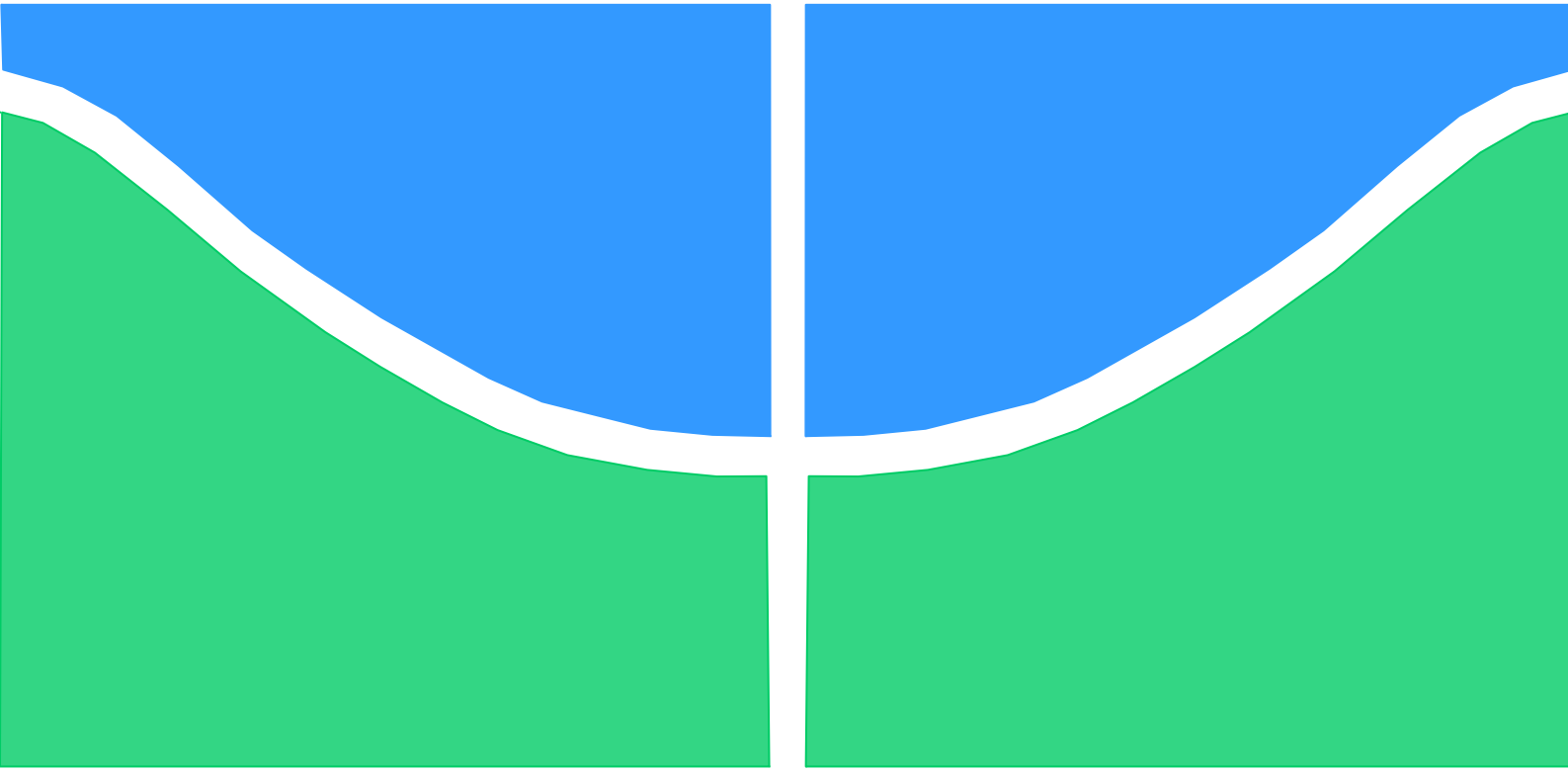

TRABALHO DE CONCLUSÃO DE CURSO

Processamento de Vídeo Aplicado a Recuperação de Vibrações Mecânicas

Por,
Lorena Pinheiro Castro

Brasília, Dezembro de 2017

UNIVERSIDADE DE BRASILIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASíLIA
Faculdade de Tecnologia
Departamento de Engenharia Elétrica

TRABALHO DE CONCLUSÃO DE CURSO

PROCESSAMENTO DE VÍDEO APLICADO A RECUPERAÇÃO DE VIBRAÇÕES MECÂNICAS

Por,
Lorena Pinheiro Castro

Trabalho de Conclusão de Curso submetido ao curso de Graduação em Engenharia Elétrica da Universidade de Brasília como requisito parcial para obtenção do Título de Bacharel em Engenharia Elétrica

Banca Examinadora

Prof. Leonardo R. A. X. Menezes, UnB/ ENE (Orientador) _____

Prof. João Paulo Leite, UnB/ ENE _____

Prof. Stefan Michael Blawid, UnB/ ENE _____

Brasília, Dezembro de 2017

FICHA CATALOGRÁFICA

CASTRO, LORENA PINHEIRO

Processamento de Vídeo Aplicado a Recuperação de Vibrações Mecânicas [Distrito Federal] 2017.

xvii, 60p., 210 x 297 mm (ENE/FT/UnB, Engenheiro, Engenharia Elétrica, 2017).

Trabalho de Graduação – Universidade de Brasília. Faculdade de Tecnologia.

Departamento de Engenharia Elétrica.

1. Vibrações	2. Microfone Visual
3. Ampliação de Vídeo	4. Processamento de Vídeo
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

CASTRO, L. P. (2017). Processamento de Vídeo Aplicado a Recuperação de Vibrações Mecânicas. Trabalho de Graduação em Engenharia Elétrica, Publicação FT.TG-nº , Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 60p.

CESSÃO DE DIREITOS

AUTORA: Lorena Pinheiro Castro.

TÍTULO: Processamento de Vídeo Aplicado a Recuperação de Vibrações Mecânicas.

GRAU: Engenheiro ANO: 2017

Lorena Pinheiro Castro
Rua 21 S, Lt 08, Bl. E. – Águas Claras.
71925-540Brasília – DF – Brasil.

DEDICATÓRIA

À Deus por tudo.

Lorena Pinheiro Castro.

AGRADECIMENTOS

A Deus que me guiou em todos os momentos da minha vida.

A toda minha a família, em especial minha mãe Lenilda e minha avó Cila, que sempre acreditaram nesse sonho e oraram a Deus por mim, e o meu pai Josemir que me apoiou em todos os momentos.

Ao meu filho, Pedro, que me trouxe mais alegria do que imaginei poder sentir e que me mostra ser mais forte e determinada do que acreditava ser.

Ao meu orientador, Professor Leonardo R.A.X, pela paciência, pelo companheirismo, carinho e conselhos que não me permitiram desfalecer durante esses árduos anos.

Aos meus professores de graduação que me ensinaram muito mais do que imaginam.

Aos meus colegas de graduação que me impulsionaram nos momentos mais difíceis e compartilharam os momentos de alegria.

A todos que de alguma forma compartilham comigo esta maravilhosa conquista.

Lorena Pinheiro Castro.

"O temor do Senhor é o princípio da sabedoria, e o conhecimento do Santo é entendimento."

Provérbios 9:10

RESUMO

O presente texto apresenta uma análise de vibrações, sob uma nova óptica, expondo como objetos reagem quando são atingidos por som e de qual modo é possível observar os acontecimentos do cotidiano extrapolando os movimentos sutis, quase imperceptíveis.

Palavras Chave: vibrações, microfone visual, ampliação de vídeo, processamento de vídeo.

ABSTRACT

The current text shows the analysis of vibrations, in a new perspective, exposing how objects react when they are hit by sound and ways of observing everyday events by extrapolating subtle movements, almost imperceptible.

Keywords: vibrations, visual microphone, video magnification, video processing.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A Engenharia é uma ciência inesgotável, e um modo de conceber novos projetos é observar os fenômenos conhecidos sob nova óptica. Neste sentido, as vibrações em objetos serão analisadas através de métodos pouco convencionais, fundamentados nas conclusões dos Doutores Abe Davis [3] e Michael Rubinstein [17].

Como Isaac Newton afirmou em 1687, em sua obra *Princípios Matemáticos da Filosofia Natural* [13], todos os corpos estão sujeitos a diferentes tipos de forças, gerando movimento, ainda que pareçam estacionárias.

Estes movimentos vibracionais podem ser causados por diferentes fontes. Quando causados por ondas mecânicas sonoras, a amplitude do movimento no objeto é pequena e sutil, porém agrega informações a respeito da estrutura do mesmo e da onda incidente.

Dessa forma, o objeto de estudo neste trabalho é obter as informações a respeito da fonte sonora e de quais formas essas vibrações mecânicas influenciam os objetos próximos, além de observar e estudar eventos pouco perceptíveis a olho nu, por meio de uma extrapolação da amplitude dos movimentos sutis.

As informações a respeito da onda sonora serão capturadas por câmeras filmadoras, traduzidas em sinais, descritas como vetores matemáticos, analisadas visualmente e novamente traduzidas em sons, associando as variações de amplitude do movimento do vídeo obtido com as características do sinal advindo da fonte [3, 4].

Em um sistema ideal, em que não há interferência de ruído na comunicação, a vibração mecânica, obtida do vídeo e descrita como vetor posicional no tempo, é traduzida em um sinal elétrico puro.

Então, como não há sobreposição de sinal ruidoso, a saída do sistema é dada pela subtração do vetor posicional entre um ponto de referência e outro ponto do meio, *pixel* a *pixel* do vídeo, em cada faixa de frequência do canal, o que é novamente traduzido em sinal sonoro, através de uma integração dos sinais parciais.

Entretanto, como o sinal sonoro apresenta informação em diversas escalas de frequência, é necessário que sua análise seja realizada em vários níveis, para tando, escolheu-se uma técnica de decomposição do sinal multi-escalar, variando a quantidade de escalas dependendo da necessidade.

Como o sistema apresenta distorções, tanto do sinal sonoro quanto na análise das imagens do vídeo, o processo de diferenciação vetorial precisa ser adaptado. Por isso, decompõe-se o sinal em sub-bandas de uma pirâmide escalonada, a fim de tentar isolar as distorções, e filtra-se as faixas de frequenciais desnecessárias para a análise do problema, o que gera uma melhor resposta espacial e temporal do sinal de entrada.

Por fim, há outra aplicação interessante da análise de vibrações em vídeos, por exemplo, identificar e ampliar processos físicos comuns porém pouco visualizados, através da manipulação computacional de um conjunto de imagens.

A fim de que a visualização destes fenômenos seja possível, é interessante que o segmento estudado seja primeiramente descrito matematicamente, então, filtrado em sua faixa de frequência de útil. Depois, o grupamento é decomposto espacialmente no tempo e o sinal é processado e ampliado, apresentando o resultado desejado [17, 27].

Pode-se aplicar esta metodologia em movimentos sutis que se quer extrapolar ou na mudança de cores de uma sequência de imagens, com o intuito de apresentar uma outra forma de observação do comportamento.

Por fim, é apresentado um esboço de uma aplicação dos conceitos e métodos apresentados neste trabalho, ainda que generalista.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão expostos os principais conceitos que regem o projeto. A teoria de processamento de imagens e de sons será apresentada inicialmente a fim de se entender como será realizada a análise dos vídeos.

2.1 INTRODUÇÃO

Apresentar-se-á a teoria matemática que descreve a análise de vibrações em vídeo e o processamento de sinais. Para isso, alguns conceitos devem ser explicados.

O som é uma onda mecânica que, quando aplicada sobre uma superfície rígida, causa pequenas vibrações. As vibrações geram um padrão de deslocamento que varia com o tempo. Uma câmera filmadora que capta esse conjunto de deslocamentos da superfície o converte em movimentos de *pixels* de um vídeo de alta velocidade, cerca de 1kHz a 20kHz.

Para simplificar o estudo, a movimentação dessas ondas será descrito como um vetor tridimensional, *V (x, y, t),* e particionar-se-á a análise do movimento vibracional em sinal local e global, em que o primeiro é relacionado ao movimento de cada *pixel* de imagem e o segundo, ao movimento como um todo, os quais podem ser mensurados em diferentes escalas e orientações ou linearmente.

Para estudar os sinais de movimento local e global, uma ferramenta de análise de imagens é necessária. Usar-se-á a pirâmide orientável complexa (*complex steerable pyramid* – CSP) [1, 16, 19, 20]. Esta técnica consiste em, a partir de uma imagem com várias orientações e escalas, decompô-la em um vetor linear, a fim de aplicá-lo em processamento de imagens.

Já que se trata de sons, principalmente de voz, deve-se apresentar conceitos de processamento de sinais sonoros, como a subtração espectral [15, 10] e relação sinal – ruído [12].

Além do mais, outra proposta de obter informação a partir de movimentos sutis é extrapolar estes movimentos, por meio do processamento computacional que utiliza técnicas eulerianas [17], através da decomposição de um sinal em faixas de frequências espaciais, filtragem temporal, ampliação e sobreposição a uma pirâmide espacial.

2.2 PROCESSAMENTO DE IMAGENS

Processamento de imagem é toda forma de processamento de dados em que a entrada e saída são imagens. Uma imagem é observada de acordo com sua textura, cores e forma. Se essa imagem é composta por objetos grandes e simples, uma visão grosseira é suficiente para examiná-la; se os objetos são pequenos, é necessária examiná-la em alta resolução. Porém, se há os dois tipos de objetos na imagem, então, deve-se analisá-la em várias resoluções, ou seja, deve-se utilizar a característica de multiresolução [22].

Dessa forma, é importante diferenciar os conceitos de resolução espacial e resolução em profundidade, ambos utilizados no processo. A resolução espacial de uma imagem está intrinsecamente relacionado ao seu número de *pixels*, ao passo que a resolução em profundidade está relacionada com o número de bits utilizado para representar os seus valores de um *pixel*.

Para analisar uma imagem, é interessante decompô-la em subpartes, a fim de não perder informações. Seguindo esta ideia, a teoria de pirâmides de imagens [22] apresenta uma forma de realizar essa decomposição em múltiplos níveis de resolução.

Suponha que há um conjunto de representações de uma imagem em resoluções espaciais diferentes, empilhadas uma a cima da outra. A imagem de maior resolução está na base da pilha e as imagens subsequentes situam-se sobre ela, em ordem decrescente, formando uma estrutura piramidal, como a Fig. 3.1 mostra.

Esse tipo de decomposição piramidal possui características multi-escalar, pois varia o número de escalas do sinal; e de multiresolução, uma vez que a resolução em profundidade das imagens varia a cada nível da pirâmide.

A fim de que o sinal original seja decomposto e reconstruído sem que haja perdas de informação, duas condições são necessárias, os operadores de análise e de síntese. A aplicação do primeiro garante que um sinal seja decomposto em sinais de mesma características. Já o segundo garante a reconstrução perfeita do sinal, sem perda de informação [22].

O sinal original, descrito como x_j, é decomposto em sinais que podem ser vistos como uma aproximação ou simplificação do sinal original, x_{j+1}, e o sinal que contém o erro, y_{j+1}, descartado a fim de obter-se o sinal simplificado mais próximo do original. A decomposição de um sinal de entrada em várias resoluções é dada por:

$$y_{j+1} = x_j - x_{j+1}, \tag{2.1}$$

com $j = 0, 1, \ldots, k-1$.

