UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Vinícius Leonardo Gadioli da Silva

ANÁLISE DO EFEITO LOMBARD NO MOVIMENTO FACIAL POR MEIO DE OPTICAL FLOW

VIÇOSA - MG
2013

VINÍCIUS LEONARDO GADIOLI DA SILVA

ANÁLISE DO EFEITO LOMBARD NO MOVIMENTO FACIAL POR MEIO DE OPTICAL FLOW

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 490 - Monografia e Seminário e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.
Orientadora: Prof[a]. Dr[a]. Ketia Soares Moreira

VIÇOSA - MG
2013

VINICIUS LEONARDO GADIOLI DA SILVA

ANÁLISE DO EFEITO LOMBARD NO MOVIMENTO FACIAL POR MEIO DE OPTICAL FLOW

Monografia apresentada ao Departamento de Engenharia Elétrica do Centro de Ciências Exatas e Tecnológicas da Universidade Federal de Viçosa, para a obtenção dos créditos da disciplina ELT 490 - Monografia e Seminário e cumprimento do requisito parcial para obtenção do grau de Bacharel em Engenharia Elétrica.

Aprovada em 23 de Agosto de 2013.

COMISSÃO EXAMINADORA

Profª. Drª. Ketia Soares Moreira - Orientadora
Universidade Federal de Viçosa

Prof. Dr. Gustavo Fernandes Rodrigues - Membro
Universidade Federal de São João Del Rei

Prof. Dr. André Gomes Torres - Membro
Universidade Federal de Viçosa

À minha mãe, POR TUDO.

Agradecimentos

Minha sincera gratidão,

À Deus, por me abrir as portas e permitir assim que os próximos agradecimentos possam ser feitos.

À minha mãe pelo amor, esforço e dedicação.

Minha irmã Gabriela e Cristiano pelo apoio nessa grande etapa.

Aos meus familiares, por acreditarem em mim.

À professora Ketia pela oportunidade de trabalhar e aprender desde o início da graduação.

Aos grandes amigos que fiz ao longo desse curso, principalmente os primeiros.

À Lais pelo amor, companheirismo e paciência.

À todos aqueles que passaram pela minha vida e de alguma forma contribuíram na minha formação como profissional e como ser humano.

À FAPEMIG por financiar minhas bolsas de iniciação científica e à CAPES pela bolsa do Ciência sem Fronteiras.

E por último e não menos importante, meus sinceros agradecimentos a todas as MARRETAS que me ajudaram durante o curso.

Marreta never ends!!

Vinícius Leonardo Gadioli da Silva

"...Grandes coisas fez o Senhor a estes.Grandes coisas fez o Senhor por nós, pelas quais estamos alegres."

Salmos 126:2-3

Resumo

Movimento é uma poderosa característica em sequências de imagens, determinando a dinâmica da cena de acordo com relação espacial das características da imagem com a variação do tempo. Uma importante técnica para estimar movimento em sequência de imagens é o chamado Optical Flow (Fluxo Óptico). Fluxo Óptico é a distribuição da velocidade aparente do movimento dos padrões de intensidade em uma imagem. O objetivo principal deste trabalho é estudar, por meio de Fluxo Ótico, o movimento facial na produção acústica, analisando as diferenças existentes do movimento na fala com e sem o Efeito Lombard. O cálculo de Fluxo Óptico será feito utilizando-se o método matemático de Horn & Shunk. Este método requer condições de suavização, restrição, minimização, estimações e simplificações, que se forem criteriosamente respeitadas podem gerar um cálculo eficiente de movimento. No estudo do movimento na produção da fala, o Optical Flow também pode acrescentar informações, sendo utilizado em recursos de leitura labial. Esta técnica, que consiste na identificação visual dos sons produzidos através dos movimentos dos lábios do locutor, é inerente ao processo de comunicação entre pessoas. Tal recurso é mais explorado em ambientes ruidosos. Nestes locais, os interlocutores falam mais alto, articulando claramente as palavras ficando mais atentos as informações visuais. Essa tendência natural é conhecida como Efeito Lombard. Assim, quando pessoas estão se comunicando em ambientes ruidosos ou não, utiliza-se da informação visual em maior ou menor resolução como complemento para um melhor entendimento das mensagens. A grande expectativa é, ao fim do trabalho, ter criado uma metodologia que possa ser seguida para a continuidade dos trabalhos envolvendo Processamento Audiovisual, Fluxo Óptico e Efeito Lombard. Espera-se também que os resultados deste trabalho demonstrem possibilidades de melhorias nos resultados de trabalhos futuros, com a finalidade de fornecer contribuições para a área de Visão Computacional que possam ser a cada dia melhoradas.

Sumário

1 Introdução

Movimento é uma poderosa característica em sequências de imagens, determinando a dinâmica da cena de acordo com relação espacial das características da imagem de acordo com a variação do tempo. A tarefa de análise de movimento é um desafio constante e um problema fundamental em inteligência computacional. A dimensão temporal em processamento visual é importante por duas razões: O movimento aparente de objetos sobre um plano de imagem é uma forte sugestão visual para entender a estrutura e movimento 3D.[1]

Uma importante técnica para estimar movimento em sequência de imagens é o *Optical Flow* (Fluxo Óptico). Fluxo Óptico é a distribuição da velocidade aparente do movimento dos padrões de intensidade em uma imagem. Fluxo Óptico pode surgir de um movimento relativo de objetos e vistas e pode gerar informações importantes sobre o movimento dos objetos vistos e a taxa de movimento desses objetos. [2]

No estudo do movimento na produção da fala, o FO também pode acrescentar informações, podendo ser utilizado na leitura labial. A identificação visual dos sons produzidos através dos movimentos dos lábios do locutor é inerente ao processo de comunicação entre pessoas, entretanto este recurso é mais explorado em ambientes ruidosos. Nestes locais, os interlocutores falam mais alto, articulando claramente as palavras, e ficando mais atentos às informações visuais. Esta tendência natural do indivíduo aumentar o volume da sua voz, articulando melhor os lábios é conhecida como *Efeito Lombard* [3]. Assim, quando pessoas estão se comunicando em ambientes ruidosos, ou não, se utiliza da informação visual em maior ou menor resolução como complemento para um melhor entendimento das mensagens .[4]

O ser humano utiliza a informação sonora contida na componente acústica da fala durante sua comunicação. Porém, a informação percebida pelo interlocutor, além da componente acústica, possui também uma componente visual. A informação visual representada pelo movimento da cabeça, lábios, mandíbula e bochechas facilita e complementa

a percepção da fala [5]. As pessoas, quando submetidas a ambientes ruidosos, necessitam mais da parcela visual da comunicação. Espectadores, ao assistir ao vídeo de um locutor falando na presença de ruído progressivamente mais intenso, dividiam sua atenção entre a região dos olhos e da parte inferior da face do locutor (onde se localiza a maior parte da informação visual da fala). À medida que o ruído aumentava, crescia também a atenção sobre a região inferior da face do locutor [6]. Porém, mesmo na presença de níveis de ruído elevados, parte da atenção mantinha-se sobre os olhos do locutor. Concluíram, desta forma, que a detecção da informação visual contida na fala acontece em baixa resolução temporal, que tal informação não se restringe à região da boca e que a atenção do ouvinte varia de acordo com a necessidade do ambiente.[7]

A relação existente entre as componentes visual e a acústica da fala vem sendo explorada no meio científico, sendo aplicada na criação de movimentos realísticos de *talking faces* [8], na estimação do movimento da face por meio da acústica da fala e vice-versa [9], e em sistemas de comunicação audiovisual [10].

É importante salientar que, em algumas aplicações, a leitura labial pode ser o único recurso para o reconhecimento da mensagem expressa. Este fato demonstra a importância de estudos nesta área, pois técnicas e tecnologias para a comunicação envolvendo pessoas com necessidades especiais ou dificuldades na produção ou na audição da fala estão em constante desenvolvimento.[11]

1.1 Cálculo da Equação de restrição do Fluxo Ótico

Os métodos para a computação do Fluxo Óptico podem ser classificados em três grandes grupos principais: Técnicas diferenciais, Técnicas de correlação e Técnicas baseadas em frequência de energia. Nas técnicas diferenciais, a hipótese inicial para a computação do fluxo ótico é a de que a intensidade entre quadros diferentes em uma sequência de imagens é aproximadamente constante em um intervalo de tempo pequeno, ou seja, em um pequeno intervalo de tempo o deslocamento será mínimo.[12]

Seja $I(x, y, t)$ a intensidade num dado pixel (x, y) num instante de tempo t. Inicialmente, considera-se que o intervalo de tempo dt entre duas imagens consecutivas é muito curto e que intensidade da imagem não se altera nesse intervalo de tempo. Assim:

$$I(x, y, t) = I(x + \partial x, y + \partial y, t + \partial t) \tag{1.1}$$

A Equação 1.1 pode ser expandida pela série de Taylor e reescrita como:

$$I(x + \partial x, y + \partial y, t + \partial t) = I(x, y, t) + \frac{\partial I}{\partial x}\partial x + \frac{\partial I}{\partial y}\partial y + \frac{\partial I}{\partial t}\partial t + O^2 \tag{1.2}$$

Aplicando-se a Equação 1.1 na Equação 1.2 e eliminando-se O^2 (que são termos de alta ordem), tem-se:

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}\partial x + \frac{\partial I}{\partial y}\partial y + \frac{\partial I}{\partial t}\partial t \tag{1.3}$$

Simplificando a Equação 1.3:

$$\frac{\partial I}{\partial x}\partial x + \frac{\partial I}{\partial y}\partial y + \frac{\partial I}{\partial t}\partial t = 0 \tag{1.4}$$

Dividindo-se todos os termos da Equação 1.4 por ∂t:

$$\frac{\partial I}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial I}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial I}{\partial t} = 0 \tag{1.5}$$

Deste modo, são encontrados os componentes do vetor velocidade $\overline{v}$ e do gradiente da função imagem nas direções x e y,∇I :

$$\overline{v} = \frac{\partial x}{\partial t} + \frac{\partial y}{\partial t} \tag{1.6}$$

$$\nabla I = \frac{\partial I}{\partial x} + \frac{\partial I}{\partial y} \tag{1.7}$$

Deste modo, a equação de restrição do Fluxo Óptico torna-se:

$$\nabla I \cdot \overline{v} + I = 0 \tag{1.8}$$

1.2 Modelo de Horn & Shunk

1.2.1 Suposições

Para evitar variações no brilho que ocorrem devido a efeitos de sombras, Horn E Shunk assumem que a superfície a estudada é plana e que a iluminação sobre a superfície

é uniforme. Assim, pode-se dizer que a iluminação num ponto qualquer da imagem é proporcional à reflexão na superfície correspondente naquele ponto no objeto. Assume-se também que a reflexão varia suavemente e não possui descontinuidade .[12]

1.2.2 Restrições

Horn & Shunck derivam a equação que retrata a mudança na iluminação em uma imagem a um ponto para o modelo de movimento da iluminação. Para realizar isto, algumas restrições são estabelecidas .[12]

1.2.2.1 Restrição de Iluminação Constante

Seja a iluminação de uma imagem no ponto *(x,y)*, no plano da imagem no tempo t descrito por *E(x,y,t)*. Considerando-se o que acontece quando o modelo se move, a iluminação de um ponto particular no modelo é constante, então:

$$\frac{\partial x}{\partial t} \tag{1.9}$$

Que pode ser expandida pela regra da cadeia:

$$\frac{\partial E}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial E}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial E}{\partial t} = 0 \tag{1.10}$$

E assumindo que:

$$u = \frac{\partial x}{\partial t} \quad v = \frac{\partial y}{\partial t} \tag{1.11}$$

Tem-se uma única equação linear com duas variáveis desconhecidas u e v.

$$E_x u + E_y v + E_t = 0 \tag{1.12}$$

A Figura 1 ilustra a influência da fonte de iluminação na computação do FO.

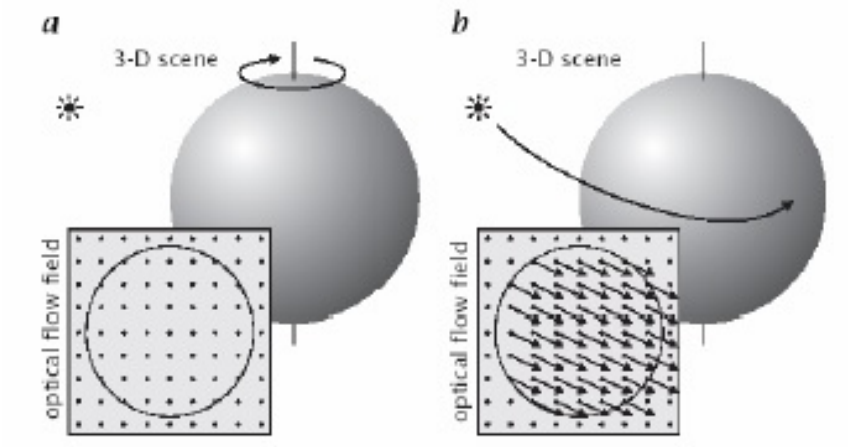

Figura 1: Limitação de Horn & Shunck: a) Um giro da esfera com a iluminação fixa determina um fluxo ótico=0. b) Um movimento da fonte de iluminação causa um campo de fluxo ótico aparente sem movimento da esfera. Fonte: [1].

