Francisco Maia Aleixo Bettencourt Neves

Algoritmos de deteção de altura: enquadramento científico e comparação experimental de métodos

Orientador: professor doutor João Marques Carrilho

Universidade Lusófona de Humanidades e Tecnologias

Lisboa

2023

Francisco Maia Aleixo Bettencourt Neves

Algoritmos de deteção de altura: enquadramento científico e comparação experimental de métodos

Tese defendida em prova pública na Universidade Lusófona de Humanidades e Tecnologias no dia 17/07/2023 para obtenção do grau de mestre em Produção e Tecnologias do Som perante o júri nomeado pelo Despacho de Nomeação n.º 296/2023 de Junho de 2023, com a seguinte composição:

Presidente: professor doutor Gonçalo Gato

Vogais: professor doutor Pedro Duarte Pestana (arguente)

Orientador: professor doutor João Marques Carrilho

Universidade Lusófona de Humanidades e Tecnologias
Cinema e Artes dos Media

Lisboa
2023

Agradecimentos

Ao João Marques Carrilho pela orientação e inspiração.

Ao Pedro Pestana pela brilhante arguição.

Ao Gonçalo Gato pela moderação da defesa pública.

Ao Bruno Santos pela revisão do primeiro capítulo.

Aos meus pais e à Maria da Paz pelo apoio.

Resumo

A altura é um fenómeno psicoacústico fundamental na música, no discurso e na análise de cena auditiva e depende de vários fatores, não apenas da frequência do som percecionado. A deteção de altura é uma vertente do processamento áudio que procura identificar a altura de tons musicais de acordo com a maneira como os humanos a percecionam, através de algoritmos otimizados para tal, que podem operar no domínio temporal, espetral ou noutros domínios. Na área da música, este tipo de processamento traz vantagens em diversas práticas, tais como na *performance*, na produção musical e no ensino. Quatro algoritmos de deteção de altura são testados e comparados entre si no SuperCollider, de modo a averiguar as suas precisões de deteção e consequentemente as suas viabilidades de aplicação.

Palavras-chave: altura, deteção, algoritmo, música, psicoacústica.

Abstract

Pitch is a psychoacoustical phenomenon that is fundamental in music, speech and auditory scene analysis and it depends on many factors, not just the frequency of the perceived sound. Pitch detection is a type of audio processing that seeks to find the pitch of musical tones in agreement with the way humans perceive it, through the use of optimized algorithms that operate on temporal, spectral or other domains. In music, this kind of processing brings advantages in many practices like performance, music production and teaching. Four pitch detection algorithms are tested and compared to each other in SuperCollider, in order to figure out their precisions and consequently their application viabilities.

Keywords: pitch, detection, algorithm, music, psychoacoustics.

Índice

Introdução

Esta dissertação tem como principal objetivo o enquadramento científico e a investigação da deteção de altura, que é uma vertente do processamento áudio que procura identificar a altura de sons musicais ou de discurso através de diversas estratégias e que tem inúmeras aplicações tanto na música como na análise de discurso. Este enquadramento assenta na investigação da altura como fenómeno psicoacústico, na categorização e investigação dos principais métodos de deteção de altura desenvolvidos até hoje e na exploração das suas aplicações na área da música. Com o intuito de expandir o conhecimento acerca destes métodos, a investigação original que completa esta dissertação incide sobre quatro algoritmos de deteção de altura implementados no ambiente de síntese e processamento áudio SuperCollider que são testados e comparados experimentalmente.

No primeiro capítulo, o conceito de altura é explorado e clarificado detalhadamente, uma vez que é fundamental na deteção de altura. Embora esta seja frequentemente associada à frequência de um tom musical, por vezes numa relação de equivalência, esta não é a única variável que define o fenómeno complexo e psicoacústico da altura: outros aspetos como o espetro, a intensidade e até a duração de um som podem influenciar a maneira como um humano perceciona a sua altura. De modo a facultar ao leitor as noções necessárias para o entendimento destes fenómenos, o subcapítulo 1.1. explica uma série de conceitos fundamentais relativos à acústica, à psicoacústica e à psicoacústica musical, baseando-se largamente na obra *Acústica Musical*, de Luís Henrique. Neste capítulo são também abordadas questões psicoacústicas fundamentais como a definição de timbre e a diferença entre tom musical e ruído.

No segundo capítulo, os principais métodos de deteção de altura são elencados e descritos detalhadamente, para além de serem categorizados consoante o domínio no qual operam, que pode ser temporal, espetral, misto ou com base na modelação do sistema auditivo humano, e consoante a sua capacidade de identificar dois ou mais tons musicais simultaneamente, que define se são monofónicos ou polifónicos. As questões que se colocam relativamente a estes algoritmos são, naturalmente, quão precisos são os seus resultados, ou seja, quão fidedignamente conseguem traduzir ondas sonoras para a linguagem psicoacústica da perceção humana, mas também quão eficientes são ao fazê-lo, ou seja, quão intenso é o seu consumo de recursos computacionais. Estas informações são recolhidas de diversas publicações científicas, publicadas em revistas como *The Journal of the Acoustical Society of America* e

abordadas em conferências como *IEEE International Conference on Acoustics, Speech, and Signal Processing*. De modo a clarificar o funcionamento dos diferentes algoritmos, as equações matemáticas que os descrevem são frequentemente apresentadas e esclarecidas. O subcapítulo 2.3 explora as aplicações destes métodos na área da música, mais especificamente na *performance*, na produção musical, no ensino, na transcrição e no entretenimento, recorrendo frequentemente a exemplos de produtos comercializados atualmente.

Por fim, no terceiro capítulo é elaborada uma experiência no SuperCollider que visa clarificar a precisão de quatro algoritmos de deteção monofónicos: ZeroCrossing, Pitch, Qitch e Tartini, cujos funcionamentos são descritos detalhadamente no subcapítulo 2.1. Esta precisão é medida através da deteção de altura de duas pequenas melodias que são apresentadas nos timbres de variados instrumentos musicais. Os resultados obtidos são comparados diretamente com as alturas reais, sabidas a priori, e expressos em percentagens que indicam o rácio entre deteção correta e deteção incorreta. Esta componente experimental visa expandir o conhecimento relativo à deteção de altura no SuperCollider, clarificando quais os métodos mais viáveis para aplicações musicais generalizadas.

Esperançosamente, ao ler esta dissertação, um leitor, mesmo que tenha poucos conhecimentos de acústica e matemática, obterá um entendimento generalizado acerca da deteção de altura.

Todas as citações diretas em inglês foram traduzidas para português pelo autor. Os ficheiros de áudio utilizados na experiência que contêm as melodias tocadas pelos instrumentos musicais virtuais foram gerados pelo autor no XPand!2. As melodias vocais utilizadas na experiência foram cantadas e gravadas pelo autor. As partituras foram criadas pelo autor no MuseScore 4. Os gráficos sem referenciação foram criados pelo autor no Praat ou no SuperCollider.

1. Enquadramento científico da altura como fenómeno psicoacústico

A altura é uma das sensações auditivas primárias e tem um papel fundamental na música, no discurso e na análise de cena auditiva (Oxenham, 2012). Na música, sequências e combinações de alturas têm papéis fundamentais na definição de melodia e harmonia. No discurso, contornos ascendentes e descendentes de altura contribuem para a definição da prosódia e, em linguagens tonais, como o mandarim, do significado das palavras. Por fim, em ambientes acústicos complexos, diferenças de altura ajudam o ouvinte a discernir fontes sonoras.

A sensação de altura é definida como a característica psicológica que está relacionada diretamente com a frequência do estímulo e traduz a sensação auditiva que nos permite ordenar os sons do grave ao agudo. Embora esteja relacionada principalmente com a frequência, esta não é a única variável em questão: outros fatores como a intensidade, o espetro, a duração, o contexto e a presença de outros sons têm também um papel fundamental na sua definição (Henrique, 2002).

Sendo uma característica psicológica, a sensação de altura é um fenómeno psicoacústico, que é um ramo da psicofísica que estuda a relação entre os estímulos acústicos e as sensações auditivas (Roederer, 2009), e deve ser definida e analisada com isso em conta.

De modo a clarificar conceitos e ilustrar o comportamento das ondas mecânicas, assim como as ferramentas com as quais são tipicamente analisadas, quer de uma perspetiva física, quer de uma perspetiva psicológica, o subcapítulo seguinte procura esclarecer uma série de conceitos fundamentais das áreas da acústica, da psicoacústica e da psicoacústica musical.

1.1. Conceitos fundamentais

1.1.1. Acústica

A acústica é o ramo da física que lida com ondas mecânicas, ou seja, com vibrações que se propagam através de meios, sejam eles gasosos, líquidos ou sólidos. Como tal, dada a sua definição abrangente, tem ramificações em diversas áreas de investigação, tais como nas geociências, nas engenharias, nas ciências da vida e nas artes.

As ondas mecânicas, ou sonoras, são propagações de uma perturbação de pressão localizada que criam uma pressão sonora cuja unidade SI é o pascal e são usualmente descritas como variações de pressão em função do tempo.

Em ondas periódicas, ou seja, que descrevem um comportamento que se repete em intervalos de tempo constantes, ou ciclos, o seu valor máximo de desvio de pressão em relação à pressão de equilíbrio é a sua amplitude e o valor temporal entre o início de um ciclo e o início do seguinte é o seu período, medido em segundos (figura 1), cujo inverso é a sua frequência, expressa em hertz (Hz), que se define como o número de ciclos efetuados na unidade de tempo. O seu comprimento de onda corresponde ao valor, em metros, da distância entre o início de cada ciclo e é inversamente proporcional à frequência. Uma oscilação periódica é também descrita pela sua fase, que consiste na "fração de um período ou ciclo entre um ponto de referência e outro qualquer ponto de uma sinusoide" (Henrique, 2002), sendo, portanto, um ângulo expresso em radianos ou graus.

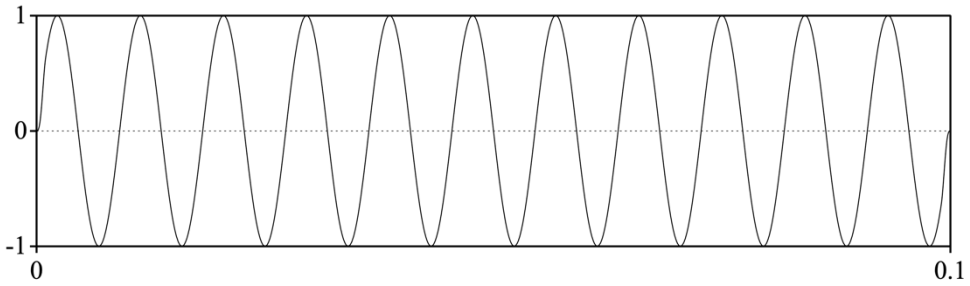

Figura 1: representação temporal (amplitude por segundo) de uma onda sinusoidal de 110 Hz.

É possível categorizar um som como simples ou complexo. Um som simples, de acordo com Ohm, descreve um movimento oscilatório simples, ou seja, um movimento de acordo com a lei do movimento pendular simples. Um som complexo é qualquer som que possa ser dividido e analisado como um conjunto de oscilações simples, portanto, que seja uma soma de sons simples. Esta definição leva à conclusão de que um som complexo "contém" outros sons e que estes sons descrevem uma relação simples, ou harmónica, entre eles (Schaeffer, 1966).

Uma onda pode ser representada e analisada num domínio temporal ou num domínio espetral ou frequencial (figura 2). A representação temporal de um sinal consiste no traçado da variação de uma determinada grandeza em função do tempo, geralmente da sua amplitude, enquanto no domínio espetral são evidenciadas as frequências nas quais o sinal contém energia, assim com as periodicidades existentes (Henrique, 2002). A operação matemática que permite

a decomposição de um som complexo em sons simples é a transformada rápida de Fourier, utilizada em inúmeras áreas da tecnologia. Através dela, é possível converter um sinal do domínio temporal para o domínio espetral, no qual são revelados todos os movimentos oscilatórios simples que compõem o som complexo analisado. Um som pode ainda ser representado bidimensionalmente num domínio temporal e espetral em simultâneo.

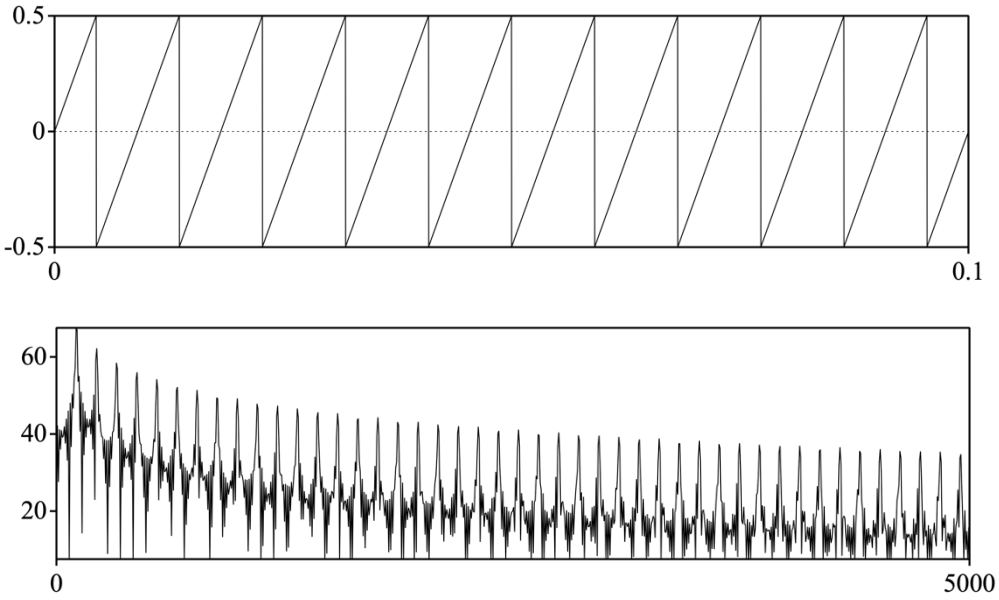

Figura 2: representação temporal e frequencial (nível de pressão sonora por frequência) de uma onda dente de serra de 110 Hz.

Num som, podem considerar-se três períodos de duração distintos: o transitório, o período de estabilidade e o decaimento (figura 3) (Henrique, 2002). O transitório, ou ataque, corresponde à componente de grande amplitude e curta duração que ocorre no início da sonância da onda e contribui largamente para a identificação do timbre de um instrumento musical. O período de estabilidade é o período intermédio e é nele que se fixam a altura e a intensidade. Por fim, o decaimento consiste na transição da sonância para o silêncio.

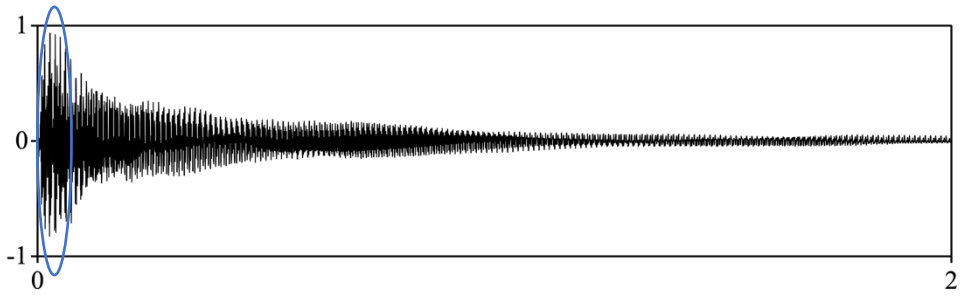

Figura 3: transitório (assinalado a azul) da nota fá 2 de um *kayagum* (cordofone tradicional coreano) (*kayageum1_F2*, 2006).

Duas ondas sonoras podem interagir de um modo construtivo ou destrutivo. A interferência construtiva ocorre quando duas ondas descrevem uma deslocação na mesma direção e faz com que a onda resultante da sua soma descreva uma deslocação maior do que as deslocações das primeiras. Por outro lado, a interferência destrutiva ocorre quando duas ondas têm deslocações em direções opostas e resulta numa onda com uma deslocação inferior à onda de maior amplitude ou num cancelamento das mesmas, se tiverem a mesma amplitude. A filtragem em pente (*comb filtering*) é um fenómeno que ocorre quando uma onda sonora é somada com uma onda idêntica desfasada por um curto intervalo de tempo, por consequência de uma reflexão, por exemplo, dando origem a uma série de cancelamentos causados por interferências destrutivas ao longo dos seus espetros.

1.1.2. Psicoacústica

A psicoacústica trata o estudo científico da perceção sonora, ou seja, como os humanos respondem psicologicamente aos fenómenos físicos descritos pela acústica.

O fenómeno auditivo é conseguido, essencialmente, através da transformação de energia mecânica em energia elétrica (estímulos neurais), à semelhança de um microfone. O ouvido é formado por três partes: ouvido externo, médio e interno. O ouvido externo compreende o pavilhão auditivo e o meato auditivo externo, fechado no seu interior pelo tímpano. O ouvido médio contém uma cadeia de pequenos ossos articulados uns nos outros (martelo, bigorna e estribo), que se designam por ossículos, um canal que faz ligação com a cavidade nasal (trompa de Eustáquio) e duas aberturas cobertas por membranas (janela oval e janela redonda) que fazem ligação com o ouvido interno. Por fim, o ouvido interno é uma estrutura dividida em três partes: cóclea ou caracol, vestíbulo e canais semicirculares. A cóclea está, por sua vez, dividida em três compartimentos (duto vestibular, duto timpânico e duto coclear) preenchidos por

líquidos (perilinfa e endolinfa) e separados por membranas (membrana de Reissner e membrana basilar). A membrana basilar contém o órgão de Corti, no qual as células ciliadas produzem sinais eletroquímicos de acordo com o seu movimento, causado pelas perturbações no líquido coclear, que por sua vez são causadas por todos os processos de transdução da energia mecânica da onda sonora que ocorrem nas partes do sistema auditivo referidas anteriormente. O ouvido interno está conectado ao cérebro através do nervo acústico, que transmite os impulsos nervosos gerados a fim de serem descodificados no córtex auditivo (Henrique, 2002).

O alcance da audição humana, ou espetro audível, é aproximadamente de 16 Hz a 20 kHz (Henrique, 2002), embora o limite superior num adulto médio seja à volta de apenas 16 kHz, uma vez que diminui com a idade (Zemlin, 1988 apud Henrique, 2002). É importante salientar que, no entanto, apenas os sons simples entre 30 Hz e 4 kHz causam uma sensação de altura definida o suficiente para transportar informação melódica (Attneave & Olson, 1971 apud Oxenham, 2012). Valores de frequência abaixo de 20 Hz dão origem a infrassons e valores acima de 20 kHz dão origem a ultrassons.

Como referido anteriormente, a audição humana não é linear e a sua resposta depende de diversos fatores. Em primeira instância, o volume de um som, ou seja, a intensidade percecionada está associada ao nível de pressão sonora, medido em decibéis (dB), que é uma transformação logarítmica da pressão sonora relativa a um valor de referência que visa compensar um desequilíbrio do ouvido, que é capaz de distinguir com mais facilidade diferenças de pressões sonoras baixas do que diferenças de pressões sonoras altas. O valor de 0 dB é considerado o limiar da audição, abaixo do qual não é possível detetar um som, enquanto o limiar da dor, acima do qual um ouvinte pode sentir dor ou sofrer danos auditivos, é compreendido aproximadamente entre os valores de 115 e 140 dB, embora ambos dependam principalmente da frequência do som e o último dependa também do tempo de exposição do ouvinte.

Em segundo lugar, a intensidade sonora percecionada também é influenciada pela frequência das ondas sonoras que chegam ao ouvido. Os contornos de volume igual (figura 4), definidos em ISO 226:2003 (International Organization for Standardization, 2003) e baseados no estudo experimental que está na base da publicação de *Loudness, Its Definition, Measurement and Calculation*, de Harvey Fletcher e Wilden Munson (1933), são medidas de nível de pressão sonora em função da frequência que demonstram valores que levam a uma perceção de volume constante. Estes contornos indicam que “todos os pontos sobre uma mesma

curva representam sons que produzem a mesma sensação de intensidade, (...). Assim nasce uma nova escala que quantifica a sensação de intensidade em níveis – a escala de fones – (...)" (Henrique, 2002). A unidade em questão, o fone, foi proposta por Markhausen (Hartmann, 1998 apud Henrique, 2002) e coincide com o decibel na frequência de 1 kHz. Os contornos demonstram ainda que o ouvido humano tem uma sensibilidade agudizada nas frequências médias, uma vez que descrevem um vale nas frequências entre aproximadamente 2 e 5 kHz. Foi também desenvolvida a escala de sones, que se baseia no facto de que um aumento de 10 fones resulta geralmente numa sensação de duplicação de volume: cada aumento de 10 fones corresponde a uma duplicação de sones, sendo a equivalência inicial entre 40 fones e 1 sone.

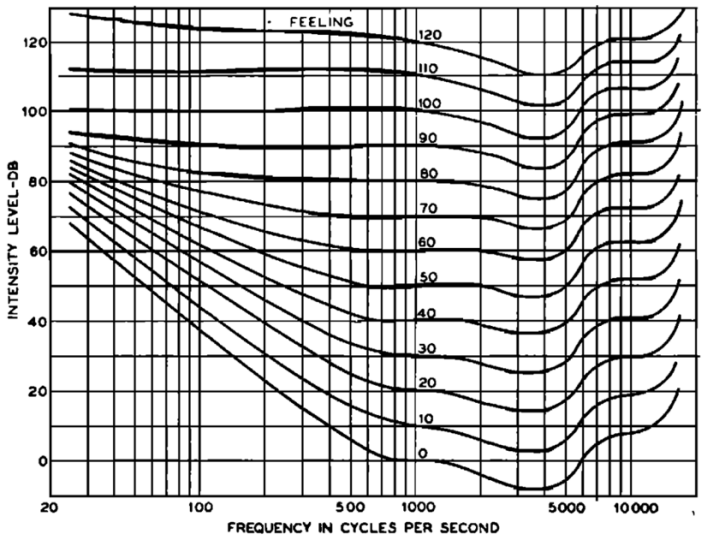

Figura 4: contornos de volume igual (nível de pressão sonora percecionado por frequência) (Fletcher & Munson, 1933).

Como descrito anteriormente, um som complexo é composto por vários sons simples. Num determinado som complexo, o som simples de frequência mais baixa é chamado fundamental, enquanto os sons simples com frequências superiores são chamados parciais, que podem ser harmónicos, se as suas frequências forem iguais ou muito próximas dos múltiplos da frequência da fundamental, ou inarmónicos, se isso não se verificar. A diferente proporção energética entre a fundamental e os parciais de um som dá origem à sensação auditiva chamada timbre, que é a "característica subjetiva do som que nos permite diferenciar dois sons de altura e intensidade iguais" (Henrique, 2002).