Esse processo é denominado *transformação pirâmide* de x_0 e a decomposição pode ser feita recursivamente:

$$x_0 \rightarrow \{x_1, y_1\} \rightarrow \{x_2, y_2, y_1\} \rightarrow \cdots \rightarrow \{x_k, y_k, y_{k-1}, \ldots, y_1\}$$

A reconstrução do sinal x_0 a partir dos sinais x_k e $y_1, y_2, \ldots, y_k$ é denominada transformação pirâmide inversa e dada pelo seguinte esquema recursivo de síntese:

$$x_j = y_{j+1} + x_{j+1}, \tag{2.2}$$

com $j = k\text{-}1, k\text{-}2, \ldots, 0$.

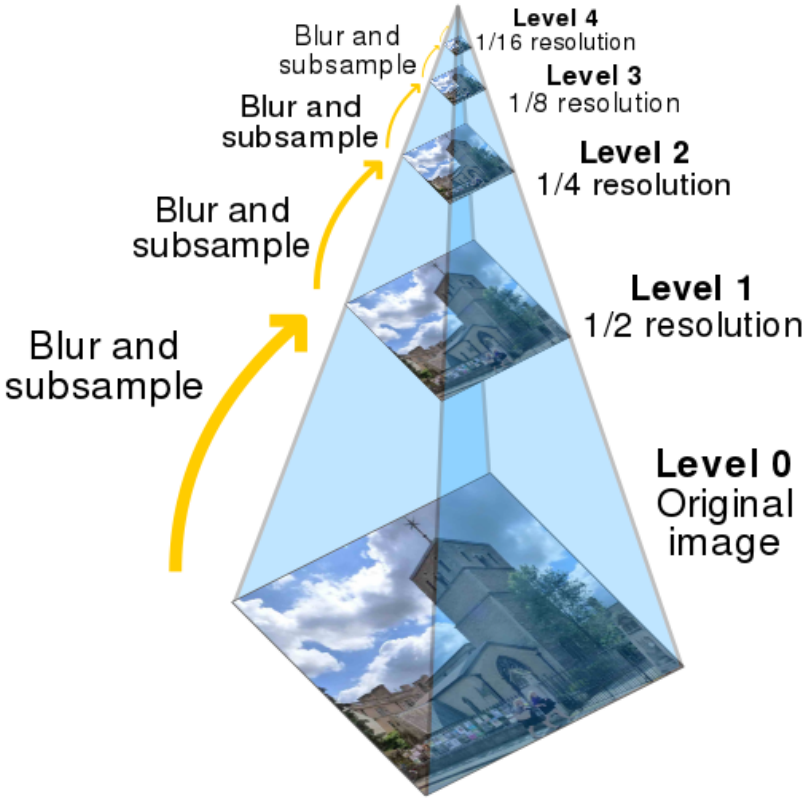

Figura 2.1: Representação de uma Pirâmide de Imagens com cinco níveis.

Se os operadores que definem uma pirâmide forem lineares, esta será uma pirâmide linear. Da mesma forma, se a decomposição for feita utilizando operadores não-lineares, temos uma pirâmide não-linear.

Uma vez explicado com que o sinal original pode ser decomposto e reconstruído, a partir do processo recursivo, utilizando os operadores de análise e de síntese, observa-se que o volume de dados deve ser reduzido a cada nível da pirâmide para que os algoritmos de multiresolução ofereçam vantagem segundo a visão computacional.

Assim, para que a representação de um sinal através de uma estrutura piramidal seja oportuna, é necessário um método de compressão, já que a quantidade de dados utilizados no sinal de saída deve ser menor do que o sinal original.

Para obter uma imagem de menor resolução, realiza-se uma filtragem e uma amostragem a cada nível, que consiste em gerar uma nova imagem composta por um subconjunto dos *pixels* da imagem original. Ou seja, esse procedimento substitui um quadrante da imagem origina por um único *pixel* de saída, assim, o número de *pixels* da imagem resultante é aproximadamente um quarto do número de *pixels* da imagem de entrada. Esse tipo de amostragem é conhecido como diádica.

2.3 ARQUITETURA *COMPLEX STEERABLE PYRAMID – CSP*

A arquitetura CSP têm se mostrado um eficiente método de decomposição de imagens em sub-níveis de escala e orientação, assim, é possível capturar a variação de textura em intensidade e orientação.

A proposta dessa arquitetura é combinar a decomposição multi-escalar com medidas diferenciais, assim, uma pirâmide de várias escalas é construída e, em seguida, operadores diferenciais (diferenças de *pixels* vizinhos) são aplicados aos sub-níveis ou sub-bandas da pirâmide [1, 16, 19, 20].

Já que a decomposição da pirâmide e a operação derivada são lineares e invariantes por mudança, podemos combiná-las em uma única operação, dessa forma, as primícias dessa arquitetura são os operadores derivativos direcionais de qualquer ordem desejada, já que pode-se obter uma pirâmide de quantos níveis forem necessários.

Inicialmente, a imagem é separada em sub-bandas de passagem baixa e alta, usando filtros L_0 e H_0, por exemplo. A sub-banda de passagem baixa, L_0, é, então, dividida em um conjunto de sub-bandas orientadas, B_0, …, B_k, e uma sub-banda de passagem baixa, L_1. Esta sub-banda de passagem baixa, L_1, é sub-amostrada nas direções X e Y. O esquema da decomposição da pirâmide recursiva é mostrada abaixo, na Fig. 2.2.

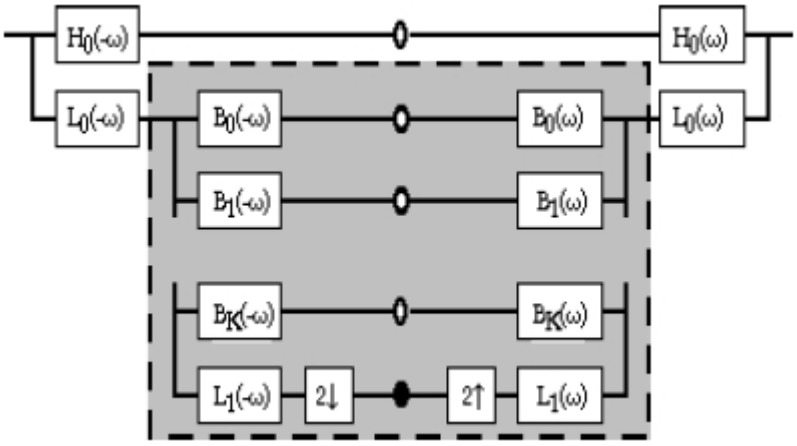

Figura 2.2: Esquema da decomposição das sub-bandas da pirâmide.

A condição necessária para que uma base de filtro seja orientável é que tenha capacidade de sintetizar um filtro de qualquer orientação a partir de uma combinação linear de filtros em orientações fixas e a condição necessária para que uma base de filtros de escala seja utilizada é que deve ser restringido por um diagrama de sistema recursivo [1]. O exemplo mais simples de um filtro orientável é a primeira derivada de filtros gaussianos, em 0° e 90°:

$$\curvearrowright \alpha_1 = 0^\circ, \alpha_2 = 90^\circ \tag{2.3}$$

A equação direcional será:

$$G_1^{\alpha}(x,y) = \cos(\alpha) G_1^{0^\circ}(x,y) + \sin(\alpha) G_1^{90^\circ}(x,y) \tag{2.4}$$

Dessa forma, é possível sintetizar um filtro com qualquer orientação a partir da combinação linear dos filtros $G_1^{0^\circ}$ e $G_1^{90^\circ}$, então, a partir das imagens com esses filtros, pode-se gerar uma imagem resultante da combinação linear de sua convolução. Ou seja, sejam as seguintes convoluções dos respectivos filtros e imagens:

$R_1^{0^\circ} = G_1^{0^\circ} * I$ e $R_1^{90^\circ} = G_1^{90^\circ} * I$

A imagem resultante será:

$$R_1^{\alpha}(x,y) = \cos(\alpha) G_1^{0^\circ}(x,y) + \sin(\alpha) G_1^{90^\circ}(x,y) \tag{2.5}$$

A Figura 2.3 ilustra este processo.

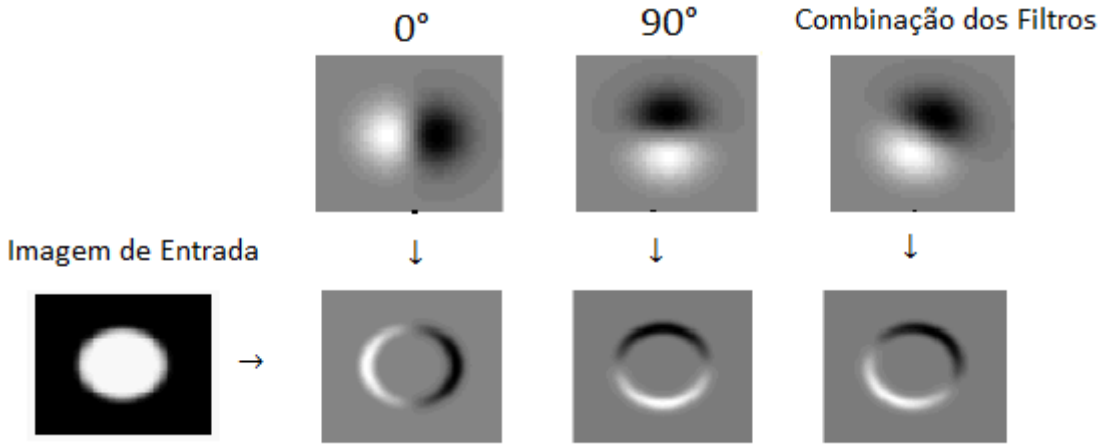

Figura 2.3: Imagem resultante da combinação de filtros orientáveis.

A decomposição é mais facilmente entendida no domínio da Fourier, em que é idealmente separável em polar. Escreve-se a magnitude de Fourier do i-ésimo filtro de passagem na notação polar:

$$B_i(\vec{w}) = A(\theta - \theta_i)\, B(w), \tag{2.6}$$

em que $\theta = \tan^{-1}(w_y / w_x)$, $\theta_i = \frac{2\pi}{k}$ e $w = |\vec{w}|$.

A componente angular da decomposição é determinada pela ordem da derivada desejada. Uma operação de derivação direcional no domínio espacial corresponde à multiplicação por uma função rampa linear no domínio de Fourier, que pode ser reescrita nas coordenadas polares, na direção x, da seguinte forma:

$$-jw_x = -jw\cos(\theta) \tag{2.7}$$

Não se considera a constante imaginária e o fator de w é absorvido pela componente radial da função.

Assim, em uma derivação direcional de ordem N, a componente angular do filtro é $\cos(\theta)^N$, uma vez que, no domínio de Fourier, corresponde à multiplicação pela função rampa elevada a N-ésima potência.

A componente radial obedece a característica piramidal, ou seja, recursiva. Como demonstrado na Figura 2.2, os filtros H_0 e L_0 são utilizados no processamento da imagem em preparação para o sistema recursivo, em que este decompõe um sinal em duas partes, passa alta

e passa baixa. O elemento de passa baixa participa do processo recursivo e geralmente escolhe-se $L_0 = L_1$, de modo que a componente inicial será a mesma usada na recursividade [1].

2.4 SINAL DE MOVIMENTO LOCAL

Sabendo que uma imagem pode ser mensurada em diferentes orientações e escalas, pode-se definir o s*inal de movimento local* como o movimento de cada *pixel* de um vídeo [4].

As funções básicas da arquitetura CSP são escalonadas e orientadas com componentes cossenoidais e senoidais, sendo utilizadas no domínio de Fourier. A textura em torno de cada *pixel* da imagem é a representada pelas componentes de amplitude e fase da Transformada de Fourier. Entretanto, a amplitude mede o contraste em uma imagem e mudanças nesta textura resultam em mudanças na componente de fase, denominada fase local.

Um vídeo será representado por um vetor, *V(x, y t)*, que relaciona o deslocamento espacial em um intervalo de tempo. E, então, cada quadro de entrada de *V* será expresso como uma imagem complexa, ou seja, em uma escala *r* e uma orientação θ . Isto pode ser escrito em função da amplitude e fase locais:

$$A(r,\theta,x,y,t)e^{i\phi(r,\theta,x,y,t)} \tag{2.8}$$

A partir disto, as variações na fase ao longo do tempo estão relacionadas linearmente com as mudanças na textura da imagem.

Agora, para quantificar o movimento, tomam-se as fases locais, ϕ , de cada quadro e subtrai-se da fase local do quadro de referência, que pode ser o primeiro quadro do vídeo, t_0, gerando a seguinte variação de fase:

$$u(r,\theta,x,y,t)=\phi(r,\theta,x,y,t)-\phi(r,\theta,x,y,t_0). \tag{2.9}$$

Referir-se-á como *deslocamento local* para o vetor $u(r,\theta,x,y,t)$; e *sinal de movimento local* para o sinal $u_l(t)$, que é um sinal de deslocamento local ao longo do tempo, com índice *l* denotando uma localização, orientação e escala particular.

Para movimentos ínfimos, as variações de fase são proporcionais aos deslocamentos de estruturas de imagens ao longo da orientação e escala correspondentes [6].

Em uma imagem com regiões com muito contraste, os sinais de movimento locais dados pela Eq. (2.9) são uma boa medida do movimento em vídeo. Porém, quando a imagem possui regiões com pouco contraste, a informação da fase local é facilmente confundida com o ruído, assim, essa qualidade não é própria para medir o movimento.

As amplitudes locais $A(r,\theta,x,y,t)$ promovem uma boa medida de contraste da imagem, as quais podem ser usadas como um valor de confirmação, a fim de prever o ruído em sinais de movimento. Dessa forma, os pontos das imagens com muito contraste possuem alta amplitude local e pouco ruído, logo, muita confiabilidade.

Já os pontos das imagens com pouco contraste possuem muito ruído e baixa confiabilidade. Como mostrado na Fig. 2.4, o ponto com menor amplitude (P1) têm maior variação em fase do que o ponto com maior amplitude (P2), que é observado no sistema de eixo complexo. Os pontos da imagem com contraste serão denominados contraste local.

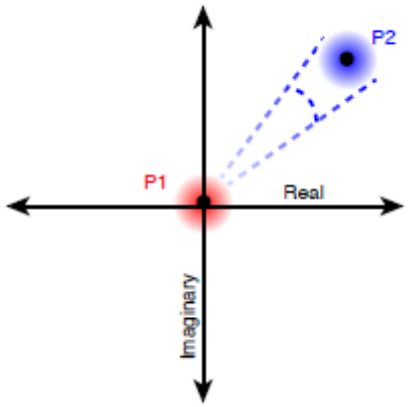

Figura 2.4: Representação de dois pontos no sistema de eixo complexo.

Retirado da referência [3].

2.5 SINAL DE MOVIMENTO GLOBAL

O sinal de vídeo registra o movimento em diversos pontos e dimensões no espaço, gerando um grande volume de dados redundantes, a depender da aplicação. Assim, uma forma de contornar este problema é calcular a média de movimentos locais em um único *sinal de movimento global* [4].

Inicialmente, obtêm-se os sinais de movimentos locais ponderadas, em que a medida de confiança é o quadro de contraste local, ou seja:

$$u(r,\theta,x,y,t)=A(r,\theta,x,y,t)^2 u(r,\theta,x,y,t) \tag{2.10}$$

Para cada orientação e escala, calcula-se a soma dos sinais de movimentos locais ponderados, a fim de gerar um único sinal de movimento intermediário, $a(r,\theta,t)$:

$$a(r,\theta,t)=\sum_{x,y} u(r,\theta,x,y,t) \tag{2.11}$$

É importante notar que antes de calcular a média dos sinais, deve-se alinhar os sinais de movimento local que correspondem a diferentes orientações, para evitar interferências destrutivas.