1.2.2.2 Restrição de Suavização

Outra restrição colocada na solução para Fluxo Óptico no trabalho de it Horn E Shunck é a restrição de suavização. Nesta restrição observa-se que os pontos vizinhos em um objeto em movimento possuem velocidades similares, consequentemente o padrão de iluminação na imagem varia suavemente por ela. A utilização desta restrição de suavização é feita através dos Mínimos quadrados da magnitude do gradiente.

$$\nabla^2 u = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 \tag{1.13}$$

$$\nabla^2 v = \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \tag{1.14}$$

1.2.3 Estimações

1.2.3.1 Estimação das Derivadas Parciais

É preciso estimar as derivadas da iluminação do conjunto discreto de imagens medidas disponíveis. É essencial que a estimação de Ex, Ey e Et sejam consistentes, isto é, todas elas devem ser referenciadas ao mesmo ponto na imagem ao mesmo tempo [12]. Para realizar esta estimação *Horn & Shunck* utilizam um ponto no centro de um cubo formado por oito medições cuja relação entre espaço e tempo entre estas medições é mostrado na Figura 2. A estimação é calculada pela média das quatro primeiras diferenças em duas regiões adjacentes da imagem:

$$E_x \approx \frac{1}{4}[E_{ii+1,jj+1,kk} - E_{ii+1,jj,kk} + E_{ii,jj+1,kk} - E_{ii,jj,kk} \ldots + E_{ii+1,jj+1,kk+1} \\ -E_{ii+1,jj,kk+1} + E_{ii,jj+1,kk+1} - E_{ii,jj,kk+1}] \quad (1.15)$$

$$E_y \approx \frac{1}{4}[E_{ii,jj,kk} - E_{ii+1,jj,kk} + E_{ii,jj+1,kk} - E_{ii+1,jj+1,kk} \ldots + E_{ii,jj,kk+1} \\ -E_{ii+1,jj,kk+1} + E_{ii,jj+1,kk+1} - E_{ii+1,jj+1,kk+1}] \quad (1.16)$$

$$E_z \approx \frac{1}{4}[E_{ii+1,jj,kk+1} - E_{ii+1,jj,kk} + E_{ii,jj,kk+1} - E_{ii,jj,kk} \ldots + E_{ii+1,jj+1,kk+1} \\ -E_{ii+1,jj,kk+1} + E_{ii,jj+1,kk+1} - E_{ii,jj+1,kk}] \quad (1.17)$$

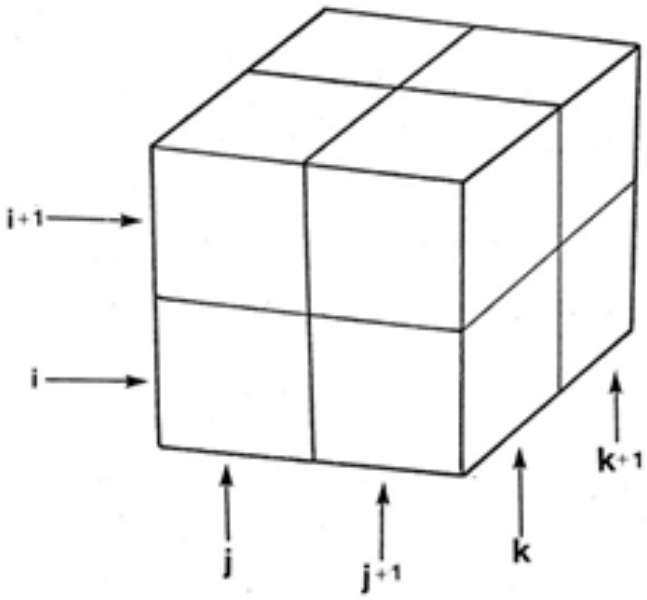

Figura 2: As três derivadas parciais da iluminação da imagem ao centro do cubo são estimadas pela média das primeiras diferenças nas quatro bordas paralelas do cubo. A coluna índice j corresponde a direção x na imagem e a coluna índice i a direção y. Enquanto k representa a direção do tempo. Fonte: [1]

1.2.3.2 Estimação do Laplaciano do Fluxo de Velocidade

É necessário aproximar o Laplaciano de u e v. Uma convincente aproximação é dada pela fórmula:

$$\nabla^2 u \approx k(\overline{u}_i, j, k - u_i, j, k) \tag{1.18}$$

$$\nabla^2 v \approx k(\overline{v}_i, j, k - v_i, j, k) \tag{1.19}$$

Onde $\overline{u}$ e $\overline{v}$ são médias locais dos vetores de velocidade. Eles são estimados pela subtração do valor em um ponto a uma média ponderada dos valores vizinhos. Daí, as equações de $\overline{u}$ e $\overline{v}$ são expressas, respectivamente, pelas Equações 1.20 e 1.21:

$$\begin{aligned} \overline{u}_{i,j,k} &\approx \frac{1}{6}[u_{ii-1,jj} + u_{ii,jj+1} + u_{ii+1,jj} + u_{ii,jj-1}] \ldots \\ &+\frac{1}{12}[u_{ii-1,jj-1} + u_{ii-1,jj+1} + u_{ii+1,jj+1} + u_{ii+1,jj-1}] \end{aligned} \tag{1.20}$$

$$\begin{aligned} \overline{v}_{i,j,k} &\approx \frac{1}{6}[v_{ii-1,jj} + v_{ii,jj+1} + v_{ii+1,jj} + v_{ii,jj-1}] \ldots \\ &+\frac{1}{12}[v_{ii-1,jj-1} + v_{ii-1,jj+1} + v_{ii+1,jj+1} + v_{ii+1,jj-1}] \end{aligned} \tag{1.21}$$

1.2.4 Minimização de Erros

O problema então é minimizar a soma dos erros nas equações para a taxa de mudança da iluminação da imagem,

$$\varepsilon_b = E_x u + E_y v + E_t \tag{1.22}$$

E a medida das saídas de suavização na velocidade do fluxo,

$$(\varepsilon_c)^2 = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2 \tag{1.23}$$

Na prática, a medida da iluminação da imagem será corrompida pelo erro de quantização e ruído, de uma maneira que não se pode esperar a ser igual a zero. Este valor tenderá ter uma magnitude de erro que é proporcional ao ruído na medição [1]. A minimização a ser alcançada achando valores satisfatórios para a velocidade do fluxo ótico (u, v).Usando o calculo de variação será obtido:

$$E_x^2 + E_x E_y = \alpha^2 \nabla^2 u - E_x E_t \tag{1.24}$$

Usando a aproximação do Laplaciano das Equações 1.18 e 1.19:

$$(E_x^2 + \alpha^2)u + E_x E_y v = \alpha^2 \overline{u} - E_x E_t \tag{1.25}$$
$$(E_y^2 + \alpha^2)v + E_x E_y u = \alpha^2 \overline{v} - E_x E_t \tag{1.26}$$

O determinante da matriz de coeficiente é igual a:

$$\alpha^2(E_x^2 + E_y^2 + \alpha^2) \tag{1.27}$$

Resolvendo as Equações 1.25 e 1.26 a partir da Equação 1.27 para u e v encontra-se:

$$(E_x^2 + E_y^2 + \alpha^2)u = (\alpha^2 + E_y^2)\overline{u} - E_x E_y \overline{v} - E_x E_t \tag{1.28}$$
$$(E_x^2 + E_y^2 + \alpha^2)v = (\alpha^2 + E_y^2)\overline{v} - E_x E_y \overline{u} - E_x E_t \tag{1.29}$$

1.2.4.1 Solução Iterativa

Uma solução direta para a restrição de minimização necessita de um elevado recurso computacional, portanto uma solução interativa pode ser sugerida. Este método calcula um novo conjunto de velocidades estimadas $u^{(n+1)}, v^{(n+1)}$, baseada nas derivadas estimadas e a média da velocidade [1]. A solução iterativa pode ser expressa por:

$$u^{n+1} = u^{-n} - E_x \frac{E_x u^{-n} + E_y v^{-n} + E_t}{E_x^2 + E_y^2 + \alpha^2} \tag{1.30}$$
$$v^{n+1} = v^{-n} - E_x \frac{E_x u^{-n} + E_y v^{-n} + E_t}{E_x^2 + E_y^2 + \alpha^2} \tag{1.31}$$

1.3 PCA - *Principal Component Analysis*

A Análise de Componentes Principais representa uma técnica estatística poderosa utilizada na redução do número de variáveis, fornecendo as ferramentas adequadas para identificar as variáveis mais importantes no espaço das componentes principais. A técnica consiste em reescrever as variáveis originais, através da transformação de coordenadas. [13]

A transformação de coordenadas é um processo trivial quando feito usando matrizes.

As novas variáveis são denominadas componentes principais, geradas através de uma transformação estatística realizada sobre as variáveis originais. Assim, cada componente principal é uma combinação linear destas. [13]

Em geral, a explicação de toda a variabilidade do sistema determinado por **p** variáveis só pode ser efetuada por **p** componentes principais. No entanto, uma grande parte dessa variabilidade pode ser explicada por um número r menor de componentes, $\boldsymbol{r \leq p}$. [14]

No método de componentes principais, primeiramente é necessário encontrar a matriz de covariância **S** da sequência **X** de **M** vetores coluna de dimensão **N** [15]. Nesta análise, a sequência de entrada é constituída de imagens (matrizes bi-dimensionais), que devem ser convertidas em vetores coluna unidimensionais. Isso pode ser feito varrendo a imagem linha por linha ou coluna por coluna. Uma imagem contendo L linhas e C colunas, produz, então, um vetor coluna contendo $\mathbf{N = L \text{ x } C}$ linhas. Tomando uma matriz **Xo** na qual cada vetor coluna é a diferença entre os vetores da matriz **X** e a média desses mesmos vetores, a matriz de covariância pode ser obtida pela seguinte relação:

$$S = X_0 X_0' \tag{1.32}$$

O objetivo do método de componentes principais é encontrar um sistema alternativo de coordenadas **Z** para a sequência de entrada **X** no qual todos os elementos fora da diagonal principal da matriz de covariância $\boldsymbol{S_z}$ (o índice **Z** denota o novo sistema de coordenadas) sejam zero. Se os autovetores da matriz de covariância são conhecidos, e U representa a matriz desses autovetores:

$$S = U^t S U \tag{1.33}$$

O cálculo dos autovetores envolve operações com a matriz de covariância **S**. Mesmo com pequenas imagens, essa matriz pode ser muito grande para que se trabalhe com ela. Entretanto, se o número de imagens na sequência **X** é consideravelmente menor, o que geralmente ocorre, é possível reduzir o esforço computacional através da aplicação de decomposição em valores singulares (*Singular Value Decomposition* - SVD). Essa decomposição permite expressar os autovetores da matriz $\boldsymbol{S = XoXo'}$ como combinação linear de uma matriz $\boldsymbol{C = Xo^t Xo}$. Como a matriz C tem dimensão M x M, o custo computacional para encontrar os autovetores da matriz S é bastante reduzido. A transformação para o espaço de componentes principais pode ser expressa pela relação:

$$Z = U^t X_0 \tag{1.34}$$

onde **Z** é a matriz de componentes principais. A transformada inversa é obtida pela relação:

$$X_0 = UZ \tag{1.35}$$

(4) Quando uma imagem qualquer é transformada para o espaço de componentes principais, e transformada de volta usando a relação 1.35, a diferença vai indicar a distância entre a imagem analisada e as imagens utilizadas para gerar o espaço de componentes principais, ou seja, o quanto essas imagens são semelhantes.

1.4 Redes Neurais Artificiais - RNA

Redes Neurais Artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento, já o cérebro de um mamífero pode ter muitos bilhões de neurônios. As sinapses podem existir entre dois neurônios, entre célula sensorial e neurônio ou entre neurônio e órgão efetor (músculo ou glândula). Quando a célula efetora é um músculo, o local. [16] Através de suas terminações, os neurônios entram em contrato e transmitem impulsos a outros neurônios e às células efetuadoras; estes locais de contato são denominados, respectivamente, sinapses interneuronais e sinapses ou junções neuroefetuadoras.

