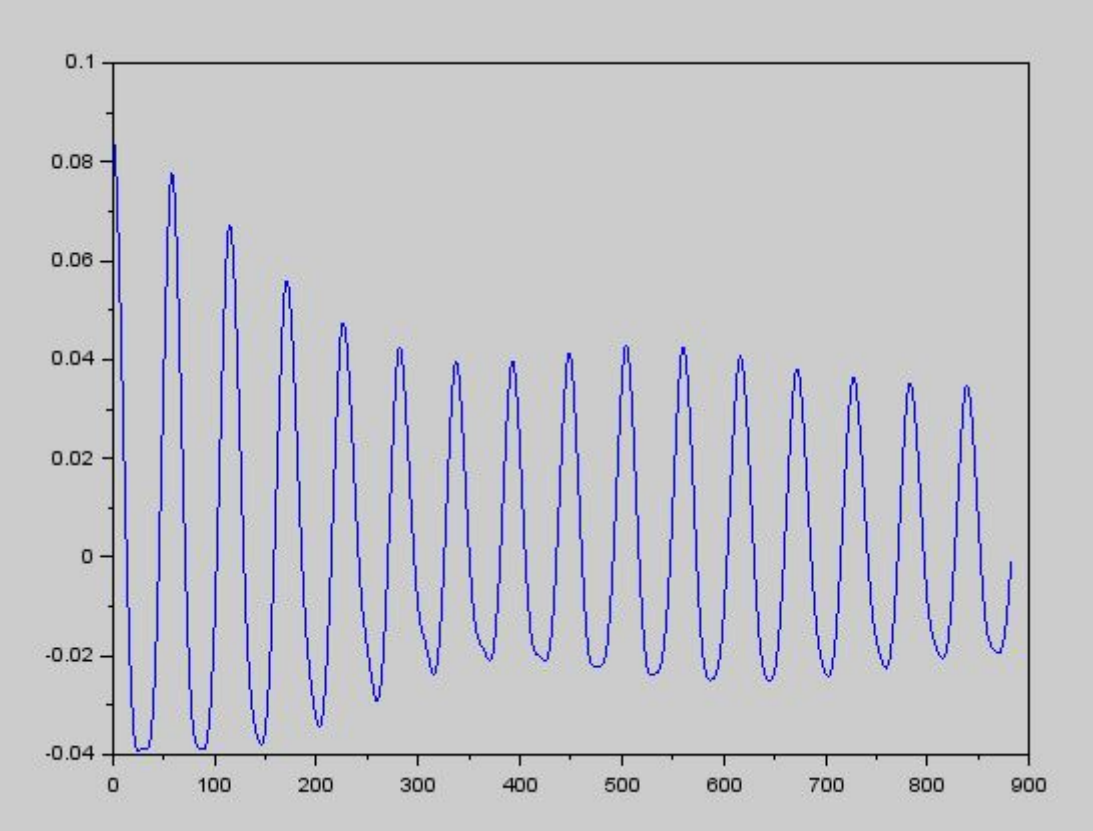

Figura 6.15 Primeira janela do teste 4 com Autocorrelação

A figura 5.15 acima mostra o sinal de autocorrelação no primeiro segmento detectado pelo algoritmo. Através deste sinal de autocorrelação, é possível obter a frequência por meio do período do pico. Já percemos diferenças notórias em relação à voz masculina neste sinal. Naturalmente, a estrutura fisiológica feminina confere mais harmônicos de alta frequência como um geral, o que deixa as regiões graves com menos amplitude.