Os sinais alinhados são dados por $a_i(t-t_l)$, tal que:

$$t_l=\underset{t_l}{argmax}\, a_0(t)^T a_l(t-t_l), \tag{2.12}$$

em que $a_0(t)$ é um sinal de movimento local ponderado de referência e *l* é um índice que conecta os pares de escala e orientação (r,θ).

O sinal de movimento global será, então:

$$\hat{s}(t)=\sum_l a_l(t-t_l). \tag{2.13}$$

2.6 SUBTRAÇÃO ESPECTRAL

Em processamento de sinais de voz, a presença do ruído pode diminuir a qualidade ou a inteligibilidade do sinal. Por conseguinte, algorítmos específicos para essa atuação foram desenvolvidos. Eles são baseados na teoria de subtração espectral [2, 10, 18].

A abordagem de subtração espectral é um método eficaz de suprimir ruído de um sinal sonoro e sua manipulação é realizada no domínio da frequência. Este artifício baseia-se na propriedade de que o espectro na frequência do sinal é expresso como a soma do espectro de voz e espectro do ruído. Ou seja:

$$y(n)=x(n)+v(n), \tag{2.14}$$

tal que *y(n)*, *s(t)* e *r(t)* é o sinal com ruído, o sinal sem ruído e o ruído, respectivamente.

Calculando a Transformada Discreta de Fourier (DFT) da Eq. (3.14), tem:

$$Y(e^{jw_k})=X(e^{jw_k})+V(e^{jw_k}). \tag{2.15}$$

Uma vez obtido o espectro do sinal com ruído $Y(e^{jw_k})$ e uma aproximação da média do espectro do ruído $\mu(e^{jw_k})$, obtido dos trechos do sinal que predomina o silêncio, pode-se estimar o espectro do sinal de voz:

$$|\hat{X}(e^{jw_k})|=|Y(e^{jw_k})|-|\mu(e^{jw_k})|. \tag{2.16}$$

Como a intenção é obter este sinal estimado no domínio do tempo, o espectro de magnitude desse sinal é conciliado com a fase do sinal com ruído, e, então, calculada a Transformada Discreta de Fourier Inversa (IDFT), logo:

$$\hat{x}(n)=\sum_{k=0}^{N-1}|\hat{X}(e^{jw_k})|e^{j\theta_Y(e^{jw_k})}e^{-jw_k n}, \tag{2.17}$$

em que $w_k=\frac{2\pi}{N}k$ é a frequência discreta da transformada.

Portanto, estes conceitos visam reduzir o erro espectral do sinal de voz.

2.7 RELAÇÃO SINAL – RUÍDO

Frequentemente abreviada por SNR (*signal-to-noise ratio*, do inglês), a relação sinal – ruído de um sinal em um meio com ruído é uma métrica regularmente utilizada para comparar o nível do sinal desejado com o nível de ruído presente no meio [12].

Definida como a razão entre a potência de um sinal e a potência do ruído sobreposto ao sinal, a relação sinal – ruído é obtida da seguinte forma:

$$SNR=\frac{P_{sinal}}{P_{ruído}}=\left(\frac{A_{sinal}}{A_{ruído}}\right)^2. \tag{2.18}$$

As potências dos sinais analisados podem ser aferidas do valor quadrático médio (RMS) de suas respectivas amplitudes, medidas em pontos equivalentes em um sistema de comunicação e respeitando a largura de banda do canal.

Assim, quanto mais alta for a relação – ruído, menor será o efeito do ruído sobre a medição do sinal, ou seja, para que a comunicação seja entendida, é necessário que haja um valor mínimo da SNR no receptor, portanto, um grande valor da potência na entrada assegura esse valor mínimo ao final da comunicação.

2.8 CONCLUSÃO

Foi exposto neste capítulo os principais conceitos que regem o projeto, como a teoria de processamento de imagens e de sons, a fim de se entender a análise dos vídeos e a manipulação matemática que descreve a análise de vibrações em vídeos.

A fim de simplificar a metodologia que rege o microfone visual, os sinais de entrada do processo foi descrito como um vetor tridimensional e o movimento vibracional da partícula foi caracterizado como sinal local e global, em que o primeiro é relacionado ao movimento de cada *pixel* de imagem e o segundo, ao movimento como um todo, os quais podem ser mensurados em diferentes escalas e orientações ou linearmente.

Os sinais de movimento local e global foram analisados a partir da arquitetura da pirâmide orientável complexa (CSP), a qual decompõe uma imagem em um vetor linear, em várias orientações e escalas.

Os sinais sonoros exigem que haja uma manipulação especial, já que seu efeito é diferente em diversas escalas, por isso, os conceitos de subtração espectral e relação sinal – ruído foram apresentados nesse contexto.

Os próximos capítulos apresentarão e explicarão as técnicas a respeito do microfone visual e da ampliação de vídeo euleriana, além da manipulação matemática que rege o respectivo sistema e exemplos de experimentos realizados com seus resultados.

3 MICROFONE VISUAL

3.1 INTRODUÇÃO

A onda sonora é uma variação de pressão que se movimenta através de um meio físico. Sempre que ela atinge uma superfície, o objeto pode se mover com o meio circundante ou deformar de acordo com seus modos de vibração.

Em ambos os casos, o padrão de movimento, observado com o auxílio de uma câmera filmadora de alta qualidade, contém informações úteis que podem ser usadas para recuperar o som ou aprender sobre a estrutura do objeto [3, 4, 17, 24, 26, 28].

A proposta neste momento é incidir um sinal áudio – por exemplo a voz humana ou uma música – sobre um objeto, gravar suas vibrações e convertê-las novamente em um outro sinal de áudio. Esta gravação das vibrações deve ser um sinal visual que será usado para recuperar parcialmente o som incidente [3].

Recuperar sons a partir de vídeo depende de alguns fatores, como as propriedades físicas dos objetos analisados, a qualidade da câmera e sua posição.

Apesar de o som recuperado não ser idêntico ao som original, esse método tem como principal vantagem a coleta de um sinal de áudio sem a necessidade de acesso físico ao meio de circulação do som.

Dessa forma, este capítulo apresenta detalhadamente como ocorre o processo de recuperação do som, a partir do conjunto de imagens de vibrações obtidas, mostrando a manipulação matemática necessária para descrever os sinais como vetores posicionais no tempo.

Inicialmente algumas restrições a respeito do objeto e do ambiente são expostas, a fim de fornecer uma visão geral de possíveis experimentos. Além disso, explica-se a técnica de processamento do sinal para obter o sinal de áudio final.

3.2 METODOLOGIA

Os microfones tradicionais necessitam de membranas vibrantes e componentes eletromecânicos para traduzir sinais sonoros em elétricos. No microfone visual, o som é capturado medindo a distância entre os pontos de impacto causados pela pressão de ar na superfície de um objeto.

Este conjunto de movimentos de *pixels* é a entrada do processo, *V (x, y, t)*, e o movimento relativo entre o objeto e a câmera, causado por vibrações do som, é $s(t)$.

As imagens são, então, processadas por um algorítimo, a fim de gerar um sinal de saída, proporcional ao sinal de entrada. A Fig. 3.1 apresenta esse processo.

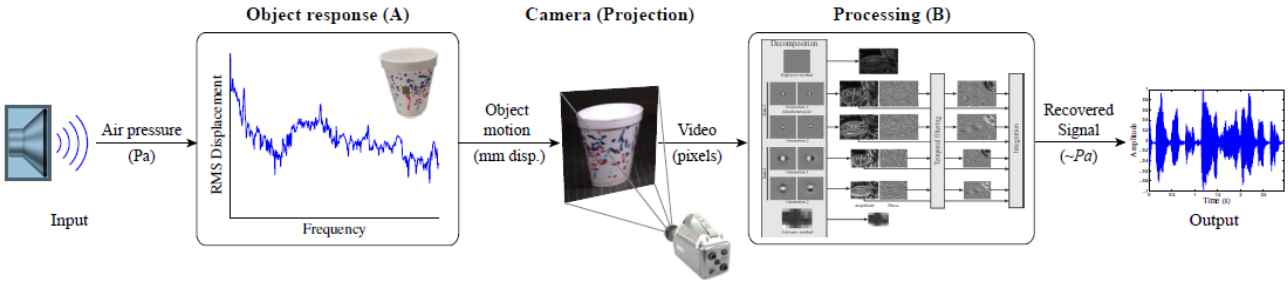

Figura 3.1: Modelo do processo de operação do microfone visual. Retirado da referência [4].

A Fig. 3.1 apresenta uma sequência de fases entre os sinais de entrada e de saída. O movimento apresentado a partir da incidência do som chama-se resposta do objeto, A. E o encadeamento do sinal utilizado é o algorítimo de processamento, B.

Como o objetivo é recuperar o som que atinge um objeto a partir do vídeo, então, uma possibilidade é relacionar o sinal de movimento local com o som, sem adicionar o ruído, ou seja, utilizar as vibrações ressonantes do objeto.

Assim, da resposta do objeto, presente na fase A, o interessante é obter o sinal de movimento global $\hat{s}(t)$ nas dimensões de uma pirâmide orientável complexa construída no vídeo. Este sinal local é alinhado, depois calcula-se a sua média em um único sinal de movimento de uma dimensão que capta o movimento global do objeto ao longo do tempo e o faz passar por um filtro de áudio.

É importante observar que inicialmente a modelagem da resposta A é um evento físico, em que se pega as variações de pressão no ar, medidas em Pascal, e as quantifica como deslocamento físico do objeto, medido em milímetros. Somente depois disso que o sistema físico é convertido em elétrico, quando os deslocamentos físicos são representados como movimentos de *pixels* das imagens filmadas.

O filtro de áudio utilizado tem como principais função atenuar as distorções e melhorar a qualidade da comunicação, quantificada pela relação entre o sinal com informação e o ruído presente.

Uma métrica quantitativa padrão em processamento de sinais é obtida através da razão entre a potência do mesmo e a potência do ruído sobreposto ao sinal, a relação sinal – ruído (SNR). Assim, quanto mais alta for a SNR, menor será o efeito do ruído de fundo sobre a detecção ou do sinal [12].

Este método produz uma medida espacial do som, ou seja, um sinal de áudio estimado em cada *pixel* no vídeo, sendo possível usá-lo para analisar deformações induzidas por um sinal de áudio de um objeto.

Para tanto, a implementação desde projeto utiliza apenas uma câmera de alta velocidade e um objeto, dispensando qualquer iluminação ativa, sensores adicionais, módulos de detecção diferenciais, superfície vibratória retrorreflectora ou especular e restrições à orientação da câmera.

Enquanto o sinal sonoro não tem restrições quanto ao seu meio de propagação, o objeto a ser utilizado tem, já que a qualidade da propagação da onda sonora e da recuperação do som depende de fatores estruturais do objeto, como a sua forma, densidade e compressibilidade.

Algumas etapas devem ser seguidas para recuperar o sinal sonoro, compondo a etapa do processamento B, melhor visualizada na Fig. 3.2.

Primeiramente, decompõe-se o vídeo de entrada em sub-bandas espaciais correspondentes às respectivas orientações e escalas.

A seguir, calcula-se os sinais de movimento local em cada *pixel*, orientação e escala, em que se utiliza variações de fase em uma representação piramidal orientável complexa do vídeo, $V(x, y, t)$.

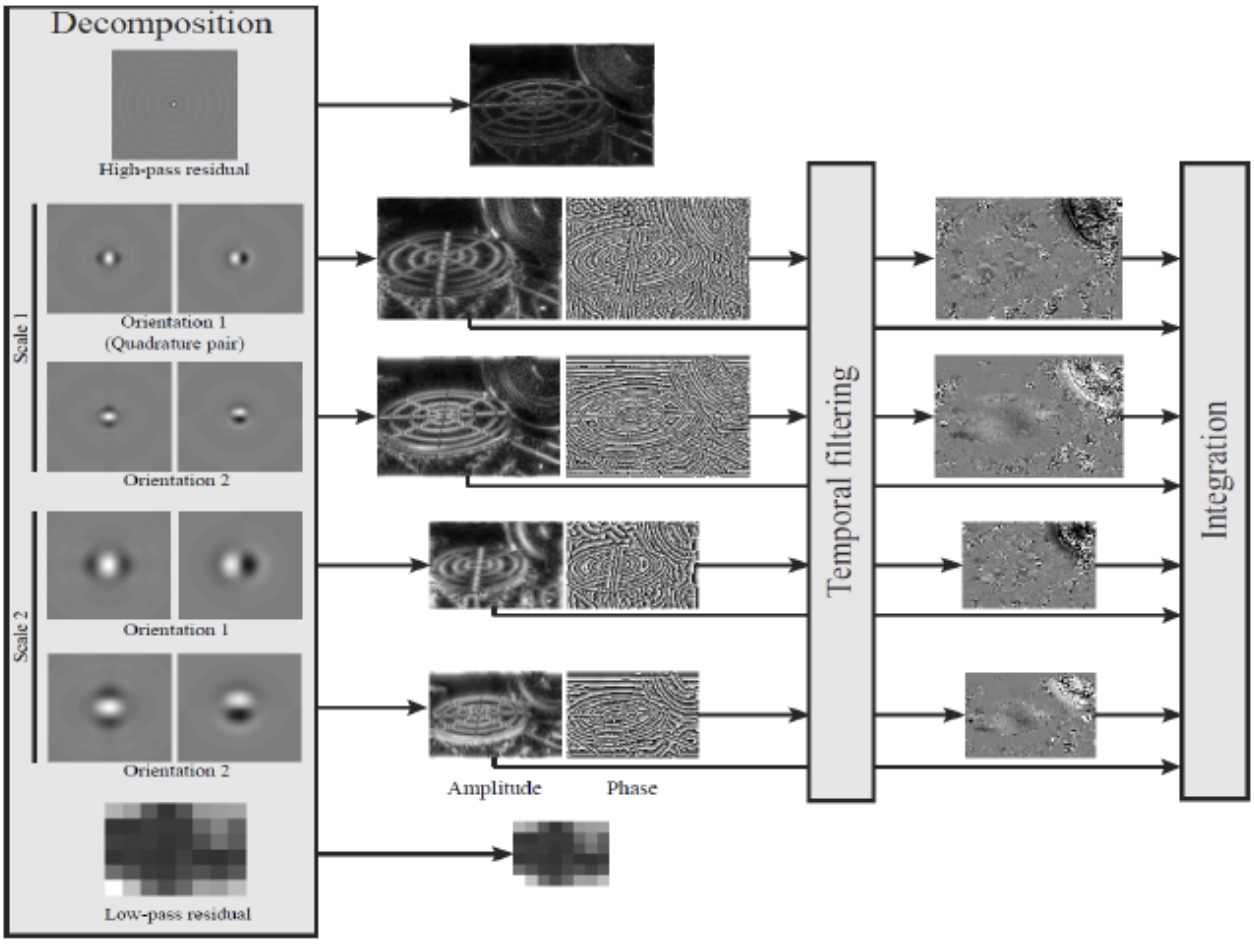

Figura 3.2: Algoritmo de processamento B. Retirado da referência [4].

Como descrito no Capítulo 2, a arquitetura piramidal orientável complexa (CSP) é um conjunto de filtros que desmembra cada quadro do vídeo em sub-bandas complexas para diferentes escalas e orientações, sendo que cada escala, r, e orientação, θ, pode ser escrita em termos de amplitude e fase:

$$A(r,\theta,x,y,t)e^{i\phi(r,\theta,x,y,t)} \tag{3.1}$$

Percebe-se da Fig. 3.2 que a amplitude e a fase são bem distintas em cada escala.