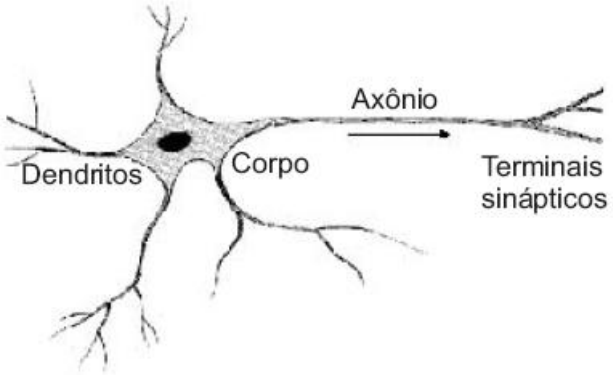

Figura 3: Representação de um neurônio biológico. Fonte: [17]

1.4.1 Características das RNA's

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede. Segundo BRAGA, a operação de uma unidade de processamento, proposta por$McCullock\&Pitts$ em 1943, pode ser resumida da seguinte maneira:

- Sinais são apresentados à entrada;
- Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade;
- É feita a soma ponderada dos sinais que produz um nível de atividade;
- Se este nível de atividade exceder certo limite (threshold) a unidade produz uma determinada resposta de saída.

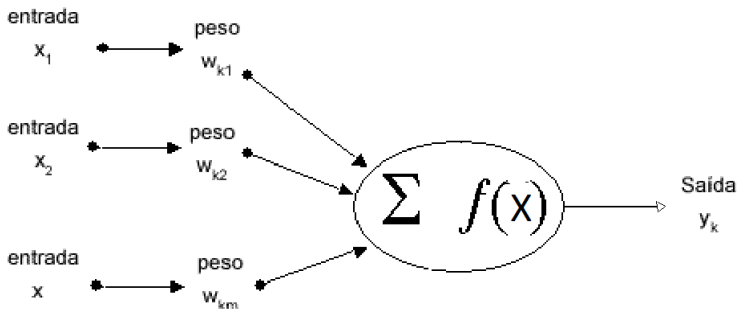

Figura 4: Neurônio Artificial proposto por *McCulloch & Pitts*.

A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem através de exemplos. Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior.

Usualmente as camadas são classificadas em três grupos:

- Camada de Entrada: onde os padrões são apresentados à rede;

- Camadas Intermediárias ou Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características;

- Camada de Saída: onde o resultado final é concluído e apresentado.

Uma rede neural é especificada, principalmente pela sua topologia, pelas características dos nós e pelas regras de treinamento. A seguir, serão analisados os processos de aprendizado.

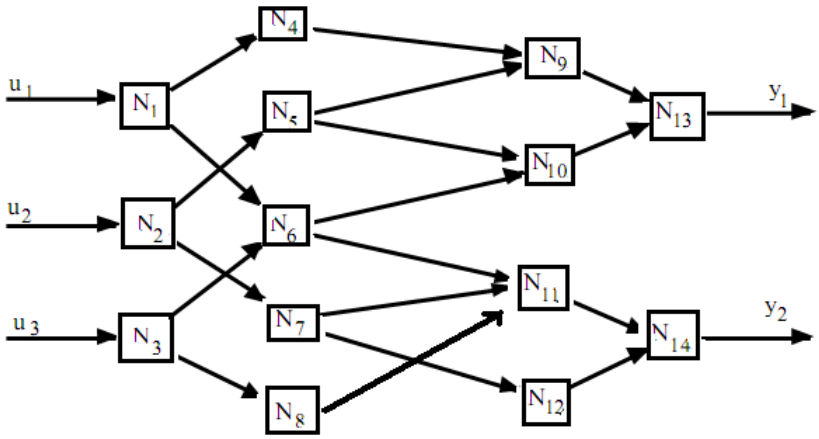

Figura 5: Rede Neural Artificial - Blocos.

1.4.2 Processos de Aprendizado

A propriedade mais importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas. Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados. (HAYKIN, S.). Outro fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente. Nesse contexto existem os seguintes paradigmas de aprendizado (HAYKIN, S.):

- **Aprendizado Supervisionado**: quando é utilizado um agente externo que indica à rede a resposta desejada para o padrão de entrada;

- **Aprendizado Não Supervisionado (auto-organização)**: é quando não existe um agente externo indicando a resposta desejada para os padrões de entrada;
- **Reforço**: quando um crítico externo avalia a resposta fornecida pela rede.

Denomina-se ciclo uma apresentação de todos os N pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos num ciclo pode ser executada de dois modos [16]:

1. Modo Padrão: A correção dos pesos acontece a cada apresentação à rede de um exemplo do conjunto de treinamento. Cada correção de pesos baseia-se somente no erro do exemplo apresentado naquela iteração. Assim, em cada ciclo ocorrem N correções.
2. Modo Batch: Apenas uma correção é feita por ciclo. Todos os exemplos do conjunto de treinamento são apresentados à rede, seu erro médio é calculado e a partir deste erro fazem-se as correções dos pesos.

1.5 Técnicas de Aquisição Acústica da Fala

O sinal acústico da fala pode ser conseguido sem maiores dificuldades por meio de um microfone e de um processo de conversão A/D (análogico-digital). Após a conversão A/D, o sinal é reamostrado a uma taxa apropriada para sua análise. Neste trabalho, seguem-se os passos descritos em [18, 9, 19, 10, 20], em que o sinal acústico da fala é transformado em parâmetros LSP [21], que são fortemente relacionados com a geometria do trato vocal. [7]

1.5.1 Representação Paramétrica Acústica da Fala

Para a representação do sinal acústico da fala utiliza-se de coeficientes LSP (*Line Spectrum Pairs*) [21] que são eficientes por estarem ligados às freqüências de ressonância do trato vocal, os formantes. Esta representação é justificada porque os formantes são determinados pela geometria do trato vocal, e o formato do trato vocal tem forte influência sobre os movimentos realizados na face [9]. Os parâmetros LSP são fundamentados no princípio de conservação da envoltória do espectro da fala que, segundo [21], é suficiente para assegurar a inteligibilidade do sinal.[7]

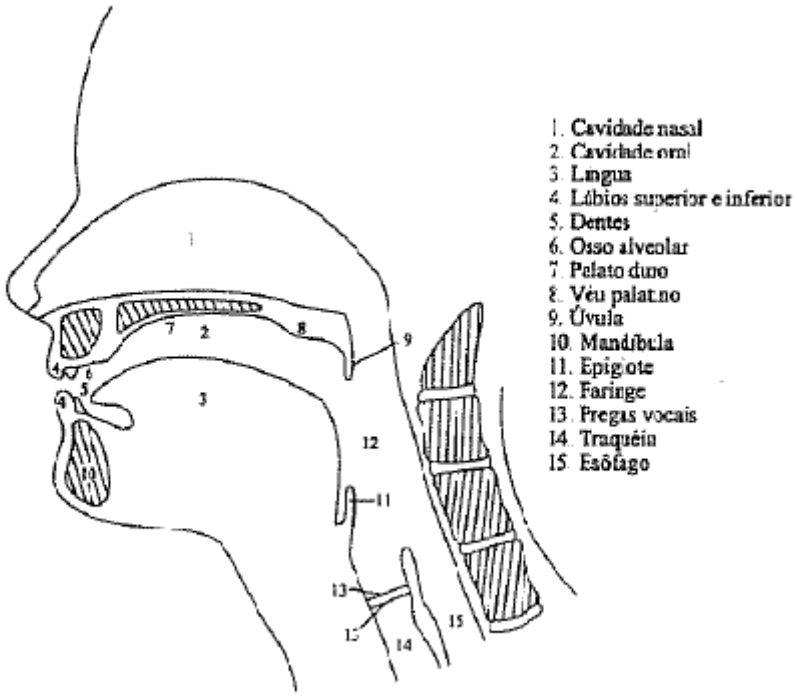

Figura 6: Partes principais do trato vocal humano *all pole*. Fonte: [18]

O sinal acústico da fala pode ser modelado como a saída de um filtro linear variante no tempo excitado por pulsos quase-periódicos no caso de fala sonora ou por ruído branco no caso de fala surda [21], [18], [22]. Em pequenos segmentos, o sinal da fala pode ser representado por um modelo fonte-filtro em que o sinal da pressão sonora é o produto: da velocidade do volume de ar gerado pela fonte, da característica de propagação dos lábios e da configuração do trato vocal. Assumindo que o processo de produção da fala seja estacionário em um curto intervalo de tempo, pode-se definir uma função de transferência para o trato vocal dentro deste intervalo. Na fala sonora, a função de transferência do trato vocal possui somente pólos, entretanto em sons surdos e nasais, normalmente, a função de transferência possui zeros e pólos, porém os zeros podem ser aproximados por pólos. Assim, o sinal da fala pode ser visto aproximadamente como um sinal de saída de um filtro de polos. [21], [18], [22]

Além da filtragem executada pelo trato vocal, a radiação labial e o fluxo glotal também contribuem para o processo de filtragem. Contudo, o fluxo de volume glotal durante um único período de vibração possui apenas pólos; e a radiação do som ao sair da boca possui zeros que por sua vez pode ser aproximado em pólos [18]. Assim, a função de transferência, no domínio z, do fluxo de volume glotal e da radiação labial pode ser representada aproximadamente como:

$$G(z)R(z) = \frac{K_1 K_2 (1 - z_{-1})}{(1 - z_a z_{-1})(1 - z_b z_{-1})}, \tag{1.36}$$

onde $G(z)R(z)$ é transformada z da contribuição conjunta do fluxo de volume glotal e da radiação labial. K_1 é uma constante relacionada com a amplitude do fluxo glotal e z_a e z_b são pólos relacionados com o fluxo glotal localizados no eixo real dentro do círculo unitário para que o filtro seja estável. K_2 é uma constante relacionada com a amplitude do fluxo de volume nos lábios e a distância dos lábios ao microfone.

Um modelo funcional do processo de produção da fala com base no modelo de pólos, em que as contribuições conjuntas do fluxo glotal, do trato vocal e da radiação labial são representadas por um único filtro auto-regressivo, linear de ordem p no instante j-ésimo, fundamentado na predição linear do sinal da fala (LPC), é descrito por:

$$\hat{s}(j) = -\sum_{i=1}^{p} \alpha_i s(j - i), \tag{1.37}$$

em que $\hat{s}(j)$ é o valor do sinal predito, $s(j - i)$ são os valores passados observados e α_i, $i = 1, ..., p$, são os coeficientes de predição linear que respondem pela ação de filtragem executada pelo trato vocal, pela radiação labial e pelo fluxo glotal. E a função de transferência do filtro de pólos é:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}}. \tag{1.38}$$

A análise por predição linear (LPC) objetiva uma boa estimação das propriedades espectrais do sinal. Para isto, os coeficientes de predição (α_1, ...,α_p) são obtidos em cada fala correspondente aos quadros, em que p é ordem do filtro usada na análise. Para uma taxa de amostragem de $f_s = 8$ kHz, há aproximadamente 4 frequências de ressonância até a frequência de Nyquist (4 kHz), implicando necessidade de 8 coeficientes de predição linear (2 coeficientes para cada par de pólos conjugados). Além disso, verificou-se ser útil empregar um par extra de coeficientes para representar a inclinação espectral determinada pela influência do pulso glotal e pela carga de irradiação nos lábios. Assim, utiliza-se um filtro de predição de ordem $p = 10$. [22], [23]

Equivalente aos parâmetros LPC, no domínio da freqüência, um novo conjunto de parâmetros chamados de LSP *(Line Spectrunm Pairs)* é definido $(w_1, \theta_1, ..., w_{p/2}, \theta_{p/2})$. Estes parâmetros LSP (w_i, θ_i) são obtidos a partir de um filtro de pólos estável:

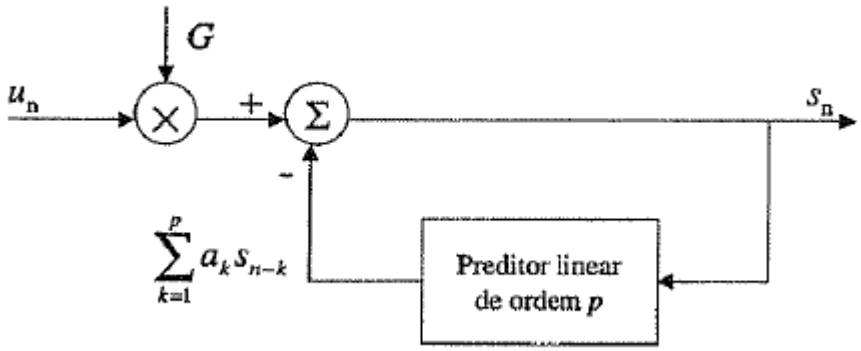

Figura 7: Diagrama de blocos de um modelo funcional do processo de produção da fala com base no modelo *all pole*. Fonte: [18]

$$A_p(z^{-1}) = 1 + \sum_{i=1}^{p} a_i z^{-i}. \tag{1.39}$$

O objetivo dos parâmetros LSP é de representar o polinômio $A_p(z^{-1})$ por meio de dois outros polinômios cujos zeros estão sobre a circunferência unitária:

$$P(z^{-1}) = A_p(z^{-1}) - z^{-(p+1)}A_p(z) = 1 + (a_1 - a_p)z^{-1} + \cdots + (a_p - a_1)z^{-p} - z^{-(p+1)}, \tag{1.40}$$

$$Q(z^{-1}) = A_p(z^{-1}) + z^{-(p+1)}A_p(z) = 1 + (a_1 + a_p)z^{-1} + \cdots + (a_p + a_1)z^{-p} + z^{-(p+1)}, \tag{1.41}$$

reconstruindo:

$$A_p(z^{-1}) = \frac{1}{2}[P(z^{-1}) + Q(z^{-1})]. \tag{1.42}$$

Considerando que todas as raízes do polinômio, (i.e e^{+jw} e e^{-jw}), estejam sobre o círculo unitário, ou seja, o filtro LSP é estável, e expressando o polinômio como um produtório, obtêm-se:

Para p par:

$$P(z^{-1}) = (1 - z^{-1}) \prod_{i=1}^{p/2} (1 - 2\cos w_i z^{-1} + z^{-2}), \tag{1.43}$$

$$Q(z^{-1}) = (1 + z^{-1}) \prod_{i=1}^{p/2} (1 - 2\cos \theta_i z^{-1} + z^{-2}). \tag{1.44}$$

Para p ímpar:

$$P(z^{-1}) = (1 - z^{-1}) \prod_{i=1}^{(p-1)/2} (1 - 2\cos w_i z^{-1} + z^{-2}), \tag{1.45}$$

$$Q(z^{-1}) = \prod_{i=1}^{(p+1)/2} (1 - 2\cos \theta_i z^{-1} + z^{-2}). \tag{1.46}$$

Assim, um conjunto de parâmetros (w_i, θ_i) é obtido e convertido em (f_i, g_i) em Hertz:

$$f_i = w_i(2\pi T) \tag{1.47}$$

e

$$g_i = \theta_i(2\pi T). \tag{1.48}$$

em que T é o período de amostragem. Os parâmetros LSP são:

$$\mathbf{f} = [f_1, g_1, f_2, g_2, \cdots, f_{\frac{p}{2}}, g_{\frac{p}{2}}]. \tag{1.49}$$

1.6 Objetivos

O objetivo principal deste trabalho é o estudo do comportamento das falas audível e visível sob influência do *Efeito Lombard*. Deste modo, espera-se caracterizar a fala através de ferramentas de processamento digital de sinais e modelagem matemática. A fala visível ou movimento facial será estudada através de Fluxo Óptico (*Optical Flow*). A fala audível por sua vez será estudada através de métodos, tais como: análise de amplitude, frequência fundamental, estimadores LPC estimadores LSP. Espera-se encontrar uma relação entre os parâmetros extraídos áudio-visualmente, a fim de criar um modelo que seja capaz de gerar parâmetros de áudio a partir de parâmetros visuais. Neste trabalho os parâmetros visuais a ser utilizado na definição de padrões será o Fluxo Óptico e os parâmetros de áudio serão estimadores LSP. A grande expectativa é , ao fim do trabalho, ter criado uma metodologia que possa ser seguida para a continuidade dos trabalhos envolvendo Processamento Audiovisual, Fluxo Óptico e Efeito Lombard. Espera-se também que os resultados deste trabalho demonstrem possibilidades de melhorias nos resultados de trabalhos futuros, com a finalidade de fornecer contribuições para a área de Visão Computacional que possam ser a cada dia melhoradas.

2 Metodologia

2.1 Modelo Baseado em Simplificações do Método de Horn & Shunk

Tomando como referência o artigo de *Horn & Shunck*, algumas simplificações matemáticas foram feitas simplificações no livro *Fundamental of Computer Vision* [24], com a finalidade de chegar mais rápido no cálculo do fluxo Óptico. Seja a função 3D $f(x, y, t)$, onde x, y são as coordenadas espaciais e t o tempo que representam uma sequência de imagem. Então $f(x1, y1, t1)$ é o nível de cinza na coordenada $x1, y1$ no tempo $t1$. Assumindo que uma pequena mudança dx, dy e dt em x, y e t não irá mudar o nível de cinza, isto é:

$$f(x, y, z) = f(x + \partial x, y + \partial y, t + \partial t) \tag{2.1}$$

Pela expansão da Série de Taylor, a Equação 2.1 se torna:

$$f_x \partial x + f_y \partial y + f_t \partial t = 0 \tag{2.2}$$

Dividindo-se cada termo da Equação 2.2 por ∂t, tem-se:

$$f_x u + f_y v + f_t = 0 \tag{2.3}$$

Onde $u = \frac{\partial x}{\partial t}$ e $v = \frac{\partial y}{\partial t}$ são as componentes horizontal e vertical do Fluxo Óptico.

Existem duas incógnitas, u e v, na Equação 2.3, de modo que a solução desta equação não é fácil. Assim, a Equação 2.3 pode ser rescrita como:

$$u = -\frac{f_x}{f_y} v - \frac{f_t}{f_y} \tag{2.4}$$

Esta (Equação 2.4) é a equação da linha reta no espaço $u - v$. Existem diversas possíveis soluções para ela. Estas soluções podem ser encontradas em qualquer lugar na linha mostrada na Figura 8. Um dos primeiros métodos para calcular o Fluxo Óptico foi proposto por *Horn&Shunck*. Este método, visto na seção anterior, propõe para estimar os valores de (u, v) seguindo a função de erro, E, que é minimizada por:

$$E(x, y) = (f_x u + f_y vy + f_t)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2) \tag{2.5}$$

O primeiro termo da Equação 2.5 é a restrição do Fluxo Óptico e o segundo termo corresponde a suavidade do FO. Para um correto fluxo ótico o primeiro termo deve ser próximo de zero, ou o quadrado do primeiro termo deve ser muito pequeno. Partindo do pressuposto que o movimento de objetos no mundo real é suave, o segundo termo força a restrição de suavidade. Diferenciando-se E com os respectivos u e v, e igualando a zero tem-se:

$$\frac{\partial E}{\partial u} = (f_x u + f_y vy + f_t) f_x + \lambda(u_x x + u_y y) = 0 \tag{2.6}$$
$$\frac{\partial E}{\partial v} = (f_x u + f_y vy + f_t) f_y + \lambda(v_x x + v_y y) = 0 \tag{2.7}$$

Substituindo $\Delta^2 u = u_x x + u_y y$ e $\Delta^2 v = v_x x + v_y y$:

$$\frac{\partial E}{\partial u} = (f_x u + f_y vy + f_t) f_x + \lambda(\Delta^2 u) = 0 \tag{2.8}$$
$$\frac{\partial E}{\partial v} = (f_x u + f_y vy + f_t) f_y + \lambda(\Delta^2 v) = 0 \tag{2.9}$$

Seja $\Delta^2 u = u - u_v \alpha$, onde $u_v \alpha$ é a média de u componentes de FO sobre os quatro vizinhos mais próximos do pixel. Seja também $\Delta^2 v = v - v_v \alpha$, assim:

$$\frac{\partial E}{\partial u} = (f_x u + f_y vy + f_t) f_x + \lambda(u - u_v \alpha) = 0 \tag{2.10}$$
$$\frac{\partial E}{\partial v} = (f_x u + f_y vy + f_t) f_y + \lambda(v - v_v \alpha) = 0 \tag{2.11}$$

Deste modo, as Equações 2.10 e 2.11 podem ser resolvidas para *(u,v)*.

$$u = \overline{u} - f_x \frac{P}{D}, (f_x = E_x) \tag{2.12}$$

$$v = \overline{v} - f_y \frac{P}{D}, (f_y = E_y) \tag{2.13}$$

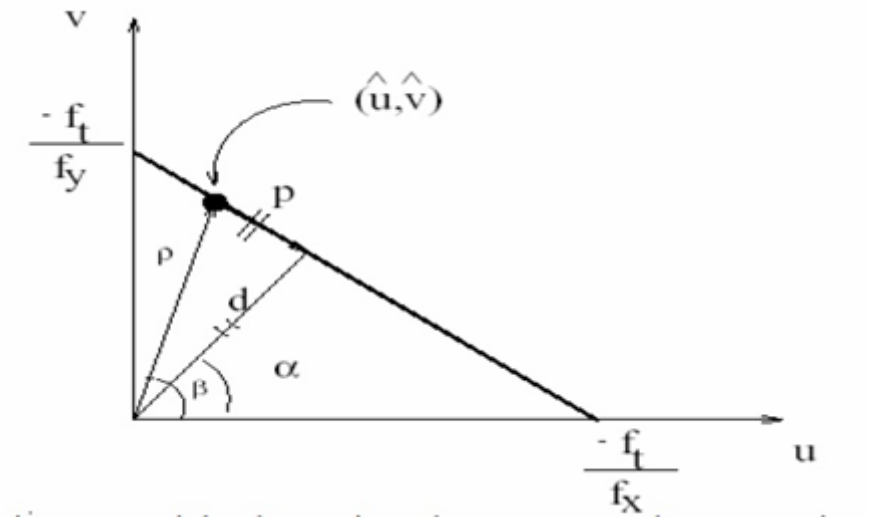

Figura 8: Linha de restrição do espaço $u - v$. d é o comprimento da perpendicular da origem até a linha, α é o ângulo entre a perpendicular e o eixo x. (u, v) é uma das possíveis soluções. Este vetor do fluxo ótico pode ser dividido em duas componentes: p, que esta ao longo da linha de restrição e d que esta perpendicular à linha de restrição. Fonte: [12]

Onde:

$$P = E_x\overline{u} + E_y\overline{v} + E_t \tag{2.14}$$

$$D = (E_x)^2 + (E_y)^2 + \lambda^2 \tag{2.15}$$

Deste modo, foi possível calcular os vetores de movimento baseados em Fluxo Óptico.

2.2 Aquisição de Dados

Por meio de uma câmera digital foram feitas filmagens no formato MP4 em HD de um mesmo interlocutor submetido a diferentes níveis de ruído. O experimento foi realizado da seguinte maneira:

- Foram feitas ao interlocutor cinco perguntas sem conhecimento prévio das mesmas;

- O interlocutor foi submetido á 5 níveis de ruído diferentes, através de um it headphone, denominados: Baixo, Médio, Alto, Muito Alto e nenhum ruído em 5 perguntas diferentes;
- A fim de se evitar interferências, as perguntas foram escritas em folha A3 e o interlocutor lia as perguntas para responder;
- O interlocutor foi submetido, tanto aos níveis de ruído quanto as perguntas, de maneira aleatória para evitar ao máximo interferência por prévio conhecimento das perguntas;
- O interlocutor só soube da finalidade do experimento após este ser realizado, a fim de manter a integridade do trabalho;
- Detalhes dos arquivos de vídeo:
  - Largura do Quadro: 1280;
  - Altura do Quadro: 720;
  - Taxa de dados: 12690 Kbps;
  - Taxa de bits total: 12830 Kbps;
  - Taxa de quadros: 59 quadros/s.

Feito o experimento, os arquivos de vídeo foram convertidos para AVI e WAV, para serem direcionados aos programas implementados para processamento de áudio e vídeo.

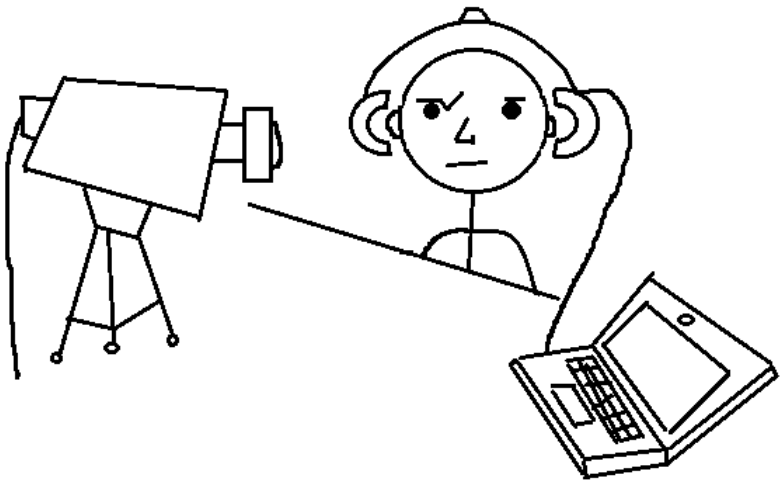

Figura 9: Aquisição da Base de Dados Audio/Video. Fonte: Willian Ramos do Carmo, Cartunista

Num primeiro momento foram feitas filmagens de movimentos de abertura e fechamento da boca conforme a Figura 9 (Detalhes na seção 3). Num segundo momento, foram estudados os comportamentos de alguns parâmetros tanto na fala visível quanto na audível. Neste segundo momento, foram feitas 5 perguntas ao voluntário e seguidos os passos descritos anteriormente.