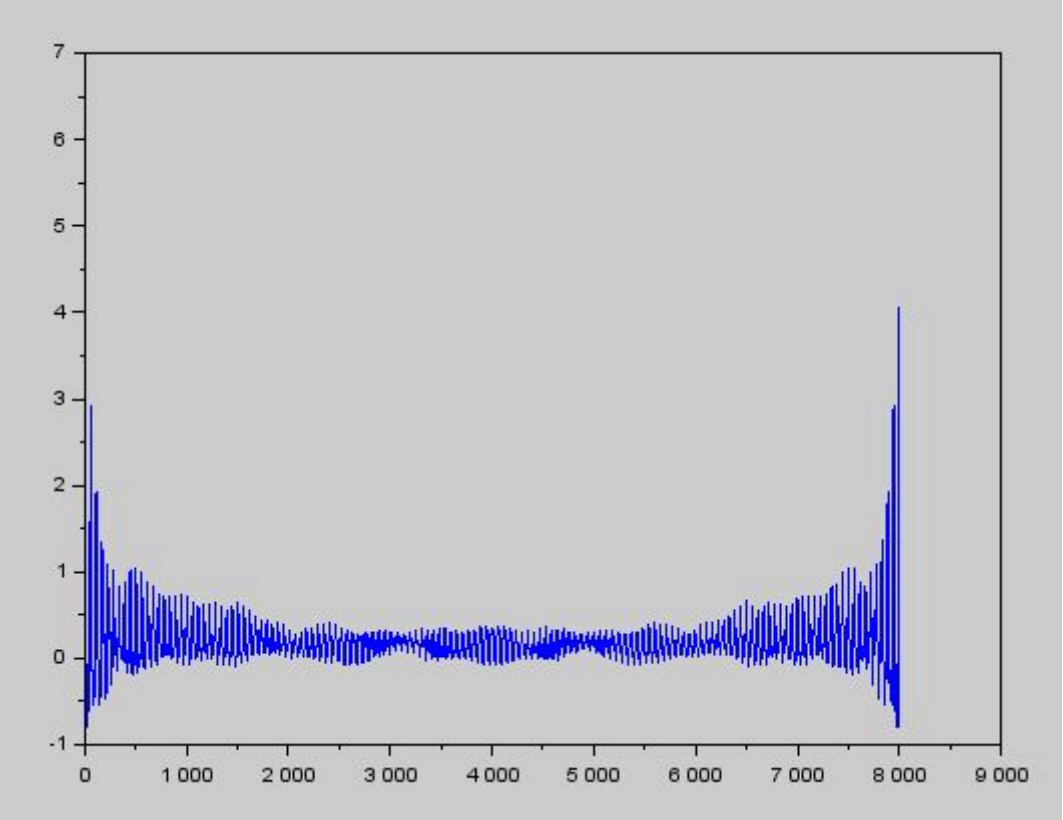

Figura 6.16 Primeira janela do teste 4 com Spectrum

As mesmas observações são pertinentes para este caso, a perturbação pelos harmônicos muito agudos, entretanto, causou uma taxa de diferença maior entre as notas encontradas nos métodos. Diferença esta que, como dito, se resolve com uma análise do vibrato, pois se mantém a no máximo meio tom.

Novamente, o algoritmo, apesar de acertar a nota, errou a oitava.

Tabela 6.4 Notas encontradas no Teste 4 de afinação

# Nota	Autocorrelação	Cepstrum
Nota 1	G3	G3
Nota 2	F#3	F3
Nota 3	F3	F#3
Nota 4	F3	F3
Nota 5	G3	G3
Nota 6	F#3	F#3
Nota 7	G3	F#3
Nota 8	F#3	G3
Nota 9	G3	G3
Nota 10	F#3	G3
Nota 11	F#3	F#3

6.1.5 Teste 5

- Arquivo: SB4FSEQ.wav
- Cantor: Soprano
- Descrição: Escala ascendente e descendente com a vogal lul.
- Objetivo: Analisar o comportamento do algoritmo numa situação de mudança de notas.