A fim de obter as variações de fase, subtrai-se das fases locais, φ, um quadro de referência, t_0:

$$u(r,\theta,x,y,t)=\phi(r,\theta,x,y,t)-\phi(r,\theta,x,y,t_0). \tag{3.2}$$

Então, combina-se esses sinais de movimento local através de uma sequência de operações de média para produzir um único sinal de movimento global para o objeto, em que, para isso, é preciso obter uma média espacial ponderada dos sinais de movimento local:

$$a(r,\theta,t)=\sum_{x,y} u(r,\theta,x,y,t)=\sum_{x,y} A(r,\theta,x,y,t)^2 u(r,\theta,x,y,t) \tag{3.3}$$

Lembrando que é importante alinhar os pontos temporariamente para evitar interferências destrutivas, por isso, as mudanças nas fases de nossa orientação x e y serão negativamente correlacionadas, somando sempre a um sinal constante. Então, os sinais alinhados são dados por $a_l(t-t_l)$, em que:

$$t_l=\underset{t_l}{argmax}\, a_0(t)^T a_l(t-t_l), \tag{3.4}$$

tal que $a_0(t)$ é um sinal de movimento local ponderado de referência e *l* é um índex que conecta os pares de escala e orientação (r,θ).

O sinal de movimento global será, então:

$$\hat{s}(t)=\sum_l a_l(t-t_l). \tag{3.5}$$

Finalmente, aplicamos técnicas de recuperação e filtragem de áudio ao sinal de movimento do objeto para, então, realizar a composição dos sinais, a fim de obter o som de saída do processo, descritas como o filtro temporal e integração, respectivamente, na Fig. 3.2.

A integração dos sinais parciais é realizada pela soma dos seus respectivos vetores, em que os sinais são ponderados com pesos quadráticos de amplitude, para cada escala e orientação da pirâmide. A fração do código que realiza esta operação é descrita a seguir:

```
for j = 1:nScales
    bandIdx = 1 + (j-1)*nOrients + 1;
    curH = pind(bandIdx,1);
    curW = pind(bandIdx,2);
    for k = 1:nOrients
       bandIdx = 1 + (j-1)*nOrients + k;
       amp = pyrBand(pyrAmp, pind, bandIdx);
       phase = pyrBand(pyrDeltaPhase, pind, bandIdx);
```

```
        phasew = phase.*(abs(amp).^2);
        sumamp = sum(abs(amp(:)));
        ampsigs(j,k,q)= sumamp;
       signalffs(j,k,q)=mean(phasew(:))/sumamp;
    end
  end
```

Do código acima, percebe-se que a integração dos sinais pode ocorrer em todo o espectro de frequência, desde o ajuste mais mais fino, com maior frequência espacial; ao mais grosso, com menor frequência.

No caso do microfone visual, observa-se que as distorções ocorrem com maior intensidade nas frequências mais baixas, nas faixas que não correspondem ao áudio de entrada. Para atenuar essas distorções, aplica-se um filtro passa-alta para bloquear as frequências mais baixas, respeitando o critério de utilizar uma frequência de corte correspondente a 1/20 da frequência de Nyquist.

Uma vez que o microfone visual opera com sinais sonoros, principalmente de voz, o algoritmo a ser implementado que integra os sinais parciais e converte-os em sinal sonoro deve focar na percepção da fala humana, a fim de garantir a inteligibilidade do sinal resultante, para tanto, é possível utilizar conceitos de subtração espectral do sinal recuperado [2, 10].

A partir da observação do processo, percebe-se que os principais fatores que interferem na qualidade do sinal de saída serão essas fases do processo.

Um exemplo de aplicação é, em um ambiente iluminado com luz solar, uma embalagem plástica é filmada através de um vidro insonorizado, enquanto uma pessoa pronuncia uma frase como demonstrado na Fig. 3.3.

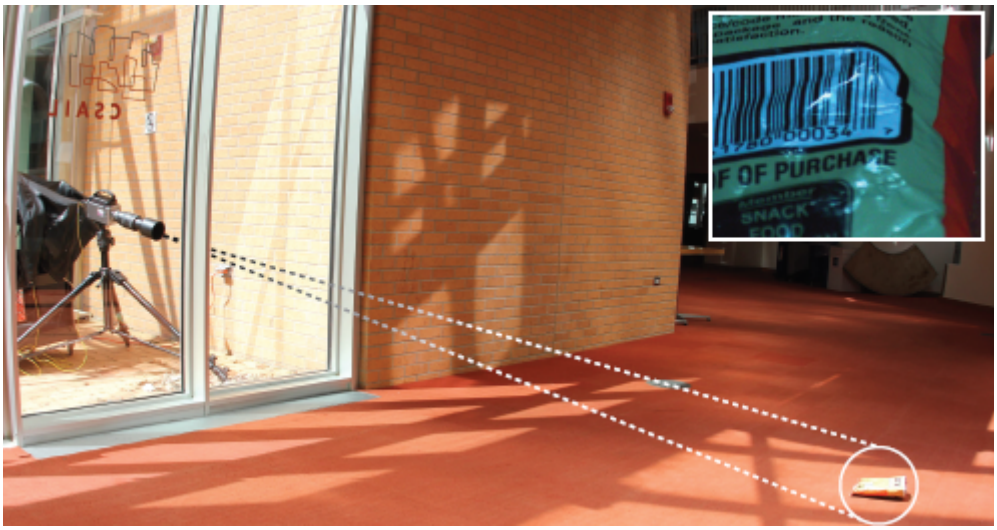

Figura 3.3: Modelo de implementação. Retirado da referência [4].

A Figura 3.3 esboça de uma possível aplicação do microfone visual, que será mais explorado no Capítulo 4, e que expõe genericamente uma possibilidade de aplicar o modelo estudado neste capítulo.

3.3 CONCLUSÃO

Este capítulo apresentou como ocorre a recuperação de um sinal de áudio, a partir de um grupo de imagens vibracionais de um objeto, explicando a manipulação matemática necessária e as técnicas utilizadas para obter o sinal de áudio integrado resultante.

No próximo capítulo apresentar-se-á a aplicação detalhada do microfone visual, assim como as respostas a algumas modificações no cenário.

4 IMPLEMENTAÇÃO DO MICROFONE VISUAL

4.1 INTRODUÇÃO

Utilizando os conceitos apresentados no Capítulo 3, estudar-se-á agora quais característica do objeto e do sinal sonoro interferem na sua recuperação e de que forma cada peculiaridade afeta o resultado do processo, através de um conjunto de experimentos.

Para tanto, alguns elementos foram modificados em cada momento, como a frequência do som incidente e a estrutura do objeto utilizado e analisou-se o sinal resultante do processo, a fim de comparar as diferentes respostas.

Esses ensaios estão descritos com detalhes na tese do doutor Abe Davis [3].

4.2 EXPERIMENTOS

Em um primeiro momento, foi realizado o experimento como disposto na Fig. 4.1.

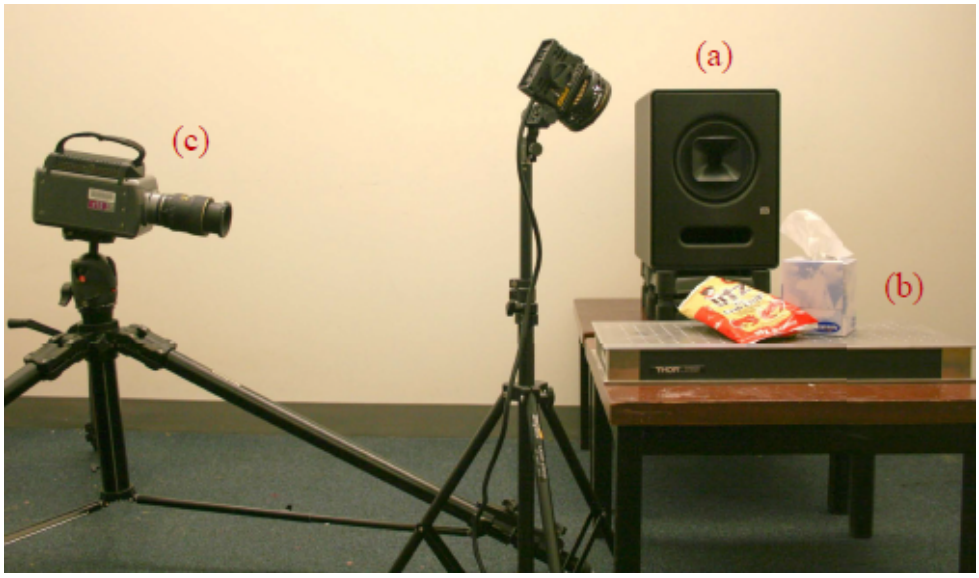

Figura 4.1: Esquema da implementação do experimento. Uma fonte sonora (a) excita um objeto (b) e uma câmera (c) grava o processo. Retirado da referência [4].

Em um ambiente fechado, um objeto foi iluminado com lâmpada de fotografia e posto a uma distância de aproximadamente 1 metro de uma câmera de alta velocidade. Ao passo que os resultados foram obtidos, algumas alterações no cenário foram realizadas, que serão explicadas ao decorrer do capítulo.

Com o intuito de reduzir ruídos externos e vibrações de contato, o ambiente escolhido é livre de sons externos e o alto-falante foi posicionado separado da superfície do objeto.

Os sons incidentes variam entre 80 dB e 110 dB. Os vídeos apresentaram taxas de quadros na faixa de 2 kHz a 20 kHz, com resoluções entre 192x192 *pixels* e 700x700 *pixels*.

Os vídeos foram processados utilizando pirâmides orientáveis complexas, com três escalas e duas orientações [17]. Os códigos utilizados são processados no Matlab e estão disponíveis no Anexo 1 e no site do projeto [21], porém algumas partes serão apresentadas ao longo deste capítulo.

Para que o processamento dos vídeos seja compreensível, deve-se ter consciência de que a manipulação dos sinais é realizada por meio de funções, preexistentes ou desenvolvidas, portanto, as etapas do algorítimo que correspondem à tradução do movimento em vetores, à manipulação matemática descrita no capítulo anterior, ao processo de filtragem temporal e à integração do sinal parcial que forma o sinal sonoro de saída estão particionadas em grupos afins, acessíveis no Anexo 1.

Assim, primeiramente estudou-se como a frequência interfere na resposta do sinal. Para tanto, a taxa da frequência foi variada, ao passo que se recuperou o sinal de diversos objetos, através da incidência de um sinal senoidal que aumentou de frequência linearmente ao longo do tempo.

A Figura 4.2(a) exibe a curva de densidade espectral do sinal de entrada, que varia entre 100 Hz e 1 kHz em 5 segundos. Além disso, as curvas de densidade espectral dos sinais recuperados de cinco objetos são mostradas na Fig. 4.2(b), considerando os vídeos com taxa de quadros de 2,2 kHz.

Os objetos estão dispostos de acordo com a qualidade do sinal recuperado. O primeiro item estudado foi um tijolo. Percebe-se que o sinal recuperado é de baixa qualidade, o que é esperado, já que ele é rígido e pouco compressivo. A fração de sinal recuperado que corresponde a baixa frequência pode advir do movimento do tijolo ou de alguma interferência na gravação do vídeo.

A segunda estrutura estudada foi a água, que possui uma resposta um pouco melhor do que o tijolo, porém o seu resultado ainda é de difícil observação, principalmente nas faixas de

frequência mais altas, a partir de 400 Hz. Uma possibilidade de se explicar este fato é que as vibrações de sua estrutura possuem amplitudes cada vez menores, quanto mais distantes estão do centro da incidência do som, além de que o movimento superficial da água possui características de refração e reflexão especular espacial ao longo do tempo.

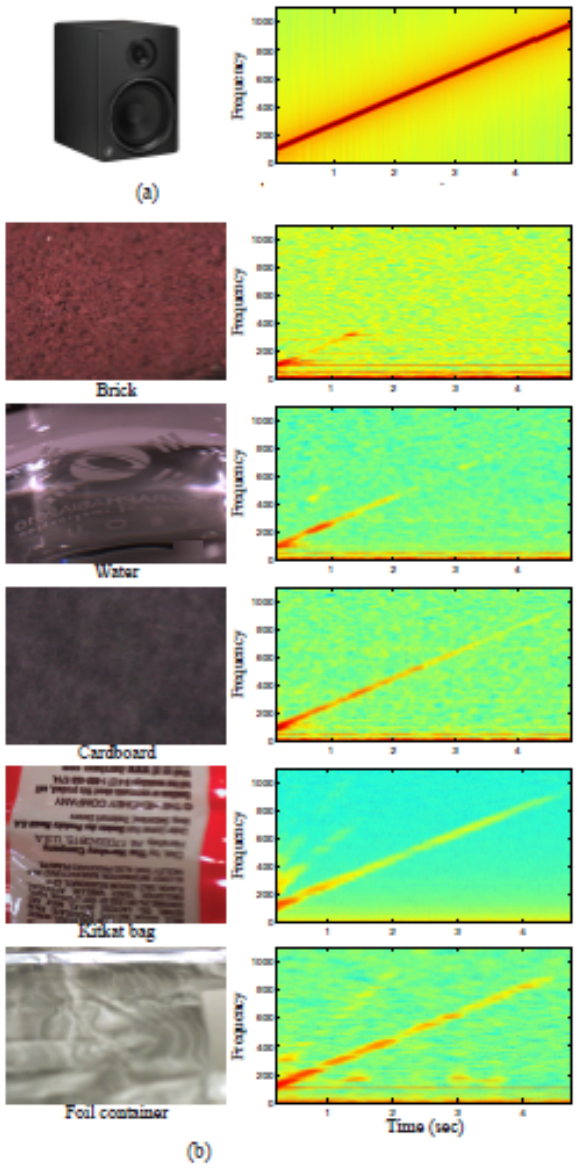

Figura 4.2: Sinal de entrada (a) e sinal reconstruído de diferentes materiais (b).

Retirado da referência [4].

O terceiro item estudado foi um cartão, que possui sinal reconstruído de boa qualidade, em uma faixa limitada de frequência, entretanto, há muito ruído nas frequências altas.

O quarto objeto foi um saco plástico, apresentando o melhor sinal reconstruído. Observa-se que, por ser mais maleável, o sinal recuperado é bem representado em todas as frequências e devido sua estrutura flexível, contém menos ruído que as demais. No início da curva de densidade espectral do seu sinal reconstruído, percebe-se um sinal de ruído de frequência um pouco maior, isto pode ser devido ao movimento que o próprio material possui quando sua estrutura é agitada, ainda que imperceptível ao ouvido humano.

O recipiente de alumínio foi o último objeto e apresenta uma boa resposta ao experimento, contudo, a faixa de frequência mais baixa possui mais interferência que as demais. Uma explicação plausível para isto é sua essência menos elástica, o que gera uma menor possibilidade de propagação das vibrações.

A partir dos resultados deste experimento, é possível aferir que o sinal recuperado é mais fraco em altas frequências, possivelmente devido a maior atenuação da vibração e menor deslocamento espacial nesse intervalo. Por isso, é possível aplicar um filtro temporal nesta faixa de frequência.

Depois, a partir dos resultados anteriores, comparou-se as respostas de áudio que vídeos de uma planta e de um saco plástico apresentam quando um sinal sonoro de 2.2 Hz de frequência gravado incide sobre eles, na configuração presente na Fig. 4.1.

Os resultados estão mostrados nas Figuras 4.3 e 4.4. E, ao contrário dos resultados da Fig. 4.2, neste momento o sinal sonoro de entrada possui tons em todas as frequências, por isso, a curva de densidade espectral de saída apresenta resposta em todas as frequências do gráfico.

Além do espectograma dos sinais recuperados, mostra-se como o sinal se comporta após passar pelo filtro passa-alta ao longo do tempo e como a sua densidade espectral se distribui na escala logarítmica.