Um som é percecionado como tendo uma altura definida quando os seus parciais são coerentes com a série dos harmónicos em relação à sua fundamental, ou seja, quando são parciais harmónicos. Caso se verifique que os parciais de um som são inarmónicos, este é percecionado como não tendo uma altura definida. O conceito de inarmonicidade é então relativo ao desvio das frequências dos parciais de um som em relação às frequências da série dos harmónicos da sua fundamental e faz-se notar em diferentes magnitudes em diferentes corpos ressonantes, como numa corda, que tem mais inarmonicidade quanto maior for o seu diâmetro. Posto isto, e a título de exemplo, é possível concluir que, num piano, as cordas de maior diâmetro, que correspondem às notas graves, têm mais inarmonicidade do que as cordas de menor diâmetro, que correspondem às notas agudas. O entendimento e domínio deste fenómeno é crucial para o desenvolvimento de algoritmos de deteção de altura de música precisos.

Quando dois ou mais sons soam simultaneamente, pode acontecer um fenómeno designado por efeito de máscara, que causa a subida do limiar de audibilidade do som que é mascarado (Henrique, 2002), ou seja, torna um dos sons inaudíveis por consequência da presença do outro. Este fenómeno pode ocorrer como consequência de os sons em questão terem uma diferença considerável de intensidade ou estarem contidos na mesma banda crítica, que é definida como a banda de frequências na qual um segundo tom interfere com a perceção do primeiro devido aos filtros auditivos criados pela cóclea (uma explicação mais detalhada pode ser encontrada no subcapítulo 1.2.5).

1.1.3. Psicoacústica musical

No que toca à relação entre a psicoacústica e a música, é fundamental esclarecer a diferença entre ruído e tom musical. Esta distinção é particularmente importante no desenvolvimento de algoritmos de deteção de altura, uma vez que qualquer som musical complexo contém também uma determinada quantidade de ruído, que tem de ser reconhecido e tratado devidamente. Para Helmholtz, embora ruídos e tons musicais tenham regiões de interseção em diversos aspetos, os seus extremos são largamente distintos: geralmente, um ruído é acompanhado por uma alteração rápida e irregular de diferentes tipos de sensações sonoras e, por outro lado, um som musical é percecionado como uniforme, sem alteração de

sensações, resultando numa única sensação simples e regular, que se mantém inalterada durante a sua sonância (Helmholtz, 2009).

Os sons produzidos por instrumentos musicais são acompanhados de ruídos característicos, como por exemplo o som irregular produzido pela fricção do arco nas cordas de um violino ou pelo fluxo de ar no bocal de uma flauta transversal. Estes são fundamentais para que um ouvinte seja capaz de distinguir e identificar diferentes instrumentos musicais numa massa complexa de sons, tal como reconhecê-los quando ouvidos individualmente.

Em música, uma escala é um conjunto de notas musicais organizadas por altura, cuja estrutura construtiva se repete a cada oitava, que é a distância musical obtida pela duplicação ou redução para metade da frequência fundamental. A escala cromática, por exemplo, contém as doze notas separadas por meios-tons obtidas pela divisão da oitava em doze partes iguais: dó, dó sustenido ou ré bemol, ré, ré sustenido ou ré bemol, mi, fá, fá sustenido ou sol bemol, sol, sol sustenido ou sol bemol, lá, lá sustenido ou lá bemol e si. Por outro lado, um intervalo corresponde à distância entre duas notas musicais e pode ser de segunda (maior ou menor), terceira (maior ou menor), quarta (perfeita ou aumentada), quinta (perfeita ou diminuta), sexta (maior ou menor), sétima (maior ou menor) ou oitava. A quinta diminuta é enarmónica da quarta aumentada, ou seja, soam idênticas, mas escrevem-se de maneira diferente, podendo também designar-se por trítono. A escala maior, que é a base estrutural da música tonal, é construída de acordo com a seguinte sequência de intervalos de segunda a partir da nota que dá nome à escala, ou fundamental: maior, maior, menor, maior, maior, maior e menor.

Na acústica, um intervalo é encarado como a relação, ou quociente, entre as frequências dos sons que o constituem. Posto isto, o intervalo entre as notas dó e sol, que corresponde a um intervalo musical de quinta, pode ser também descrito pela relação 2/3, por exemplo. É fundamental notar que a distância musical de um intervalo não corresponde diretamente à diferença entre as frequências dos sons que o constituem, uma vez que os intervalos são percecionados de acordo com uma escala logarítmica de frequências. Este fenómeno e as suas implicações serão abordadas no subcapítulo 1.2.4.

O sistema dos cêntimos (*cents*) foi desenvolvido por Alexander Ellis, proposto pela primeira vez no artigo *Equal Semitones as a Measure of Relative Pitch* publicado em *Journal of the Society of Arts Vol. 28* em 1880, e surge com o objetivo de quantificar pequenas diferenças entre intervalos e afinações, permitindo a divisão do meio tom temperado em 100 partes iguais (cêntimos), cujo limiar mínimo de perceção depende largamente da frequência, da

amplitude, do timbre do som, da experiência musical do ouvinte e do contexto no qual é percecionado. Na zona de maior sensibilidade do ouvido, entre dó 5 (523 Hz) e dó 8 (4186 Hz), a diferença mínima percetível é de aproximadamente 6,6 cêntimos (Fyk, 1987), embora em situações ideais de laboratório este valor possa diminuir até 0,5 cêntimos (Rakowski, 1978 apud Fyk, 1987). Nas zonas de menor sensibilidade, de dó 1 (32 Hz) a dó 2 (65 Hz) e dó 8 a dó 9 (8372 Hz), a diferença mínima percetível aumenta para 50 cêntimos.

Os intervalos musicais formados entre os parciais harmónicos de um som são chamados naturais ou puros, porque descrevem uma relação extremamente aproximada das relações matemáticas dos intervalos acústicos. Por outro lado, os intervalos musicais baseados no temperamento igual consistem na divisão da oitava em doze partes iguais, comprometendo a precisão da afinação dos instrumentos musicais em relação à série dos harmónicos (figura 5), mas tornando-a mais versátil, para que possam ser tocados consonantemente em todas as tonalidades da escala cromática ocidental. Como exemplo do comprometimento referido, de acordo com Luís Henrique (2002), a diferença entre o intervalo natural e o intervalo temperado de terceira maior é de 14 cêntimos, que é um valor inequivocamente percetível.

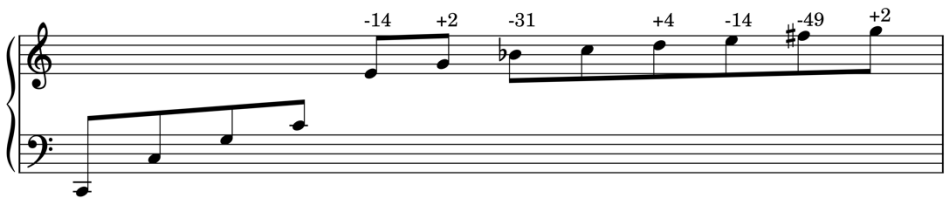

Figura 5: série dos harmónicos de dó 2 (os valores acima do pentagrama indicam a diferença em cêntimos entre o intervalo acústico e o intervalo temperado).

1.2. Relação entre altura, frequência e outras variáveis

O estudo científico da relação entre altura, frequência e outras variáveis é desenvolvido principalmente a partir do séc. XIX por físicos como Helmholtz, Ohm e Seebeck. A partir de 1950, com o desenvolvimento de sons gerados por computador, começam a ser elaborados estudos mais detalhados sobre a perceção da altura (Suits, 2019).

Como referido anteriormente, a sensação de altura é a consequência percetual de uma série de características da onda sonora, da maneira como é resolvida pelo ouvido e pelo cérebro humano e do contexto no qual é percecionada. Neste subcapítulo, procura-se expor o

conhecimento científico que existe sobre a relação entre a altura e o conjunto de fatores referidos.

1.2.1. Relação entre altura e espetro: altura virtual e timbre

Um fenómeno fundamental no estudo da perceção de altura é a altura virtual, ou ausência da fundamental, que "(…) consiste na perceção da frequência fundamental de um som, sem que ela esteja presente" (Henrique, 2002), ou seja, mesmo que a frequência fundamental de um som seja filtrada, o ouvido continua a percecionar a sua altura como se ela estivesse presente.

Consideremos por exemplo um som periódico complexo com uma frequência fundamental de 110 Hz, que corresponde à nota musical lá 2. Mesmo que esta seja filtrada, fazendo com que o seu segundo harmónico (220 Hz, lá 3) se torne o som de frequência mais baixa do seu espetro, o ouvido humano continua a percecionar a altura correspondente à nota lá 2. O mesmo acontece se este segundo harmónico for também filtrado, levando assim à conclusão de que a frequência fundamental de um som não é necessária para a perceção da altura do mesmo. Em instrumentos musicais, este fenómeno é largamente observável. A nota dó 1 corresponde à frequência fundamental 32,7 Hz. Analisando o espetro frequencial da sua sonância num piano acústico (figura 6), chega-se imediatamente à conclusão de que o nível de energia da sua fundamental é praticamente nulo e que o nível de energia do seu primeiro harmónico (65,4 Hz) é muito baixo, sendo ainda assim percecionada como dó 1.

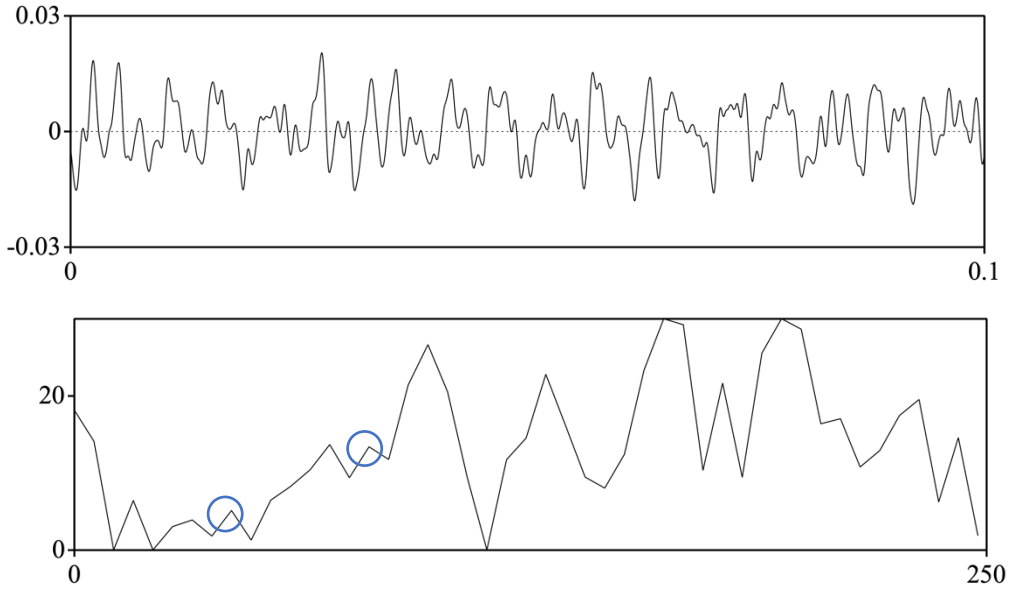

Figura 6: representação temporal e frequencial (nível de pressão sonora por frequência) da nota dó 1 de um piano Steinway & Sons (*Piano.mf.C1*, 2001). Os círculos assinalam a localização da energia da fundamental e do primeiro harmónico, respetivamente.

Schaeffer conclui que, em situações nas quais a fundamental não está presente ou tem muito pouca energia, o ouvido não a ouve, mas infere-a através da perceção das relações entre os harmónicos, ou rede harmónica, do som (Schaeffer, 2017). Partindo desse raciocínio e voltando ao exemplo do som complexo anterior, uma frequência fundamental de 110 Hz tem os seguintes quatro primeiros harmónicos: 220, 330, 440 e 550 Hz. Filtrando as frequências de 110 e 220 Hz, o ouvido perceciona as relações entre os harmónicos de 330, 440 e 550 Hz e infere a frequência fundamental de 110 Hz. O som não é confundido com outro som de frequência fundamental de 330 Hz porque este contém os harmónicos 660, 990 e 1320 Hz.

Quanto mais alta for a frequência fundamental de um som, menor a ordem dos harmónicos que se podem filtrar sem comprometer a sensação de altura que causa. Por exemplo, em frequências fundamentais até 200 Hz é possível filtrar até ao quarto ou quinto harmónicos, enquanto em frequências fundamentais a partir de 2500 Hz deixa de se poder filtrar a fundamental (Henrique, 2002).

A inferência da fundamental foi pensada inicialmente como sendo o produto da substituição da fundamental por uma distorção aural causada pelos processos mecânicos do ouvido, embora certas experiências tenham vindo a comprovar que esta não se verifica. Atualmente, é aceite que o ouvido infere a fundamental através das relações entre os parciais do som, como referido anteriormente. De modo a esclarecer este fenómeno, será necessário

descrever mais detalhadamente o processo auditivo e a relação entre a altura e as transformações auditivas nele presentes (consultar subcapítulo 1.2.5).

Não será de admirar que o fenómeno da ausência da fundamental tenha implicações profundas no desenvolvimento de algoritmos de deteção de altura, uma vez que não é suficiente identificar a frequência mais baixa de um som para prever a sensação de altura que causará.

Outro fator espectral que influencia a sensação de altura que um determinado som causa é o seu timbre. De acordo com certos estudos, muitas pessoas têm dificuldade em ignorar mudanças de brilho, ou seja, mudanças de conteúdo espectral, quando fazem julgamentos de altura (Moore and Glasberg, 1990 apud Oxenham, 2012) e mesmo ouvintes treinados musicalmente têm dificuldade em discernir pequenas diferenças de altura entre sons com timbres muito distintos (Borchert et al., 2011 apud Oxenham, 2012).

1.2.2. Relação entre altura e tempo: duração, repetição e modulação

A duração de um estímulo sonoro tem também uma influência na perceção de altura do mesmo. De acordo com Josephs (1967 apud Fyk 1987) e Piazza e Giulio (1982 apud Fyk 1987), o limite temporal de perceção de altura é de cerca de 60 ms para as frequências de 50 Hz e 10 ms para as de 1000 Hz. No artigo *Duration of Tones Required for Satisfactory Precision of Pitch Matching* de Janina Fyk, publicado em 1987 no *Bulletin of the Council for Research in Music Education No. 91*, a autora descreve uma experiência que visa clarificar a duração mínima de vários tons de diferentes frequências que causam uma sensação de altura definida e também averiguar o efeito da prática em treino auditivo na precisão de identificação de altura dos sujeitos envolvidos. Os tons escolhidos para a experiência foram sons complexos com espetros ricos e frequências fundamentais entre 110 e 1000 Hz, uma vez são os valores mais abundantemente presentes em música, com durações entre 6 e 96 ms. Os sujeitos eram alunos de música da Polónia e foram divididos em dois grupos, um com alunos no início do curso de treino auditivo e outro com alunos já com três anos e meio de experiência em treino auditivo. Estes foram propostos a fazer corresponder a altura de um estímulo sonoro à de outro através do ajuste da sua frequência. Os tons foram apresentados aos pares através de fones de ouvido e com um intervalo de 1 segundo entre cada par. A duração dos estímulos fixos foi de 6, 12, 48 ou 96 ms e a dos estímulos de correspondência foi de 1 segundo. Os estímulos fixos tiveram frequências fundamentais de 110, 220, 440 e 1000 Hz e a intensidade de todos os sons foi de

70 dB. A experiência em questão levou a uma série de conclusões. Primeiramente, demonstrou que a duração de um estímulo requerida para a correspondência de alturas com uma precisão de 25 cêntimos decresce com o aumento da frequência e que essa precisão aumenta com o aumento da duração, embora a partir dos 48 ms não haja uma melhoria na precisão dos tons agudos (Fyk, 1987). Demonstrou também que os alunos com mais experiência em treino auditivo conseguiram resultados mais precisos e que essa precisão é no geral superior em tons agudos e inferior em tons graves.

Conclui-se que a experiência é um fator determinante na capacidade de perceção de altura e está associada à repetição, que leva à familiarização do ouvinte com um determinado cenário auditivo e à adaptação mental a um conjunto de tarefas que implicam o discernimento de alturas. Numa experiência de Heinz Werner, o psicólogo demonstra que, ao fazer soar dois tons com frequências muito próximas cinco vezes consecutivas, embora os ouvintes não percecionem um intervalo ao início, este torna-se cada vez mais claro a cada repetição, reiterando a importância do contexto e do treino na perceção de altura (Schaeffer, 2017).

A maioria dos estudos experimentais acerca da perceção de altura utiliza sons com alturas fixas e constantes, embora na natureza e na música as alturas sejam frequentemente variáveis no tempo. Ainda assim, embora o processo de atribuição de altura se torne menos linear, os humanos conseguem, em algumas circunstâncias, percecionar uma única altura em tons cuja frequência fundamental oscila no tempo, como no caso do *vibrato* na música (figura 7) (Mesz & Eguia, 2009), que consiste na modulação periódica da frequência fundamental de uma nota musical que tipicamente não ultrapassa a extensão (desvio máximo acima e abaixo da frequência média) de um quarto de tom, ou 50 cêntimos. O valor de altura atribuído a este tipo de sons foi inicialmente reportado como sendo o valor médio da frequência fundamental (Shonle & Horan, 1980 apud Mesz & Eguia, 2009). No entanto, a altura percecionada não coincide com a média em modulações assimétricas. Gockel, Moore e Carlyon (2001 apud Mesz & Eguia, 2009) propuseram um modelo no qual as porções lentas de variação da fundamental contribuem mais para a altura percecionada do que as rápidas, que designaram de ponderação sensível à estabilidade (*stability-sensitive weighting*) e que sugere que a estimativa de altura se torna menos precisa quando a frequência varia rapidamente.

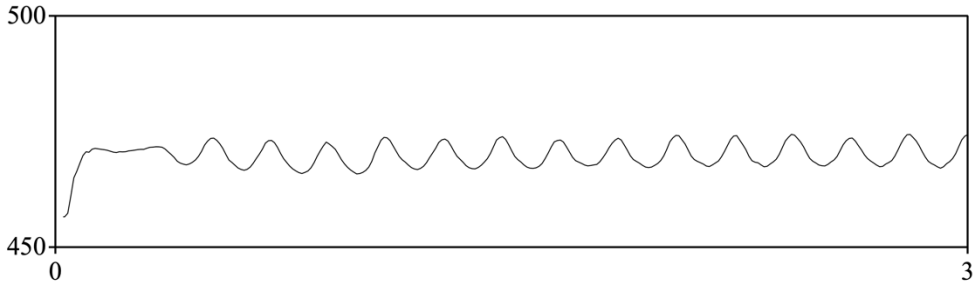

Figura 7: oscilação da frequência fundamental da nota lá sustenido 4 de um violino tocada com *vibrato* (*violinarcovibA#4*, 2008) em função do tempo. A extensão é de aproximadamente 3 Hz, que nesta banda de frequências equivale a aproximadamente um oitavo de meio tom ou 12,5 cêntimos.

1.2.3. Relação entre altura e intensidade

Uma das propriedades de uma onda mecânica que afeta a perceção de altura é a variação de intensidade (Stevens, 1935). O efeito Stevens, descrito por S. S. Stevens, demonstra que um som grave cuja intensidade aumenta se torna percetualmente mais grave e que um som agudo cuja intensidade aumenta se torna percetualmente mais agudo. Este fenómeno foi comprovado através de uma experiência que consistiu no posicionamento de um observador numa sala com tratamento acústico com a sua cabeça fixada entre um par de pequenos altifalantes perto dos seus ouvidos e na reprodução alternada de um tom padrão e de um tom de comparação com uma frequência ligeiramente diferente. O observador teve então de igualar a altura dos dois sons através da regulação da intensidade do tom de comparação, a partir de um botão que lhe foi disponibilizado inicialmente. A intensidade absoluta dos tons foi determinada por um microfone suspenso posicionado muito perto do ouvido direito do observador. Este procedimento foi repetido dez vezes para cada par de tons, sendo que a diferença de frequência entre os dois tons nunca foi superior a um valor que não permitisse a igualação das suas alturas por um ajuste de intensidade do tom de comparação superior a 25 dB. Participaram na experiência três observadores excecionalmente eficazes na discriminação de altura. Os resultados deram origem ao gráfico apresentado (figura 8), que mostra os contornos da dependência da altura em função da intensidade através da representação da percentagem de mudança de altura em função do nível de intensidade em decibéis, com a indicação da frequência do som referente a cada curva. Quanto maior o declive da curva, mais a perceção de altura do som com a frequência em questão depende da intensidade. O declive ascendente ou descendente determina se a altura é percecionada como mais aguda ou mais grave,

respetivamente, face a um aumento da intensidade. Como se pode observar, este efeito é mais acentuado nas frequências altas e baixas, sendo pouco notável nas médias.

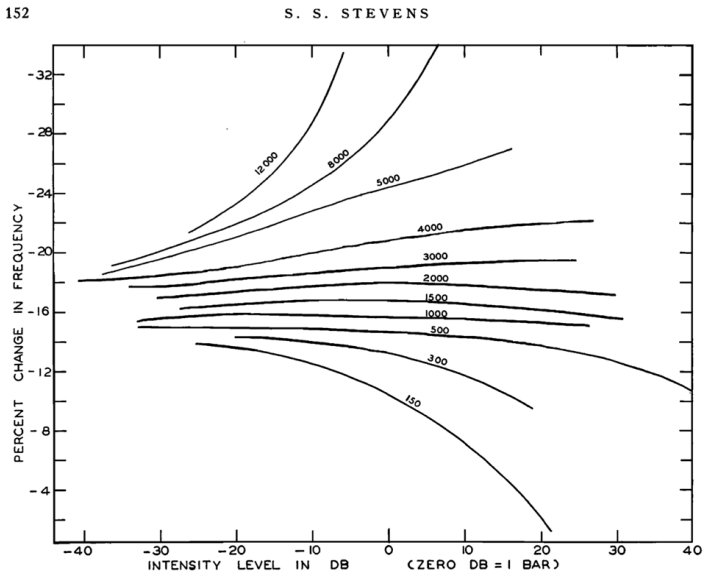

Figura 8: percentagem de mudança de altura por nível de pressão sonoro por frequência (Stevens, 1935).

Teorias anteriores admitiam que estas alterações de perceção de altura em função da intensidade da onda existiam apenas numa direção e procuraram explicá-las pela assunção de que as frequências naturais da membrana basilar aumentam com a amplitude (Zurmuhl, 1930 apud Stevens, 1935) ou de que um aumento de intensidade causa uma sensação de aumento de altura porque resulta numa maior proporção de harmónicos subjetivos (Troland, 1930 apud Stevens, 1935). Os resultados desta experiência, que evidenciam uma mudança de declive negativo para positivo no registo das frequências médias, sugerem que a altura de um tom se afasta da região de maior sensibilidade auditiva quando a intensidade é aumentada e que se aproxima quando é diminuída.