2.3 Pré-processamento do vídeo

Um vídeo pode ser ponderado como uma sequência de imagens conectadas *frames*. Deste modo, se faz necessário separar essas imagens para realizar as iterações necessárias para o cálculo do FO pelo algoritmo desenvolvido. A base de dados de vídeo foi feita em HD, de modo que foi necessário converter os vídeos para o formato AVI a uma taxa de 29 $quadros/s$.

Assim, o vídeo é carregado e a sequência de imagens é salva num vetor do tipo *struct*. Isso permite que as imagens sejam acessadas durante a execução do programa.

Além disso, quando o código é executado é aberta uma guia com a primeira imagem (ou primeiro frame do vídeo) e permite que o usuário trace um retângulo, especificando assim a região de interesse da imagem na qual será aplicado o algoritmo de Fluxo Óptico. Feito isto, o programa recorta todos os frames do vídeo na mesma região retangular em que usuário escolheu no primeiro frame. Neste trabalho, a região de interesse é a boca.

2.4 Algoritmo para a Computação do Fluxo Óptico

O algoritmo implementado neste trabalho foi desenvolvido baseado nos modelos matemático anteriormente explicados e nos trabalhos do **CEFALA** (Centro de Estudos da Fala, Acústica, Linguagem e música) - UFMG e da *University of British Columbia* . O código desenvolvido utiliza o método de $Horn\&Shunk$ da seguinte maneira: [10]

Primeiramente é realizado o pré-processamento. O vídeo resultante desse pré-processamento é submetido ao modelo simplificado. Começando do frame $k = 1$ são feitas iterações aos pares de frames (o primeiro com o segundo, o segundo com o terceiro e assim por diante). O número de iterações realizadas é o número total de frames do vídeo menos 1. Isso ocorre devido ao tipo de janelamento utilizado.

Cada iteração retorna uma matriz u e uma matriz v contendo o fluxo ótico correspon-

dente ao movimento horizontal e vertical dos pixels da região de interesse. Em seguida é efetuada a soma de todos os valores de cada matriz u e v. Essa soma gera um coeficiente(valor arbitrário) de gradiente de deslocamento na região de interesse. Assim, são gerados o número de frames-1 coeficientes podem ser amostrados tanto graficamente, quanto utilizados para análise estatística. Neste trabalho avaliou-se os coeficientes de deslocamento horizontal e vertical no estudo nos resultados da Seção 3.1, e o módulo do Fluxo Óptico nas definições de padrões e modelagem das redes neurais.

Este método de coeficientes foi escolhido por ser mais viável computacionalmente. Caso fosse escolhido salvar todas as matrizes de fluxo ótico de todas as iterações, seria necessário um custo computacional elevado para tal aplicação.

2.5 Processamento dos Vetores de Movimento (Fluxo Óptico)

Os vetores de movimento u e v devem ser agora devidamente computados e armazenados, a fim de gerar um banco de dados que possibilitará o mapeamento acústico visual da fala.

Deste modo, foram criados dois bancos de dados com os *structs* de FO: o primeiro é o resultado da média aritmética de 7 em 7 vetores de movimento. Esse valor de janelamento foi escolhido por se tratar do tempo médio de duração de um fonema na língua portuguesa. Através da PCA dos vetores resultantes deste janelamento criou-se o segundo banco de dados

2.6 Processamento de Áudio

Neste trabalho, o sinal de voz é amostrado a uma taxa de 8040 amostras/s e dividido em 29 quadros/s. Cada quadro é multiplicado por uma janela de $Hamming$, para reduzir efeitos causados pelo janelamento. A cada quadro foi aplicada análise LPC de ordem 10. Os coeficientes LPC foram convertidos em coeficientes LSP, usados para representar cada quadro acusticamente. A Figura 10 mostra a parametrização de um trecho do sinal de voz, em que a envoltória espectral é a resposta em freqüência do filtro LPC. Os parâmetros LSP são representados nessa figura pelas linhas verticais. Por sua vez, a Figura 11 mostra as trajetórias dos parâmetros LSP ao longo do tempo.

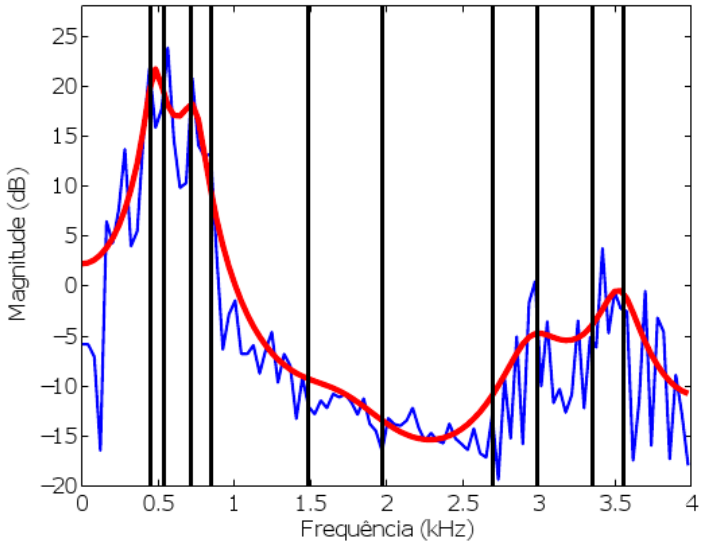

Figura 10: Parametrização do sinal acústico, em que a envoltória espectral é a resposta em frequência do filtro LPC. Os parâmetros LSP são representados pelas linhas verticais. Os pares dos parâmetros LSP são usados para representar cada quadro acusticamente. Fonte: [7]

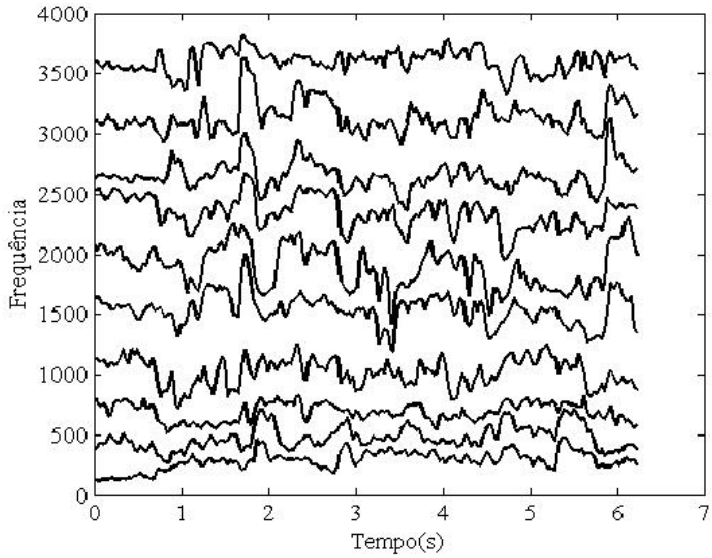

Figura 11: Exemplo das trajetórias dos dez parâmetros LSP ao longo do tempo. Estes parâmetros LSP representam o sinal acústico em cada quadro. Fonte: [7]

2.7 Análise das Componentes Principais

Utilizando-se do método da PCA no grupo de informações, obtiveram-se novas matrizes tanto para os sinais de Fluxo Óptico e LSP, onde se poderia escolher o número de componentes principais. Analisando as Figuras 12 e 13 , que representam a variância por componente principal e a variância acumulada das PCA's, pode-se perceber que a primeira componente principal da Figura 12 representam cerca de 60% da variabilidade

total nas avaliações padronizadas, enquanto as 4 primeiras componentes principais da Figura 13 representam 90% da variabilidade total nas avaliações padronizadas. Na definição de padrões do Fluxo Óptico foi escolhida somente a primeira componente principal por uma questão de redução de volume de dados. Para ambos os casos (Fluxo Óptico e LSP), a Análise das Componentes Principais é um caminho matemático para reduzir as dimensões, a fim de visualizar os dados.

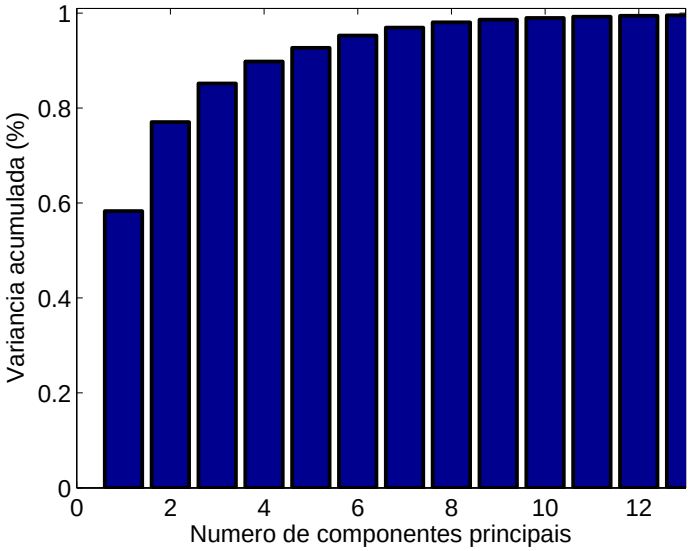

Figura 12: PCA dos vetores de Fluxo Óptico.

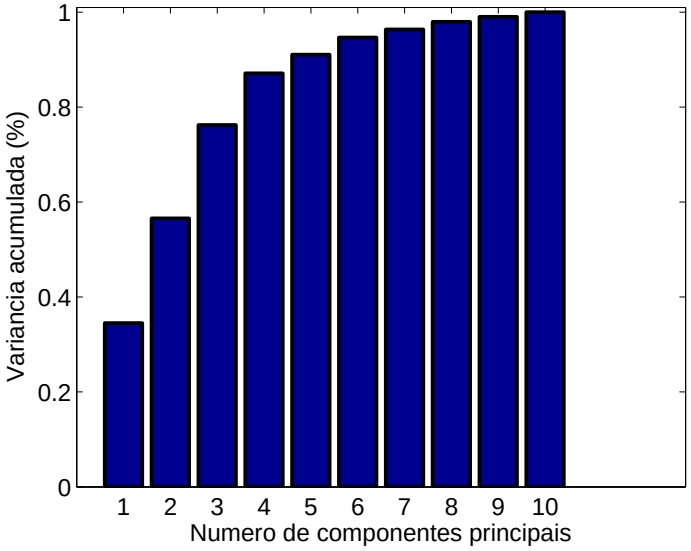

Figura 13: PCA dos parâmetros LSP.

2.8 Redes Neurais Artificiais

No processo de classificação, adotou-se uma composição de rede simples, *feedforward network*, cuja função de formação foi baseada na otimização de *Levenberg-Marquardt*, no qual se fez sucessivas simulações, adotando uma camada neural escondida com 5 neurônios, cuja função transferência adotada foi a hiperbólica sigmóide, seguida de uma camada de saída que foi ativada por uma função de transferência linear [25]. A entrada da rede foram as PCA's do Fluxo Óptico e a saída as PCA's dos parâmetros LSP. As PCA's de Fluxo Óptico e LSP de entrada do treinamento foram calculadas nos vídeos onde o voluntário foi submetido a ruídos, ou seja, nos vídeos onde houve a presença do *Efeito Lombard*. Já os dados utilizados na validação foram calculados nos vídeos que não contaram com a presença de ruídos, ou seja, sem a presença do *Efeito Lombard*. Isto porque espera-se que o *Efeito Lombard* realce tanto o áudio quanto o vídeo, tornando assim o treinamento mais eficiente. O acerto da rede foi calculado utilizando correlação.

Os parâmetros LSP de áudios com e sem *Efeito Lombard* para $p = 10$ seguem nas Figuras 14 e 15

O coeficiente de correlação $\rho_{X,Y}$ entre duas variáveis aleatórias X e Y com valores esperados μ_X e μ_Y e desvios padrão s_X e s_Y é definida como:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2.16}$$

Onde *cov* é a covariância que mede a dependência linear entre duas variáveis aleatórias. O valor da correlação fica entre 0 e 1, de modo que pode-se, assim, calcular a correlação entre a saída da rede treinada e o valor esperado.