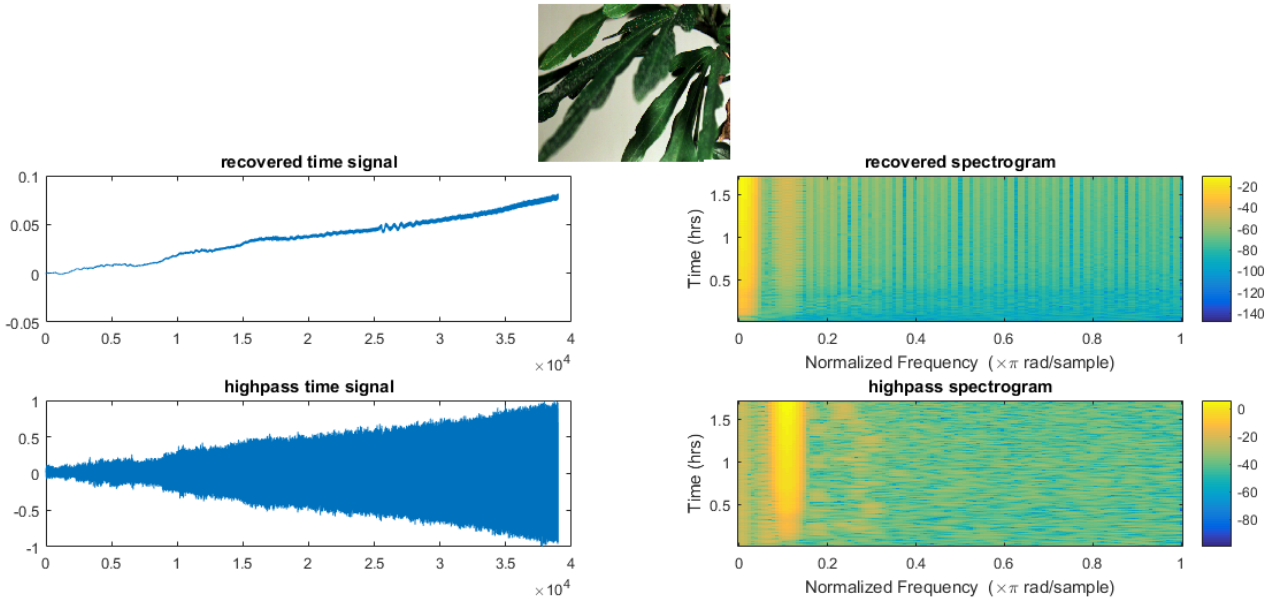

Figura 4.3: Sinal recuperado de uma planta, a partir de uma gravação.

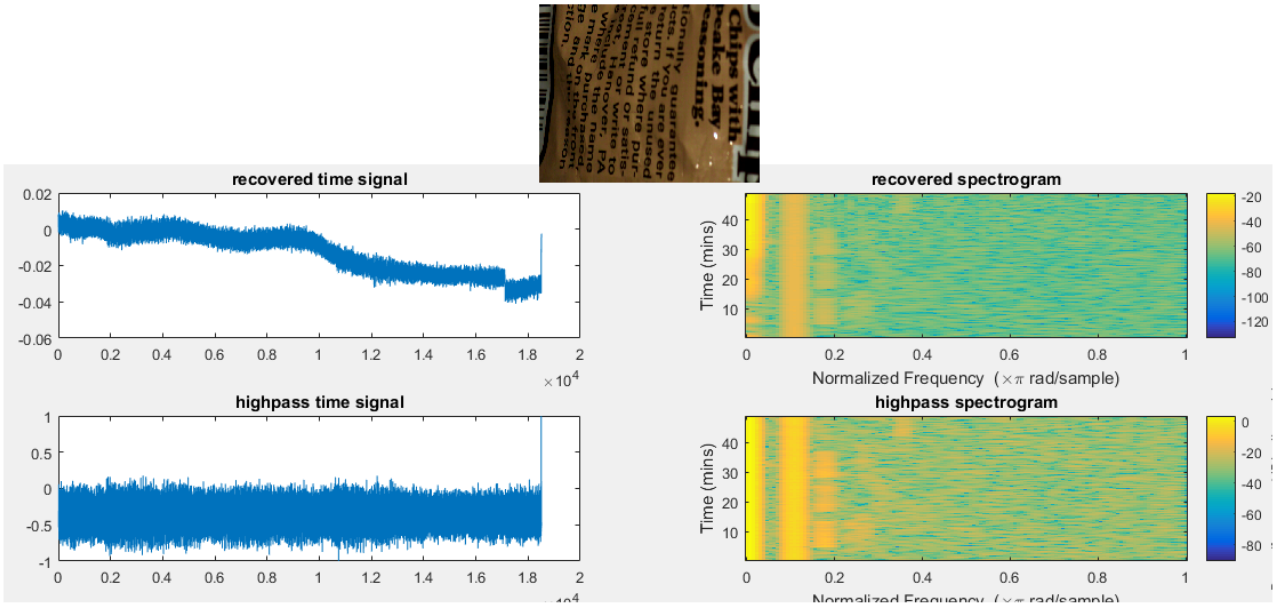

Figura 4.4: Sinal recuperado de um saco plástico, a partir de uma gravação.

A partir da análise das Figuras 4.3 e 4.4, percebe-se que o som recuperado de um saco plástico é mais nítido do que o som advindo da movimentação de uma planta e que, após passar pelo filtro, o sinal é condizente com o sinal de voz, portanto, o som responde melhor ao processo quando incide sobre a superfície de saco plástico do que sobre uma planta, o que corrobora com os resultados obtidos na comparação de cinco objetos (Fig. 4.2).

Então, como o saco plástico apresentou o melhor sinal reconstruído, ele foi mais estudado no segundo momento, em que se observou como ele responde a um conjunto diferente de sons incidentes. Para isso, reproduziu-se gravações de três oradores pronunciando duas frases diferentes através do alto-falante, para assim, estudar a resposta do sistema à frequência do sinal de entrada.

Nestes casos, os vídeos tinham resolução espacial de 700 x 700 *pixels* e 2.200 quadros por segundo.

Os gráficos de densidade espectral dos sinais de entrada e saída dessa configuração estão dispostos na Fig. 4.5.

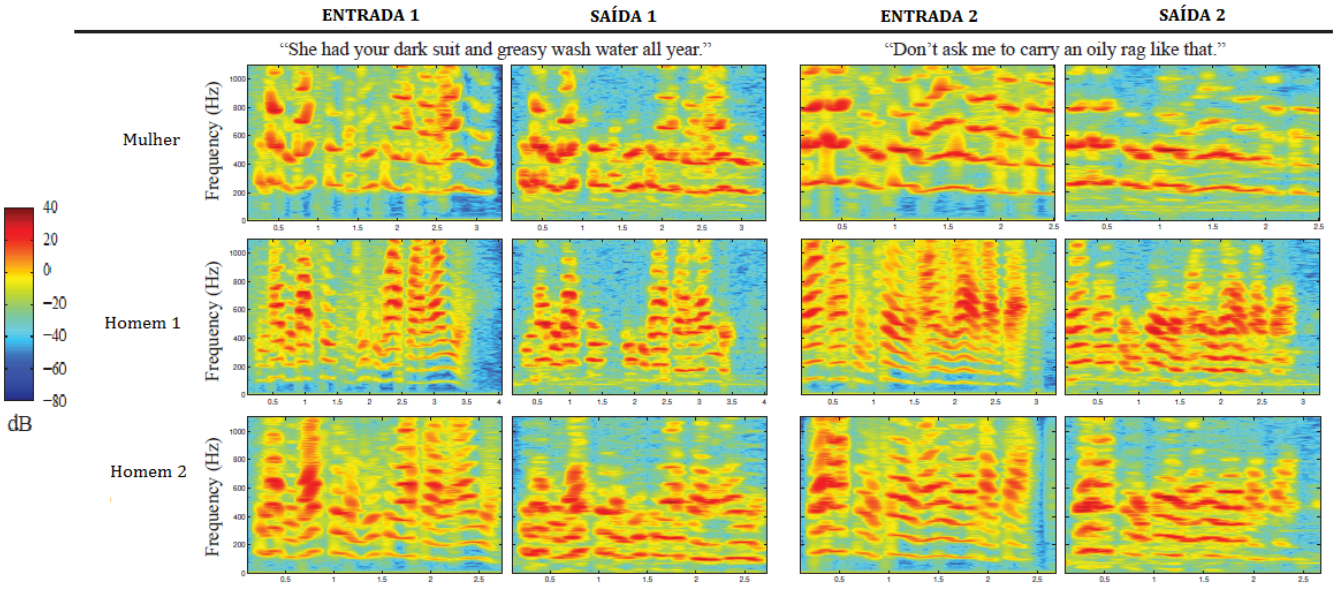

Figura 4.5: Sinal recuperado de um saco plástico, a partir da gravação de sons de três oradores diferentes. Retirado da referência [4].

Da Figura 4.5, percebe-se que, dentre os oradores, há uma mulher e dois homens. E dos sinais de entrada, retirado das suas falas gravadas, entende-se que um homem (Homem 2) possui tom de voz predominantemente na faixa de frequência mais baixas do que o outro (Homem 1). Além disso, dos sinais de saída, observa-se que os melhores resultados estão presentes nas faixas de frequências mais baixas e, ao passo que as frequências aumentam, o sinal reconstruído torna-se mais fraco.

Um forma de avaliar a inteligibilidade do sinal recuperado é ouvir os sinais de entrada e saída e compará-los, entretanto, essa forma de avaliação é subjetiva. Então, também optou-se por avaliar o resultado utilizando a relação sinal – ruído (SNR) [7] ao longo do tempo. A Tabela 4.1 apresenta a avaliação do SNR dos sinais de saída [4].

Tabela 4.1: Avaliação Da Relação Sinal – Ruído dos Sinais de Resposta de um Saco Plástico.

Gênero	Mulher (Entrada 1)	Mulher (Entrada 2)	Homem 1 (Entrada 1)	Homem 1 (Entrada 2)	Homem 2 (Entrada 1)	Homem 2 (Entrada 2)
SNR [dB]	24,5	28,7	20,4	23,2	23,3	25,5

A Tabela 4.1 mostra que os sinais recuperados da segunda frase possuem melhor qualidade, quando comparados com a primeira frase. Isto é perceptível observando a Fig. 4.5, em que o conjunto de curvas de densidade espectral da segunda frase assemelha-se mais com os sinais de entrada do que o conjunto da primeira frase, principalmente nas faixas de frequências mais baixas.

A seguir, outro teste foi realizado. O grupo de sinais estudado foi o sinal de voz ao vivo, em que o alto-falante da Fig. 4.1 foi substituído por uma pessoa. Os vídeos, entretanto, possuíam divergências na qualidade também, em que um conjunto de vídeos tinha resolução espacial de 700 x 700 *pixels* e 2.200 quadros por segundo e o outro conjunto, 192 × 192 *pixels* e 20.000 quadros por segundo. Os resultados estão apresentados nas Figuras 4.6, 4.7 e 4.8.

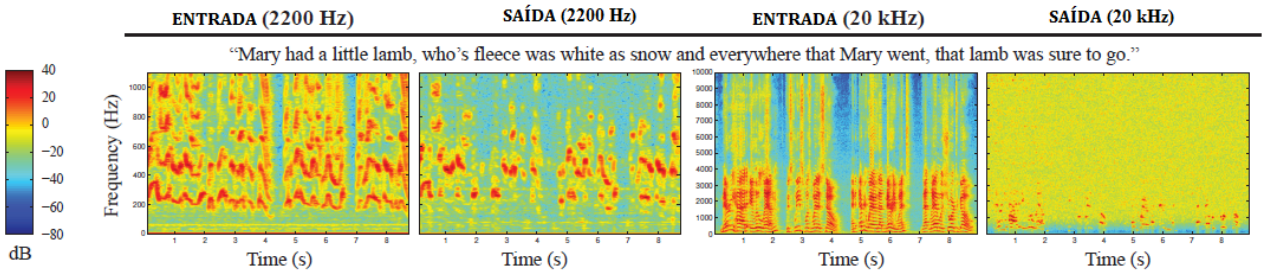

Figura 4.6: Sinal recuperado de um saco plástico de um som incidente ao vivo. Retirado da referência [4].

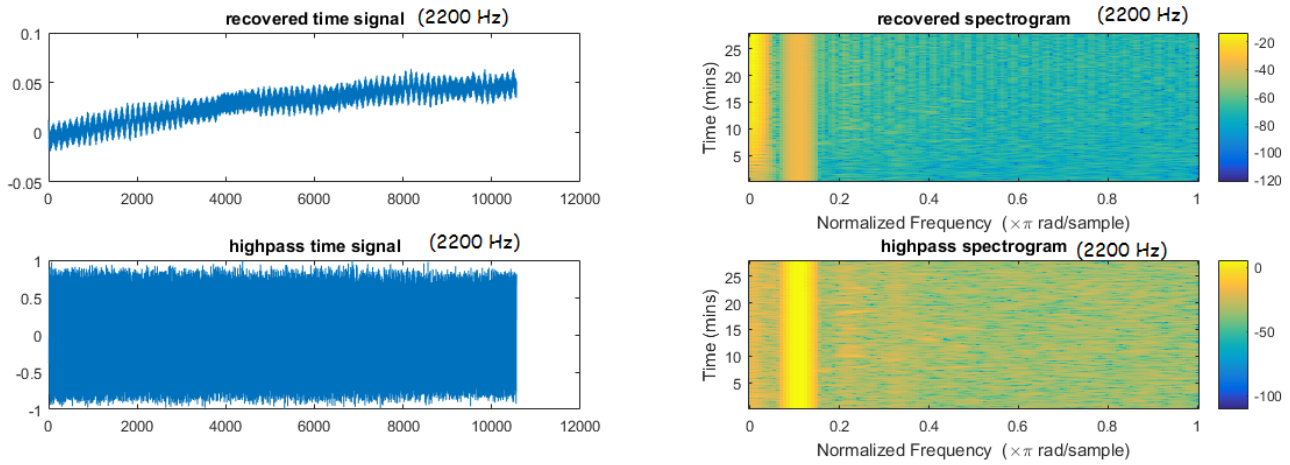

Figura 4.7: Sinal recuperado de um saco plástico, com frequência de 2.200 Hz.

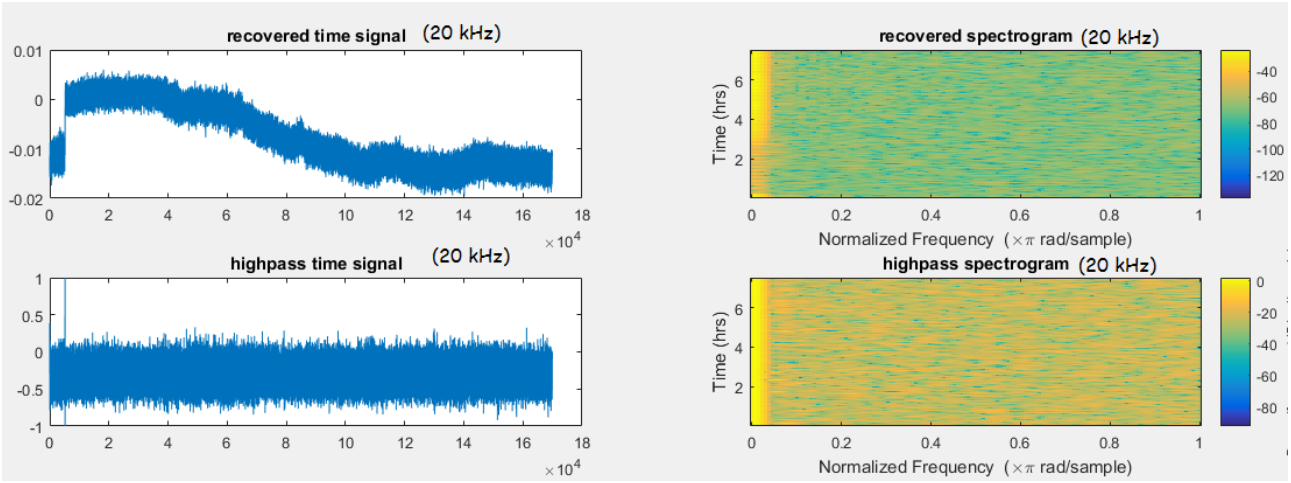

Figura 4.8: Sinal recuperado de um saco plástico, com frequência de 20 kHz.