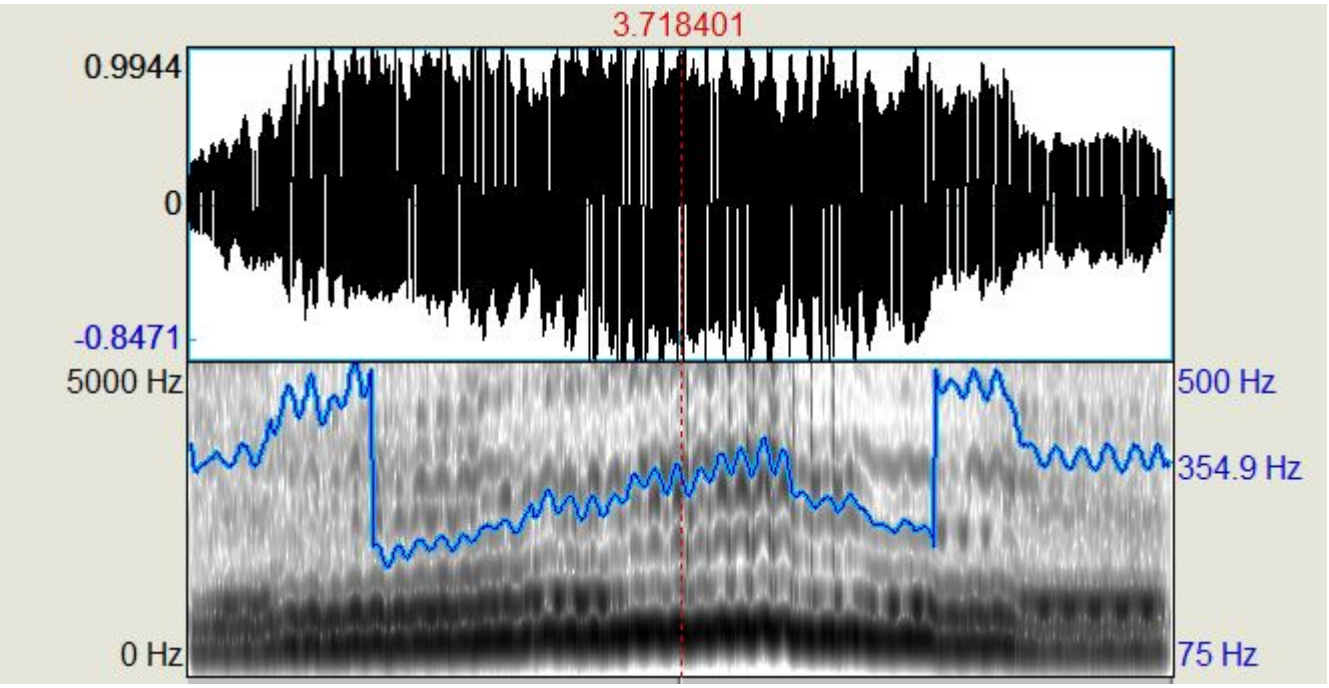

Figura 6.17 Espectrograma e frequência fundamental teste 5 pelo PRAAT

A identidade na amostra da amplitudo do sinal por amostra se mantém em todas os testes, apesar da versão mostrada na figura 5.18 parecer mais comprimida.