1.2.4. Relação entre altura e música: circularidade, intervalos musicais e contexto musical

De acordo com Stevens e Volkmann (1940), a altura é uma das características de um tom e difere da frequência na medida em que é determinada pela resposta psicológica do ouvinte humano, enquanto a última é medida com o auxílio de instrumentos, ou seja, é um

fenómeno físico independente do ouvinte. A altura é fundamental na música na medida em que define diretamente as notas musicais (que podem ou não pertencer ao sistema temperado ocidental), enquanto intervalos de altura definem intervalos musicais e contornos de altura definem uma série de efeitos musicais, como o *vibrato* ou o *glissando*.

Uma altura pode ser descrita por duas dimensões: croma e posição, que existem por consequência da circularidade da sensação de altura, ou seja, do facto de que quando um som de frequência ascendente atinge um intervalo de oitava, que corresponde à relação matemática 1/2, é novamente identificada como a nota musical de partida e assim sucessivamente em todas as mudanças de oitava, ascendentes ou descendentes. O croma é, então, referente à sensação de nota musical de um tom independentemente da oitava na qual se encontra e a posição é referente apenas à oitava na qual se encontra (Henrique, 2002). As notas dó 2 e dó 3, por exemplo, partilham o mesmo croma, mas têm posições diferentes. As notas dó 2 e sol 2 partilham a mesma posição, mas têm cromas diferentes. A escala de Shepard (Shepard, 1964 apud Henrique, 2002) é uma demonstração que prova esta circularidade. Esta consiste numa sequência de tons musicais que dá a ilusão de ascender em altura indefinidamente, sem nunca sair do espetro audível. Este fenómeno ocorre porque os tons da escala cromática utilizados são sons complexos com 10 parciais, cada um à distância de uma oitava do seguinte, que têm sempre mais energia nos parciais médios do que nos extremos, causando uma série de aparecimentos e desaparecimentos graduais dos mesmos à medida que se aproximam e afastam desse registo frequencial. Jean-Claude Risset, compositor francês, criou uma versão contínua desta escala chamada Glissando Shepard-Risset, na qual as alturas dos tons cromáticos da escala estão conectadas continuamente.

Existe uma relação logarítmica entre a banda de frequências e a distância dos intervalos musicais percecionados: quanto mais alta é a banda de frequências, mais próximos são os intervalos musicais nela contidos. Tomemos por exemplo as notas dó 2, sol 2, dó 5 e sol 5, que correspondem aproximadamente às frequências, respetivamente, de 65, 98, 523 e 784 Hz. Os pares dó 2 e sol 2 e dó 5 e sol 5 descrevem ambos um intervalo musical de quinta e uma relação de 2/3, embora apresentem distâncias de frequência diferentes entre si: 33 e 261 Hz, respetivamente.

De modo a clarificar este fenómeno, foi desenvolvida uma escala que tem por unidade o mel (figura 9), que "corresponde a uma tentativa de encontrar uma escala semelhante à escala de tons. (...) A escala é estabelecida de tal modo que duplicando o número de mel duplica a

sensação subjetiva de altura" (Henrique, 2002). Como tal, um som de 1 kHz corresponde a 1000 mel e a escala varia de 0 a aproximadamente 3500, correspondendo às frequências de 20 Hz a 20 kHz.

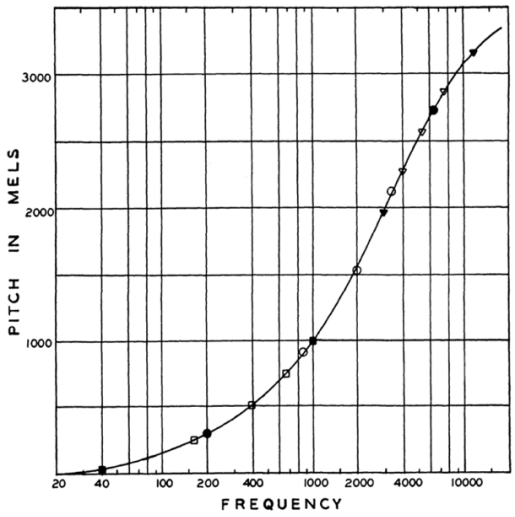

Figura 9: escala de mels (altura em mels por frequência) (Stevens & Volkmann, 1940).

Os dados apresentados foram determinados através dos métodos experimentais descritos no artigo *The Relationship of Pitch to Frequency: a Revised Scale*, publicado no *The American Journal of Psychology* (Stevens & Volkmann, 1940). Foi desenvolvido um piano elétrico com vinte teclas, que causam a produção de 20 tons simples num altifalante, e com vinte potenciómetros, que permitem a regulagem das frequências dos tons produzidos. Na experiência em questão, foram escolhidas cinco teclas: a tecla mais grave produzia uma frequência de 200 Hz e a mais aguda produzia uma frequência de 6500 Hz. O ouvinte foi então proposto a ajustar as frequências dos tons referentes às três teclas intermédias de modo a criar intervalos de distância de altura iguais entre todos os tons. O mesmo método foi repetido para as frequências extremas de 40 a 1000 Hz e de 3 a 12kHz.

Ainda sobre intervalos musicais, embora a teoria musical proponha uma categorização estanque da consonância e dissonância dos mesmos, há uma série de fatores contextuais e psicoacústicos que influenciam esta categorização, sendo um dos mais óbvios o registo e consequentemente a informação espetral do par de tons em questão, como explica Sethares (1998 apud Henrique, 2002): "as sensações de consonância e dissonância não são qualidades inerentes aos intervalos; elas dependem do espetro e do timbre". Na teoria musical, os intervalos de oitava, quinta e quarta são considerados consonâncias perfeitas (embora a quarta seja por

vezes considerada ambígua: consonante em alguns contextos, dissonante noutros) os intervalos de terceira maior ou menor e sexta maior ou menor são considerados consonâncias imperfeitas e os intervalos de segunda maior ou menor, quarta aumentada/quinta diminuta e sétima maior ou menor são considerados dissonâncias. Por outro lado, de um ponto de vista psicoacústico, dois sons são considerados consonantes quando a diferença entre as frequências dos mesmos é maior do que a banda crítica em questão e dissonantes quando é menor (Rasch & Plomp apud Henrique, 2002). Posto isto, uma quinta perfeita soada no registo mais grave de um piano, como lá 0 e mi 1, por exemplo, será tendencialmente mais dissonante que uma terceira maior soada no registo médio, como dó 4 e mi 4, por consequência das suas distintas distribuições de parciais.

No que toca à influência do contexto musical na perceção de altura, foram desenvolvidos vários estudos. Na publicação *Familiar Tonal Context Improves Accuracy of Pitch Interval Perception*, por exemplo, Graves e Oxenham (2017) propõem que o contexto tonal, ou seja, as hierarquias tonais entre as alturas discretas de uma tonalidade musical têm uma influência na precisão da perceção de intervalos de altura, que pode ser o resultado das expetativas musicais adquiridas pelo ouvinte por ter sido exposto a música tonal durante toda a sua vida. Estas podem manifestar-se de duas formas: uma altura que é expectada por uma hierarquia tonal pode causar um tempo de resposta mais rápido, uma vez que é preciso menos tempo para reagir a um acontecimento previsível, ou porque pode haver uma ativação antecipada das alturas expectadas. Analisando os resultados da experiência, os autores concluem que "os resultados sugerem que embora os contextos tonais consigam gerar expetativas fortes, não produzem aprimoramentos substanciais nas representações percetuais de altura e intervalos de altura" (Graves & Oxenham, 2017).

De acordo com Francès (apud Schaeffer, 2017), um vetor musical, ou seja, a direção de uma frase melódica tem também uma influência considerável na perceção dos intervalos musicais. Uma nota musical abaixada em alguns cêntimos em relação à escala temperada será mais bem tolerada pelo ouvinte se constituir a nota alvo de uma apogiatura ou frase melódica descendente, ao contrário de uma ascendente, uma vez que no primeiro caso acompanha o vetor musical, enquanto no segundo este é contrariado. Note-se que a cultura musical do ouvinte terá também uma larga influência na tolerância de notas que não estejam afinadas de acordo com o temperamento igual.

1.2.5. Relação entre altura e transformações auditivas e neurológicas

O processo auditivo introduz uma série de não-linearidades nos sinais sonoros percecionados, uma vez que se trata de um conjunto de processos mecânicos e neurais. Neste subcapítulo, procura-se explicar detalhadamente a relação entre estes mecanismos e as distorções aurais que criam.

Quando um som entra na cóclea, esta é excitada em diferentes regiões ao longo da sua extensão, de acordo com as frequências do som em questão. Este processo é conhecido por tonotopia e corresponde a um mapeamento de frequência e localização (Oxenham, 2012). Este termo fisiológico estende-se também ao mapeamento das regiões do cérebro nas quais diferentes frequências são processadas. O cientista húngaro Georg von Békésy recebeu o Prémio Nobel da Fisiologia ou Medicina em 1961 pelo seu extenso trabalho de investigação acerca da tonotopia, comprovando-a experimentalmente. As consequências percecionais deste fenómeno são largamente observáveis e passam principalmente pela existência de filtros auditivos, que têm na sua base o processo de filtragem coclear (Shera et al., 2002 apud Oxenham, 2012). Estes filtros podem ser representados por um padrão de excitação, que é, essencialmente, uma "(...) representação esquemática da ativação mecânica da partição coclear ou atividade neural em função da frequência característica" (Glasberg & Moore, 1990 apud Oxenham, 2012) e são, essencialmente, uma consequência da seletividade de frequências do sistema auditivo, ou seja, da propriedade que o torna afinado para responder mais eficazmente a algumas frequências do que a outras. Posto isto, "as primeiras etapas do processamento auditivo são geralmente descritas como consistindo num conjunto de filtros auditivos com diferentes frequências centrais" (American Psychological Association, n.d.).

Quando o parcial de um som complexo é captado pela região da cóclea correspondente, este pode ser resolvido ou não-resolvido. Um parcial resolvido é aquele que pode ser representado exclusivamente por um único filtro e dá origem a uma onda filtrada semelhante ao som puro da mesma frequência, enquanto um parcial não-resolvido interage com outros parciais dentro do mesmo filtro e dá origem a uma onda complexa que reflete essa interação (Oxenham, 2012). Uma série de fenómenos relacionados com a perceção de altura podem ser explicados por esta resolução harmónica, ou resolução de parciais.

O conceito de banda crítica advém do facto de que uma oscilação que excita uma determinada região da membrana basilar abrange um número de terminações nervosas em torno

da região de maior excitação, que corresponde ao máximo de amplitude da onda. Esta é então a banda de frequências à qual correspondem os pontos da membrana basilar afetados e varia de acordo com as frequências em questão (Henrique, 2002). Quando dois sons puros se encontram dentro da mesma banda crítica, dão origem a batimentos e rugosidades. O batimento é o fenómeno percetual que causa uma aparente modulação de amplitude no sinal cuja frequência é mais alta quanto maior for a diferença de frequências entre os sons envolvidos. A rugosidade é a "sonoridade dura, agressiva, semelhante a um ruído" (Henrique, 2002) descrita por intervalos acústicos pequenos, mas grandes o suficiente para não descreverem batimentos. Os *clusters* (intervalos musicais de segunda menor ou maior) apresentam, tendencialmente, esta sonoridade. Será importante salientar que, se cada som for ouvido por cada um dos ouvidos em separado, esta rugosidade desaparece, uma vez que cada som excita apenas a membrana basilar do seu respetivo lado, o que leva a concluir que há interferência na perceção apenas quando os sinais se sobrepõem.

Zwicker (1961) propõe uma escala que relaciona as 24 bandas críticas do ouvido humano com a frequência, conhecida por escala de barks (figura 10), nomeada em homenagem ao físico Heinrich Barkhausen, que indica a banda de frequências contida em cada banda crítica. Uma diferença de 1 bark corresponde à largura de 1 banda crítica e também ao intervalo de aproximadamente 100 mels. Será importante salientar que embora as bandas críticas apresentem uma largura fixa, a sua posição na escala de frequências é variável.

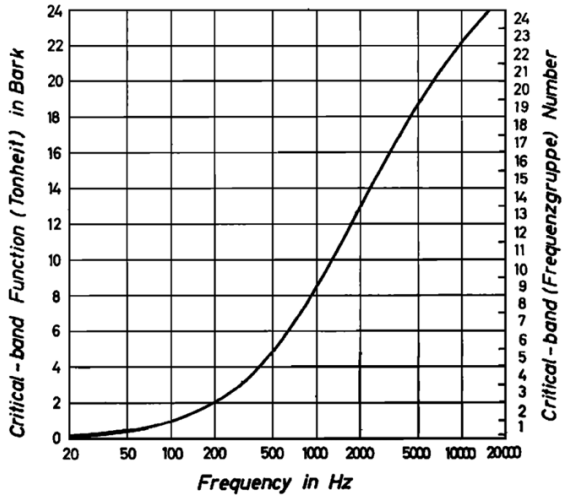

Figura 10: escala de barks (função de banda crítica em barks por frequência) (Zwicker, 1961).

No que toca aos processos neurais envolvidos na audição, também estes induzem não-linearidades na perceção humana da altura de um som. Citando Oxenham (2012), "os neurónios no nervo auditivo têm uma maior probabilidade de disparar numa fase do ciclo da onda do que noutras fases", originando o fenómeno designado por fixação de fase (*phase locking*), que é responsável pela capacidade de percecionar pequenas diferenças temporais entre os dois ouvidos, que é por sua vez fulcral na localização de fontes sonoras no espaço. Expecta-se que a fixação de fase tenha também um papel na codificação da periodicidade de um estímulo e, consequentemente, na perceção da sua altura: a perda da capacidade de reconhecimento de melodias em sons simples acima de 4 a 5 kHz foi interpretada como a consequência do degradamento da fixação de fase nas frequências altas (Attneave and Olson, 1971; Moore, 1973 apud Oxenham, 2012). Esta interpretação sugere que a perceção de altura depende de informação temporal no nervo auditivo, embora haja indicações de que esta possa não ser suficiente e de que seja também preciso informação tonotópica. Uma proposta para a explicação da maneira como a altura é percecionada consiste na análise dos padrões temporais gerados pelos harmónicos não-resolvidos dos sons (Schouten et al., 1962 apud Oxenham, 2012).

De acordo com uma série de estudos, a resposta de seguimento de frequência, ou seja, a medida de atividade de fixação de fase no cérebro, que reflete a precisão de codificação de altura, é mais forte em músicos do que em pessoas sem treino musical (Wong et al., 2007 apud Oxenham, 2012) e em falantes de línguas tonais do que em não-falantes (Krishnan et al., 2005 apud Oxenham, 2012).

2. Algoritmos de deteção de altura

Os algoritmos de deteção de altura são algoritmos que têm como objetivo estimar a altura de um sinal sonoro através de diferentes estratégias de processamento de sinal que visam identificar ou inferir a frequência fundamental do mesmo. São utilizados principalmente nas áreas da fonética e análise de discurso e na música e recuperação de informação musical (consultar subcapítulo 2.3).

Um tom musical pode ser monofónico ou polifónico, isto é, pode conter apenas uma altura em cada instante, como uma linha melódica tocada numa flauta ou num clarinete, ou conter duas ou mais alturas em simultâneo, como um intervalo ou um acorde tocado num piano ou numa guitarra. Como tal, os algoritmos podem separar-se em dois grupos, monofónicos e polifónicos, de acordo com o tipo de sinal que estão otimizados a analisar.

Outra categorização destes algoritmos está relacionada com o domínio no qual operam, que pode ser temporal, espetral, misto, que tem simultaneamente componentes de análise temporal e espetral, ou através da modelação do sistema auditivo humano.

Para além da precisão dos resultados de um algoritmo de deteção de altura, há que ter em conta também os recursos computacionais que este utiliza para funcionar. Se um método de deteção for computacionalmente caro, ou seja, exigir uma grande quantidade de computações, terá um tempo de resposta mais lento que o pode tornar desvantajoso em situações nas quais uma resposta rápida seja imperativa, como no seguimento de altura em tempo real, por exemplo.

Uma questão fundamental na identificação de altura de melodias cantadas com letra ou vocalizos, ou seja, quando existe uma componente fonética associada, é a distinção entre os elementos vozeados e desvozeados de um sinal. Os elementos vozeados são fonemas que são produzidos através da vibração das cordas vocais e transportam toda a informação relevante acerca da altura do som cantado ou falado, enquanto os elementos desvozeados são essencialmente ruído, na definição espetral do termo, que faz com que não tenham uma altura definida e sejam, por isso, negligenciáveis na identificação da mesma. Na língua portuguesa, as vogais e algumas consonantes como [b], [d] e [g] são vozeadas, enquanto outras como [p], [t] e [c] são desvozeadas (Veloso, 1997). Portanto, um algoritmo otimizado para a deteção de altura da voz humana, quer no discurso, quer na música, deverá conseguir identificar e negligenciar as componentes desvozeadas do sinal.

Note-se que os algoritmos referidos operam digitalmente e são, por isso, aplicados em sinais discretos com uma determinada taxa de amostragem, que corresponde ao número de amostras por segundo do sinal. Um sinal com uma taxa de amostragem de 44.1 kHz, que corresponde à taxa de amostragem do áudio num *compact disc* (CD), terá 44100 amostras por segundo, por exemplo.

2.1. Algoritmos monofónicos

2.1.1. Domínio temporal

2.1.1.1. Taxa de cruzamento de zero (*zero-crossing rate*)

A taxa de cruzamento de zero é um dos algoritmos monofónicos mais simples e consiste na análise do número de vezes que o sinal cruza o nível de referência de amplitude zero (figura 1). Esta técnica é computacionalmente barata, ou seja, utiliza poucos recursos de processamento, mas não é muito precisa. Ao lidar com sinais com muito ruído, com parciais mais fortes que a fundamental ou com oscilações à volta do eixo zero, este método tem resultados fracos (Amado & Filho, 2008). Por isso, pode-se concluir que este algoritmo é confiável apenas na análise de sinais muito simples, como ondas sinusoidais, e não em tons musicais, que descrevem uma maior complexidade e podem apresentar uma grande disparidade de rácios energéticos de parciais ou cruzar o zero mais do que duas vezes ao longo do seu ciclo.

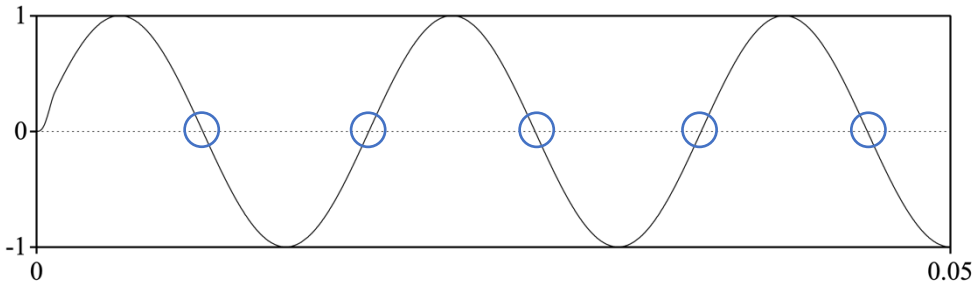

Figura 1: pontos de cruzamento de zero (assinalados pelos círculos) de um sinal sinusoidal de 55 Hz.

Sendo *s(n)* um sinal discreto de duração *N* e $\coprod\{x\}$ um operador que devolve 1 se o argumento *x* for verdadeiro e 0 se for falso, a taxa de cruzamento de zero *zcr* é dada por:

$$zcr = \frac{1}{N}\sum_{n=0}^{N-1}\coprod\{s(n).s(n-1) < 0\}$$

A frequência do sinal *F* é então calculada pela seguinte equação na qual *fs* é a taxa de amostragem do sinal (o segundo membro é dividido por 2 porque são precisos dois cruzamentos de zero para que um sinal periódico reinicie um ciclo):

$$F = \frac{zcr.fs}{2}$$

De acordo com Amado e Filho (2008), uma maneira de tornar este algoritmo mais preciso para a deteção de altura de sons musicais passa por deslocar o sinal de acordo com um valor de amplitude calculado a partir da média de amplitude do sinal, ou limiar. Esta deslocação é determinada separadamente para a parte positiva e para a parte negativa do sinal. Outra otimização consiste na contabilização de apenas as transições do sinal de amplitude negativa para positiva. Posto isto, o algoritmo proposto é dado pelas seguintes equações, nas quais *L* é o limiar, $s_p(n)$ o sinal com deslocação positiva, $s_n(n)$ o sinal com deslocação negativa, zcr_p a taxa de cruzamento de zero de $s_p(n)$, zcr_n a taxa de cruzamento de zero de $s_n(n)$, F_{sp} a frequência do sinal $s_p(n)$, F_{sn} a frequência do sinal $s_n(n)$ e F_s a frequência determinada do sinal:

$$L = 1.2\sum_{n=0}^{N-1}|s(n)|$$

$$s_p(n) = s(n) - L \qquad 0 \le n \le N-1$$

$$zcr_p = \frac{1}{N}\sum_{n=0}^{N-1}\coprod\{s_p(n-1) < 0 < s_p(n)\}$$

$$F_{sp} = zcr_p.fs$$

$$s_n(n) = s(n) + L \qquad 0 \leq n \leq N-1$$

$$zcr_n = \frac{1}{N}\sum_{n=0}^{N-1} \coprod \{s_n(n-1) < 0 < s_n(n)\}$$

$$F_{sp} = zcr_n . fs$$

$$F_s = \frac{F_p + F_n}{2}$$

Outra solução de otimização consiste na aplicação de um filtro passa-banda no sinal que exclua todas as frequências acima ou abaixo da banda que contém as suas frequências fundamentais. Esta estratégia contorna o problema que a estratégia descrita no parágrafo anterior visa resolver, mas requer conhecimento a priori acerca do sinal e dos resultados esperados.

O SuperCollider tem na sua biblioteca uma classe chamada ZeroCrossing que consiste numa versão não otimizada deste algoritmo.

2.1.1.2. Filtragem em pente (*comb filtering*) e autocorrelação

A deteção de altura através de filtragem em pente ou de autocorrelação é um método que consiste na comparação de um sinal com uma cópia dele mesmo desfasada por consecutivos intervalos de tempo, ou atrasos. Ao ser desfasado consecutivamente, um sinal tem um maior grau de semelhança consigo mesmo quando o valor do atraso corresponde ao seu período.