Neste caso, a diferença entre os sinais de saída a diferentes frequências é bastante significante.

Quando se estuda o sinal de entrada com taxa de quadros por segundo de 2.2 Hz, percebe-se que a densidade espectral do sinal reconstruído assemelha-se mais com o sinal incidente do que o sinal de entrada com taxa de 20 kHz. Ao passo que, este possui densidade espectral do sinal de saída pouco perceptível, ou seja, apresenta muito ruído.

As Figuras 4.7 e 4.8 também mostram como os sinais se comportam ao longo do tempo após passarem por um filtro e os respectivos espectros de frequência.

Analisando os resultados dos vídeos com taxa de quadros diferentes, mostrado nas Figuras 4.6, 4.7 e 4.8, percebe-se que a qualidade do vídeo tem grande interferência no resultado. O vídeo com taxa de quadros mais alta (20 kHz) gerou um tempo menor de exposição, logo, mais ruído na imagem foi observado no mesmo intervalo de tempo.

4.3 CONCLUSÃO

Este capítulo apresentou os resultados de um conjunto de experimentos a respeito do microfone visual. Analisou-se a resposta para cada parâmetro que interfere na recuperação do sinal sonoro, tal como a estrutura do objeto utilizado, a melhor faixa de frequência da onda incidente e a qualidade do vídeo gravado.

Dentre os cinco tipos de materiais (tijolo, água, cartão, saco plástico e recipiente de alumínio) avaliados quanto as suas estruturas, percebeu-se que o saco plástico apresenta o melhor resultado. Isto se deve às suas qualidades intrínsecas, como maleabilidade, flexibilidade, elasticidade, compressibilidade e rigidez, que oportunizam uma melhor propagação vibracionais sobre sua superfície.

Também se contrastou os resultados que uma planta e um saco plástico apresentam quando são atingidos por uma música gravada e aferiu-se que o som obtido do saco plástico se aproxima mais do esperado do que quando é adquirido da movimentação de uma planta. Principalmente, quando se compara os sinais após passar pelo filtro passa-alta, que no melhor caso possui aspecto muito mais próximo ao sinal de entrada; e, no pior caso, apresenta muito mais ruído, logo, o erro é melhor isolado quando se estuda a resposta do saco plástico.

Em um segundo momento, estudou-se como o algoritmo responde a três diferentes tons de voz, ou seja, três oradores gravaram duas frases distintas, que foram processadas e avaliadas, quanto às suas faixas de frequências.

Deste caso, obteve-se melhores resultados em baixas frequências, enquanto, em frequências altas, constatou-se mais ruído no sinal resultante, logo, a voz dos homens responde com mais acurácia ao experimento.

Depois trocou-se as gravações por sinais de voz ao vivo, tornando a qualidade da gravação divergente. No caso em que o vídeo possui taxa de quadros por segundo de 2.2 Hz, o sinal reconstruído é muito mais semelhante do que no caso em que a taxa é de 20 kHz.

Além disso, aferiu-se que há uma relação linear entre a frequência dos quadros do vídeo e o ruído presente no sinal de saída, assim, quanto maior a frequência dos quadros de vídeo, mais ruído o sinal reconstruído apresentará, em um mesmo intervalo de tempo.

O próximo capítulo explorará outra utilização da análise de vibrações mecânicas, em que o objetivo também é obter informações do vídeo, é a ampliação de vídeo euleriana.

5 AMPLIAÇÃO DE VÍDEO EULERIANA

5.1 INTRODUÇÃO

Outra interessante aplicação da técnica de extrair informações de movimentos extremamente sutis de um vídeo está concentrada na sua ampliação e melhor visualização.

Para que isso ocorra, deve-se traduzir o movimento extraído de vídeos de acordo com a teoria matemática que melhor relaciona o objetivo desejado com as facilidades computacionais, a fim de manipular as variações temporais sequenciais de vídeos e observar os comportamentos mecânicos interessantes presentes.

Nesse caso, aplicaram-se os conceitos de Euler, que é um método de integração matemática utilizado para solucionar equações diferenciais ordinárias com um valor inicial conhecido. Ele tem como objetivo medir quão rapidamente um problema converge para a solução analítica ao passo que as etapas da integração numérica diminuem [27]. Entretanto, é interessante utilizá-lo somente em intervalos infinitesimais do problema.

Visando uma abordagem mais dinâmica, o que permite uma análise mais delicada das pequenas amplitudes das características que evoluem com o tempo, o método de Euler se mostrou eficiente, considerando sua facilidade de implementação e desempenho.

Então, com o propósito de visualizar estas ações, foi desenvolvido um método chamado de ampliação de vídeo euleriana. Esta ferramenta tem como entrada um sinal de vídeo sequencial, que é decomposto espacialmente e que sofre uma filtragem temporal de seus quadros. O sinal de saída é amplificado, a fim de revelar as informações úteis desejadas [17].

Tal abordagem de manipulação temporal, que aponta os movimentos espaciais de baixa amplitude, traduz um sinal de intensidade temporal em movimento espacial de vídeos, a depender de uma aproximação linear relacionada com a conjectura de constância de brilho utilizada no fluxo óptico [8]. Porém, essa aproximação é aplicada a movimentos muito pequenos, justamente o que viabiliza a utilização dos conceitos eulerianos.

Uma vez que o sistema perceptível humano é limitado, certos padrões são importantes fontes de informação quando mudados. Assim, é interessante analisar alguns fenômenos em tempo real, a depender da frequência estudada. Por exemplo, a alteração no fluxo de sangue de uma pessoa que gera uma variação na coloração da pele, tem, como um possível resultado, a taxa de pulsação do indivíduo [14, 15, 23].

Há outras formas de solucionar estas questões, como criar um movimento exagerado, a partir do movimento estudado [26], ou abordar o tema sob a perspectiva de dinâmica de fluidos, em que a trajetória das partículas é rastreada ao longo do tempo [9]. Entretanto, estas abordagens necessitam de uma estimativa acurada do movimento, o que torna o processo computacional dispendioso.

Portanto, a perspectiva Euleriana apresentada neste contexto tem se mostrado mais eficiente e menos onerosa. Pois, as propriedades de um fluido evoluem com o tempo são ampliadas *pixel* a *pixel*, de maneira espacial e multi-escalar, já que as mudanças do vídeo e suas consequências ocorrem diferentemente em diversas escalas.

Contudo, não se estima explicitamente o movimento, foca-se nos agrupamentos espaciais, amplia-se as mudanças em posições fixas, filtra-se o sinal em banda passante e obtêm-se o sinal final, utilizando as aproximações diferenciais.

A próxima seção apresenta e esclarece as etapas que são utilizadas no método de ampliação de vídeo euleriana.

5.2 METODOLOGIA

A fim de evidenciar sutis mudanças temporais e espaciais em um vídeo, combina-se o processamento espacial e temporal, utilizando análise computacional e seguindo o esquema da Fig. 5.1 [17].

Inicialmente decompõe-se o sinal de entrada de vídeo em diferentes faixas de frequência espacial. Uma vez que tais bandas podem ser manuseadas diferentemente, a depender das relações sinal – ruído [6,7] que cada faixa possui e da aproximação linear usada em cada banda de frequência espacial.

Nos momentos em que o objetivo do processamento é aumentar a relação sinal – ruído temporal, agrupa-se múltiplos *pixels* e faz este conjunto passar por um filtro passa-baixa, a fim de diminuir as amostras de quadros do vídeo.

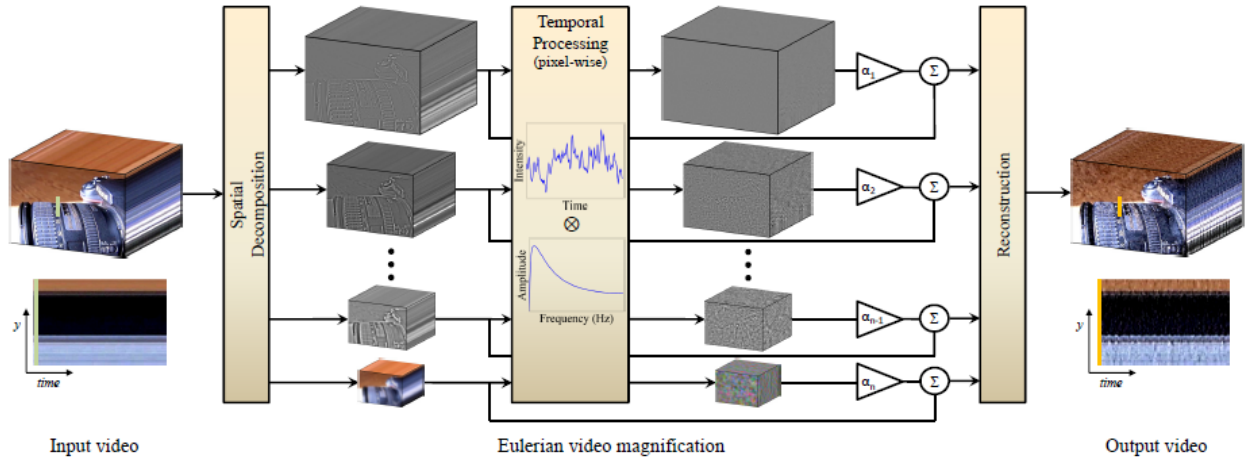

Figura 5.1: Modelo do processamento de ampliação de vídeo euleriano.

Retirado da referência [28].

A seguir, em cada banda espacial, há o processamento temporal, o qual é uniforme para todos os níveis espaciais e para todos os *pixels* dentro de cada nível. Neste instante, a série temporal obtida é correlacionada ao valor de um *pixel*.

Então, aplica-se o filtro temporal, com a finalidade de extrair as faixas de frequências interessantes, o sinal de banda passante é multiplicado por um fator de ampliação α, e por fim, adiciona-se o sinal ampliado ao original e sobrepõe-se à pirâmide espacial para obter o vídeo final.

A respeito da amplificação de pequenos movimentos, o processamento temporal utiliza uma análise que se baseia nas expansões da série Taylor de primeira ordem comuns em análises de fluxo óptico [8, 11].

A intensidade da imagem na posição *x* e no tempo *t* será o vetor unidimensional $I(x,t)$. Como a imagem possui movimento de translação, expressa-se os vetores de intensidade como uma função de deslocamento no tempo $\delta(t)$, tal que $I(x,t)=f(x+\delta(t))$ e $I(x,0)=f(x)$ é a intensidade da imagem inicial. O objetivo é ampliar o movimento para um valor específico de, ou seja, obter o sinal:

$$\hat{I}(x,t)=f(x+(1+\alpha)\delta(t)). \tag{5.1}$$

Supondo que a imagem pode ser descrita com uma expansão da série de Taylor de primeira ordem, então:

$$I(x,t) \approx f(x) + \delta(t)\frac{\partial f(x)}{\partial x}. \tag{5.2}$$

Fazendo o sinal *I(x,t)* passar por um filtro passa-faixa em cada posição *x*, exceto *f(x)*, resultará no sinal *B(x,t):*

$$B(x,t) = \delta(t)\frac{\partial f(x)}{\partial x}. \tag{5.3}$$

Assim, amplificando o sinal em α e somando-o novamente à *I(x,t)*:

$$\widetilde{I}(x,t) = I(x,t) + \alpha B(x,t). \tag{5.4}$$

Combinado as Equações 6.2, 6.3 e 6.4, tem-se:

$$\widetilde{I}(x,t) \approx f(x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{\partial x}. \tag{5.5}$$

Assumindo que esta expansão é válida para a maior perturbação, $(1+\alpha)\delta(t)$, então, relacionando a amplificação do sinal na banda passante com a ampliação do movimento:

$$\widetilde{I}(x,t) \approx f(x + (1+\alpha)\delta(t)). \tag{5.6}$$

Ou seja, o processamento amplia o movimento $\delta(t)$ da imagem local $f(t)$ em um fator de $(1+\alpha)$.

Em exemplo desse processo é mostrado na Fig. 5.2.

Um deslocamento pequeno $f(x+\delta)$ é aplicado a um sinal senoidal de baixa frequência $f(x)$ e, então, descrito como a expansão da série de Taylor $f(x) + \delta\frac{\partial f(x)}{\partial x}$. Filtrando-o, ampliando-o em α e somando-o novamente ao sinal senoidal, tem-se o sinal resultante transladado.

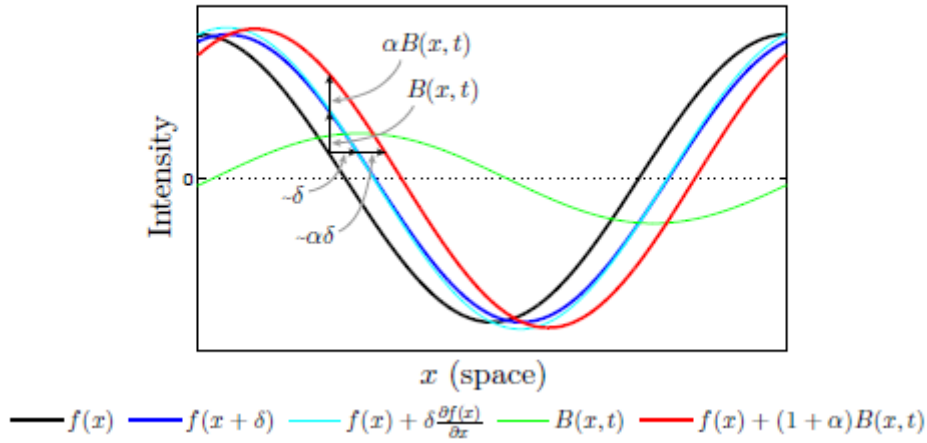

Figura 5.2: Exemplo de filtragem temporal que gera um sinal transladado. Retirado da referência [28].

Agora, supondo que o deslocamento temporal $\delta(t)$ não está limitado na faixa de frequência do filtro temporal, então, há um conjunto $\delta_k(t)$ que representa os diferentes componentes espectrais temporais e cada item deste conjunto será atenuado pelo filtro temporal de fator γ_k. Isto dará o sinal:

$$B(x,t)=\sum_k \gamma_k \delta_k(t)\frac{\partial f(x)}{\partial x}. \tag{5.7}$$

Como a atenuação depende da frequência temporal, que pode ser entendido que a ampliação do movimento se dá por $\alpha_k=\gamma_k\alpha$, então, o sinal de saída do movimento ampliado será:

$$\tilde{I}(x,t)\approx f\left(x+\sum_k (1+\alpha_k)\delta_k(t)\right). \tag{5.8}$$

Portanto, a modulação linear dos componentes espectrais do sinal torna o fator de ampliação local α_k em fator de amplificação do sinal δ_k para cada banda de movimento do mesmo.

Os resultados são válidos para sutis movimentos ou suaves transições no vídeo. É necessário também que haja uma manipulação para os demais casos, em que as mudanças bruscas ocorrem em altas frequências espaciais e a aproximação da série de Taylor gera um erro não desprezível.

Já que a frequência espacial, w, é um fator que influencia a análise de sinais de movimento, então, ele será o argumento limitante para o fator de amplificação do movimento α de um movimento $\delta(t)$.