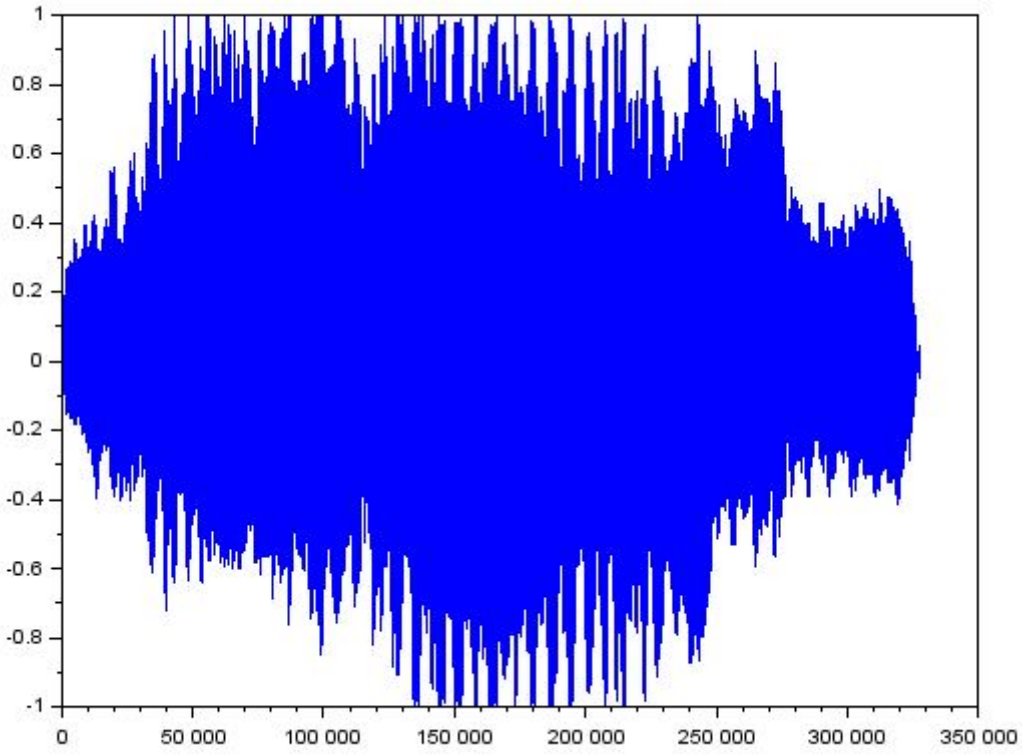

Figura 6.18 Amplitude por amostra de sinal, teste 5.

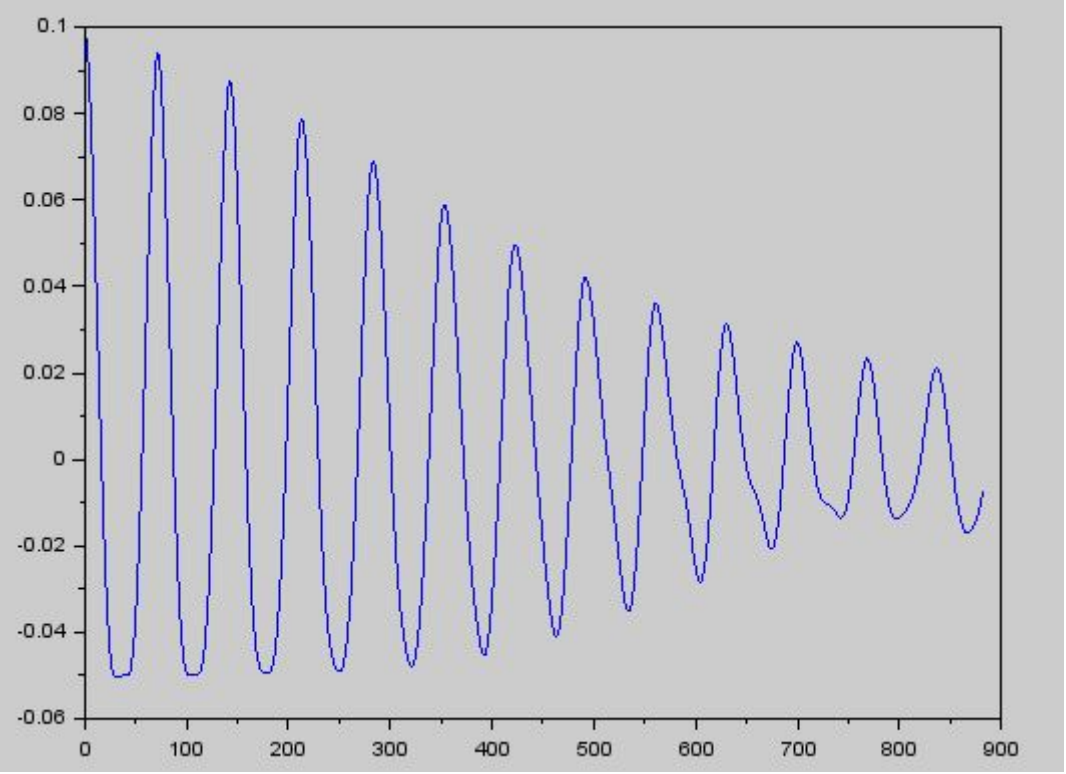

Figura 6.19 Primeira janela do teste 5 com Autocorrelação

A figura 5.19 acima mostra o sinal de autocorrelação no primeiro segmento detectado pelo algoritmo. Através deste sinal de autocorrelação, é possível obter a frequência por meio do período do pico.