A função de filtragem em pente descreve mínimos nos valores que correspondem ao período do sinal. No caso de haver mais do que um mínimo devido à presença de harmónicos mais energéticos do que a fundamental, o mínimo que assinala o período do sinal é geralmente o mais pontudo, ou seja, aquele cujos valores adjacentes são mais afastados do mesmo. A função de filtragem *f(Δt)* do sinal *s(t)* é dada pela seguinte equação, na qual *Δt* é o valor do atraso em segundos (Haken, 2020):

$$f(\Delta t) = \sum |s(t + \Delta t) - s(t)|$$

A função de autocorrelação *r(Δt)* (figura 2) é bastante similar ao processo descrito anteriormente, mas existe agora uma multiplicação em vez de uma subtração:

$$r(\Delta t) = \sum s(t + \Delta t).s(t)$$

Como consequência desta mudança, a função procura agora um máximo em vez de um mínimo e é também mais eficiente computacionalmente, uma vez que os microprocessadores tipicamente utilizados no processamento digital de sinal são otimizados para calcular multiplicações (Haken, 2020).

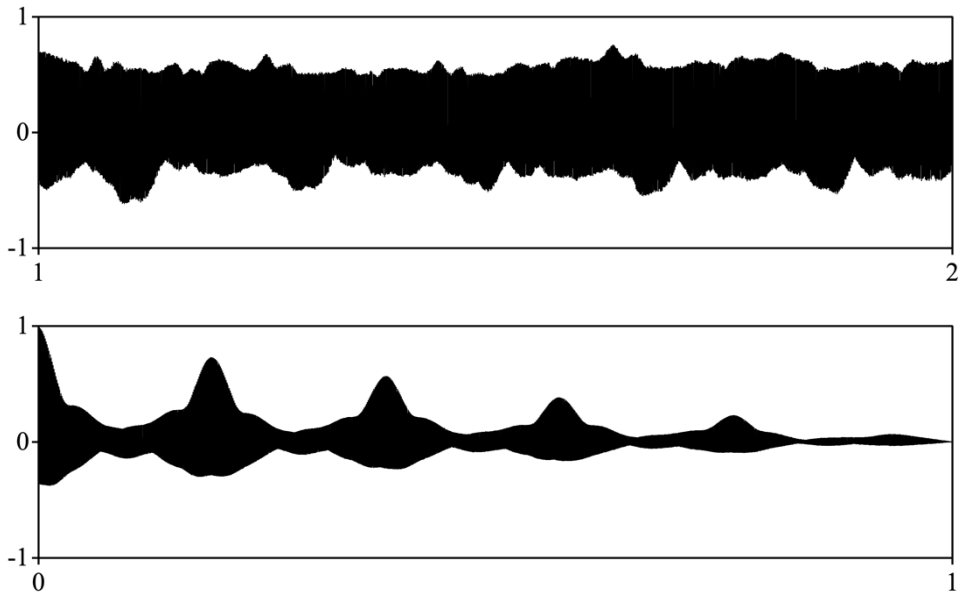

Figura 2: representação temporal de parte do sinal da nota lá sustenido 4 de um violino (*violinarcovibA#4*, 2008) e representação da sua função de autocorrelação (coeficiente de autocorrelação por tempo), na qual os máximos assinalam os múltiplos do período da fundamental.

Embora estes métodos tenham tendencialmente resultados mais precisos que a taxa de cruzamento de zero, dão ainda origem a algumas imprecisões, como erros de oitavas, ou seja, a identificação correta de cromas, mas incorreta de posições, para além de serem incapazes de detetar alturas corretamente se a fundamental estiver ausente e de serem potencialmente muito caros computacionalmente, uma vez que calculam um somatório para vários valores de *(Δt)* dentro de um determinado intervalo, que pode ser bastante largo, dependendo do problema a resolver. Como tal, um dos desafios latentes à utilização deste algoritmo é a escolha de um intervalo de valores de *(Δt)* com um tamanho apropriado, que deve conter idealmente 2 a 3

períodos completos do sinal a analisar (Rabiner, 1976), mas não deve ser demasiado extenso para não atrasar a computação.

Ainda assim, a função de autocorrelação serve de base para inúmeros algoritmos de deteção de altura, incluindo alguns que serão descritos nos subcapítulos seguintes. O sucesso deste método deve-se ao facto da sua computação ser feita diretamente na onda, de ser relativamente simples, ainda que demorada, porque requer apenas um multiplicador e um acumulador como elementos computacionais, e também de ser amplamente insensível a distorções de fase no sinal (Rabiner, 1976).

A classe Pitch da biblioteca do SuperCollider é um exemplo de um algoritmo de deteção de altura baseado na autocorrelação.

2.1.1.3. Estimativa de semelhança máxima (*maximum likelihood estimate*)

Noll (1969) descreve um algoritmo baseado na autocorrelação, derivado por David Slepian, no qual um sinal *r(t)* de duração *T* e período *τ0* é dividido em *N* intervalos de duração τ de modo que $T = N\tau + b$ (figura 3):

$$r(t,\tau) = \begin{cases} \dfrac{1}{N+1}\displaystyle\sum_{n=0}^{N} r(t+n\tau) & 0 \le t < b \\ \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1} r(t+n\tau) & b \le t < \tau \end{cases}$$

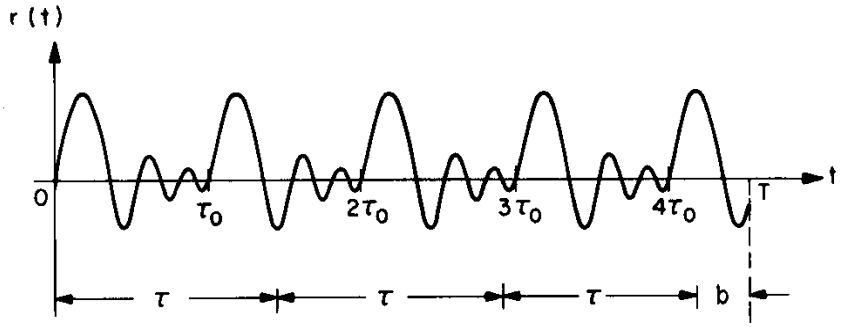

Figura 3: representação temporal de uma onda sonora dividida de modo que $T = N\tau + b$ (Noll, 1969).

A função de estimativa de semelhança máxima *J(τ)* é máxima quando τ é igual ao período do sinal *τ0* e é dada por:

$$J(\tau) = (N+1)\int_0^b r^2(t,\tau)\,dt + N\int_b^\tau r^2(t,\tau)dt$$

Um dos problemas deste algoritmo é a existência de vários máximos de *J(τ)* que tornam a identificação de *τ0* menos clara, uma vez que a função também maximiza quando $\tau = m\tau 0$ se *N* for múltiplo de *m*.

2.1.1.4. YIN e pYIN

Tanto no discurso como na música, um sinal sonoro é raramente perfeitamente periódico, pelo que um algoritmo de deteção de altura eficaz tem de conseguir lidar com desvios de periodicidade de um modo consistente, que podem ser, por exemplo, modulações de amplitude ou frequência. O método YIN, apresentado por De Cheveigné e Kawahara (2002), é essencialmente um algoritmo de autocorrelação com uma série de otimizações que visam evitar possíveis erros do processo de deteção.

A primeira otimização consiste na utilização de uma função diferencial $d_t(\tau)$. Sendo r_t a função de autocorrelação de um sinal x_t de período τ, esta pode ser dada por:

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau)$$

De acordo com os resultados experimentais, a utilização desta função diferencial, em vez da função de autocorrelação inalterada, fez com que a taxa de erro decrescesse de 10% para 1,95% na identificação das alturas de uma base de dados de discurso humano (De Cheveigné et Kawahara, 2002). Este aumento de precisão é devido ao facto da função de autocorrelação ser bastante sensível a mudanças de amplitude, uma vez que um aumento da amplitude de um sinal ao longo do tempo faz com que os picos desta função cresçam com o atraso e com que o algoritmo escolha, portanto, um máximo incorreto e cometa o erro de identificar uma altura abaixo da correta. Por outro lado, a função diferencial é imune a estas alterações de amplitude.

De modo a contornar o problema da possível existência de mínimos relativos ao primeiro harmónico de um sinal mais profundos que os mínimos relativos à sua fundamental, a função diferencial proposta anteriormente é substituída por uma função de diferença cumulativa mediana normalizada (*cumulative mean normalized difference*) $d'_t(\tau)$, dada por:

$$d'_t(\tau) = \begin{cases} 1, & se\ \tau = 0 \\ d_t(\tau) \Big/ \left[(1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & senão \end{cases}$$

Esta difere da anterior porque tem a sua origem em 1 em vez de 0, mantém-se com valores grandes em valores de atraso pequenos e apenas assume valores inferiores a 1 quando $d_t(\tau)$ assume valores abaixo da média. Face a esta alteração, a taxa de erro desce de 1,95% para 1,69% (De Cheveigné et Kawahara, 2002).

A proposta seguinte consiste em definir um limite absoluto e escolher o valor mais pequeno de τ que corresponde a um mínimo *d'* inferior a este limite ou escolher o mínimo global caso não seja encontrado nenhum valor de τ que satisfaça esta condição, reduzindo a taxa de erro para 0,78% na utilização de um limite de 0,1.

Se o período do sinal não for múltiplo do seu período de amostragem, o algoritmo pode estimar um valor errado até metade do período de amostragem. Para resolver este problema, o sinal é interpolado parabolicamente: cada mínimo de $d'_t(\tau)$ e os seus valores adjacentes traçam uma parábola que é por sua vez utilizada na seleção do mínimo. Neste caso, e como consequência desta interpolação, a taxa de erro baixou apenas 0,01%, embora tenha obtido melhores resultados em sons sintéticos com frequências fundamentais altas (De Cheveigné et Kawahara, 2002).

Finalmente, através de um processo de seleção da melhor estimativa local, a taxa de erro baixou de 0,77% para 0,5%. Este consiste na procura do mínimo de $d'_{\theta}(T_{\theta})$ para cada valor de tempo t onde θ assume um valor entre $t - T_{max}/2$ e $t + T_{max}/2$, onde T_{θ} corresponde à estimativa do período em θ e onde T_{max} é o maior período esperado. Com base nesta estimativa inicial, o algoritmo é executado novamente num intervalo de valores restringido.

O método pYIN, ou YIN probabilístico, é uma modificação do algoritmo descrito anteriormente desenvolvido por Mauch e Dixon em 2014. Uma das desvantagens do método YIN é o facto de estimar apenas uma frequência fundamental por amostra. O algoritmo pYIN

transforma o anterior num algoritmo que produz várias frequências candidatas com probabilidades associadas, em vez de uma única estimativa de frequência. A estimativa final é de seguida obtida a partir dessas probabilidades num modelo oculto de Markov (modelo estatístico). Comparado com YIN, pYIN consegue obter resultados mais precisos de acordo com os testes feitos pelos mesmos autores (figura 4) (Mauch et Dixon, 2014).

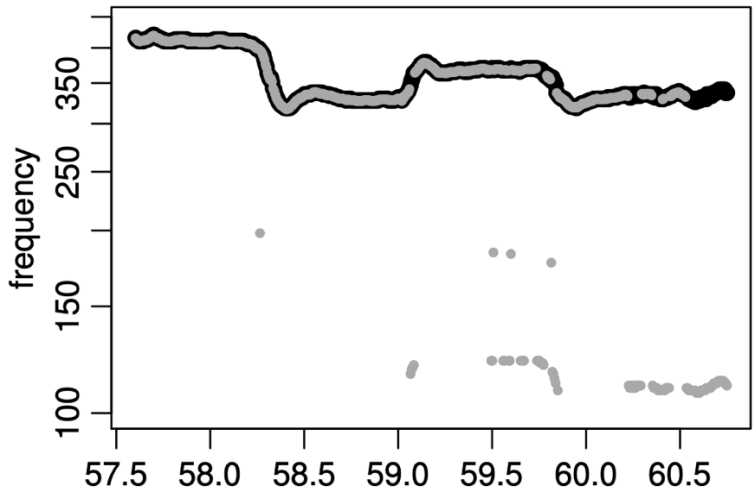

Figura 4: estimativas de altura (frequência fundamental por tempo) dos algoritmos pYIN a preto e YIN a cinzento em canto humano: últimas quatro notas do tema "Parabéns" (Mauch et Dixon, 2014). Como se pode observar neste caso, o algoritmo YIN descreveu erros de oitavas, enquanto o pYIN não.

2.1.1.5. Tartini

Em 2005, Philip McLeod e Geoff Wyvill propuseram um algoritmo de deteção de altura chamado Tartini, em homenagem ao violinista e compositor Guiseppe Tartini, no artigo *A Smarter Way to Find Pitch*. Este foi de seguida implementado na biblioteca de extensões do SuperCollider.

Este algoritmo iguala os níveis do sinal de acordo com a sensibilidade do ouvido humano, ou seja, de acordo com os contornos de volume igual (descritos no capítulo 1). No que toca ao seu processo de deteção, o algoritmo Tartini consiste numa versão normalizada da função de diferença quadrada (figura 5), que é uma função semelhante à de autocorrelação, seguida de um algoritmo de escolha de máximos.

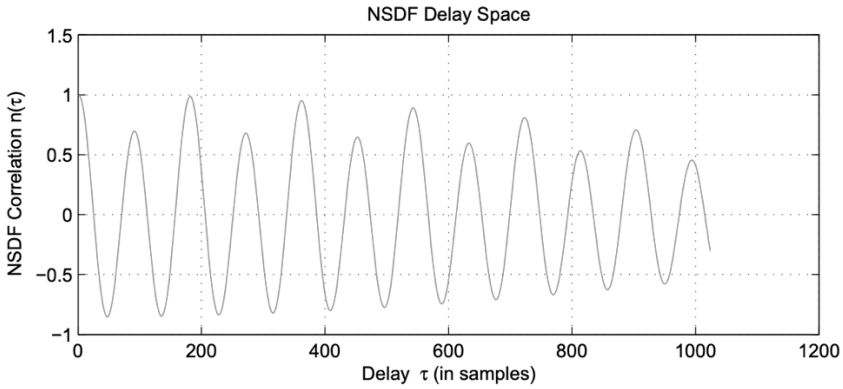

Figura 5: representação da função de diferença quadrada normalizada de um sinal cuja frequência fundamental tem um período de 190 amostras (McLeod et Wyvill, 2005).

A função de diferença quadrada $m_t(\tau)$ e a função de diferença quadrada normalizada $n_t(\tau)$, na qual $r_t(\tau)$ é a função de autocorrelação do sinal, são dadas por:

$$m_t(\tau) = \sum_{j=t}^{t+W-\tau-1} (x_j^2 + x_{j+\tau}^2)$$

$$n_t(\tau) = \frac{2r_t(\tau)}{m_t(\tau)}$$

Por fim, a escolha de máximos consiste na seleção dos máximos mais prováveis de representar a frequência fundamental. Para isso, em primeiro lugar, apenas o maior máximo entre cada cruzamento de zero positivo e cada cruzamento de zero negativo é contabilizado e o máximo referente ao atraso zero é excluído. É utilizada interpolação parabólica para determinar as posições dos máximos com maior precisão. De seguida, a partir dos máximos válidos, define-se um limite que assume o valor do maior máximo multiplicado por uma constante, que tem de ser um valor alto o suficiente para evitar a escolha de máximos causados por harmónicos muito energéticos em relação à fundamental, e escolhe-se o primeiro máximo que ultrapassa este limite, cujo valor de atraso é atribuído ao período da frequência fundamental do sinal.

2.1.1.6. CREPE

O algoritmo CREPE, acrónimo de Convolutional Representation for Pitch Estimation, proposto por Kim, Salamon, Li e Bello em 2018, é baseado numa rede neural convolucional que opera no domínio temporal de um sinal sonoro e apresenta resultados igualmente bons ou melhores que o algoritmo pYIN. Este algoritmo consegue uma precisão de acima de 90% na estimativa de altura, aquando da admissão uma margem de erro de 10 cêntimos (Kim, Salamon, Li et Bello, 2018).

Uma rede neural convolucional é um tipo de rede neural que utiliza convolução, que é uma operação matemática aplicada a duas funções que é definida pela integral do produto das duas após a reflexão em relação ao eixo das ordenadas e o desfasamento de uma delas em pelo menos em uma das camadas da sua arquitetura.

Um algoritmo de deteção de altura tem de ter uma determinada resistência a ruído para conseguir lidar eficazmente com certas situações, tais como a análise de discurso por comunicação telefónica ou a análise de música num contexto de performance ao vivo, ambas particularmente ruidosas. O algoritmo CREPE foi comparado, neste aspeto, aos algoritmos pYIN e SWIPE (descrito no subcapítulo 2.1.2.5): os resultados podem ser consultados abaixo (figura 6). CREPE teve consistentemente a melhor prestação, à exceção da experiência com ruído castanho (ruído cuja intensidade espetral diminui 6 dB por oitava). No que toca à sua precisão de deteção de altura, este também conseguiu melhores resultados que os anteriores, especialmente na deteção de sons com timbres complexos (Kim, Salamon, Li et Bello, 2018).

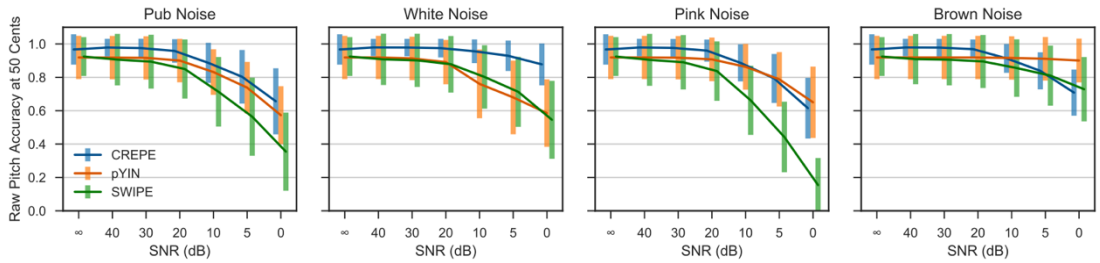

Figura 6: precisão da deteção de altura, de 0 a 1, dos algoritmos CREPE, a azul, pYIN, a laranja, e SWIPE, a verde, em função da intensidade de diferentes tipos de ruído adicionados ao sinal: *pub* (ruído de fundo gravado num *pub* cheio), branco, rosa e castanho (Kim, Salamon, Li et Bello, 2018).

2.1.2. Domínio espetral

2.1.2.1. Análise cepstral

O procedimento típico numa análise espetral consiste na divisão do sinal em pequenas amostras e na obtenção da transformada de Fourier dessas amostras. Se um sinal for periódico, a sua transformada irá descrever picos nos múltiplos da sua frequência fundamental, ou seja, na própria e nos seus harmónicos. Um algoritmo de deteção de altura que opera neste domínio procura identificar o pico espetral que corresponde à frequência fundamental do sinal ou calculá-la.

Noll, em 1967, propõe a seguinte equação para calcular a frequência fundamental de um sinal através de um método no domínio espetral que designa por análise cepstral (Haken, 2020):

$$c_x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X(e^{j\omega})\right| e^{jwn} d\omega$$

A característica definidora deste método é a existência de uma transformada de Fourier de uma transformada de Fourier. Seria de esperar que, dado que esta operação matemática é a sua própria inversa, uma segunda transformada de Fourier devolvesse as amostras do sinal inicial. No entanto, na análise cepstral existe uma componente de magnitude logarítmica associada a estas transformadas, pelo que o resultado não devolve o sinal inicial. Ao resultado obtido dá-se o nome de *cepstrum* ou cepstro, palavra proposta pelo mesmo autor que consiste na reorganização das letras da palavra *spectrum* (espetro, em português).

Na análise de um espetro de frequências de um sinal, obtido através de uma transformada de Fourier, o harmónico mais energético muitas vezes não corresponde à frequência fundamental do sinal. Este fenómeno pode ocorrer por uma série de razões, ora pelas características do próprio som, ora pela maneira como foi captado. Neste caso, a identificação do pico mais energético do espetro do sinal vai conduzir a um resultado errado na deteção de altura. Mesmo considerando o pico correspondente à frequência mais baixa do espetro como a fundamental, há a possibilidade de se obter imprecisões, uma vez que se extrai apenas um ponto de todo o espetro. Uma abordagem global, e não local, como a descrita anteriormente, deverá

conseguir resultados mais precisos. A análise ceptral procura então a periodicidade do espetro de um sinal de modo a obter um resultado mais global e preciso, recorrendo para isso a uma segunda transformada de Fourier. De acordo com Haken (2020), esta componente de magnitude logarítmica torna mais fortes os harmónicos menos energéticos de um sinal, homogeneizando o conteúdo espetral do mesmo.

O sinal resultante deste processo descreve um pico num determinado valor que equivale ao período da frequência fundamental do sinal (figura 7). Os valores no eixo horizontal do cepstro de um sinal são referentes à *quefrency* ou quefrência do sinal, grandeza novamente apelidada por Noll que resulta da reorganização das letras de *frequency* (frequência, em português), que corresponde ao período da frequência fundamental do mesmo. Os primeiros coeficientes cepstrais até às quefrências de aproximadamente 15 segundos correspondem ao envelope espetral do sinal, que indica as suas ressonâncias.

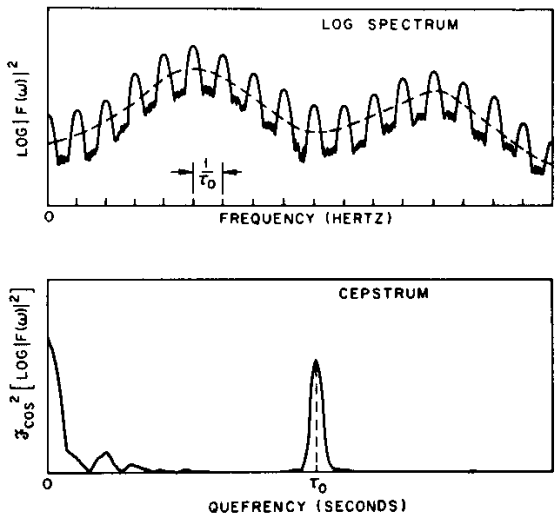

Figura 7: representações do espetro logarítmico e do cepstro de um sinal, cujo valor de quefrência assinala o período da sua frequência fundamental (Noll, 1969).