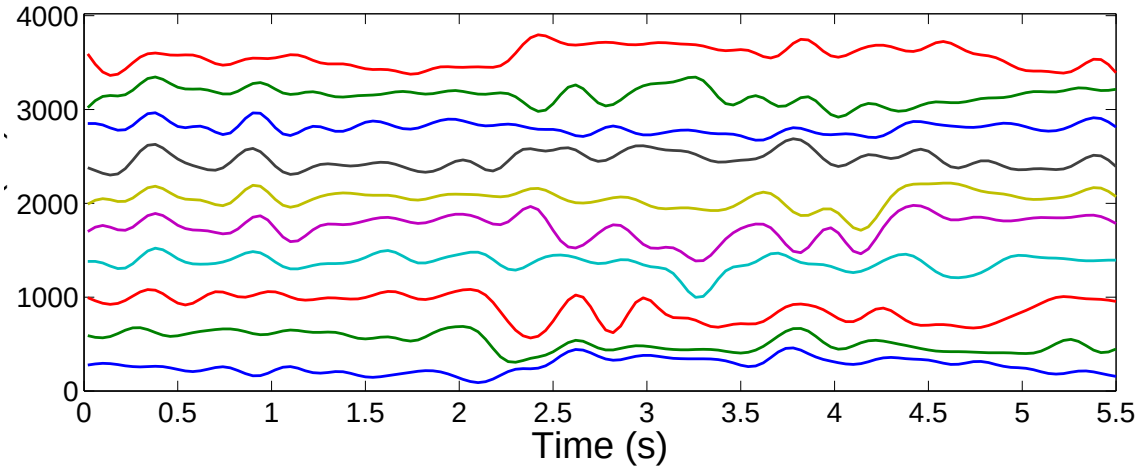

Figura 14: Parâmetros LSP com *Efeito Lombard*.

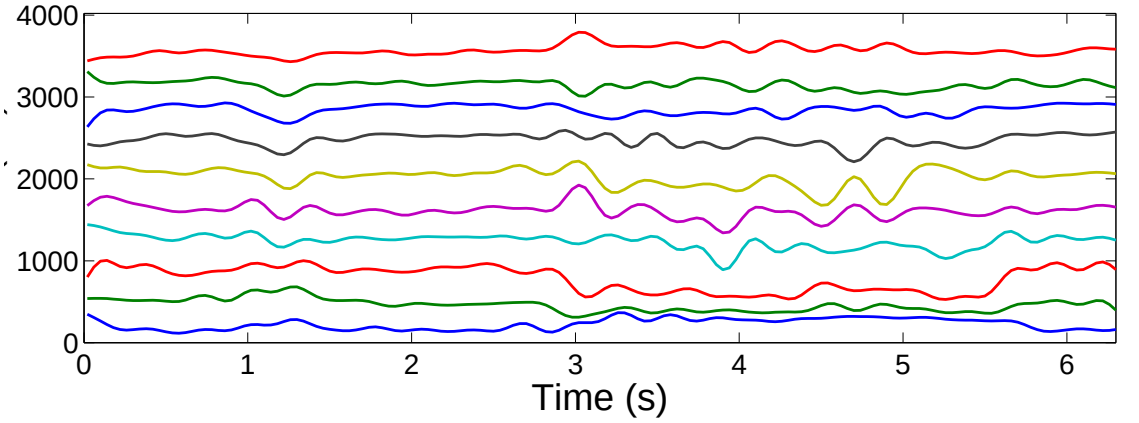

Figura 15: Parâmetros LSP sem *Efeito Lombard*.

3 Resultados e Discussões

3.1 Estudo Qualitativo e Computação do Fluxo Óptico no Movimento Facial

Nesta seção serão analisdos os resultados da computação do Fluxo Óptico, conforme descrito na seção 2.4.

A região de interesse do vídeo a ser processado é escolhia pelo usuário através de uma guia conforme a Figura 16

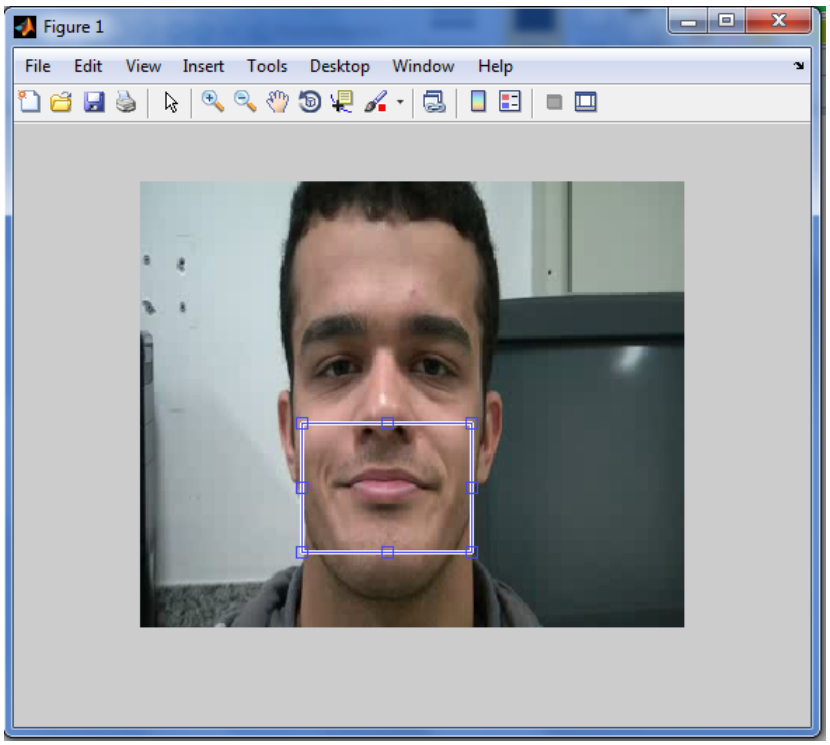

Figura 16: Interface com o Usuário

O estudo prévio do Fluxo Óptico, em termos qualitativos, do movimento da boca demonstra o quanto a região que vai das narinas à parte mediana do queixo se movimenta na horizontal, na vertical e em módulo. Nesta seção, serão analizados três tipos de

movimentos de articulação facial, mais especificamente na região da boca, denominados:

- Movimento de abertura Normal da boca: O voluntário movimentou a boca com sucessivas aberturas e fechamentos, simulando um ambiente de diálogo onde não há ruídos, ou seja, não há a necessidade de articular a musculatura facial como um todo demasiadamente. Este movimento simula a fala sem a presença do *Efeito Lombard*.

- Movimento de abertura Horizontal da boca: Trata-se de um movimento de sorriso, no qual as extremidades da boca incidem perpendicularmente em relação à bochecha. Este movimento também foi realizado repetidamente.

- Movimento de abertura Total da boca: Trata-se de um movimento de bocejo, no qual o voluntário abriu a boca, articulando o maxilar, o máximo que conseguiu.Este movimento simula a fala com a presença do *Efeito Lombard*.

As Figuras 17, 18 e 19 estão divididas em (a) e (b). A parte (a) mostra o comportamento do movimento da região de interesse (boca e bochecha) na horizontal, vertical e em módulo. A parte (b) mostra um zoom no primeiro ciclo dos movimentos repetitivos, onde só aparecem o comportamento do movimento horizontal e vertical.

O gráfico da Figura 17 representa o movimento baseado em Fluxo Óptico do movimento de abertura normal da boca, ou seja, sem a influência do *Efeito Lombard*. A Figura 18 representa o fluxo óptico para uma movimentação horizontal da boca. Esta movimentação horizontal ocorre com a boca fechada e suas laterais pressionando perpendicularmente a bochecha. A Figura 19 mostra o fluxo óptico em um movimento de abertura total, em que o voluntário movimentou a boca de modo a ter a maior articulação possível. As Figuras 20, 21 e 22 mostram o experimento propriamente dito.

Como é possível visualizar, tanto na Figura 17 quanto nas Figuras 18 e 19 há uma grande predominância no movimento vertical. Na Figura 17, o movimento vertical começa negativo, o que implica que o vídeo começou com a boca do voluntário fechada, o que faz com que o gradiente de deslocamento seja negativo, uma vez que a movimentação do maxilar é para baixo. Na Figura 18, existe maior movimento vertical. Mesmo que o experimento tenha sido feito com a movimentação horizontal da boca, é importante observar que neste movimento a musculatura da bochecha se move para cima e para baixo. Na Figura 19, o movimento vertical inicia positivo, pois a boca do voluntário está fechada no início da filmagem. Estes aspectos demonstram a importância de se conhecer os parâmetros de referência da filmagem ao se trabalhar com Fluxo Óptico.