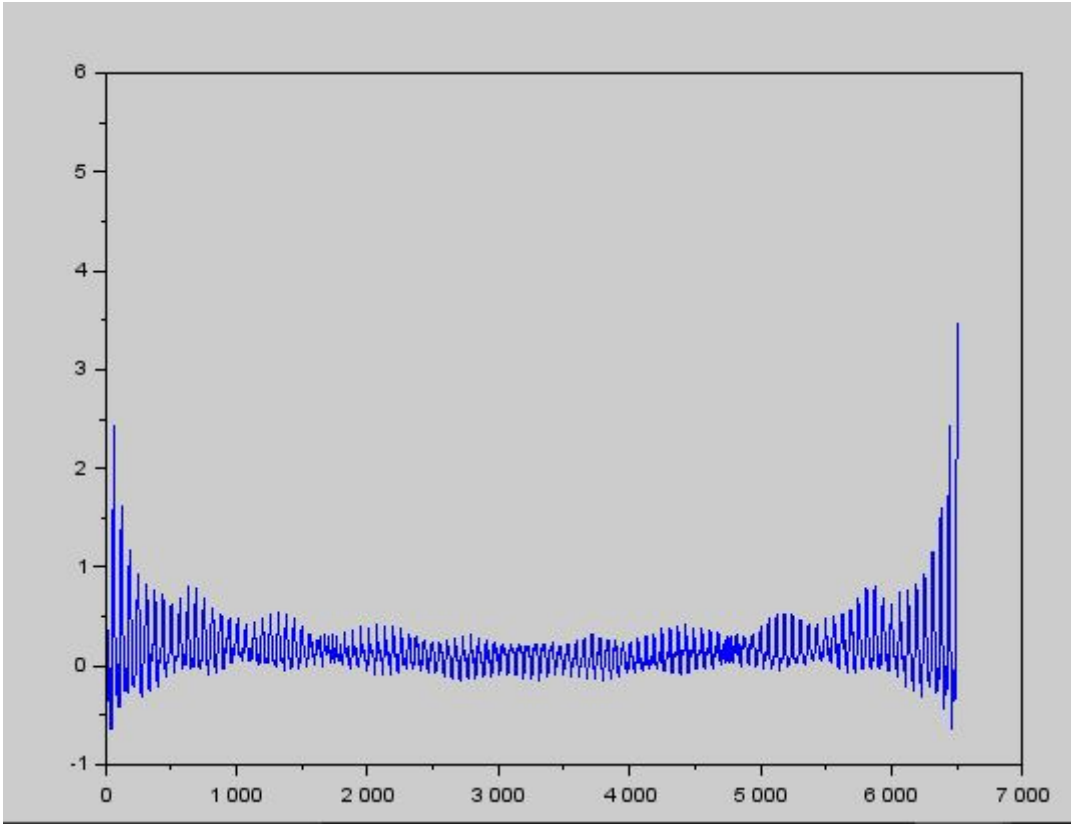

Figura 6.20 Primeira janela do teste 5 com Spectrum

Na tabela 5.5 abaixo, percebemos que as variações, apesar de ocorrem até mesmo porque o sinal de áudio tem maior duração, são menos frequentes do que em um sinal com a emissão de uma nota apenas. Isso pode acontecer porque as mudanças de notas são mais evidentes e apesar do vibrato, se sobresaem. Ainda, a cantora não empregou tanto vibrato quanto se estivesse emitindo uma nota apenas. Então, os resultados seguiram como o esperado.

Tabela 6.5 Notas encontradas no Teste 5 de afinação

# Nota	Autocorrelação	Cepstrum
Nota 1	F3	F3
Nota 2	F#3	F3
Nota 3	G3	G3
Nota 4	A3	A3
Nota 5	A3	A3
Nota 6	A3	A3
Nota 7	A#3	A#3
Nota 8	B3	B3
Nota 9	C#3	C#3
Nota 10	D3	D3
Nota 11	D3	D#3
Nota 12	D#3	D#3
Nota 13	E3	E3
Nota 14	E3	E3
Nota 15	F3	F#3
Nota 16	F3	F#3
Nota 17	F3	F3
Nota 18	F#3	G3
Nota 19	D#3	D#3
Nota 20	D#3	D#3
Nota 21	C#3	C#3
Nota 22	A#3	A#3
Nota 23	A#3	A#3
Nota 24	G3	G#3
Nota 25	F#3	F#3
Nota 26	F#3	F#3
Nota 27	F#3	F#3
Nota 28	F#3	G3
Nota 29	F3	F#3

Mais uma vez, percebemos que o algoritmo fica "preso"a uma única oitava e não identifica a transição de maneira correta, apesar de acertar as notas.