A fim de que o sinal processado, $\tilde{I}(x,t)$, seja aproximadamente igual ao sinal de movimento ampliado, $\hat{I}(x,t)$, algumas condições devem ser respeitadas:

$$\tilde{I}(x,t) \approx \hat{I}(x,t),$$

$$f(x) + (1+\alpha)\delta(t)\frac{\partial f(x)}{(\partial x)} \approx f(x + (1+\alpha)\delta(t)). \tag{5.9}$$

Supondo que o sinal estudado seja cossenoidal, $f(x) = \cos(wx)$, e fazendo $\beta + 1 + \alpha$, tem-se da Eq. (5.9):

$$\cos(wx) - \beta w \delta(t) \sin(wx) \approx \cos(wx + \beta w \delta(t)), \tag{5.10}$$

$$\cos(wx) - \beta w \delta(t) \sin(wx) \approx \cos(wx)\cos(\beta w \delta(t)) - \sin(wx)\sin(\beta w \delta(t)). \tag{5.11}$$

Para que a Eq. (5.11) seja verdadeira, então:

$$\cos(\beta w \delta(t)) \approx 1, \tag{5.12}$$

$$\sin(\beta w \delta(t)) \approx \beta w \delta(t). \tag{5.13}$$

A Equação 5.13 somente é válida para ângulos pequenos, ou seja, o argumento da função senoidal deve ser menor que 10% de $\beta w \delta(t)$. Além disso, para que a Eq. 5.12 também seja verdadeira, então:

$$\beta w \delta(t) < \pi/4. \tag{5.14}$$

Escrevendo em termos de comprimento de onda do sinal, $\lambda = 2\frac{\pi}{w}$, resulta em:

$$(1+\alpha)\delta(t) < \frac{\lambda}{8}. \tag{5.15}$$

O resultado apresentado na Eq. (5.15) fornece orientação acerca dos parâmetros que regem o sinal do movimento, a saber, o fator de amplificação do sinal α, o deslocamento do movimento $\delta(t)$ e o comprimento de onda do sinal espacial λ.

Os sinais à esquerda correspondem ao movimento real e possuem comprimento de onda $\lambda = 2\pi$ e posição inicial $\delta(1) = \pi/8$. Enquanto que os sinais à direita correspondem aos sinais processados com $\lambda = \pi$ e $\delta(1) = \pi/8$.

A Figura 5.3(a) apresenta uma comparação do sinal $I(x,0)$ deslocado pelo fator $(1+\alpha)\delta(t)$, variando de 0,2 (onda azul) até 3,0 (onda vermelha), como também em (b) mostra o deslocamento ampliado pelo filtro temporal correspondente à variação em (a).

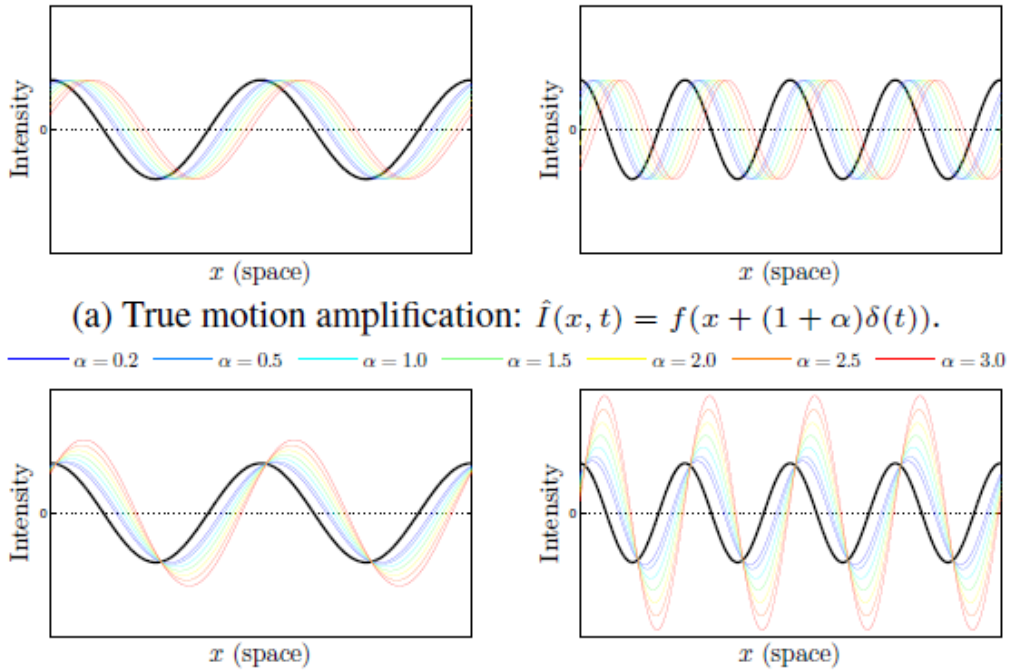

Figura 5.3: Amplificação do movimento para diferentes frequências espaciais e valores de α. Retirado da referência [28].

Observa-se que a Fig. 5.3 exibe os erros de ampliação do sinal para grandes valores de α. Por isso é importante que o processo computacional ocorra em várias escalas, uma vez que o fator de ampliação α deve ser escolhido para cada faixa de frequência espacial.

5.3 CONCLUSÃO

Este capítulo descreveu os conceitos que circundam a ampliação de vídeo euleriana, assim como quais limites devem ser respeitados para que esse método seja aplicável a um conjunto de movimentos mecânicos e venham a ser estudados mais detalhadamente e ampliados.

Apesar de ser uma metodologia eficiente e com boa vantagem computacional, apresenta duas desvantagens, por ser uma ferramenta linear, o ruído amplifica-se linearmente também ao passo que a amplificação do movimente ocorre, e a limitação de poder ser usado somente em movimentos sutis gera uma grande descontinuidade do movimento em altas frequências espaciais.

O próximo capítulo expõe como as ideias aqui apresentadas podem ser postas em prática em situações simples e quais as principais dificuldades encontradas nos experimentos.

6 IMPLEMENTAÇÃO DA AMPLIAÇÃO DE VÍDEO EULERIANA

6.1 INTRODUÇÃO

Com o intuito de aplicar a modelagem apresentada no capítulo anterior, alguns experimentos foram realizados [28], utilizando o software Matlab, cujos códigos estão disponíveis no site do projeto [5].

Esse capítulo mostra como movimentos mecânicos simples podem ser observados sob uma nova óptica e quais as principais vantagens e desvantagens desse método de ampliar as imagens pouco perceptíveis.

6.2 EXPERIMENTOS

A fim de processar o vídeo usando a técnica de ampliação de vídeo eulariana, alguns parâmetros foram importantes para uma boa resposta e algumas etapas precisam ser respeitadas.

Primeiramente é preciso optar por um filtro Butterworth para construir a pirâmide espacial do vídeo com cinco níveis.

Neste momento, o filtro é aplicado com a finalidade de eleger qual faixa de sinal que será ampliado. Essa escolha depende da aplicação, assim, visando a ampliação de movimento, prefere-se um filtro de banda larga; já para a ampliação de cor, um filtro de banda passante estreita gera um sinal de saída com menos ruído.

Depois, considerando as condições limitantes do modelo, apresentadas no capítulo anterior, escolhe-se um fator de amplificação α e um limite de frequência espacial, a partir do comprimento de onda λ_c. Em alguns casos, escolhe-se valores acima do limite, a fim de exagerar no movimentos estudados ou apresentar variações de cor acentuadas.

Por fim, selecionar a forma de atenuação do fator de ampliação, que pode ser um decrescimento linear de α a zero, quando se quer ampliar o movimento; ou impor que α seja nulo para todo valor de $\lambda < \lambda_c$, quando se busca enfatizar mudanças de cor.

As experiências foram agrupadas em dois grupos, ampliação de cor e ampliação de movimento.

No experimento de análise da cor, o objetivo foi visualizar a pulsação de um indivíduo, a partir do fluxo de sangue, como exposto na Fig. 6.1. A amplificação neste caso gerou uma variação na cor do rosto, tornando-a mais vermelha à medida que a pulsação aumentou e o sangue fluiu sob a pele [14, 15, 23].

Para tanto, selecionou-se a faixa de frequência que inclui taxas cardíacas possíveis, ou seja, frequências espaciais mais baixas. Então, o sinal de entrada passou por um filtro espacial com banda passante entre 0,83 Hz e 1 Hz, o equivalente à faixa de 50-60 bpm, a fim de reduzir o ruído do sinal e o volume de dados.

Neste caso, os valores de α e λ_c escolhidos foram 100 e 1.000, respectivamente. E a forma de atenuação foi linear, relacionada à continuidade de brilho do fluxo visual. Após passar por essas etapas, o sinal foi adicionado ao sinal de entrada e pode ser observado a variação na coloração da pele, proporcional à pulsação, Fig. 6.1(b).

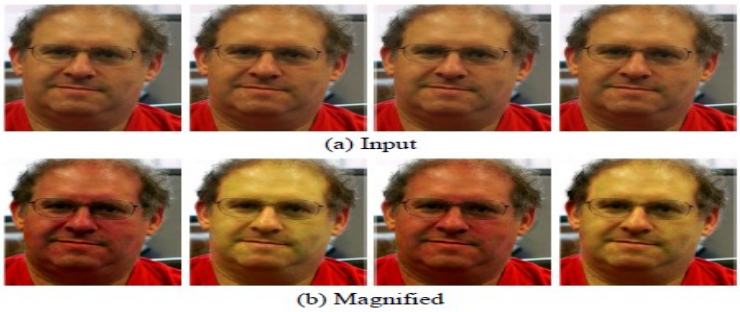

Figura 6.1: Visualização da mudança na coloração da pele de um indivíduo. (a) Quadros do vídeo de entrada. (b) Quadros do vídeo de saída. Retirado da referência [28].

Ao passo que, no experimento de amplificação de movimento, o objetivo foi estudar a pulsação de um indivíduo, agora ela será analisada a partir da movimentação suave de seus vasos sanguíneos, como mostrado na Fig. 6.2.

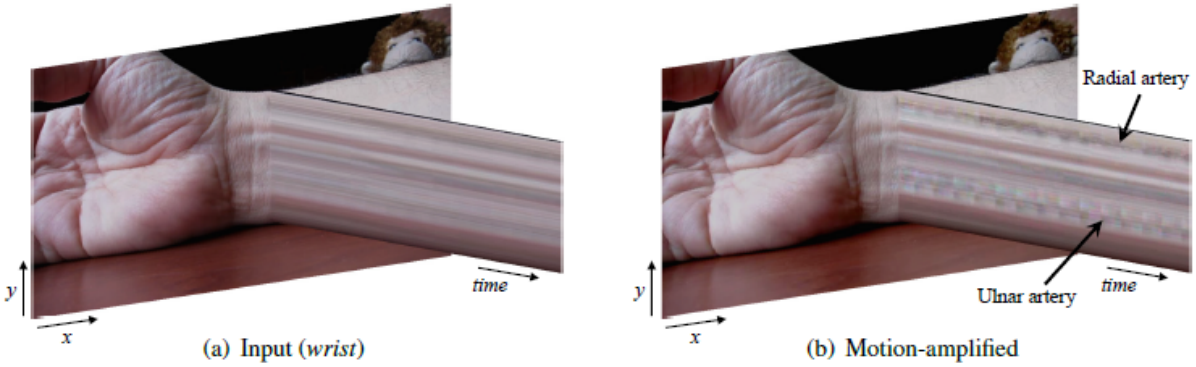

Figura 6.2: Visualização do movimento suave da pulsação de uma pessoa. (a) Quadros do vídeo de entrada. (b) Quadros do vídeo de saída. Retirado da referência [28].

Uma vez que, neste momento, o interessante é focar no movimento sutil, usou-se um filtro temporal com banda passante mais ampla, com uma taxa de frequência que inclui a frequência cardíaca – 0,88 Hz (53 bpm), o que conseguiu destacar o movimento da pulsação em vez de amplificar a mudança na cor da pele. Além disso, foi escolhido um menor valor para o fator de amplificação, $\alpha = 10$. As demais etapas do processo foram obedecidas e, então, obteve-se o resultado apresentado na Fig. 6.2(b).

6.3 CONCLUSÃO

Este capítulo expôs dois conjuntos de experimentos a respeito da ampliação de vídeo euleriana, em que movimentos com pequenas amplitudes são melhores observadas, através da extrapolação do movimento ou enfatizando a mudança na cor do vídeo.

A teoria euleriana foi escolhida neste método para estudar movimentos mecânicos suaves com uma boa eficiência computacional, para tanto há a necessidade de variar dos parâmetros dos códigos, a depender do objetivo da experiência.

No caso em que se analisou a variação da coloração do vídeo, quando a cor da pele de um indivíduo mudou de tonalidade ao passo que houve a variação na sua pulsação, utilizou-se uma faixa de frequências mais baixas, o sinal de entrada passou por um filtro espacial com banda passante entre 0,83 Hz e 1 Hz, o equivalente à faixa de 50-60 bpm de pulsação, fator de amplificação α igual a 100 e comprimento de onda do sinal λ_c igual a 1.000.

Já no segundo caso, em que se estudou a pulsação de uma pessoa a partir da extrapolação do movimento sutil de seus vasos sanguíneos, usou-se um filtro temporal com banda

passante mais ampla, com uma taxa de frequência que inclui a frequência cardíaca usual, ou seja, 53 bpm, o que equivale a 0,88 Hz, destacando a pulsação, e o valor do fator de amplificação α foi 10.

Devido à simplicidade na análise qualitativa desta metodologia, não há sobrecarga na complexidade dos algoritmos, perceptível na pequena mudança que precisa ser realizada para que experimentos diferentes sejam feitos e resultados tão divergentes sejam gerados.

7 CONSIDERAÇÕES FINAIS

Pode ser compreendido do trabalho que uma variada gama de informações estão presentes em todos os momentos do cotidiano, sendo possível capturá-las e estudá-las através de dispositivos simples e arranjos acessíveis.

O Microfone Visual é uma importante ferramenta de análise e recuperação de sons, principalmente quando a acessibilidade à fonte é comprometida.

Entretanto, ainda há algumas limitações em sua aplicação, como a complexidade em recuperar alguns sons inteligíveis e a obtenção de um sinal que traduza o comportamento do sinal de entrada.

Certos sinais sonoros são mais difíceis de serem estudados sob essa técnica, isto é caracterizado pela dificuldade no entendimento do som resultante. Embora o som ininteligível carrega informações utilizadas, a depender da finalidade, o ideal é recuperar um som aproximadamente igual ao inicial.

Outro elemento complicado de ser observado é o modo de vibração do sinal. A metodologia apresentada neste trabalho [3] é interessante, especialmente pela sua simplicidade, porém, a implementado pode ser laborioso. Uma vez que a descrição matemática de como o sinal espacial se comporta no meio físico não é trivial, embora realizável. Por isso, a maioria das técnicas utilizadas de recuperação de áudio são formas ativas.

Sob o mesmo ponto de vista, um fator limitante para esta aplicação é quão próxima a câmera está da fonte sonora. Posto que a potência do sinal estudado varia com o movimento em *pixels* e o volume de dados da imagem e que ambos sofrem mudanças de valor à medida que a distância entre a câmera e o objeto altera.

Sendo assim, as vibrações de objetos causadas pelo som podem ser extraídas de vídeos e se tornarem o meio de recuperá-lo, através da obtenção do sinal que rege matematicamente o movimento, chamado de movimento global ao longo do tempo, que sofre uma manipulação computacional apropriada, a fim de extrair o sinal recuperado.

As características físicas de tais objetos também interfere no resultado, sendo que utensílios leves respondem melhor à incidência sonora. Além disso, a frequência do som incidente afeta o som recuperado. Do observado, as frequências mais baixas geram melhores respostas no final do processo.