2.1.2.2. Espetro do produto harmónico (*harmonic product spectrum*) e espetro da soma harmónica (*harmonic sum spectrum*)

O método do espetro do produto harmónico, também descrito por Noll (1969), tem a sua origem na premissa de que os picos do espetro logarítmico de um sinal se adicionam coerentemente, dado que são todos múltiplos de uma frequência fundamental, enquanto as outras porções não são correlacionadas e, portanto, não se adicionam coerentemente. Como tal,

este método procura um máximo que indica a coincidência de todos os harmónicos do sinal num espetro logarítmico de frequência comprimida (figura 8) através da seguinte equação:

$$\pi(\omega) = \prod_{k=1}^{K} |F(k\omega)|^2$$

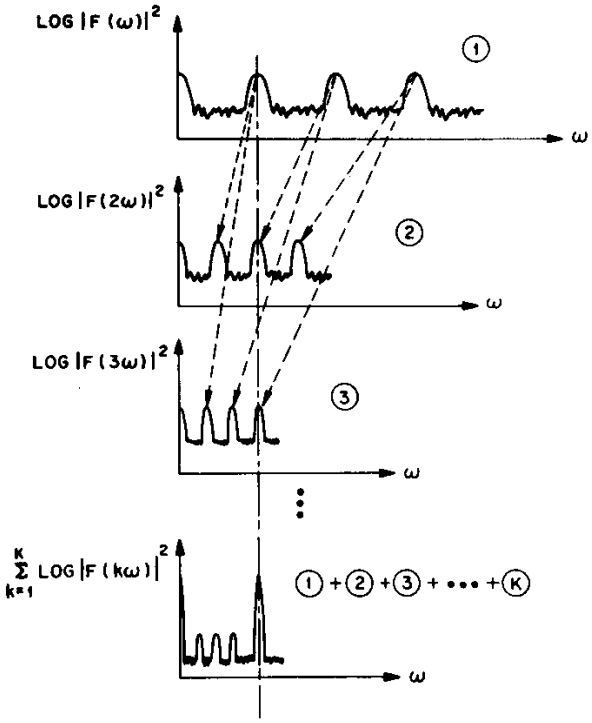

Figura 8: desenvolvimento do espetro do produto harmónico de um sinal como o antilogaritmo da soma dos espetros logarítmicos comprimidos harmonicamente (Noll, 1969).

Outra abordagem possível é através do cálculo do espetro da soma harmónica, que não utiliza o espetro logarítmico, mas uma versão de frequência comprimida do espetro do sinal (Noll, 1969), que pode ser definido por:

$$\sigma(\omega) = \sum_{k=1}^{K} |F(k\omega)|^2$$

2.1.2.3. Problema da inarmonicidade e deteção por rotulação de picos espetrais e reatribuição de tempo e frequência (*spectral peak labelling and time-frequency reassignment*)

Como foi referido anteriormente, os instrumentos musicais de cordas apresentam diferentes graus de inarmonicidade (figura 9) de acordo com os diâmetros e materiais que constituem as suas cordas. Este fenómeno causa um desvio cumulativo dos harmónicos de um tom musical, o que faz com que métodos como a deteção cepstral não funcionem com uma precisão satisfatória, uma vez que lidam com a periodicidade dos harmónicos de um sinal, que deixa de existir em cenários de elevada inarmonicidade.

Este problema pode ser contornado através da seguinte equação, na qual f_n é a frequência do harmónico *n* (sendo o primeiro harmónico a frequência fundamental, neste caso) e *B* o valor de inarmonicidade, que assume valores entre aproximadamente 0,0001 e 0,01. No mesmo instrumento musical, esta variável depende da nota tocada e da intensidade com que é tocada (Haken, 2020):

$$f_n = nf_1\sqrt{1 + Bn^2}$$

Como exemplo da influência deste valor, considere-se o caso de uma nota grave de um piano com um valor de inarmonicidade hipotético de 0,01. A nota lá 0, cuja frequência fundamental é 27,5 Hz, deveria ter um décimo sexto harmónico com uma frequência de 440 Hz, também um lá. Tendo em conta a inarmonicidade apresentada anteriormente, é possível calcular que este harmónico tem na verdade a frequência de aproximadamente 830 Hz, que é quase o dobro da esperada.

A deteção por rotulação de picos espetrais e reatribuição de tempo e frequência é um método que tem em conta o fenómeno da inarmonicidade, procurando o espetro periódico que melhor coincide com o espetro do sinal inarmónico (Haken, 2020).

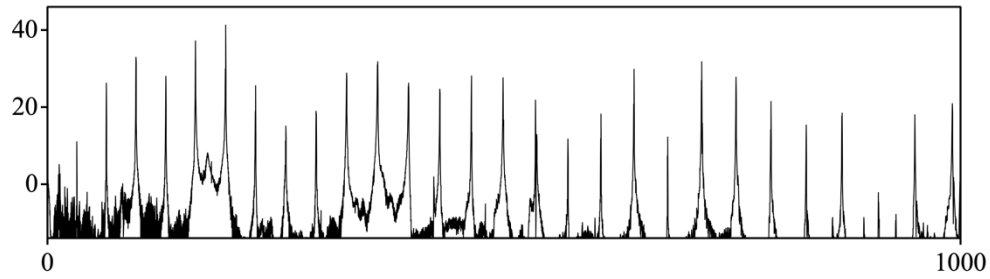

Figura 9: espetro da nota dó 1 (nível de pressão sonora por frequência) de um piano Steinway & Sons (*Piano.mf.C1*, 2001). Como se pode observar, a distância entre os primeiros harmónicos representados é inferior à distância entre os últimos, compreendidos entre 750 Hz e 1 kHz.

2.1.2.4. Qitch

Qitch é o nome de um algoritmo implementado na biblioteca de extensões do SuperCollider que atua no domínio espetral. Este algoritmo calcula uma transformada de Fourier do sinal a partir da transformada de Q constante de Brown e Puckette numa escala de quartos de tons e identifica a sua altura através de correlação espetral cruzada. Qitch é baseado nos três seguintes artigos: *An efficient algorithm for the calculation of a constant Q transform*, de Brown e Puckette, 1992, *Musical Fundamental Frequency Tracking Using a Pattern Recognition Method*, de Brown, 1992 e *A High-Resolution Fundamental Frequency Determination Based on Phase Changes of the Fourier Transform*, de Brown e Puckette, 1993.

A transformada de Q constante (figura 10) é uma técnica de transformação de um sinal do domínio temporal para o domínio espetral na qual as frequências centrais dos *bins* de frequências (intervalos entre amostras no domínio espetral) são geometricamente espaçadas e os seus fatores de Q (rácios entre a energia de uma frequência central em relação à banda de frequências na qual está inserida) são todos iguais (Schörkhuber et Klapuri, 2010).

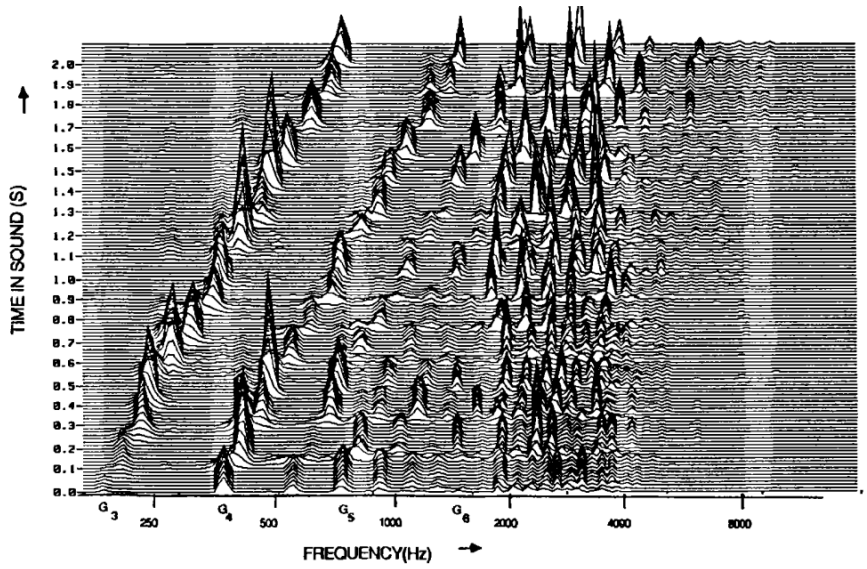

Figura 10: transformada de Q constante de um violino a tocar a escala maior de sol de sol 3 (196 Hz) até sol 5 (784 Hz) (Brown, 1991).

Esta resulta tendencialmente numa melhor representação da informação espetral de um sinal musical do que uma transformada rápida de Fourier, uma vez que consegue uma melhor resolução espetral nas frequências baixas e uma melhor resolução temporal nas agudas e também pelo facto dos tons musicais utilizados na música ocidental serem geometricamente espaçados. Por outro lado, a transformada de Fourier resulta em *bins* de frequências linearmente espaçados que não mantêm uma resolução aceitável em toda a gama de frequências audíveis.

Esta transformada é calculada pela seguinte equação, na qual $X^{cq}\left[k_{cq}\right]$ é o componente k_{cq} da transformada de Q constante, *x[n]* a função de tempo amostrada e $\omega\left[n, k_{cq}\right]$ a função de janela (*window*) de comprimento $N\left[k_{cq}\right]$ (Brown et Puckette, 1992):

$$X^{cq}\left[k_{cq}\right] = \sum_{n=0}^{N\left[k_{cq}\right]-1} \omega\left[n, k_{cq}\right] x[n] e^{-j\omega {k_{cq}}^{n}}$$

Embora este algoritmo seja tendencialmente mais preciso, há um conjunto de razões pelas quais não consegue destronar a transformada de Fourier. Primeiro, é mais caro computacionalmente. Em segundo lugar, não existe uma transformada inversa que permita a reconstrução do sinal original através dos seus resultados e, por fim, produz resultados que são mais difíceis de trabalhar do que o espetrograma produzido pela transformada rápida de Fourier,

uma vez que a resolução temporal da transformada de Q constante varia ao longo dos *bins* (Schörkhuber et Klapuri, 2010).

2.1.2.5. SWIPE

O algoritmo Sawtooth Waveform Inspired Pitch Estimator, de acrónimo SWIPE, foi proposto em 2008 por Camanho e Harris num artigo publicado em *The Journal of the Acoustical Society of America* e procura, essencialmente, "a frequência que maximiza a distância média do pico ao vale dos harmónicos dessa frequência" (Camacho et Harris, 2008).

Se um sinal for periódico, o seu espetro descreve picos, separados por vales, nos múltiplos da frequência fundamental. Posto isto, a distância média global entre os picos e os vales que os ladeiam $D_n(f)$ é dada pelas seguintes equações, nas quais *n* é o número de picos considerados, *f* é a frequência, *k* é o número do pico e $d_k(f)$ é a distância do pico *k* ao vale correspondente:

$$d_k(f) = \frac{1}{2}\left[|X(kf)| - \left|X(\left(k - \frac{1}{2}\right)f)\right|\right] + \frac{1}{2}\left[|X(kf)| - \left|X(\left(k + \frac{1}{2}\right)f)\right|\right]$$

$$D_n(f) = \frac{1}{n}\sum_{k=1}^{n} d_k(f)$$

Esta distância pode também ser expressa por uma função *kernel* $K_n(f, f')$, que depende não só da frequência, mas também do número de harmónicos considerados *n* e descreve pulsos positivos de valor 1 e pulsos negativos de valor -1 entre os pulsos positivos:

$$K_n(f, f') = \frac{1}{2}\delta\left(f' - \frac{f}{2}\right) - \frac{1}{2}\delta\left(f' - \left(n + \frac{1}{2}\right)f\right) + \sum_{k=1}^{n} \delta\left(f' - kf\right) - \delta(f' - (k - 1/2)f)$$

A primeira abordagem consiste na escolha da frequência que maximiza a distância média global entre picos e vales, embora esta só funcione se o sinal for perfeitamente harmónico. De modo a contemplar uma possível inarmonicidade de um sinal, o algoritmo SWIPE introduz uma otimização que consiste no "ofuscamento da localização dos harmónicos"

(Camacho et Harris, 2008), ou seja, na substituição dos pulsos da função *kernel* por uma onda que contemple as suas posições. Na fase inicial do desenvolvimento deste algoritmo foi utilizada, para este propósito, uma onda triangular, mas esta não era particularmente eficaz, pelo que foi substituída por uma onda cosseno.

De seguida, o uso do logaritmo do espetro foi descartado e substituído pelo uso da raiz quadrada do espetro, uma vez que esta é mais próxima à resposta do sistema auditivo humano face à amplitude e também porque produziu resultados de deteção de altura mais precisos nos testes descritos no artigo (Camacho et Harris, 2008).

Outras otimizações consistem na aplicação de um fator de decaimento da energia dos harmónicos de modo a evitar problemas causados por subarmónicos, na utilização de harmónicos com frequências até, no máximo, 3.3 kHz na deteção de altura de discurso e 5 kHz na deteção de altura de música e na deformação da escala de frequências de modo a que esta tenha uma maior resolução, obtida através de um aumento da taxa de amostragem numa certa região de frequências de acordo com a escala psicoacústica ERB (semelhante à escala de barks descrita no capítulo 1), que representa uma aproximação das bandas de frequências dos filtros auditivos humanos.

Para concluir, após as otimizações descritas, o algoritmo SWIPE calcula a semelhança entre a raiz quadrada do espetro do sinal e a raiz quadrada do espetro de uma onda dentes-de-serra (figura 11), sendo essa a justificação do nome que lhe foi atribuído.

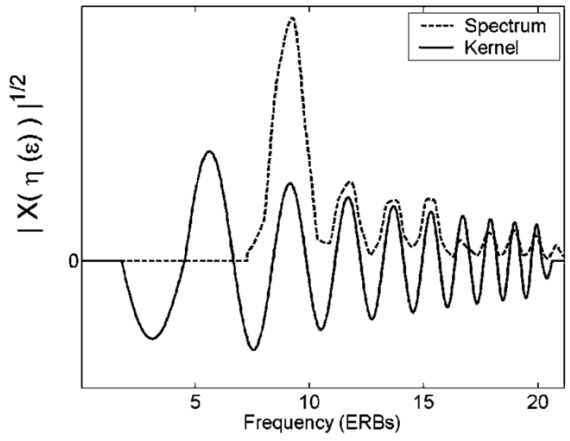

Figura 11: função *kernel* normalizada do algoritmo SWIPE (Camacho et Harris, 2008).

2.1.2.6. SPICE

SPICE, acrónimo de Self-Supervised Pitch Estimation, é um algoritmo proposto por Gfeller et al. (2020) que toma partido de uma técnica de aprendizagem automática com supervisão própria. Este método também utiliza transformada de Q constante, à semelhança do algoritmo Qitch (descrito no subcapítulo 2.1.2.4).

A base da conceção deste algoritmo assenta no facto de que para os humanos é mais fácil estimar alturas relativas do que absolutas, ou seja, identificar o intervalo entre duas alturas, ou a relação que descrevem, do que as alturas em si (Ziv et Radin apud Gfeller et al., 2020), exceto para quem possui ouvido absoluto (capacidade de identificar notas musicais sem o auxílio de um tom de referência). Posto isto, o algoritmo SPICE opera com duas versões do sinal cuja altura pretende detetar, uma delas transposta (*pitch shifted*) por um valor aleatório, mas conhecido. De seguida, é aplicada uma função que força a diferença entre os valores escalares com os quais o algoritmo opera a serem proporcionais à diferença entre as alturas dos dois sinais, estimando deste modo uma altura relativa que é calibrada através de um conjunto de dados gerado sinteticamente, de modo a ser traduzida para uma altura absoluta (figura 12). Este processo de estimação de altura é supervisionado por ele próprio e pode ser treinado sem acesso a dados etiquetados (Gfeller et al., 2020).

O SPICE recebe, como entrada, sinais transformados de acordo com a transformada de Q constante, pelo que consegue lidar sem dificuldades com sinais inarmónicos, ruidosos ou cuja frequência fundamental não tem energia. Este algoritmo está também programado para decidir se uma parte do sinal é vozeada ou desvozeada e fá-lo através de uma aprendizagem automática que se baseia no nível de confiança das estimativas de altura que produz, que também é supervisionada pelo próprio.

De acordo com os testes feitos pelos mesmos autores (Gfeller et al., 2020), o algoritmo SPICE consegue níveis de precisão semelhantes aos do algoritmo CREPE (descrito no subcapítulo 2.1.1.6).

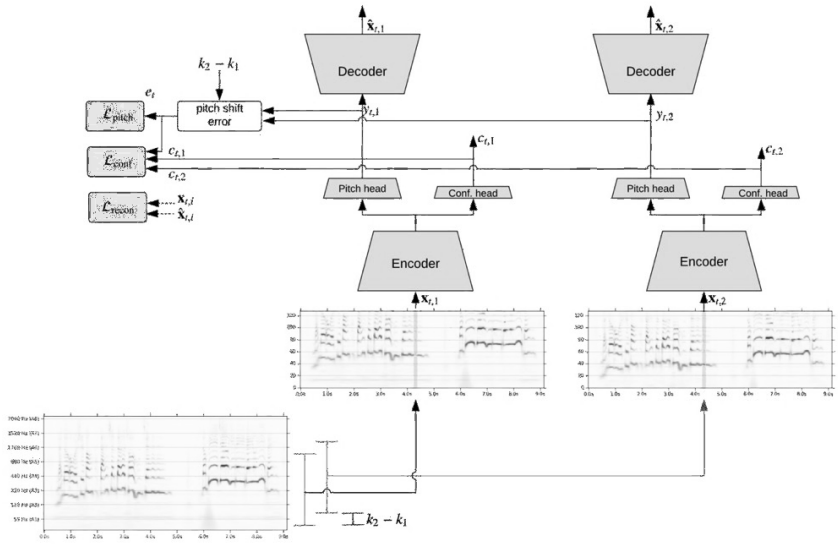

Figura 12: arquitetura do algoritmo SPICE (Gfeller et al., 2020).

2.1.3. Outros domínios: misto e modelação do sistema auditivo humano

2.1.3.1. YAAPT

Um exemplo de um método de técnica mista, ou seja, simultaneamente temporal e espectral, é o algoritmo YAAPT, acrónimo de Yet Another Algorithm for Pitch Detection, proposto por Kavita Kasi e Stephan Zahorian (2002) e desenvolvido com o objetivo de identificar a altura de discurso, tanto a partir de áudio de alta qualidade como a partir de áudio telefónico, que é particularmente difícil de conseguir, dada a qualidade degradada do sinal, que é sujeito a distorções e ruído, para além de ter geralmente frequências fundamentais muito fracas ou inexistentes.

O funcionamento deste algoritmo pode ser dividido em cinco componentes: pré-processamento, seleção da frequência fundamental através de correlação cruzada normalizada, refinação da seleção através de informação espectral, modificações baseadas na plausibilidade e limitações de continuidade e determinação do resultado através de programação dinâmica.

O primeiro ponto consiste na filtragem do sinal com um filtro passa-banda de 100 a 900 Hz, uma vez que este algoritmo se foca na deteção de altura de voz falada, e na realização de um corte central (*center clip*) no mesmo, que faz com que apenas os valores que ultrapassem um determinado limite de amplitude sejam contabilizados.

De seguida, é aplicada uma correlação cruzada normalizada *NCCF(k)*, que é essencialmente uma função de autocorrelação no domínio temporal modificada de acordo com as seguintes equações, nas quais *s(n)* é o conjunto de amostras do sinal em questão e $0 \leq n \leq N-1$:

$$NCCF(k) = \frac{\sum_{n=0}^{N-K} s(n)s(n+k)}{\sqrt{e_0 . e_k}}$$

$$e_k = \sum_{n=k}^{n=k+N-K} s^2(n), \qquad 0 \leq k \leq K-1$$

De modo a otimizar este passo, é aplicado um algoritmo de escolha inteligente de picos (*intelligent peak-picking*) que consiste, primeiro, na identificação dos picos da NCCF, que são considerados legítimos se forem maiores do que um determinado valor de limite. Em segundo lugar, são excluídos quaisquer picos que estejam mais próximos do que 2 milissegundos de outros picos maiores. Posto isto, são atribuídos valores de mérito aos picos iguais às magnitudes das NCCF para cada um deles, que são essencialmente valores que indicam se um pico é mais ou menos provável de corresponder à frequência fundamental do sinal. Por último, se um pico descrever metade do valor de atraso de outro, o valor de mérito desse pico mais baixo aumenta. Os picos restantes, cada um com o seu valor de mérito, tornam-se então candidatos à frequência fundamental do sinal.

De seguida, o sinal é analisado no domínio espetral de modo a refinar a probabilidade dos candidatos e tomar decisões no que toca à identificação das regiões vozeadas e desvozeadas do sinal, que é conseguida a partir do rácio normalizado de energia das baixas frequências *NLFER*, dado pela seguinte equação, que descreve valores altos nas regiões vozeadas e valores baixos nas desvozeadas e na qual *N* é o número total de amostras, *i* é o índice de frequência, *x(i,j)* é a magnitude logarítmica das regiões das baixas frequências do espetrograma e *j* é o índice da amostra:

$$NLFER = \frac{\sum_i x(i,j)}{\frac{1}{N}\sum_i \sum_j x(i,j)}$$

Nas amostras vozeadas do sinal, os candidatos à frequência fundamental obtidos no passo anterior são testados em relação à sua proximidade ao ponto espetral correspondente. Se forem próximos, o seu valor de mérito aumenta, se forem distantes, o seu valor de mérito diminui.

Os passos descritos dão origem a uma série de dados que servem como base para a decisão da frequência fundamental: uma matriz de candidatos à frequência fundamental, uma matriz de valores de mérito, uma curva NLFER e uma deteção de altura baseada apenas na informação espetral do sinal. Após a implementação de outras otimizações e refinamentos do processo de seleção, o algoritmo escolhe o candidato mais provável de acordo com o seu valor de mérito (figura 13).

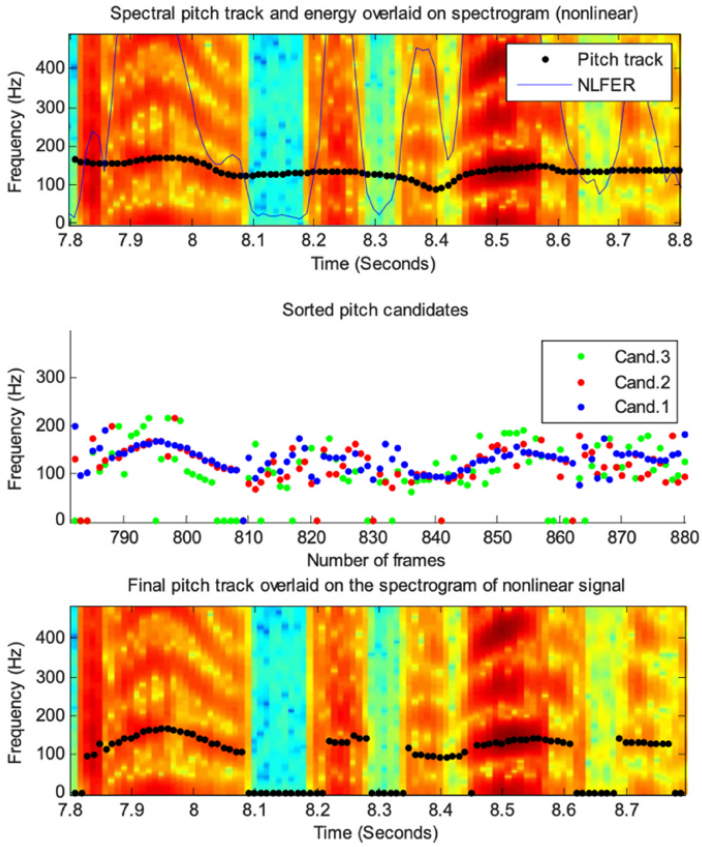

Figura 13: o primeiro gráfico representa o espetrograma de um sinal com a curva NLFER e a estimativa de altura proveniente da análise espetral sobrepostas, o segundo apresenta os candidatos à frequência fundamental obtidos no passo anterior e o último indica a estimativa final da frequência fundamental do sinal (Zahorian et Hu, 2008).