6.2 Voz Modal X Falsete

Para esta característica, conforme descrito na metodologia, pretendia-se utilizar os dados acústicos do JItter e variações e Shimmer e variações, aplicados a um método estatístico que pudesse diferenciar.

Porém, os valores obtidos para ambos os parâmetros não sartisfizeram as condições de valores mínimos desse teste de Teste Mann-Whitney de forma que pudessem ser relacionados com o processo de troca de registro entre voz modal e falsete.

6.2.1 Teste Único

Para esse teste, utilizou-se quatro arquivos de áudio:

- Cantor com voz modal
- cantora com voz modal
- cantor com voz em falsete
- cantora com voz em falsete

Os parâmetros jitter e shimmer foram calculados de cada trecho de áudio e então o teste estatístico foi realizado conforme explicado no capítulo anterior. Os resultados são mostrados abaixo:

Tabela 6.6 Tabelas com valores do teste de Mann-Whitney

Parâmetro	**Valor de U**
Jitter local	38,60
Jitter absoluto	107,23
Jitter rap	62,30
Jitter ppq5	46,97
Shimmer dB	127,21
Shimmer local	15,08
Shimmer apq3	27,36
Shimmer apq5	73,21

Como todos os valores de U são maiores do que os contido na tablea de Mann-Whitney [Birnbaum et al., 1956], a hipótese inicial de que os trechos de áudio provém da mesma população não pode ser descartada. Logo, conform o método de [Murphy, 2008], não podemos utilizar esses parâmetros para relacionar a transição de voz modal e falsete.

CAPÍTULO 7

Conclusões e Trabalhos Futuros

Certamente, a análise da voz cantada através de técnicas de processamento de sinais e MIR ainda tem um longo caminho a ser percorrido. Porém, não podemos negar os grandes avanços alcançados desde as primeiras produções de síntese de voz até grandes plataformas de entretenimento como Karaokês e outros jogos. O investimento no ramo de ensino da música principalmente no caso da voz cantada ainda não se compara a outros mercados mas vem alcançando resultados promissores dentro da academia científica. Principalmente em países em desenvolvimento como o Brasil, profissões como a do cantor não são extremamente valorizadas e talvez isso seja um dos motivos pelo qual alguns aprimoramentos demoram tanto a chegar ao mercado. Usar a tecnologia principalmente de forma não invasiva como é o caso do processamento de sinais para obter informaçõees a respeito da voz cantada é um poderoso aliado ao processo de estudo (tanto individual como em aula) que já é adotado em países mais desenvolvidos, como o Reino Unido. Em todo caso, é importante destacar que jamais este trabalho ou qualquer outro na área substitui análise e a expertise de um professor ou profissional experinente como um fonoaudiólogo para inferir diagnósticos ou padrões sobre a voz cantada.

O corpo humano, que é o grande responsável pelo processo de fonação, é além do mais antigo, o mais complexo instrumento musical principalmente no canto. O trato vocal e as características físicas que são únicos em cada pessoa, tornam ainda mais desafiador o processo de entender e analisar o fenômeno da voz. Também, se olharmos um contexto histórico, muitos mistérios já foram revelados sobre esse processo e hoje é possível saber com fatos concretos como vários mecanismos são acionados para cantar e como potencializar o uso da voz diante das mais diversas necessidades que vêm de repertório, público, estilo e até mesmo de limitações físicas.

Alcançar parâmetros acústicos é um dos fatores mais importantes para que se consiga extrair informações preciosas sobre a voz e, graças ao avanço tanto matemático quanto tecnológico, temos poderosas ferramentas como o matlab e mesmo as linguagens de programação que viabilizam a criação de softwares e sistemas de análise vocal cada vez mais precisos e abrangentes.

Relacionar as características interpretativas da voz com esses parâmetros é algo muito mais complexo do que o imaginado no início do trabalho. Mesmo a afinação, que parece ser algo abordado por praticamente todas as plataformas estudadas que são destinadas a análise vocal com foco em canto, requer esforços e atenções para que os resultados sejam fiéis e sirvam como feedback para o usuário final, que, pode ser um cantor.