Enquanto o microfone visual procura recuperar o áudio, a técnica de ampliação de vídeo euleriana pretende extrair movimentos de pequena amplitude de um vídeo e ampliá-lo exageradamente, utilizando conceitos computacionais de processamento espacial e temporal.

A metodologia apresentada analisa um conjunto de *pixels* em uma faixa de frequência, a partir de uma apropriada filtragem, e expõe aspectos do movimento que antes eram quase imperceptíveis, seja por meio de uma variação na cor da imagem ou de uma amplificação exagerada do movimento.

Contudo, a sensibilidade ao ruído é um ponto crítico no modelo. Como dito anteriormente, a potência do sinal estudado é função da amplitude do movimento. Porém, em alguns casos, esta potência é muito menor que a potência do ruído presente no sinal. Assim, o resultado apresentado não será proporcional ao movimento inicial. Uma maneira de burlar este problema é uma melhor escolha do filtro espacial, sem que sua faixa de frequência interfira negativamente na análise, ou seja, uma faixa extensa que não amplia o sinal desejado ou uma faixa pequena que não mostra diferença de cor na imagem.

Logo, percebe-se que este estudo de pequenas vibrações é uma vertente da Engenharia que compreende que o equilíbrio entre eficiência e execução deve ser sempre buscado, até o ponto que as possibilidades se aproximam da exaustão.

7.1 PROPOSTAS DE TRABALHOS FUTUROS

As observações elaboradas neste trabalho abre uma gama de opções para pesquisas futuras no campo de análise de vibrações.

Entre elas, uma aplicação interessante da ampliação de vídeo euleriana é visualizar o fluxo de leite materno que flui durante a amentação de um bebê.

Uma vez que o experimento apresentado no Capítulo 6 expôs a visualização do fluxo de sangue no rosto de uma pessoa, através da variação na coloração do mesmo; e da pulsação de vasos sanguíneos de um indivíduo, por meio da extrapolação da amplitude dos seus movimentos, então, é possível fazer um estudo a respeito da quantidade de leite materno, que flui em um intervalo de tempo e alimenta um a bebê.

Tal emprego dos conceitos estudados é atraente ao universo da maternidade, em que a alimentação de um novo indivíduo, que não sabe expressar suas necessidades básicas, traz preocupações e dúvidas aos pais. Assim, caso seja possível aferir quando o bebê se alimenta,

pode-se buscar mais tranquilidade, caso se alimente bem; ou buscar formas de contornar o problema, caso haja deficiência na alimentação, ainda que seja apenas uma estimativa.

Portando, pode-se estudar a movimentação dos músculos do seio da mulher que agem durante a amentação e extrapolar as suas amplitudes ou variar a coloração do seio, já que agrega calor durante a realização do trabalho, além de que todo o corpo da mulher sofre mudanças nesse momento.

Além de que, caso foque a análise no bebê, é factível estudar a movimentação de sua mandíbula e garganta ao sugar e engolir o líquido ou sua variação de temperatura, já que ele tende ganhar calor na atividade.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] BENJELIL, M., KANOUN, S., RÉMY, M., ALIMI, A. M. *Steerable pyramid based complex documents images segmentation*. In: International Conference on Document Analysis and Recognition, 10., 2009.

[2] BOLL, S. *Suppression of acoustic noise in speech using spectral subtraction*. Acoustics, Speech and Signal Processing, IEEE, 1979, p. 113–120.

[3] DAVIS, A.*Visual Vibration Analysis*. 2016. 113p. Tese (Doutorado). Instituto de Tecnologia de Massachusetts, Massachusetts.

[4] DAVIS, A., RUBINSTEIN, M., WADHWA, N., MYSORE G., DURAND F., AND FREEMAN, W. T. *The visual microphone: Passive recovery of sound from video*. ACM Transactions on Graph. 2014.

[5] Eulerian Video Magnification for Revealing Subtle Changes in the World. 2012. Disponível em <http://people.csail.mit.edu/mrub/evm/#code > Acesso em: 17 de novembro de 2017.

[6] GAUTAMA, T., VAN HULLE, M. , IEEE Transactions, 2002, p. 1127 – 1136.

[7] HANSEN, J. H., PELLOM, B. L. *An effective quality evaluation protocol for speech enhancement algorithms.* In *ICSLP*, vol. 7, 1998 , p. 2819–2822.

[8] HORN, B., SCHUNCK, B. *Determining optical flow.* Artificial intelligence. 1981, p. 185–203.

[9] LIU, C., TORRALBA, A., FREEMAN, W. T., DURAND, F., ADELSON, E. H. *Motion magnification.* ACM Transactions on Graphics. 2005.

[10] LOIZOU, P. C. *Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum*. Speech and Audio Processing, IEEE Transactions, 2005, p. 857 – 869.

[11] LUCAS, B. D., AND KANADE, T. *An iterative image registration technique with an application to stereo vision*. In: Proceedings of IJCAI, 1981, p. 674–679.

[12] Meca-Wiki. Relação sinal – ruído. Disponível em: <http://pt-br.mecawiki.wikia.com/wiki/Rela%C3%A7%C3%A3o_sinal-ru%C3%ADdo> Acessado em 02/12/2017.

[13] NEWTON, Isaac. *Princípios Matemáticos da Filosofia Natural*. Cambridge, Reino Unido. 1687.

[14] PHILIPS. Philips Vital Signs Camera. Disponível em <http://www.vitalsignscamera.com> Acesso em: 08 de novembro de 2017.

[15] POH, M.-Z., MCDUFF, D. J., AND PICARD, R. W. *Non-contact, automated cardiac pulse measurements using video imaging and blind source separation*. Opt. Express, 2010, p. 10762–10774.

[16] PORTILLA, J., SIMONCELLI, E. P. *A parametric texture model based on joint statistics of complex wavelet coefficients*. Int. J. Comput. Vision, 2000, p. 49–70.

[17] RUBINSTEIN, M. *Analysis and Visualization of Temporal Variations in Video*. 2014. 118p. Tese (Doutorado). Instituto de Tecnologia de Massachusetts, Massachusetts.

[18] SILVA, L., F*iltros de Kalman no tempo e frequência discretos combinados com subtração espectral*. 2007. 113p. Dissertação. Escola de Engenharia de São Carlos, Universidade de São Paulo, São Paulo.

[19] SIMONCELLI, E. P., FREEMAN, W. T., ADELSON, E. H., HEEGER, D. J. *Shiftable multi-scale transforms*. IEEE Trans. Info. Theory, 1992, p. 587–607.

[20] SIMONCELLI, E. P., FREEMAN, W. T. *The Steerable Pyramid: A Flexible Architecture For Multi-Scale Derivative Computatio.* In Image Processing (ICCP), 1995 IEEE International Conference, IEEE, 1995.

[21] The Visual Microphone: Passive Recovery of Sound from Video. 2014. Disponível em <http://people.csail.mit.edu/mrub/VisualMic/> Acesso em: 17 de novembro de 2017.

[22] VAQUERO, D. *Pirâmides de imagens.* 2004. 39p. Dissertação. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

[23] VERKRUYSSE, W., SVAASAND, L. O., NELSON, J. S. *Remote plethysmographic imaging using ambient light*. Opt. Express, 2008, p. 21434–21445.

[24] WADHWA, N., RUBINSTEIN, M., DURAND, F., FREEMAN, W. T. *Phase-based video motion processing*. ACM Transactions on Graphics (TOG). 2013.

[25] WADHWA, N., RUBINSTEIN, M., DURAND, F., FREEMAN, W. T. *Riesz pyramid for fast phase-based video magnification*. In Computational Photography (ICCP), 2014 IEEE International Conference, IEEE. 2014.

[26] WANG, J., DRUCKER, S. M., AGRAWALA, M., COHEN, M. F. T*he cartoon animation filter*. ACM Transactions on Graphics. 2006.

[27] Wikipedia. Método de Euler. Disponível em: <https://pt.wikipedia.org/wiki/M%C3%A9todo_de_Euler> Acessado em 30/11/2017.

[28] WU, H.-Y., RUBINSTEIN, M., SHIH, E., GUTTAG, J., DURAND, F., FREEMAN, W. *Eulerian video magnification for revealing subtle changes in the world.* ACM Transactions on Graphics (TOG). 2012.

ANEXO

I. MICROFONE VISUAL

Este anexo apresentará os pricipais códigos utilizados no software Matlab para desesnvolver o projeto de microfone visual. Eles também estão disponíveis no site do projetista [7].

FUNÇÃO PRINCIPAL

```
clear
clc
setPath
currentDirectory = pwd;
dataDir = 'C:\Users\Lorena\Documents\UnB\2017.2\TCC\FINAL\2CapMicrofone\VMSlim';
vidName = 'VID-20171006-WA0005';
vidExtension = '.avi';
testcasename = vidName;
nscales = 3;
norientations = 2;
dsamplefactor = 0.1; %downsample to 0.1 full size

filename = [vidName vidExtension];

vr = VideoReader(fullfile(dataDir, filename));

samplingrate = 2200;

wndw = 80;
olap = 40;
```

```
S = vmSoundFromVideo(vr, nscales, norientations, 'SamplingRate',
samplingrate,'DownsampleFactor', dsamplefactor);
S.fileName = vidName;

%%
%show spectrogram and play sound
close all;
spectrogram(S.x, 100, 50)
vmPlaySound(S)

%%compute spectral subtraction
Sspecsub = vmGetSoundSpecSub(S);

%%
close all;
spectrogram(Sspecsub.x, wndw, olap)
vmPlaySound(Sspecsub)

%%
wndw = 80;
olap = 40;

S_unfiltered = S;
S_unfiltered.x = S.aligned;
```

```
nc = 3;nr = 2;pn=1;
close all;

subplot(nc,nr,pn);pn=pn+1;
plot(S_unfiltered.x);
title('recovered time signal')
subplot(nc,nr,pn);pn=pn+1;
spectrogram(S_unfiltered.x, wndw, olap)
title('recovered spectrogram')

subplot(nc,nr,pn);pn=pn+1;
plot(S.x);
title('highpass time signal')
subplot(nc,nr,pn);pn=pn+1;
spectrogram(S.x, wndw, olap)
title('highpass spectrogram')

subplot(nc,nr,pn);pn=pn+1;
plot(Sspecsub.x)
title('spec sub time signal')
subplot(nc,nr,pn);pn=pn+1;
spectrogram(Sspecsub.x, wndw, olap)
title('spec sub spectrogram')

%%
vmWriteWAV(S, 'RecoveredSound.wav');
```

FUNÇÃO vmSoundFromVideo – Extrai Áudio de Pequenas Vibrações do Vídeo

```
function [S] = vmSoundFromVideo(vHandle, nscalesin, norientationsin, varargin)

%   Extracts audio from tiny vibrations in video.
tic;
startTime = toc;

% Parameters
defaultnframes = 0;
defaultDownsampleFactor = 1;
defaultsamplingrate = -1;
p = inputParser();
addOptional(p, 'DownsampleFactor', defaultDownsampleFactor, @isnumeric);
addOptional(p, 'NFrames', defaultnframes, @isnumeric);
addOptional(p, 'SamplingRate', defaultsamplingrate, @isnumeric);
parse(p, varargin{:});
nScales = nscalesin;
nOrients = norientationsin;
dSampleFactor = p.Results.DownsampleFactor;
numFramesIn = p.Results.NFrames;
samplingrate = p.Results.SamplingRate;

if(samplingrate<0)
    samplingrate = vHandle.FrameRate;
end
```

```
'Reading first frame of video'
colorframe = vHandle.read(1);
'Successfully read first frame of video'

if(dSampleFactor~=1)
    colorframe = imresize(colorframe,dSampleFactor);
end

fullFrame = im2single(squeeze(mean(colorframe,3)));
refFrame = fullFrame;

[h,w] = size(refFrame);%height and width of video in pixels

nF = numFramesIn;
if(nF==0)
    %depending on matlab and type of video you are using, may need to read
    %the last frame
    %lastFrame = read(vHandle, inf);
    nF = vHandle.NumberOfFrames;%number of frames
end

%%
[pyrRef, pind] = buildSCFpyr(refFrame, nScales, nOrients-1);

for j = 1:nScales
    for k = 1:nOrients
```

```
        bandIdx = 1+nOrients*(j-1)+k;
    end
end

%
totalsigs = nScales*nOrients;
signalffs = zeros(nScales,nOrients,nF);
ampsigs = zeros(nScales,nOrients,nF);

% Process
nF

for q = 1:nF
    if(mod(q,floor(nF/100))==1)
        progress = q/nF;
        currentTime = toc;
        ['Progress:' num2str(progress*100) '% done after ' num2str(currentTime-startTime) '
seconds.']
    end

    vframein = vHandle.read(q);
    if(dSampleFactor == 1)
        fullFrame = im2single(squeeze(mean(vframein,3)));
    else
        fullFrame = im2single(squeeze(mean(imresize(vframein,dSampleFactor),3)));
    end
```

```
    im = fullFrame;

    pyr = buildSCFpyr(im, nScales, nOrients-1);
    pyrAmp = abs(pyr);
    pyrDeltaPhase = mod(pi+angle(pyr)-angle(pyrRef), 2*pi) - pi;


    for j = 1:nScales
        bandIdx = 1 + (j-1)*nOrients + 1;
        curH = pind(bandIdx,1);
        curW = pind(bandIdx,2);
        for k = 1:nOrients
            bandIdx = 1 + (j-1)*nOrients + k;
            amp = pyrBand(pyrAmp, pind, bandIdx);
            phase = pyrBand(pyrDeltaPhase, pind, bandIdx);

            %weighted signals with amplitude square weights.
            phasew = phase.*(abs(amp).^2);

            sumamp = sum(abs(amp(:)));
            ampsigs(j,k,q)= sumamp;

            signalffs(j,k,q)=mean(phasew(:))/sumamp;
        end
    end
end
```

```
%avx is average x
S.samplingRate = samplingrate;

%%
sigOut = zeros(nF, 1);
for q=1:nScales
   for p=1:nOrients
      [sigaligned, shiftam] = vmAlignAToB(squeeze(signalffs(q,p,:)), squeeze(signalffs(1,1,:)));
      sigOut = sigOut+sigaligned;
      shiftam
   end
end

S.aligned = sigOut;

%sometimes the alignment aligns on noise and boosts it, in which case just
%use averaging with no alignment, or highpass before alignment
S.averageNoAlignment = mean(reshape(double(signalffs),nScales*nOrients,nF)).';

highpassfc = 0.05;
[b,a] = butter(3,highpassfc,'high');
S.x = filter(b,a,S.aligned);

%sometimes butter doesn't fix the first few entries
S.x(1:10)=mean(S.x);

maxsx = max(S.x);
```

```
minsx = min(S.x);
if(maxsx~=1.0 || minsx ~= -1.0)
    range = maxsx-minsx;
    S.x = 2*S.x/range;
    newmx = max(S.x);
    offset = newmx-1.0;
    S.x = S.x-offset;
end

%
end
```

FUNÇÃO vmGetSoundSpecSub – Recupera o Som

```
        function [ Smod ] = vmGetSoundSpecSub( S , qtl1, qtl2)
%vmGetSoundSpecSub Recover sound from the stft of S
st = vmComputeSTFT(S);

if(nargin == 3)
    newst = vmComputeSpecSub(st,qtl1,qtl2);
else
    newst = vmComputeSpecSub(st);
end

Smod = S;
Smod.x = double(real(vmSTFTResynth(newst.s,st.windowSize,st.hopSize, 0, 'hann')));
```

```
%%

Smod.x = Smod.x(1:length(S.x));


%scale to -1,1 so we can listen to it.

Smod = vmGetSoundScaledToOne(Smod);


end
```