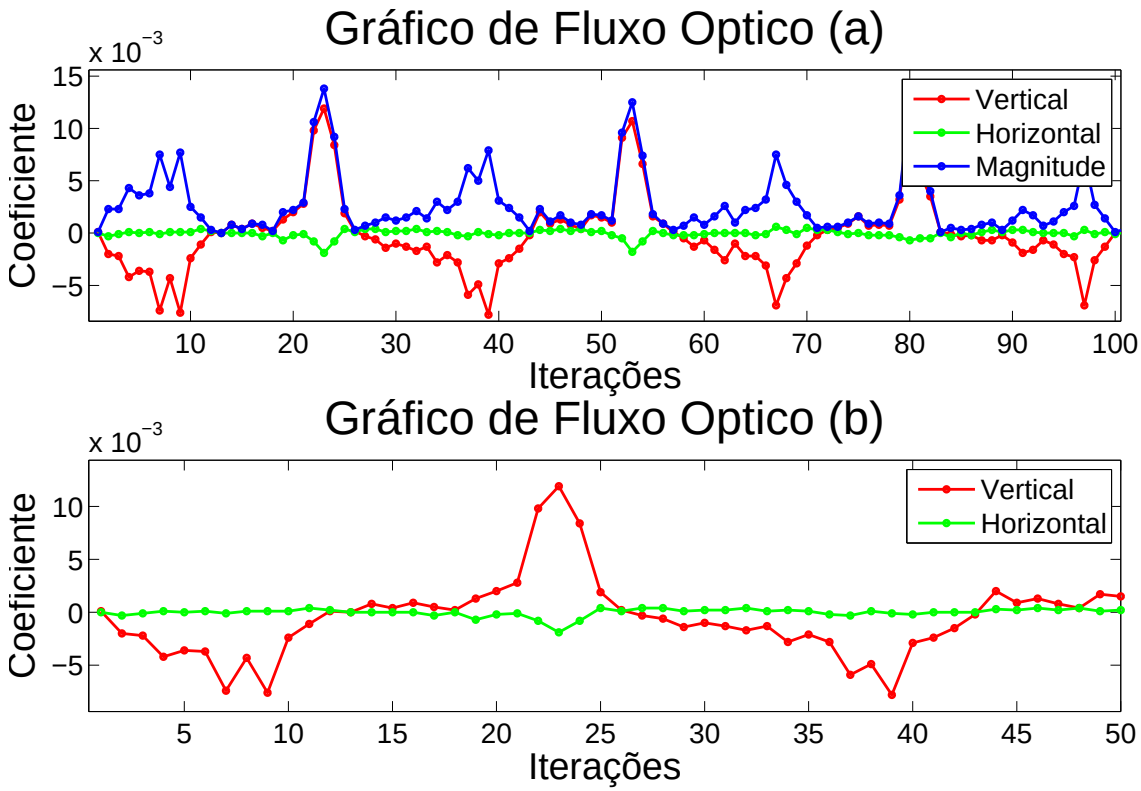

Figura 17: Gráfico do Movimento Normal da boca

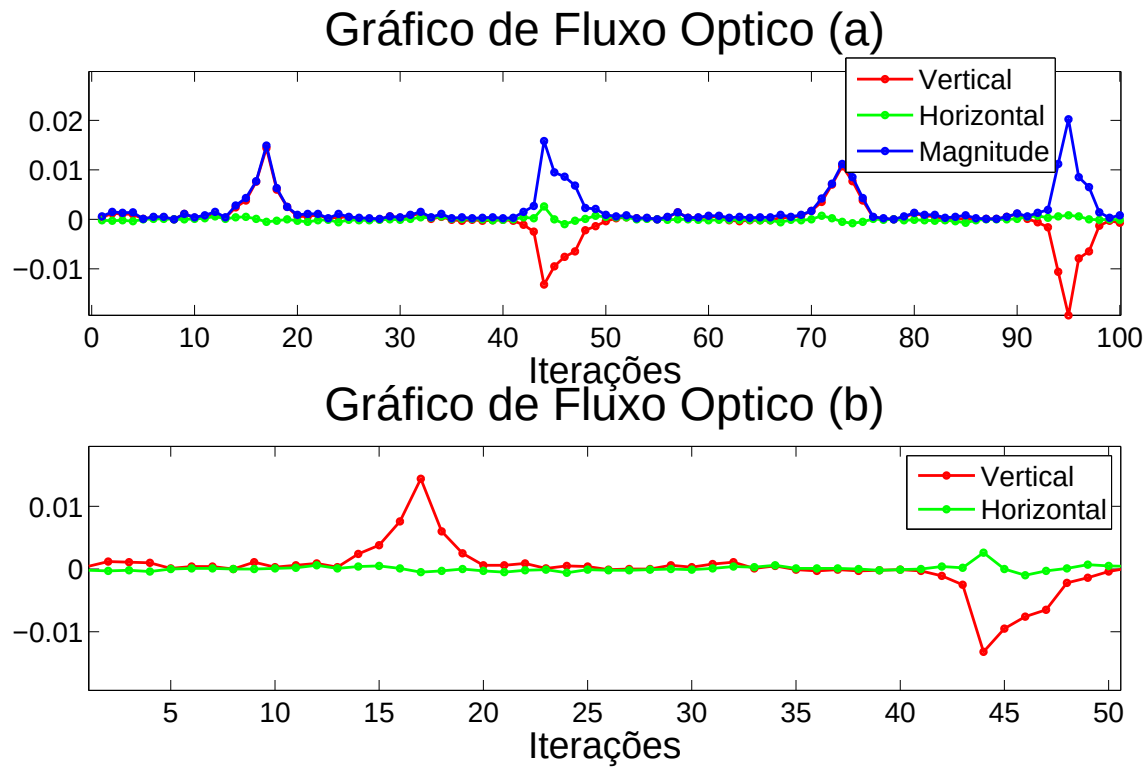

Figura 18: Gráfico do Movimento Horizontal da boca

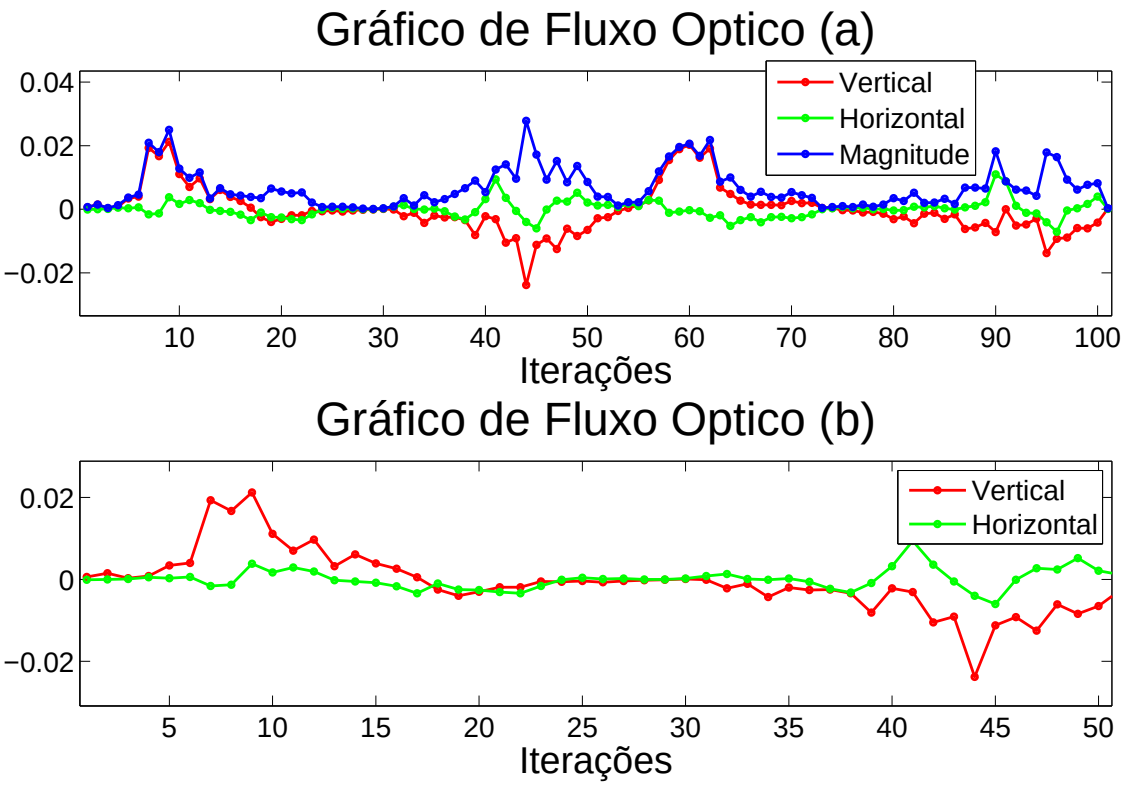

Figura 19: Gráfico do Movimento de Abertura Total da boca

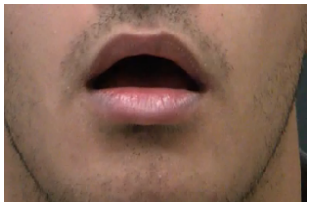

Figura 20: Movimento Normal da boca

Figura 21: Movimento Hprizontal da boca

Figura 22: Movimento Total da boca

Além disso, a diferença entre o movimento de abertura total e os demais é quase o dobro, dentro dos valores mensurados no eixo y dos gráficos. Tal resultado demonstra que há diferença na fala visível com e sem o $EfeitoLombard$.

3.2 Estudo da Fala Audível e Visível

No estudo da influência do *Efeito Lombard* sobre a fala é imprescindível a busca por relações positivas e negativas entre o audível e o visível. No campo da fala visível os resultados obtidos não geram muitas surpresas quanto ao esperado. Por isso há na fala audível um volume maior de informações a serem processadas para se poder chegar ao ponto de criar suposições, ou até mesmo afirmações.

Existem algumas características inerentes à um diálogo que não necessitam de argumentos matemáticos para saber que existem. Uma vez que existam, formulações matemáticas nos auxiliam a entender o comportamento de tais características. As perguntas feitas ao interlocutor podem ser divididas em dois grupos. Grupo 1: Resposta invariante. Grupo 2: Resposta Variante.

- Grupo 01:
  - Pergunta 01 (P1): Qual o nome da sua mãe?
  - Pergunta 02 (P2):Qual a sua data de nascimento?
- Grupo 02;
  - Pergunta 03 (P3):O que gosta de fazer?
  - Pergunta 04 (P4):O que fez hoje?
  - Pergunta 05 (P5):Quais desenhos gostava quando criança?

Características observadas ao longo dos experimentos:

- Existe um *delay* entre o interlocutor processar a pergunta e respondê-la;
- Algumas vezes existe perda e retomada de resposta para as perguntas do Grupo 02;

3.2.0.1 Amplitude do Sinal no Tempo

São apresentados os resultados da amplitude do sinal no tempo para alguns dos experimentos realizados.

Na Tabela 1, pode-se observar os valores RMS da amplitude de todos os experimentos realizados. Para o cálculo do valor RMS definiu-se um limiar de amplitude de $0,02$, ou seja, os valores de amplitude abaixo deste limiar foram considerados região de silêncio. Comparando a Tabela 1 com seus equivalentes das Figuras 23 a 32, nota-se que quando submetido a ruídos, os interlocutores tendem a emitir sinais sonoros mais intensos. Isto significa que o *Efeito Lombard* está ligado tanto à articulação dos movimentos da musculatura facial na fala visível, quanto ao comportamento do trato vocal na fala audível.