2.1.3.2. Detetor percetual de altura

O algoritmo percetual de altura, proposto em 1990 por Slaney e Lyon, é um detetor de altura que se baseia no sistema percetual dos humanos, utilizando para esse efeito um modelo auditivo. Como tal, este algoritmo consegue lidar com sinais inarmónicos ou ruidosos, da mesma maneira que um humano consegue resolver a altura de casos considerados desafiantes para a deteção de altura através de algoritmos. Este algoritmo pode ser descrito por três componentes: um modelo coclear, um correlograma e um detetor de altura.

O "(...) modelo coclear (...) converte uma onda sonora num vetor de números que representa a informação enviada para o cérebro" (Stanley et Lyon, 1990), ou seja, não pretende ser uma modelação exata da estrutura interna do ouvido, mas sim aproximar a informação gerada pelo nervo auditivo em resposta a um estímulo sonoro. Uma série de outras operações são utilizadas para mimetizar o sistema auditivo humano com maior precisão, tais como a implementação de filtros para modular a propagação do som pela membrana basilar, a modulação do facto das células ciliadas responderem apenas ao movimento da membrana basilar numa direção e, por fim, a compressão do alcance dinâmico do sinal em função da maneira como o ouvido humano responde às diferentes bandas de frequências e à intensidade do som. Este passo dá origem a uma representação multicanal do sinal que pode ser interpretada como a probabilidade do disparo instantâneo dos nervos dos humanos (Stanley et Lyon, 1990).

Em segundo lugar, esta representação multicanal é representada num correlograma, que é um gráfico obtido a partir do cálculo da autocorrelação de cada um destes canais cocleares. Por fim, o detetor de altura analisa a informação de todos os canais do correlograma e escolhe aquela que considera ser a altura correta do sinal (figura 14) através de uma série de cálculos e otimizações, como a utilização da autocorrelação estreitada (*narrowed autocorrelation*), proposta por Brown (Stanley et Lyon, 1990).

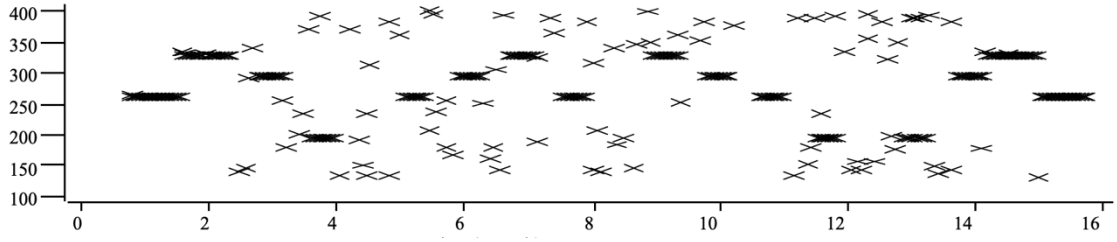

Figura 14: resultados do algoritmo percetual de altura na deteção de altura da melodia "Quartos de Westminster", expressos em frequência da fundamental por segundos. As linhas horizontais, que são na verdade conjuntos de pontos seguidos, representam as frequências fundamentais detetadas e os pontos espalhados surgem entre as alturas residuais da melodia detetada e são determinadas pelo ruído de fundo.

2.2. Algoritmos polifónicos

2.2.1. PolyPitch

Como explicado anteriormente, a deteção de altura polifónica traz uma série de vantagens na recuperação de informação musical e separação de discurso e está relacionada com a separação sonora e análise de cena auditiva, uma vez que os sinais coexistentes têm de ser separados nas suas componentes fundamentais, neste caso, nas suas alturas. A classe de extensão da biblioteca do SuperCollider PolyPitch, baseada no artigo *Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model*, de Klapuri (2008), propõe uma solução para este problema.

À semelhança do detetor percetual de altura, o algoritmo proposto neste artigo utiliza um modelo auditivo que procura replicar a maneira como os humanos percecionam a altura, uma vez que os "(...) humanos são muito bons a resolver misturas sonoras e, por isso, parece natural utilizar a mesma representação de dados que está disponível para o cérebro humano" (Klapuri, 2008). Utiliza também, como segundo passo, um método específico de análise de periodicidade no qual as frequências fundamentais são iterativamente detetadas e canceladas da mistura sonora. Este método substitui a autocorrelação por uma transformada específica com melhores resultados em sinais polifónicos.

O modelo auditivo utilizado funciona do seguinte modo: primeiro, o sinal de entrada é separado de acordo com um conjunto de filtros passa-banda lineares que mimetizam os filtros auditivos; de seguida, os sinais, separados por bandas, ou canais, são processados não-linearmente de modo a modular a atividade neural da audição humana; por fim, é feita a análise da periodicidade dos sinais obtidos no ponto anterior. As bandas de frequências utilizadas neste

algoritmo são dadas pelas seguintes expressões, nas quais f_c é a frequência central do filtro, b_c é a largura da banda, *c* é o índice da banda e ξ_0 é o número de banda crítica da banda mais baixa (ξ_0 corresponde a 2,3 e ξ_1 a 0,39):

$$f_c = 229\left[10^{(\xi_1 c+\xi_0)/21.4} - 1\right]$$

$$b_c = 0{,}108 f_c + 24{,}7$$

Em segundo lugar, na análise de periodicidade utilizada, a força de um candidato a período da fundamental é calculada como a soma das amplitudes dos harmónicos da frequência correspondente, à semelhança do algoritmo do espetro da soma harmónica de Noll (Klapuri, 2008). Este método é eficaz porque utiliza apenas os componentes espetrais relacionados com o período em questão e ignora os componentes espetrais entre os parciais, diminuindo a influência do ruído na estimativa de altura. A saliência $s_t(\tau)$, que equivale à força referida no início do parágrafo, é dada pela seguinte equação, na qual τ é o período da fundamental candidata na amostra *t*, *m* é o índice do parcial, *w(τ,m)* é a função que determina o peso do parcial *m* do período τ na soma, $k_{t,m}$ é um conjunto de *bins* de frequências próximos do harmónico *m* da frequência fundamental candidata e *Ut(k)* é o espetro da soma harmónica:

$$s_t(\tau) = \sum_{m=1}^{M} w(\tau, m) max_{k \in k_{t,m}} Ut(k)$$

O máximo da função $s_t(\tau)$ é um bom indicador de uma das frequências fundamentais do sinal. Após a sua deteção, esta é cancelada da mistura sonora para que possam ser resolvidas outras frequências fundamentais. Este processo é repetido até que todas as fundamentais sejam identificadas, ou seja, até não haver mais energia no espetro residual, daí ser considerado um método iterativo.

Os resultados dos testes de deteção de altura de música polifónica publicados no artigo em questão (figura 15) indicam que o método proposto consegue resultados mais precisos que outros métodos de referência. O gráfico da estimativa de frequências fundamentais múltiplas corresponde à taxa de erro calculada pela percentagem de frequências fundamentais identificadas incorretamente, enquanto o gráfico da estimativa de frequência fundamental

predominante corresponde à taxa de erro da estimativa de apenas uma das frequências fundamentais do sinal polifónico, ou seja, da identificação correta de pelo menos uma das suas alturas constituintes.

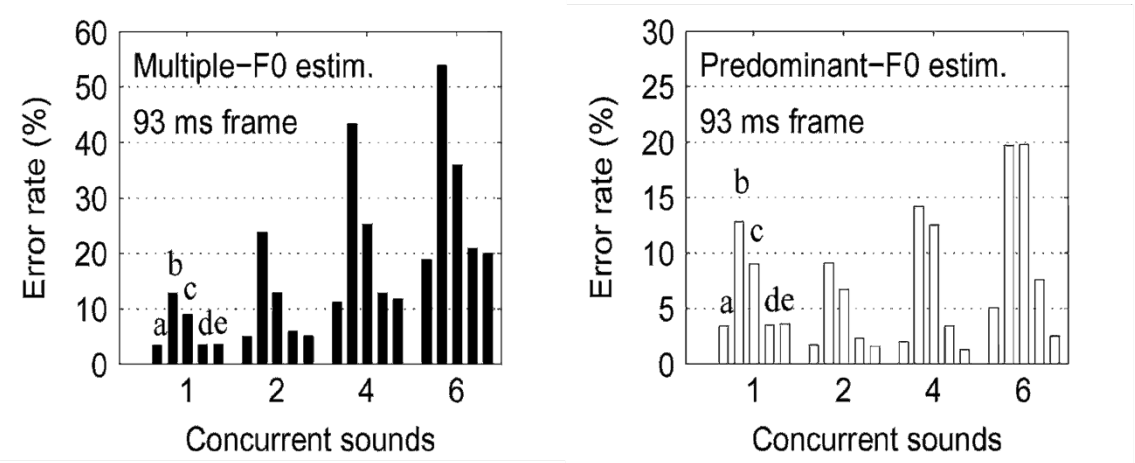

Figura 15: comparação dos resultados obtidos pelo método proposto por Klapuri, indicados pelas barras assinaladas pela letra "a", e por outros métodos de referência. O eixo vertical representa a taxa de erro, em percentagem, e o horizontal o número de alturas simultâneas do sinal polifónico analisado. O gráfico da esquerda é referente à estimativa de frequências fundamentais múltiplas e o da direita à estimativa de frequência fundamental predominante (Klapuri, 2008).

2.3. Aplicações dos algoritmos de deteção de altura na música

Os algoritmos de deteção de altura têm uma série de aplicações na música e na análise e reconhecimento de discurso, como referido anteriormente. Na área da música, estas podem separar-se nas seguintes aplicações principais: *performance*, produção, ensino, transcrição e entretenimento, embora haja bastante sobreposição entre estas cinco vertentes.

Na componente da *performance* encontram-se os algoritmos otimizados para operar em tempo real, uma vez que têm de apresentar resultados na ordem dos milissegundos de modo a acompanhar a *performance* dos músicos. Incluem os tipos de processamento de sinal que tomam partido da deteção de altura aplicados a instrumentos musicais ou tons sintetizados em tempo real, como acionadores de sintetizadores ou *samples*, que detetam a altura de um sinal de entrada e utilizam essa informação para acionar tons de sintetizadores ou reproduzir *samples* específicas de acordo com a altura detetada, à semelhança do que acontece em dispositivos MIDI (ou outros protocolos de comunicação) como teclados, que enviam sinais com valores entre 0 e 127 para dispositivos que os utilizam para gerar ou reproduzir áudio. Um exemplo deste tipo de processamento é o pedal de efeitos para guitarra elétrica Electro-Harmonix Mel 9

Tape Replay Machine, lançado em 2016. O produto da empresa nova-iorquina recebe o sinal polifónico da guitarra tal e qual como ele é transmitido pela saída do instrumento e identifica as suas alturas constituintes através de um algoritmo de deteção de altura, acionando por fim as *samples* de um dos nove instrumentos que emula de modo que as alturas dos mesmos correspondam às detetadas. Posto isto, o pedal sobrepõe ao sinal da guitarra timbres de outros instrumentos, como violoncelo ou flauta. Outros pedais desta série incluem B9 Organ Machine, Bass 9 Bass Machine, C9 Organ Machine, Key 9 Electric Piano Machine, String 9 String Ensemble e Synth 9 Synthesizer Machine. Outra aplicação comum destes algoritmos na *performance* musical consiste no auxílio na afinação de instrumentos musicais. De maneira a afinar os seus instrumentos de um modo preciso e rápido, músicos de determinados géneros e instrumentos recorrem frequentemente a ferramentas de afinação, afinadores, que tomam partido destes algoritmos para indicar a altura dos tons que produzem, com o objetivo de facilitar a sua correção ou alteração.

Na produção musical, para além do uso do mesmo tipo de processamento referido no parágrafo anterior, que neste caso não necessita de operar em tempo real, a deteção de altura é usada principalmente na correção de altura. Esta consiste num método de edição áudio que altera a intonação de um sinal para que as suas alturas equivalham ou se aproximem das notas musicais do sistema de temperamento igual. A correção de altura é muito utilizada na edição de vozes na produção musical e procura corrigir problemas de afinação, mas também pode ser utilizada noutros instrumentos, com o mesmo propósito. Esta correção pode ser feita automaticamente a partir de *software* de afinação automática (*auto-tune*) ou manualmente a partir de *software* como o que será descrito de seguida. Os processadores de afinação automática podem também ser utilizados em contextos de *performance* se conseguirem corresponder às velocidades de resposta necessárias. O *software* de correção de altura Melodyne, lançado pela empresa alemã Celemony em 2001, é largamente utilizado na produção musical profissional e permite a manipulação da altura de sinais áudio monofónicos e polifónicos e também a manipulação da duração de fragmentos dos mesmos. Devido ao facto de processar as componentes vozeadas e desvozeadas do sinal de maneiras diferentes, este *software* garante uma sonoridade extremamente natural no áudio que edita, desde que as manipulações de altura e duração não sejam extremas, como a transposição de uma oitava, por exemplo.

No ensino da música existem também ferramentas que tomam partido da deteção de altura. Estas funcionam essencialmente através da análise de altura da *performance* de um músico em formação e dão indicações em função da mesma de modo a corrigir problemas de afinação. Como tal, fazem sentido no ensino de instrumentos musicais nos quais a capacidade de afinação é fundamental, como voz ou violino. Por exemplo, o método de ensino de violino Trala, desenvolvido principalmente para auxiliar adultos na aprendizagem do instrumento, toma partido de uma tecnologia de deteção de altura na arquitetura da aplicação (*app*) que serve de pilar central do método. Esta aplicação analisa a *performance* do estudante e dá *feedback* acerca da sua afinação e tempo que não consiste apenas na indicação binária de certo ou errado, inclui também sugestões acerca de como melhorar.

Os algoritmos de deteção de altura têm também um papel no auxílio ou automatização da transcrição musical. Posto isto, pode-se afirmar que auxiliam na recuperação de informação musical, que é um campo de pesquisa emergente que tem aplicações, por exemplo, no reconhecimento de géneros musicais e nos sistemas de recomendação de música utilizados pelas plataformas de *streaming* de música. A ferramenta de transcrição automática AudioScore, incluída no *software* de notação musical Sibelius da empresa americana Avid Technology, converte áudio diretamente em notação musical. Este deteta automaticamente toda a instrumentação de um ficheiro áudio e cria pentagramas devidamente identificados para cada instrumento, sendo capaz de identificar até 16 notas musicais simultâneas e durações de notas ou pausas até à fusa (1/32 da duração do compasso).

Por fim, na área do entretenimento é possível encontrar o exemplo da série de videojogos SingStar, cujo primeiro jogo foi publicado em 2004 pela Sony Computer Entertainment e o último em 2017. Neste videojogo de competição musical, um ou dois jogadores têm como objetivo cantar corretamente as melodias vocais de diversas canções, procurando acertar a altura e o ritmo das mesmas. De modo a reconhecer se o jogador está a cantar com a afinação pretendida, o videojogo deteta a altura da sua voz e atribui-lhe pontos, ou não, de acordo com as alturas detetadas.

3. Comparação experimental de métodos de deteção de altura no SuperCollider

Este capítulo tem como objetivo a elaboração de uma experiência que visa clarificar a precisão de um conjunto de métodos de deteção de altura de música. Esta consiste na comparação de quatro tipos de deteção de altura monofónica implementados no *software* de programação áudio SuperCollider: ZeroCrossing, que se baseia na taxa de cruzamento de zero (2.1.1.1), Pitch, que se baseia na autocorrelação (2.1.1.2), Tartini (descrito no subcapítulo 2.1.1.5), e Qitch (descrito no subcapítulo 2.1.2.4). Estes quatro algoritmos serão utilizados para averiguar a altura de dois pequenos trechos de melodias no timbre de uma voz e de três instrumentos musicais diferentes e para comparar os resultados obtidos com as alturas sabidas a priori, que podem ser verificadas visualmente através dos pentagramas inseridos no próximo subcapítulo e que para os efeitos desta experiência serão designadas por alturas reais. Através destas informações, os quatro algoritmos serão comparados entre si. Os áudios utilizados foram gravados pelo autor ou gerados através da biblioteca de instrumentos virtuais do *plugin* Xpand!2, da empresa AIR Music Tech.

Como referido anteriormente, a plataforma utilizada nesta experiência é o ambiente para síntese e processamento de áudio em tempo real SuperCollider, desenvolvido por James McCartney e lançado em 1996, que conta com uma linguagem de programação própria. Este é utilizado em variadas disciplinas que vão desde a investigação acústica até ao *live coding*, a arte performativa que tem como base a programação em tempo real. A sua escolha neste contexto deve-se ao facto da sua linguagem de programação ser intuitiva e eficaz e de ter já uma série de algoritmos de deteção de altura implementados na sua arquitetura.

Note-se que a identificação de altura é um processo contínuo que não apresenta um único resultado durante a sonância completa de uma nota musical, pelo contrário, apresenta uma sequência de valores discretos que dependem da taxa de amostragem do algoritmo, pelo que uma nota musical pode ser identificada corretamente no seu período transitório, mas ter um decaimento que dá origem a um erro de oitava, por exemplo. Por essa razão, os resultados serão expressos em gráficos de deteção de altura em função do tempo e não em resultados binários de certo ou errado referentes a cada altura singular das melodias.

Com base nos gráficos obtidos, será calculada uma percentagem aproximada que representa a quantidade de deteção correta em relação a toda a extensão temporal de deteção. Para isso, cada gráfico será separado em 100 partes iguais ao longo da sua extensão temporal e

será retirado 1% à sua percentagem aproximada de deteção correta cada vez que uma deteção que não corresponde à altura real estiver contida numa dessas porções. Este valor percentual representará a precisão dos algoritmos em cada uma das deteções.

3.1. Deteção de altura a partir dos algoritmos monofónicos ZeroCrossing, Pitch, Tartini e Qitch

Nesta experiência são utilizados dois trechos de quatro compassos de melodias de canções (figura 1) cantados com letra uma oitava abaixo por uma voz masculina. Esta transposição tem como objetivo colocar as melodias num registo mais confortável para a voz do cantor. Os mesmos trechos são também reproduzidos através do Xpand!2 nos timbres de piano acústico, violino e flauta transversal, exportados em *mono* e sem reverberação. Estes áudios são posteriormente submetidos a deteção de altura através dos algoritmos referidos anteriormente. Os resultados são apresentados em gráficos gerados pelo SuperCollider que indicam a frequência da fundamental, expressa em hertz, em função do tempo, expresso em segundos.

As frequências fundamentais das notas utilizadas nesta experiência estão compreendidas entre 784 Hz (sol 5) e 523 Hz (dó 5) no caso da melodia 1 e entre 523 Hz (dó 5) e 330 Hz (mi 4) no caso da melodia 2. Quando cantadas uma oitava abaixo, estão compreendidas entre 392 Hz (sol 4) e 262 Hz (dó 4) e 262 Hz (dó 4) e 165 Hz (mi 3), respetivamente.

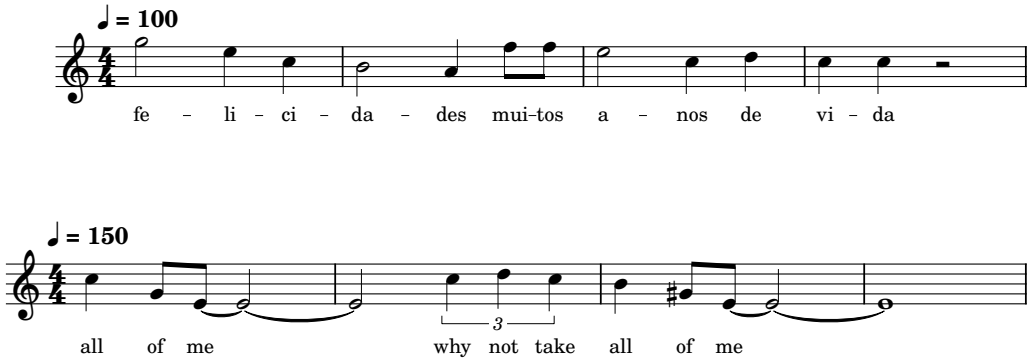

Figura 1: trechos da canção tradicional "Parabéns a Você" (melodia 1) e do tema "All of Me" de Gerald Marks e Seymour Simons (melodia 2).

3.1.1. Resultados: ZeroCrossing

Os gráficos que traçam os resultados obtidos (figuras 3 e 4) mostram frequências dos 0 aos 2000 Hz e unidades de tempo dos 0 aos 9 segundos no caso da melodia 1 e dos 0 aos 7 segundos no caso da melodia 2. A tabela abaixo (tabela 1) indica as percentagens aproximadas de deteção de altura correta de todos os áudios.