Ainda sobre a afinação, ela é muito peculiar e perceptivel principalmente para ouvidos treinados. A presença mais abundante de harmônicos e mesmo do vibrato, que são características principalmente da técnica lírica (clássica) de canto, faz com que alguns algoritmos como os es-

tudados neste trabalho necessitem de módulos extra para que não dêem resultados imprecisos. O método Cepstrum para a indentificação da frequência fundamental é mais sensível a presença dos harmônicos e principalmente do formante do cantor. Tal sensibilidade pode ser enxergada na mudança constante de notas (apesar de próximas, no máximo um semi-tom e, considerando ainda a aproximação de frequências que pode refletir poucos comas). O método de autocorrelação tende a ser mais robusto diante da presença dessas características mas, detecta com menos sensibilidade algumas informações como, por exemplo, portamento.

Já na característica referente a nota de passagem, estudada pelo binômio falsete X voz modal, o desafio é constante não apenas para o cantor mas também tecnológico: entender que parâmetros acústicos estão relacionados a este processo pode ser um bom início para que esta informação seja entregue em forma de feedback. O jitter e o shimmer, pelo método empregado, não estão relacionados a este fenômeno e isso fomenta ainda mais a análise de outros parâmetros para que se possa encontrar uma maneira de extrair essa informação de maneira que seja usada para ajudar cantores e professores.

Destaca-se como uma importante contribuição deste trabalho, a base de dados criada especificamente para este fim, com cantores treinados, áudios contento diversas características e anotações importantes feitas sobre a visualização das marcações.

Como trabalhos futuros, podemos apontar a necessidade de extração de mais características relacionadas a interpretação e técnica que possam apoiar o processo de ensino e também a visualização destas por parte dos usuários. Disseminar o conceito tecnológico no âmbito do ensino do canto pode ser um meio importante para que a oferta por esse tipo de software seja aumentada.

Referências Bibliográficas

[SS,] Sing and see. `http://www.singandsee.com/`. Accessed: 2016-10-30.

[Alemán and Carlosena, 2004] Alemán, I. A. and Carlosena, A. (2004). *Signal Processing Techniques for Singing and Vibrato Modeling*.

[Alku and Backstrom, 2004] Alku, P. and Backstrom, T. (2004). Linear predictive method for improved spectral modeling of lower frequencies of speech with small prediction orders. *IEEE transactions on speech and audio processing*, 12(2):93–99.

[Birnbaum et al., 1956] Birnbaum, Z. et al. (1956). On a use of the mann-whitney statistic. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 13–17. University of California Press Berkeley, CA.

[Blowes,] Blowes, A. F. J. M. R. Measuring voice in the clinic-laryngograph speech studio analyses.

[Boersma, 2009] Boersma, P. (2009). Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica*, 61(5):305–308.

[Bogert et al., 1963] Bogert, B. P., Healy, M. J., and Tukey, J. W. (1963). The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the symposium on time series analysis*, volume 15, pages 209–243. chapter.

[Bonada et al., 2001] Bonada, J., Celma, Ò., Loscos, À., Ortolà, J., Serra, X., Yoshioka, Y., Kayama, H., Hisaminato, Y., and Kenmochi, H. (2001). Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models. In *Proceedings of International Computer Music Conference*.

[Brandão et al., 2007] Brandão, A. S., Cataldo, E., and Leta, F. (2007). Um novo método usando autocorrelação para extração da freq¨ uência fundamental em sinais de voz. *Trends in Applied and Computational Mathematics*, 8(2):191–200.

[Cook, 1991] Cook, P. (1991). {Identification of Control Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing}.

[Cooley and Tukey, 1965] Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301.

[De Cheveigné and Kawahara, 2002] De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

[de Krom, 1993] de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266.

[de Sá Ferreira, 2012] de Sá Ferreira, J. F. T. (2012). *Tecnologia de Apoio em Tempo-Real ao Canto-Relação entre parâmetros perceptivos da voz cantada com fenómenos acústicos objectivos.* PhD thesis, Universidade do Porto.

[Donoso, 2012] Donoso, J. P. (2012). História da acústica. *Universidade de São Paulo–IFSC*.