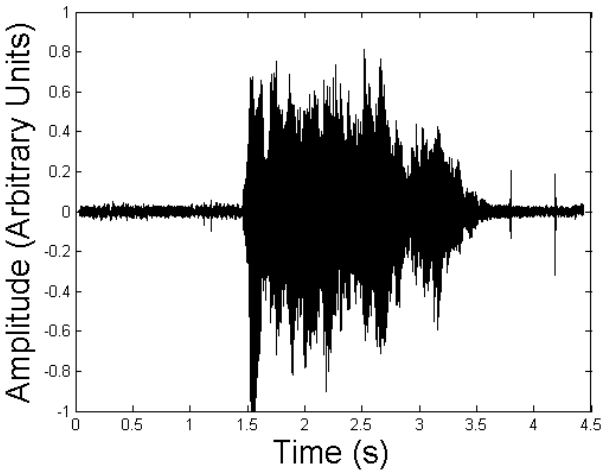

Figura 23: Amplitude P1 - Muito Alto

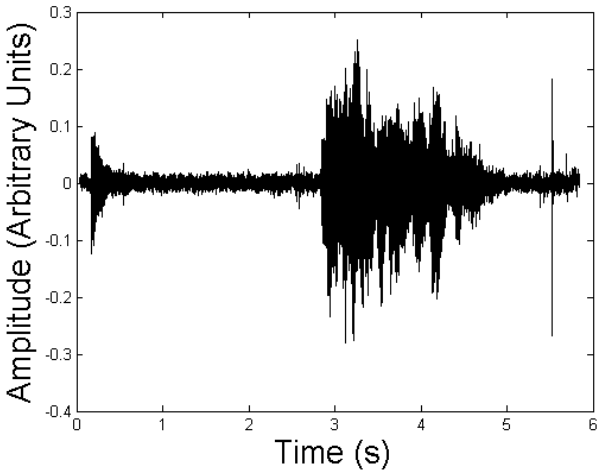

Figura 24: Amplitude P1 - Normal

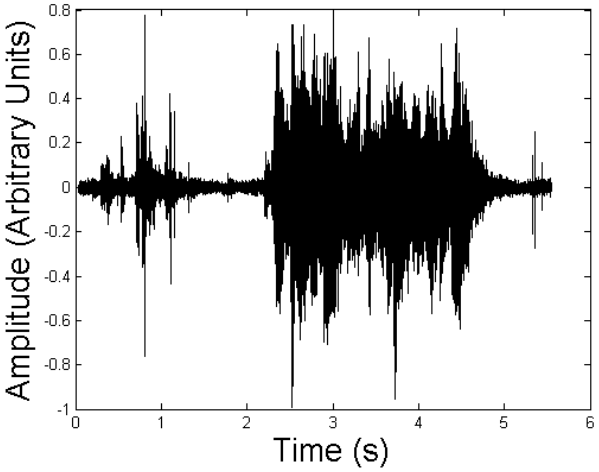

Figura 25: Amplitude P2 - Muito Alto

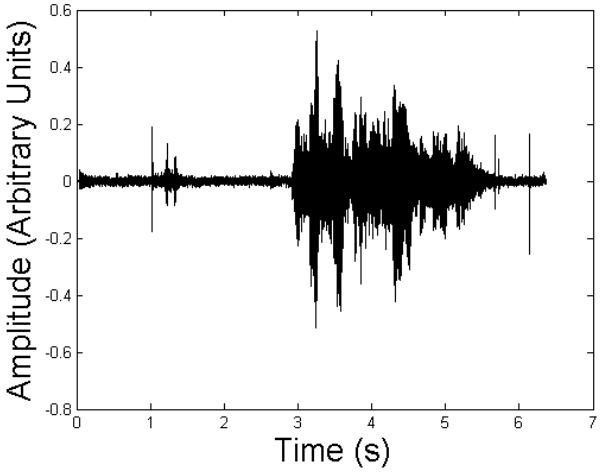

Figura 26: Amplitude P2 - Normal

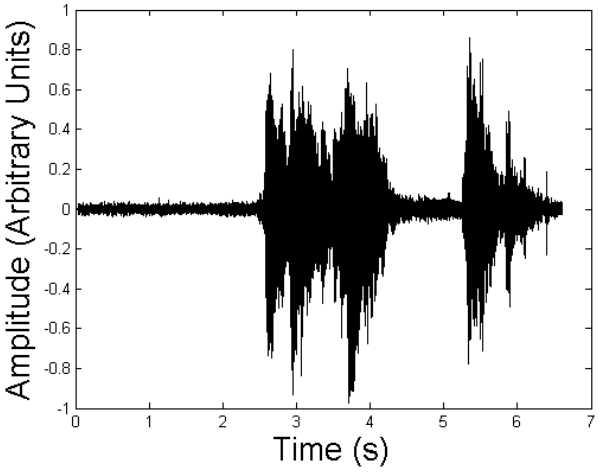

Figura 27: Amplitude P3 - Muito Alto

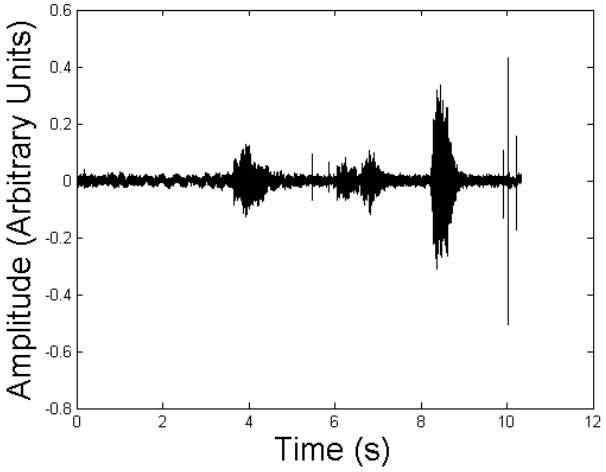

Figura 28: Amplitude P3 - Normal

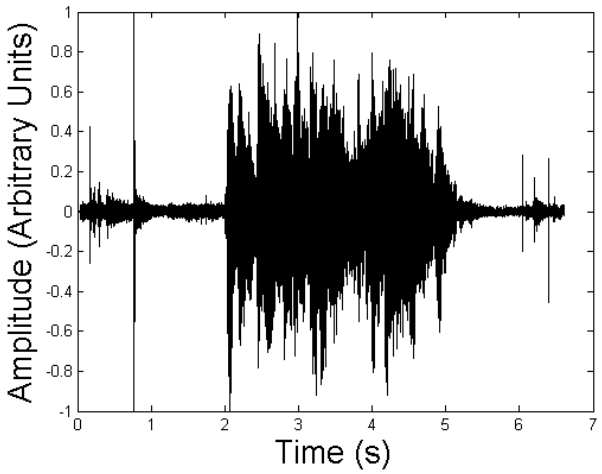

Figura 29: Amplitude P4 - Muito Alto

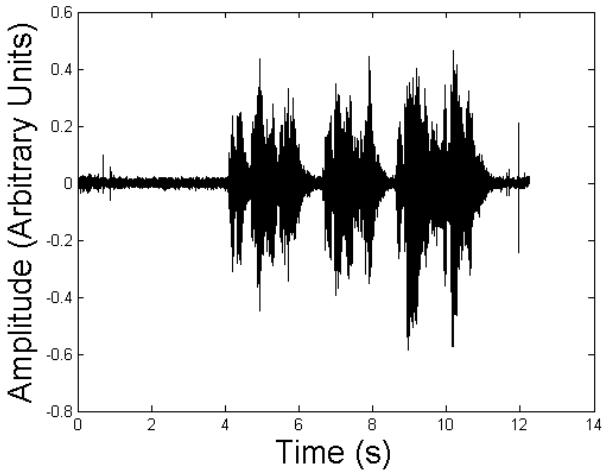

Figura 30: Amplitude P4 - Normal

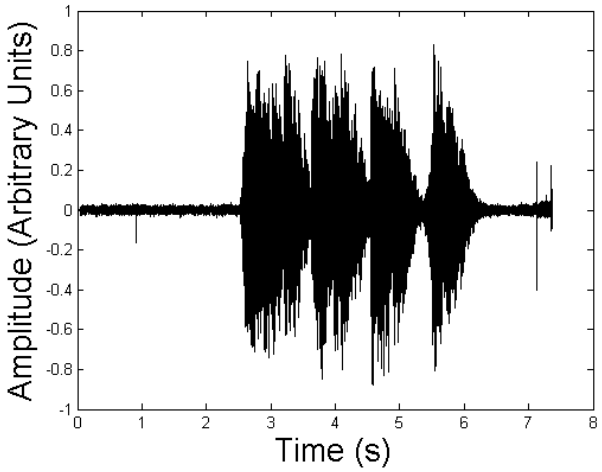

Figura 31: Amplitude P5 - Muito Alto

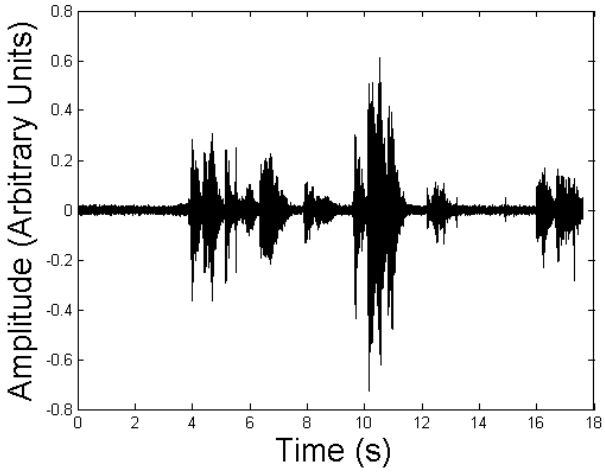

Figura 32: Amplitude P5 - Normal

3.2.1 Definição de Padrões e Classificação através de Redes Neurais

Para a classificação através de Redes Neurais, foram utilizadas as respostas das perguntas do Grupo 01, sendo que 70% dos dados foram utilizados para treinamento e 30% para validação. Os dados de entrada foram as PCA's do Fluxo óptico e os de saída as PCA's dos parâmetros LSP. O acerto obtido nesta rede foi baixo, apenas $37,75\%$.

Como já citado na Seção 2, foi utilizada apenas uma PCA de cada vetor de fluxo óptico. Esse vetor possuía uma variância acumulada de cerca de 60%. Além de uma variância baixa, este treinamento mostrou um desempenho muito baixo pois na entrada da rede havia um elevado número de pontos para cada vetor de entrada (cerca de 2100).

Tabela 1: Valores RMS da Amplitude dos sinais de áudio referentes aos 5 ruídos diferentes dos experimentos

Pergunta	Ruído	Valor RMS da Amplitude
P1	Alto	$0,1558$
P1	Baixo	$0,1432$
P1	Médio	$0,1475$
P1	Muito Alto	$0,1541$
P1	Normal	$0,1032$
P2	Alto	$0,1653$
P2	Baixo	$0,1693$
P2	Médio	$0,1613$
P2	Muito Alto	$0,1916$
P2	Normal	$0,0911$
P3	Alto	$0,1603$
P3	Baixo	$0,1659$
P3	Médio	$0,1483$
P3	Muito Alto	$0,1791$
P3	Normal	$0,1025$
P4	Alto	$0,1609$
P4	Baixo	$0,1783$
P4	Médio	$0,1483$
P4	Muito Alto	$0,1646$
P4	Normal	$0,0623$
P5	Alto	$0,1665$
P5	Baixo	$0,1436$
P5	Médio	$0,1284$
P5	Muito Alto	$0,1878$
P5	Normal	$0,0655$

4 Considerações Finais

4.1 Análise do Fluxo Óptico

Neste trabalho, pôde-se concluir que se as restrições e suposições para o método de Horn & Shunk forem rigorosamente satisfeitas, pode-se extrair o Fluxo Ótico de um vídeo ou de uma sequência de imagens de maneira satisfatória.

A utilização dos coeficientes baseados no gradiente de deslocamento da região de interesse pode trazer várias informações sobre a filmagem, podendo ser utilizada para aplicação em análises de interesse.

Na área de interesse escolhida para serem feitos os testes do algoritmo deste trabalho, não foram observadas variações de movimentos significativas em frases curtas, mas vale ressaltar que este mesmo trabalho pode ser utilizado novamente atacando-se estudos estatísticos e métodos de realização de experimentos mais complexos como, por exemplo, utilizar uma câmera atrás do interlocutor para fazer o cálculo de compensação do movimento da cabeça. Além disso, o programa pode ser também utilizado para outras áreas de interesse do vídeo. Em trabalhos que venham a tomar esse como base, também se sugere o uso de técnicas de janelamentos mais variadas, a fim de se colher maior número de informações das sequências de imagens.

4.2 Aquisição de Dados

O banco de dados utilizado, neste trabalho, contou apenas com um voluntário. Mesmo assim foi possível construir uma base sólida de trabalho criando passos a serem seguidos para a continuidade deste projeto. Sugere-se que a partir de agora, a quem for dar continuidade ao trabalho, que além de refinar os algoritmos e códigos, faça também experimentos com mais voluntários. Vale salientar que neste trabalho não foi feita a compensação do movimento da cabeça, o que pode melhorar ainda mais os resultados.

4.3 Codificação audiovisual da fala

Já existem métodos capazes de converter vídeo em áudio, mas é necessário a continuidade de novos estudos a fim de aperfeiçoar esta área. A utilização de Fluxo Óptico nesta codificação, abre uma gama de possibilidades de estudos e pesquisa na área. Neste trabalho, mostrou-se existir uma relação entre a fala audível e a fala visível com e sem *Efeito Lombard*, e que isto pode ser objeto de estudos mais aprofundados, uma vez que o *Efeito Lombard* de certo modo reforça a fala audiovisual.

4.4 Redes Neurais Artificiais

A aplicação do método de Redes Neurais Artificiais pode ser comparado ao aprendizado humano. Deste modo, para um funcionamento eficaz, ou seja, com elevado percentual de acerto, é necessário uma grande base de dados e um pré-processamento de dados que possa comprimir a base de dados comprometendo o mínimo possível sua qualidade. A modelagem matemática para compressão sem perda de qualidade deve ser um dos pontos principais a ser atacado na continuidade deste trabalho.

4.5 Comentário Final

Este trabalhou criou uma metodologia que se seguida e aperfeiçoada pode trazer grandes contribuições científicas na área da Visão Computacional. Espera-se que na continuidade deste projeto hajam novas vertentes que possibilitem melhorias e novas descobertas.

Referências

[1] FARIA, A. W. C. Fluxo Óptico. *ICEx-DCC- Visão Computacional*, 2007.

[2] JEPSON A.;BLACK, M. *Mixture Models for Optical Flow Comparation*. [S.l.].

[3] QUEDAS A. ; DUPRAT, A. C. G. G. Implicações do efeito lombard sobre a intensidade, freqüência fundamental e estabilidade da voz de indivíduos com doença de parkinson. *Revista Brasileira Otorrinolaringologia online*, v. 73, p. 675–683, Sep.-Oct 2007.

[4] ERBER, N. P. Auditory visual perception of speech. *Journal Speech and Hearing Disorders*, v. 40, p. 481–492, November 1975.

[5] VATIKIOTIS-BATESON, E. et al. The dynamics of audiovisual behavior in speech. *Speechreading by humans and machines (NATO-ASI Series F)*, D. Stork & M. Hennecke, v. 150, p. 221–232, 1996.

[6] VATIKIOTIS-BATESON, E. et al. Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, v. 60, n. 6, p. 926–940, 1998.

[7] MOREIRA, K. S. *Um Estudo sobre as Relações de Padrões do Movimento Facial com a Acústica da Fala e com a Identidade do Locutor*. 2008.

[8] BLANZ, V. et al. Reanimating faces in images and video. *Proceedings of EUROGRAPHICS 2003*, 2003.

[9] YEHIA, H. C.; RUBIN, P.; VATIKIOTIS-BATESON, E. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, v. 26, p. 23–43, October 1998.

[10] VATIKIOTIS-BATESON E.; BARBOSA, A. V. C. C. Y. O. M. T. J.; YEHIA, H. C. Audiovisual lombard speech: Reconciling production and perception. *AVSP*, p. 45–50, August-September 2007.

[11] MOURA A.; PÊRA, V. F. D. *Leitura Labial automática em sistemas de reconhecimento automático da fala: Desenvolvimento de um sistema para o Português Europeu*. [S.l.], 2005.

[12] HORN, B. K. P. B. G. Dertemining optical flow. *A. I. Memo, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA-USA*, n. 572, Aug.-Oct 1980.

[13] MORE, B. C. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions On Automatic Control*, Ac-26, n. 1, p. 17–31, 1981.

[14] FERREIRA, D. F. *Análise Multivariada*. [S.l.]: Lavras - MG, 1996.

[15] ANTOSZCZYSZYN P. M.; HANNAH, J. M. G. P. M. Tracking of the motion of important facial features in model-based coding. *Elsevier Signal Processing Nº66*, p. 249–260, 1998.

[16] HAYKIN, S. *Redes Neurais, Princípios e Prática*. [S.l.]: Bookman, 2ª Edição, 2001.

[17] FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. *Ciência da Informação, Brasília*, v. 35, p. 25–30, 2006.

[18] BARBOSA, A. V. *Um Estudo Sobre Relações Entre as Falas Audível e Visível (a Study On The Relations Between Audible And Visible Speech)*. 2005.

[19] BARBOSA, A. V. *Codificação Audio-Visual Integrada da Fala*. Dissertação (Mestrado) — CPDEE - UFMG, 2000.

[20] KURATATE T.; MUNHALL, K. G. R. P. E. V.-B. E. Y. H. C. Audio-visual synthesis of talking faces from speech production correlates. *Proc. 6th European Conference on Speech Communication and Technology (EuroSpeech'99)*, v. 3, p. 1279–1282, September 1999.

[21] SUGAMURA, N.; ITAKURA, F. Speech analysis and synthesis methods developed at ecl in ntt-from lpc to lsp. *Speech Commun.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 5, n. 2, p. 199–215, 1986.

[22] FLANAGAN, J. L. *Speech analysis, synthesis, and perception*. Third edition. [S.l.]: Springer–Verlag, 1972.

[23] ATAL, B. S.; HANAUER, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, April 1971.

[24] SHAH, M. *Fundamental of Computer Vision*. [S.l.]: University of Central Florida, Florida-USA, 1997.

[25] DOURADO D. M.; FELIX, L. B. *Redes Neurais Artificiais Para Classificação da atenção Seletiva Auditiva: Aplicação em interface Cérebro-Computador*. [S.l.], 2013.