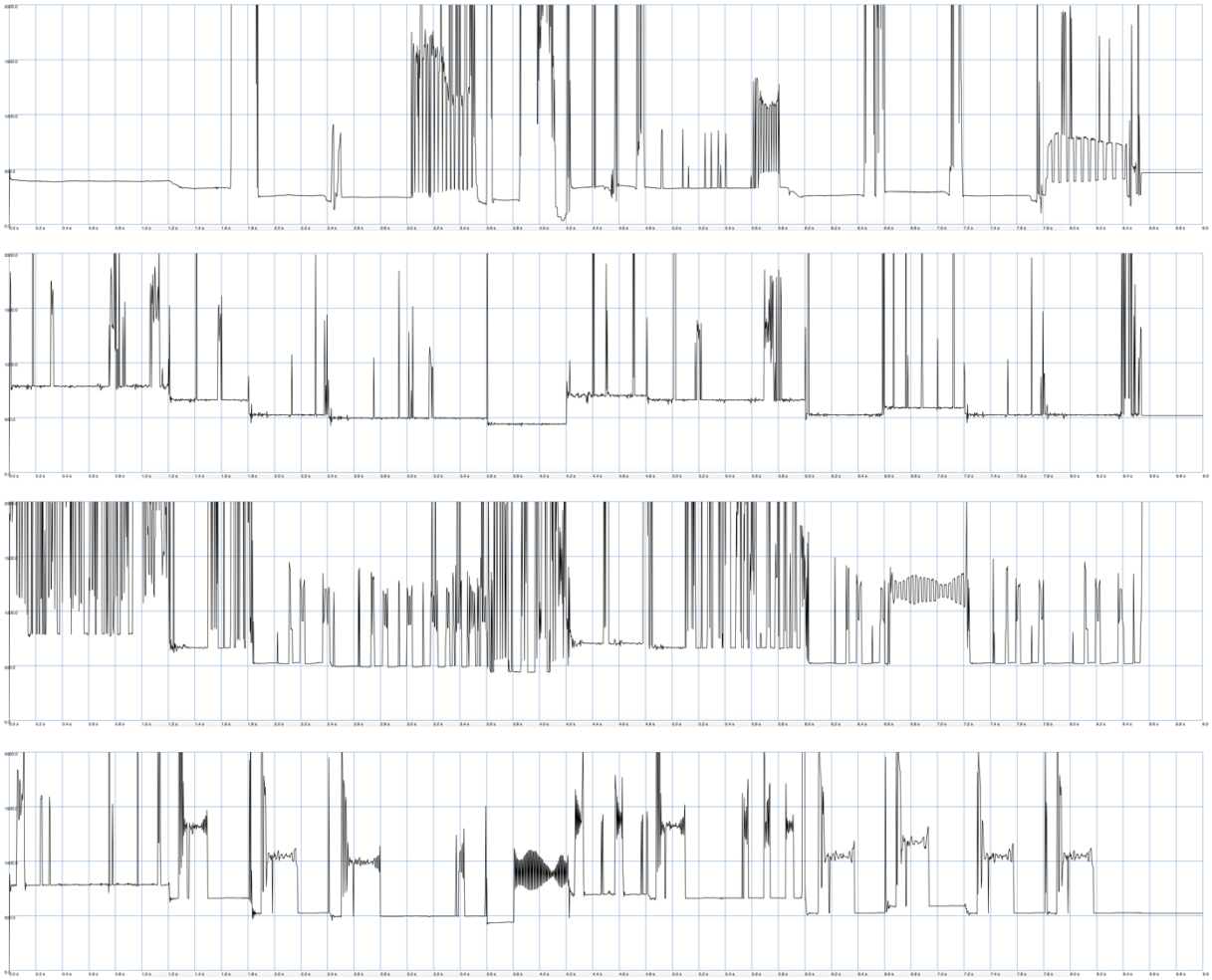

Figura 3: resultados da deteção de altura do algoritmo ZeroCrossing da melodia 1 cantada e nos timbres de piano, violino e flauta, respetivamente.

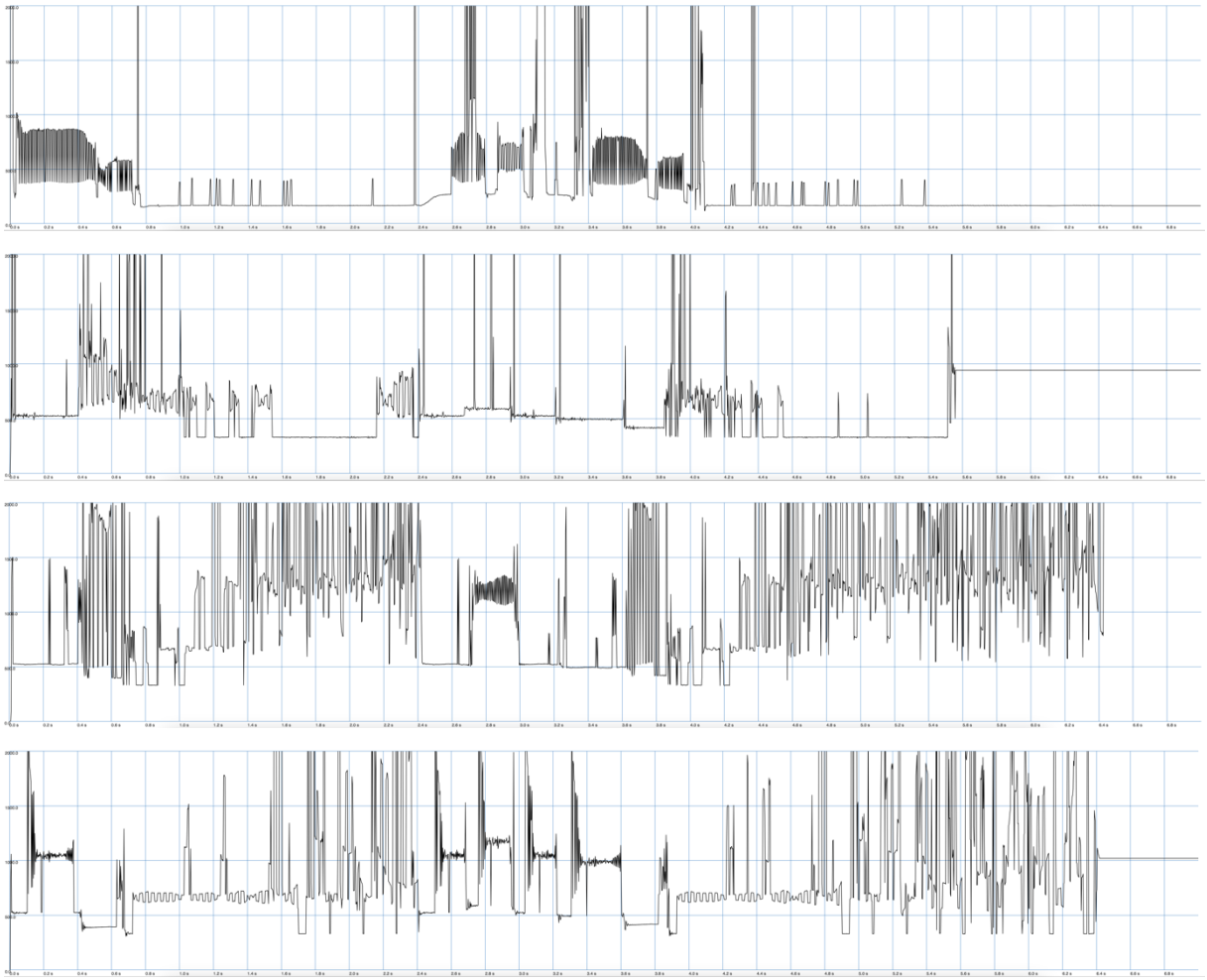

Figura 4: resultados da deteção de altura do algoritmo ZeroCrossing da melodia 2 cantada e nos timbres de piano, violino e flauta, respetivamente.

Melodia	Percentagem aproximada de deteção correta
1 voz	56%
1 piano	56%
1 violino	19%
1 flauta	36%
2 voz	44%
2 piano	42%
2 violino	11%
2 flauta	7%

Tabela 1: percentagens aproximadas de deteção de altura correta das melodias 1 e 2 pelo algoritmo ZeroCrossing.

As linhas de código introduzidas no SuperCollider para a plotagem dos gráficos anteriores, nas quais α corresponde ao número do *buffer* a analisar, ou seja, ao índice do ficheiro

de áudio e β corresponde ao valor máximo de segundos plotado nos gráficos, que no caso da melodia 1 corresponde a 9 segundos e no caso da melodia 2 corresponde a 7 segundos, são as seguintes:

```
(
{
var a;
a = PlayBuf.ar(1, α);
[a, ZeroCrossing.ar(a)]
}.plot(β, minval: 0, maxval: 2000, bounds: Rect.new(width: 5000, height: 1000));
)
```

3.1.2. Resultados: Pitch

Os gráficos obtidos (figuras 5 e 6) mostram novamente frequências dos 0 aos 1000 Hz e unidades de tempo dos 0 aos 9 ou dos 0 aos 7 segundos. A tabela abaixo (tabela 2) indica as percentagens aproximadas de deteção de altura correta de todos os áudios.

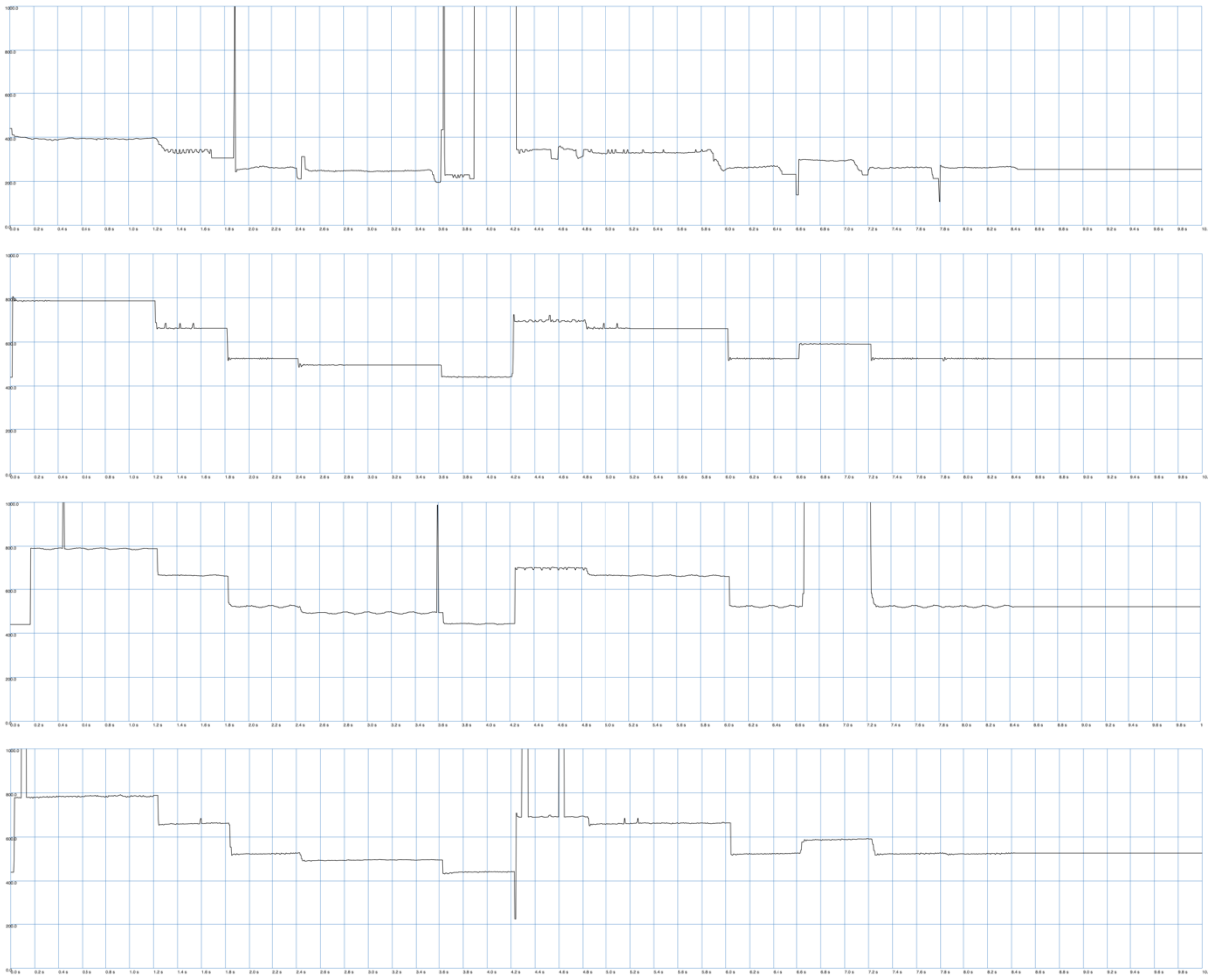

Figura 5: resultados da deteção de altura do algoritmo Pitch da melodia 1 cantada e nos timbres de piano, violino e flauta, respetivamente.

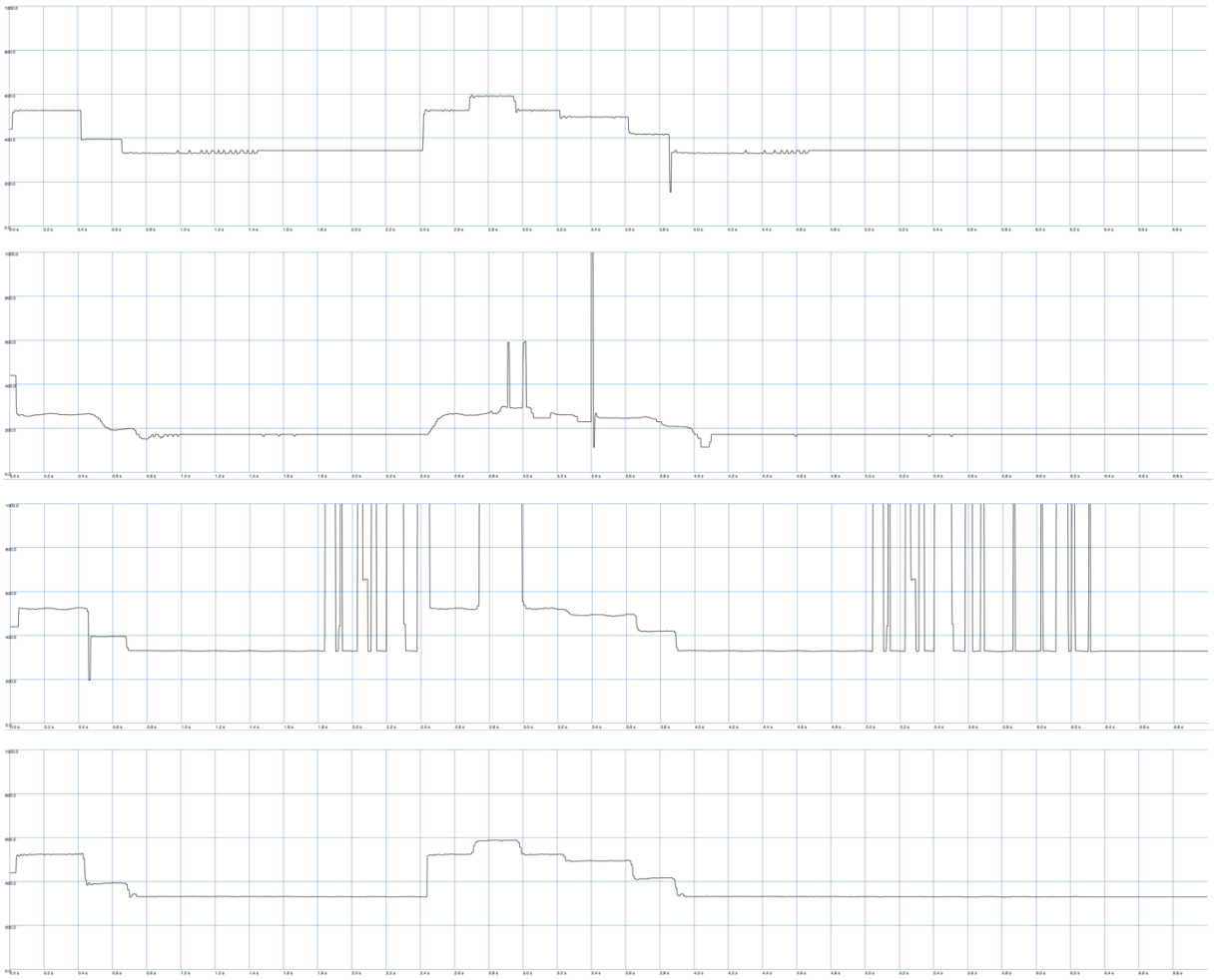

Figura 6: resultados da deteção de altura do algoritmo Pitch da melodia 2 cantada e nos timbres de piano, violino e flauta, respetivamente.

Melodia	Percentagem aproximada de deteção correta
1 voz	81%
1 piano	**100%**
1 violino	89%
1 flauta	95%
2 voz	95%
2 piano	99%
2 violino	69%
2 flauta	**100%**

Tabela 2: percentagens aproximadas de deteção de altura correta das melodias 1 e 2 pelo algoritmo Pitch.

As linhas de código introduzidas no SuperCollider para a obtenção dos gráficos anteriores são as seguintes:

```
(
{
var a, freq, hasFreq;
a = PlayBuf.ar(1, α);
# freq, hasFreq = Pitch.kr(a, minFreq: 100, maxFreq: 2000)
}.plot(β, minval: 0, maxval: 1000, bounds: Rect.new(width: 5000, height: 1000));
)
```

Os argumentos *minFreq* e *maxFreq* correspondem ao valor mínimo e máximo, respetivamente, que a frequência fundamental pode assumir, servindo por isso para restringir a deteção e evitar a obtenção de resultados errados cujos valores ultrapassam estes limites. De modo a tomar partido desta funcionalidade, mas não tornar a tarefa demasiado fácil, uma vez que um algoritmo de deteção de altura de música versátil deverá ser capaz de detetar alturas numa larga banda de frequências, foram escolhidos os valores de 100 e 2000 Hz para permitir o aparecimento de erros de oitavas.

3.1.3. Resultados: Qitch

Os gráficos obtidos (figuras 7 e 8) mostram novamente frequências dos 0 aos 1000 Hz e unidades de tempo dos 0 aos 9 ou dos 0 aos 7 segundos. A tabela abaixo (tabela 3) indica as percentagens aproximadas de deteção de altura correta de todos os áudios.

Figura 7: resultados da deteção de altura do algoritmo Qitch da melodia 1 cantada e nos timbres de piano, violino e flauta, respetivamente.

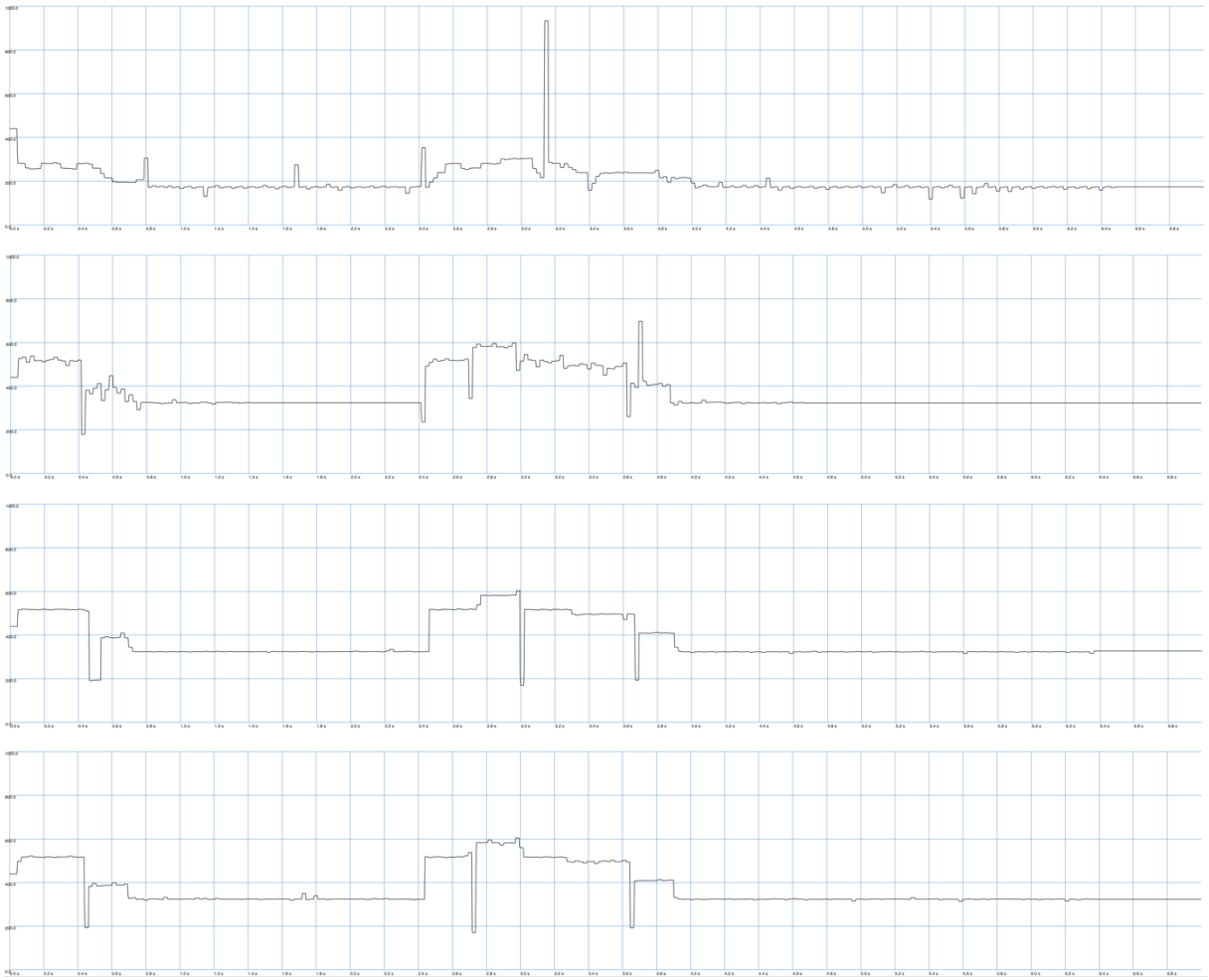

Figura 8: resultados da deteção de altura do algoritmo Qitch da melodia 2 cantada e nos timbres de piano, violino e flauta, respetivamente.

Melodia	Percentagem aproximada de deteção correta
1 voz	72%
1 piano	77%
1 violino	83%
1 flauta	84%
2 voz	73%
2 piano	81%
2 violino	93%
2 flauta	93%

Tabela 3: percentagens aproximadas de deteção de altura correta das melodias 1 e 2 pelo algoritmo Qitch.

As linhas de código introduzidas no SuperCollider para a obtenção dos gráficos anteriores são as seguintes:

```
(
{
var a, freq, hasFreq;
a = PlayBuf.ar(1, α);
# freq, hasFreq = Qitch.kr(a, databufnum: 1, minfreq: 100, maxfreq: 2000)
}.plot(β, minval: 0, maxval: 1000, bounds: Rect.new(width: 5000, height: 1000));
)
```

Os valores dos argumentos *minfreq* e *maxfreq* foram escolhidos pelas razões explicadas no subcapítulo anterior. O argumento *databufnum* refere-se ao índice do ficheiro de áudio auxiliar que define a resolução da transformada, que é neste caso de 2048 amostras, por constituir um bom compromisso entre precisão e custo computacional.

3.1.4. Resultados: Tartini

Os gráficos obtidos (figuras 9 e 10) mostram novamente frequências dos 0 aos 1000 Hz e unidades de tempo dos 0 aos 9 ou dos 0 aos 7 segundos. A tabela abaixo (tabela 4) indica as percentagens aproximadas de deteção de altura correta de todos os áudios.