[dos Santos Ventura, 2011] dos Santos Ventura, J. A. P. (2011). Biofeedback da voz cantada.

[Downie, 2004] Downie, J. S. (2004). The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28(2):12–23.

[Echternach and Richter, 2012] Echternach, M. and Richter, B. (2012). Passaggio in the professional tenor voice—evaluation of perturbation measures. *Journal of Voice*, 26(4):440–446.

[Fung, 2009] Fung, A. (2009). Consuming karaoke in china: Modernities and cultural contradiction. *Chinese Sociology & Anthropology*, 42(2):39–55.

[Grant et al., 2008] Grant, M., Boyd, S., and Ye, Y. (2008). Cvx: Matlab software for disciplined convex programming.

[Guimarães, 2007] Guimarães, I. (2007). A ciência e a arte da voz humana. *Alcoitão, Escola Superior de Saúde de Alcoitão*.

[Henrique, 2002] Henrique, L. L. (2002). *Acústica musical*.

[Högset, 2001] Högset, Johan Sundberg, C. (2001). Voice source differences between falsetto and modal registers in counter tenors, tenors and baritones. *Logopedics Phoniatrics Vocology*, 26(1):26–36.

[Jang et al., 2001] Jang, J.-S. R., Chen, J.-C., and Kao, M.-Y. (2001). Miracle: a music information retrieval system with clustered computing engines. *Bloomington, Indiana, USA: ISMIR*.

[Kedem, 1986] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493.

[Kosugi et al., 2000] Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., and Kushima, K. (2000). A practical query-by-humming system for a large music database. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 333–342. ACM.

[Li and Wang, 2007] Li, Y. and Wang, D. (2007). Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1475–1487.

[Loscos, 2007] Loscos, A. (2007). *Spectral processing of the singing voice*. PhD thesis, Citeseer.

[Mangini et al., 2013] Mangini, M. M., de Andrada, M. A., et al. (2013). Classificacao vocal: um estudo comparativo entre as escolas de canto italiana, francesa e alema. *OPUS-Revista Eletrônica da ANPPOM*, 19(2):209–222.

[Mota et al., 2009] Mota, L., Vasconcelos, J., and Cavalcanti, T. (2009). Videolaringoscopia e análise perceptivoauditiva em atores. *ACTA ORL/Técnicas em Otorrionolaringologia, 27 (2)*, pages 71–75.

[Murphy, 2008] Murphy, K. (2008). Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice.

[Oguz et al., 2007] Oguz, H., Demirci, M., Safak, M. A., Arslan, N., Islam, A., and Kargin, S. (2007). Effects of unilateral vocal cord paralysis on objective voice measures obtained by praat. *European Archives of Oto-Rhino-Laryngology*, 264(3):257–261.

[Piszczalski and Galler, 1979] Piszczalski, M. and Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. *The Journal of the Acoustical Society of America*, 66(3):710–720.

[Sundberg and Rossing, 1990] Sundberg, J. and Rossing, T. D. (1990). The science of singing voice. *the Journal of the Acoustical Society of America*, 87(1):462–463.

[Teixeira, 1995] Teixeira, J. P. (1995). *Modelização paramétrica de sinais para aplicação em sistemas de conversão texto-fala*. PhD thesis, FEUP.

[Teixeira and Fernandes, 2014] Teixeira, J. P. and Fernandes, P. O. (2014). Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices. *Procedia Technology*, 16:1228–1237.

[Titze, 1995] Titze, I. (1995). Summary statement: Workshop on acoustic voice analysis. *National Center for Voice and Speech*, pages 26–30.

[Yan et al., 2005] Yan, Y., Ahmad, K., Kunduk, M., and Bless, D. (2005). Analysis of vocal-fold vibrations from high-speed laryngeal images using a hilbert transform-based methodology. *Journal of Voice*, 19(2):161–175.

[Zeitels et al., 2002] Zeitels, S. M., Hillman, R. E., Desloge, R., Mauri, M., and Doyle, P. B. (2002). Phonomicrosurgery in singers and performing artists: treatment outcomes, management theories, and future directions. *The Annals of otology, rhinology & laryngology. Supplement*, 190:21–40.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).