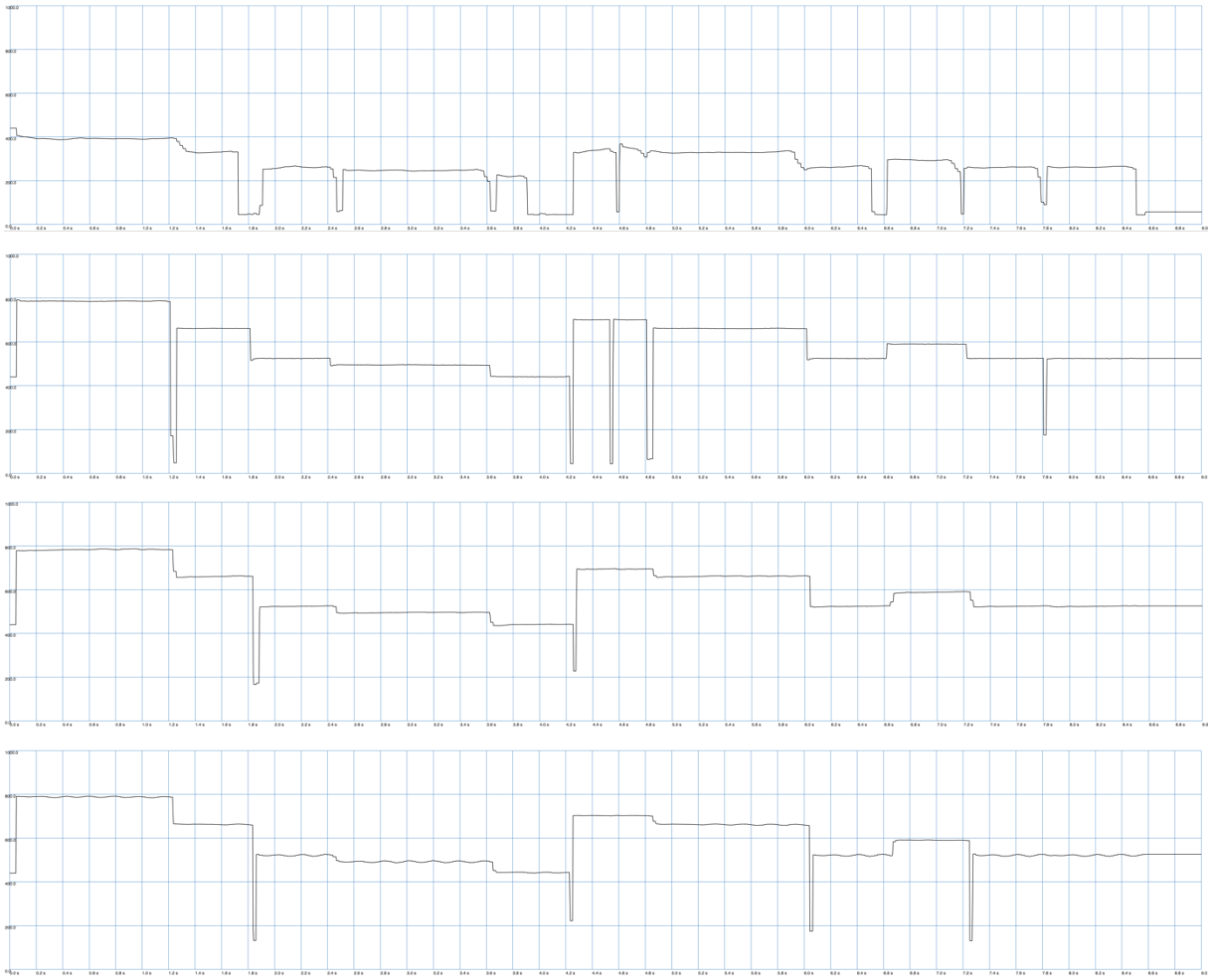

Figura 9: resultados da deteção de altura do algoritmo Tartini da melodia 1 cantada e nos timbres de piano, violino e flauta, respetivamente.

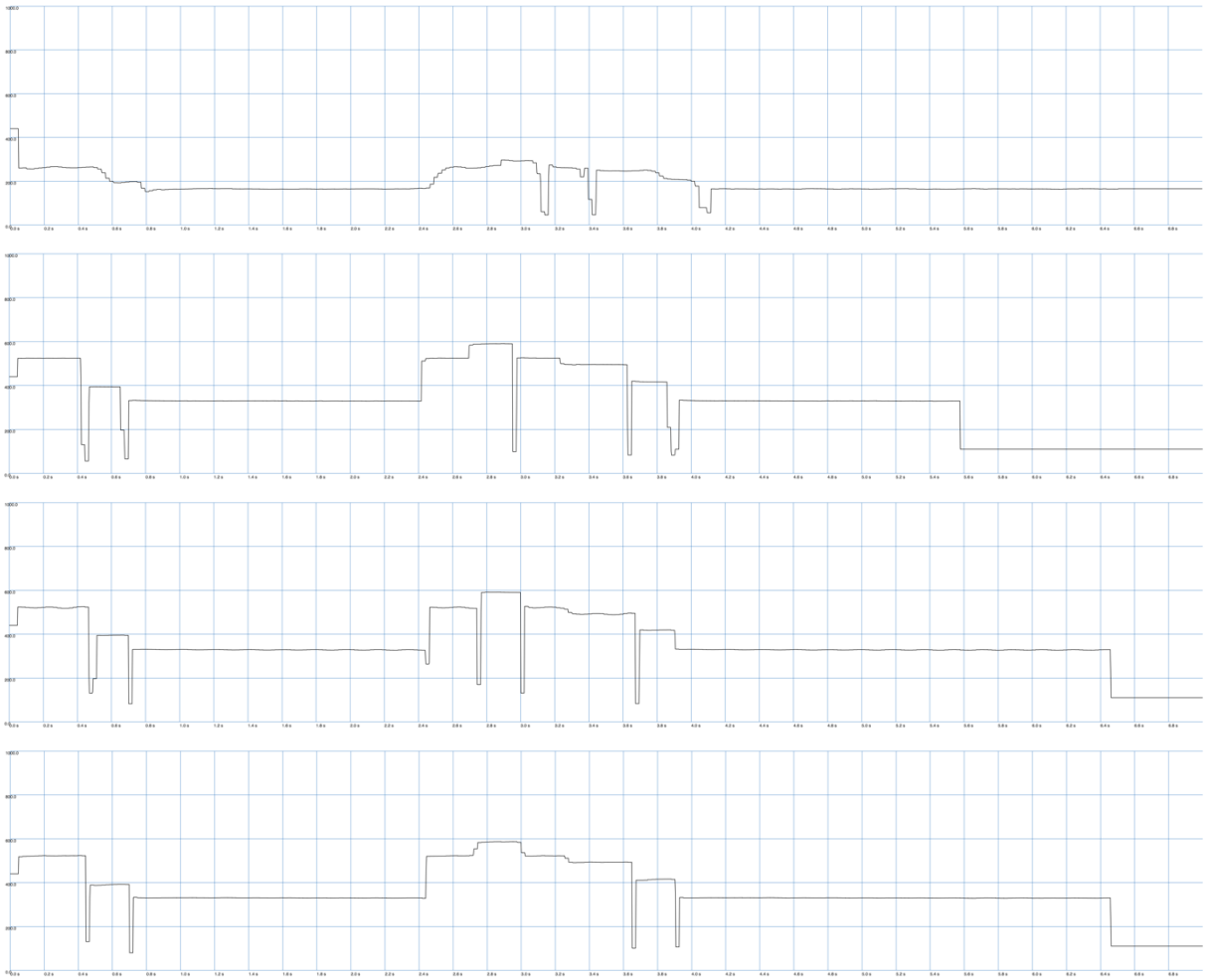

Figura 10: resultados da deteção de altura do algoritmo Tartini da melodia 2 cantada e nos timbres de piano, violino e flauta, respetivamente.

Melodia	Percentagem aproximada de deteção correta
1 voz	79%
1 piano	94%
1 violino	98%
1 flauta	96%
2 voz	94%
2 piano	80%
2 violino	94%
2 flauta	96%

Tabela 4: percentagens aproximadas de deteção de altura correta das melodias 1 e 2 pelo algoritmo Tartini.

As linhas de código introduzidas no SuperCollider para a obtenção dos gráficos anteriores são as seguintes:

```
(
{
var a, freq, hasFreq;
a = PlayBuf.ar(1, α);
# freq, hasFreq = Tartini.kr(a)
}.plot(β, minval: 0, maxval: 1000, bounds: Rect.new(width: 5000, height: 1000));
)
```

3.1.5. Discussão

A partir dos valores das percentagens aproximadas de deteção correta obtidas anteriormente, é possível construir uma nova tabela com as médias e desvios máximos dos valores referentes a cada um dos algoritmos (tabela 5). Os valores das médias podem ser interpretados como a precisão geral dos algoritmos, enquanto os valores dos desvios podem ser encarados como a consistência e confiabilidade dos mesmos, uma vez que um valor pequeno indica que existe pouca disparidade entre as precisões de deteção de altura obtidas.

Algoritmo	Média das percentagens aproximadas de deteção correta	Desvio máximo das percentagens aproximadas
ZeroCrossing	34%	49%
Pitch	**91%**	31%
Qitch	82%	21%
Tartini	**91%**	**18%**

Tabela 5: comparação das médias das percentagens aproximadas de deteção correta e dos desvios máximos das percentagens aproximadas dos algoritmos utilizados.

Como se pode observar nos gráficos e nos valores obtidos, o algoritmo ZeroCrossing, baseado na taxa de cruzamento de zero, é bastante limitado no que toca à deteção de altura. Esta limitação deve-se ao facto de que uma onda sonora complexa cruza frequentemente o zero em pontos que não correspondem ao período da mesma. Para além de ter obtido o pior resultado na precisão da deteção (34%), teve também o maior valor de desvio (49%), sendo por isso o algoritmo menos consistente.

O algoritmo Pitch obteve, tal como o algoritmo Tartini, o melhor resultado em termos de precisão (91%), provando que ambos são robustos na deteção de altura. Pitch conseguiu até dois resultados de precisão perfeita (100%) na melodia 1 tocada pelo piano e na melodia 2 tocada pela flauta, que nenhum outro algoritmo conseguiu. No entanto, no que toca ao seu valor de desvio (31%) e consistência, ficou atrás dos algoritmos Qitch e Tartini, que obtiveram um

desvio de apenas 21% e 18%, respetivamente. A deteção da melodia 2 tocada pelo violino foi particularmente repleta de erros de deteção.

De seguida, Qitch obteve resultados aceitáveis em relação à sua precisão (82%), embora esta esteja repleta de pequenos erros e inconsistências. Os erros ocorrentes durante a sonância das notas tendem a ser inferiores à oitava, enquanto os erros do algoritmo Pitch tendem a ser superiores. Os resultados referentes à melodia 1 cantada são provavelmente o melhor exemplo do modo como este algoritmo lida com os elementos desvozeados do sinal: os picos, que consistem em erros de deteção, correspondem às secções mais sibilantes do áudio cantado, neste caso às sílabas [ci] e [des] da palavra *felicidades*, [nos] de *anos* e [vi] de *vida*, uma vez que são as componentes com mais ruído espetral do mesmo.

Por fim, o algoritmo Tartini revela ser o mais robusto dos quatro testados nesta experiência, uma vez que obteve o melhor valor de precisão (91%), empatado com Pitch, mas também o menor valor de desvio (18%) de todos os algoritmos. Os vales ocorrentes nas mudanças de notas foram contabilizados como erros no cálculo da percentagem de precisão, embora a maioria tenha uma duração curta o suficiente para não afetar negativamente uma deteção de altura numa aplicação real. O erro mais flagrante deste algoritmo ocorreu na deteção da melodia 2 tocada pelo piano: aproximadamente a meio da sonância da última nota mi, a deteção assumiu um valor errado.

Conclusão

A investigação levada a cabo nesta dissertação teve como resultado a categorização e o levantamento do estado da arte da deteção de altura através de algoritmos digitais. As estratégias apresentadas servem de base para a maioria dos algoritmos de deteção de altura utilizados atualmente. Paralelamente, foram clarificados vários fenómenos psicoacústicos essenciais para a compreensão do funcionamento e dos desafios do desenvolvimento e da utilização destes algoritmos, tal como as suas aplicações na área da música.

Conclui-se que, embora haja estratégias com resultados tendencialmente mais precisos, a escolha de um algoritmo de deteção prende-se com uma série de fatores que dependem não só do áudio a analisar, mas também da aplicação da deteção, que irá necessitar de resultados mais ou menos precisos e mais ou menos rápidos, sendo esta rapidez uma consequência direta da eficiência computacional do mesmo. No que toca às suas aplicações musicais, deduz-se que este tipo de algoritmos têm já um papel relevante em várias vertentes da música, possibilitando uma vasta expansão tímbrica em *performances* musicais ou a manipulação e correção de elementos musicais na produção musical. Portanto, têm uma série de aplicações artísticas e técnicas cujas fronteiras se cruzam frequentemente.

No que toca especificamente à componente experimental desta dissertação, é possível concluir com os resultados obtidos que a deteção de altura no SuperCollider é viável e tem uma maior probabilidade de dar origem aos resultados mais precisos através da utilização do algoritmo Tartini, embora, como referido no parágrafo anterior, a escolha do algoritmo mais indicado para um determinado trabalho dependa de uma série de fatores. É importante realçar que a amostra utilizada nesta experiência pode não ser suficientemente significativa em termos tímbricos para assegurar a veridicidade da afirmação anterior. Para além disso, os algoritmos utilizados podem ter os seus parâmetros editados de modo a serem otimizados para determinadas deteções, como por exemplo a resolução da transformada de Qitch, que pode fazer divergir consideravelmente os valores obtidos. Não obstante, esta experiência indica valores representativos sob as condições padrão de utilização destes algoritmos.

Apêndice

Considera-se pertinente incluir nesta dissertação a explicação detalhada de uma aplicação artística da manipulação digital da altura de um som. O processador Northern Lights, criado no SuperCollider pelo autor no decorrer do mestrado, é inspirado em processadores digitais de efeitos para guitarra elétrica como o Helix HX Effects, da empresa Line 6. Este permite uma série de modificações de um som de entrada, que pode ser proveniente de um instrumento musical ou um ficheiro de áudio, nomeadamente a aplicação de reverberação e eco (*delay*). Permite também a modificação da altura (*pitch shifting*) desse mesmo som, embora não tome partido de algoritmos de deteção de altura. Ainda assim, será interessante demonstrar os universos sonoros que podem ser criados através destas manipulações.

Neste processador, a altura de um som de entrada pode ser alterada através de dois dos seus cinco módulos: Pitch Shifter e Harm. O primeiro permite a sua transposição para cima de acordo com os intervalos musicais de uma quinta, uma oitava ou duas oitavas ou a sua transposição para baixo de acordo com o intervalo de uma oitava. Permite também uma transposição gradual para cima e para baixo através da modulação da frequência do som de entrada a partir de uma onda dente de serra cuja frequência pode ser ajustada, dando origem a *glissandos* de diferentes velocidades. O módulo Harm é também um manipulador de altura que permite a sobreposição simultânea de três vozes transpostas cujos intervalos de transposição podem ser selecionados de uma tessitura máxima de duas oitavas com a resolução mínima de um meio-tom.

A utilização destas ferramentas de transposição, com o auxílio dos outros módulos do processador, permitem a criação de ambientes sonoros interessantes que esbatem as fronteiras entre música e design sonoro. Um vídeo que demonstra a utilização do processador Northern Lights (figura 1) pode ser acedido através da seguinte hiperligação:

https://www.youtube.com/watch?v=8OWzulCvxGc

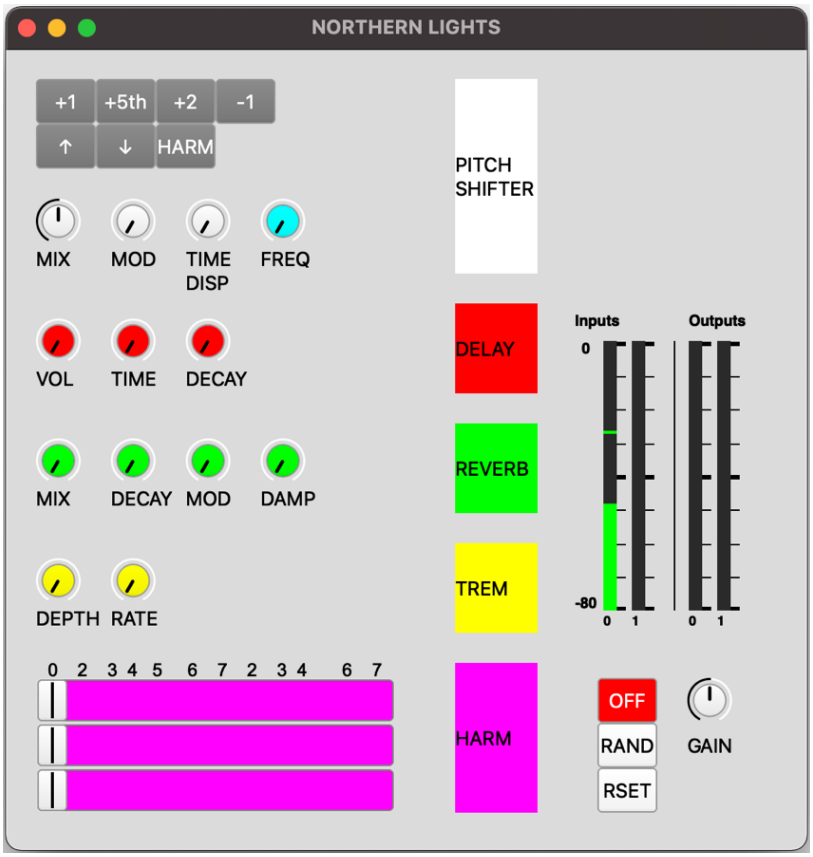

Figura 1: GUI (*graphical user interface*) do processador Northern Lights.

Bibliografia

Capítulo 1

Auditory filter. (n.d.). In APA Dictionary of Psychology. Retrieved January 14, 2023, from https://dictionary.apa.org/auditory-filter

Fletcher, H., & Munson, W. A. (1933). Loudness, Its Definition, Measurement and Calculation.

Fyk, J. (1987). Duration of Tones Required for Satisfactory Precision of Pitch Matching. Bulletin of the Council for Research in Music Education, 91, 38–44.

Graves, J. E., & Oxenham, A. J. (2017). Familiar Tonal Context Improves Accuracy of Pitch Interval Perception. Frontiers in Psychology, 8, 1753. https://doi.org/10.3389/fpsyg.2017.01753

Helmholtz, H. L. F. (2009). On the Sensations of Tone as a Physiological Basis for the Theory of Music (A. J. Ellis, Trans.; 3rd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511701801

Henrique, L. L. (2002). Acústica musical. Fundação Calouste Gulbenkian.

Mesz, B. A., & Eguia, M. C. (2009). The Pitch of Vibrato Tones: A Model Based on Instantaneous Frequency Decomposition. Annals of the New York Academy of Sciences, 1169(1), 126–130. https://doi.org/10.1111/j.1749-6632.2009.04767.x

Oxenham, A. J. (2012). Pitch Perception. Journal of Neuroscience, 32(39), 13335–13338. https://doi.org/10.1523/JNEUROSCI.3815-12.2012

Roederer, J. G. (2009). *The Physics and Psychophysics of Music*. Springer US. https://doi.org/10.1007/978-0-387-09474-8

Schaeffer, P., North, C., & Dack, J. (2017). Correlation between Spectra and Pitches. In Treatise on musical objects: essays across disciplines. University of California Press.

Stevens, S. S. (1935). The Relation of Pitch to Intensity. The Journal of the Acoustical Society of America, 6(3), 150–154. https://doi.org/10.1121/1.1915715

Stevens, S. S., & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. The American Journal of Psychology, 53(3), 329–353.

Suits, B. H. (2019). Frequency and Pitch. The Physics Teacher, 57(9), 630–632. https://doi.org/10.1119/1.5135796

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). The Journal of the Acoustical Society of America, 33(2), 248. https://doi.org/10.1121/1.1908630

kayageum1_F2. (2006). https://freesound.org/people/spt3125/sounds/24568/

Piano.mf.C1. (2001). https://theremin.music.uiowa.edu/MISpiano.html

violinarcovibA#4. (2008). https://freesound.org/people/ldk1609/sounds/55916/

Capítulo 2

Amado, R. G., & Filho, J. V. (2008). Pitch Detection Algorithms Based on Zero-Cross Rate and Autocorrelation Function for Musical Notes. *2008 International Conference on Audio, Language and Image Processing, ICALIP*.

Brown, J. C. (1991). Calculation of a Constant Q Spectral Transform. *The Journal of the Acoustical Society of America*, *89*.

Brown, J. C., & Puckette, M. S. (1992). An Efficient Algorithm for the Calculation of a Constant Q Transform. *The Journal of the Acoustical Society of America*, *92*(5).

Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, *124*(3), 1638–1652. https://doi.org/10.1121/1.2951592

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930. https://doi.org/10.1121/1.1458024

Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., & Velimirović, M. (2020). SPICE: Self-supervised Pitch Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1118–1128. https://doi.org/10.1109/TASLP.2020.2982285

Kasi, K., & Zahorian, S. (2002). *Yet Another Algorithm for Pitch Tracking*. IEEE International Conference on Acoustics, Speech, and Signal Processing.

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). *CREPE: A Convolutional Representation for Pitch Estimation*. arXiv. http://arxiv.org/abs/1802.06182

Klapuri, A. (2008). Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model. *IEEE Transactions on Audio, Speech and Language Processing*, *16*(2).

Mauch, M., & Dixon, S. (2014). *pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions*. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014).

McLeod, P., & Wyvill, G. (2005). *A Smarter Way To Find Pitch*. Proc. of Int. Computer Music Conf.

Noll, A. M. (1969). *Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum, and a Maximum Likelihood Estimate*. Symposium on Computer Processing in Communications.

Rabiner, L. R. (1977). On the Use of Autocorrelation Analysis for Pitch Detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-25*(1).

Schörkhuber, C., & Klapuri, A. (2010). *Constant-Q Transform Toolbox For Music Processing*. https://doi.org/10.5281/ZENODO.849740

Slaney, M., & Lyon, R. F. (1990). A Perceptual Pitch Detector. *1990 International Conference on Acoustics Speech and Signal Processing*, *1*, 357–360.

Veloso, J. (1997). Vozeamento, Duração e Tensão nas Oposições de Sonoridade das Oclusivas Orais do Português. *Línguas e Literaturas*, *14*, 59–80.

Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, *123*(6), 4559–4571. https://doi.org/10.1121/1.2916590

Haken. (2020). *ECE402 Lecture 20 (Pitch Detection)*.