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

LUCAS GRACIANO CARDOSO

ESTIMAÇÃO DA ACÚSTICA DA FALA ATRAVÉS DO FLUXO ÓPTICO DO MOVIMENTO FACIAL

VIÇOSA
2014

LUCAS GRACIANO CARDOSO

ESTIMAÇÃO DA ACÚSTICA DA FALA ATRAVÉS DO FLUXO ÓPTICO DO MOVIMENTO FACIAL

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.
Orientador: Profª. Drª. Ketia Soares Moreira

VIÇOSA
2014

LUCAS GRACIANO CARDOSO

ESTIMAÇÃO DA ACÚSTICA DA FALA ATRAVÉS DO FLUXO ÓPTICO DO MOVIMENTO FACIAL

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 DD de MMMM de 20YY.

COMISSÃO EXAMINADORA

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

Eng. Eletricista Gerson Ovidio Luz Pedruzi - Membro
Universidade Federal de Viçosa

Eng. Eletricista Michael de Oliveira Resende - Membro
Universidade Federal de Viçosa

‘‘Porque aos seus anjos dará ordens a teu respeito, para que te guardem em todos os teus caminhos.’’

(Salmos 91:11)

Dedico este trabalho primeiramente a Deus, por tudo que tem feito e por tudo que vai fazer em minha vida.

Dedico também aos meus amados pais Flavio Graciano e Alvina Cardoso, que sempre foram meus exemplos de vida e de onde busco forças para continuar.

Agradecimentos

Agradeço a Deus por ter me dado sabedoria em minhas escolhas e paz nas horas difíceis.

Agradeço a minha mãe por todo amor, carinho e companheirismo sempre incondicionais durante todos esses anos de caminhada.

Agradeço ao meu pai pelo exemplo de perseverança que me ajudou e me inspirou em todos os momentos, e aos conselhos, mesmo que simples, serviam de base para minhas tomadas de decisão.

Agradeço a todos os meus familiares, avós, tios e primos por sempre estarem dispostos a me ajudar não importasse qual fosse a situação, e sempre mativeram suas portas abertas nos meus dias de férias.

Agradeço aos meus amigos e a minha namorada pelo acolhimento, pelas muitas conversas as quais não faltavam palavras de ânimo e de fé que no final daria tudo certo.

Agradeço ao CNPq pelo apoio dado no desenvolvimento de minha pesquisa.

Agradeço a minha professorea e orientadora Kétia por ter me mostrado os caminhos a seguir durante cada passo do desenvolvimento deste trabalho, e também a todos os professores do departamento que me ajudaram.

Resumo

O estudo da relação existente entre o movimento facial e a acústica da fala é importante para a compreensão do processo de produção da mesma. O objetivo deste trabalho é estudar os movimentos faciais em ambientes com e sem ruído, ou seja, com e sem a presença do Efeito Lombard, e sua relação com a acústica da fala através de estimadores lineares. Na representação do movimento facial, a extração dos vetores de velocidade deste movimento é feita por meio do fluxo óptico através do método iterativo de Horn e Schunck. O fluxo óptico é obtido respeitando com rigor as restrições advindas do método utilizado, de modo a se obter os melhores vetores de movimento possíveis, isto é, vetores de velocidade que indiquem mais fielmente o movimento da face durante a fala e com o mínimo de erros. De posse dos vetores de movimento é observado que as cinco primeiras componentes principais extraídas após o uso do método iterativo representam cerca de 90% da variância observada nos dados do movimento facial e que são suficientes para representar o fluxo óptico do movimento facial. Para a acústica da fala são calculados os parâmetros LSP, que estão relacionados com a geometria do trato vocal. Técnicas de estimação são aplicadas a palavras isoladas retiradas de um discurso, onde locutores estão submetidos a diferentes tipos de ruídos. Os dados obtidos dos vídeos onde o locutor é submetido a um ruído alto foram utilizados para treinamento do modelo matemático pelo fato do mesmo movimentar mais a boca para proferir as palavras quando comparado com as outras situações de ruído. Com as cinco componentes principais são determinados os parâmetros LSP por meio dos estimadores modelados. Os resultados numéricos mostram que a predição das faixas de frequência dos parâmetros LSP feitas pelo modelo matemático obtiveram uma similaridade com o esperado que chegou a atingir 97% na faixa de frequências em torno de 2033 Hz, melhor caso; e 55% na faixa de frequências em torno de 3552 Hz, pior caso. Espera-se que tanto a metodologia utilizada neste trabalho quanto os resultados obtidos possam ajudar em trabalhos futuros e fornecer contribuições para a área de Visão Computacional.

Abstract

The study of the relationship between facial motion and speech acoustics is important for understanding the process of producing the same. The objective of this work is to study facial movements in environments with and without noise, ie, with and without the presence of the Lombard Effect, and its relation to speech acoustics using linear estimators. In the representation of facial motion, the extraction of velocity vectors of this movement is made through optical flow using the iterative method of Horn and Schunck. The optical flow is obtained respecting rigorously the constraints resulting from the method used to obtain the best possible motion vectors, i.e. vectors indicating the velocity of movement of the face during speech more accurately and with minimal errors. With the motion vectors is seen that the first five principal components extracted after the use of iterative method represent about 90% of the variance observed in the data of facial movement and are sufficient to represent the optical flow of facial movement. For speech acoustics are computed LSP parameters, which are related to the geometry of the vocal tract. Estimation techniques are applied to isolated words taken from a speech where speakers are subjected to different types of noise. The data of the videos where the speaker is subjected to a loud noise were used for training of the mathematical model because the same move over his mouth to utter the words when compared with the other noise situations. With the first five principal components are determined the LSP parameters through the modeled estimators. Numerical results show that the prediction of the LSP parameters made by the mathematical model obtained a similarity ranges with the expected which reached 97% in the frequency range around 2033 Hz, best case; and 55% in the frequency range around 3552 Hz, worst case. It is expected that both the methodology used in this study as the results obtained may help in future work and provide contributions to the field of computer vision.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Os avanços computacionais vêm, ao longo dos anos, provindo melhoras na vida das pessoas que fazem uso dos mesmos. Estes avanços tornaram o mundo pequeno do ponto de vista da comunicação, fazendo com que todos, mesmo que longe, possam estar de alguma forma conectados. A fim de que continue havendo avanços tecnológicos, várias pesquisas são feitas nas mais diversas áreas e, consequentemente, novas ideias e caminhos para significativas melhoras no que já existe vão surgindo.

Segundo **[1]** e **[2]** na área dos estudos dos movimentos, várias pesquisas comprovam que é possível correlacionar e até mesmo estimar a fala, ou os parâmetros desta, através da computação do movimento facial do indivíduo e vice-versa. Com um OPTOTRAK, **[2]** obteve os movimentos da face, e da língua do indivíduo fazendo assim a sua relação com a acústica da fala.

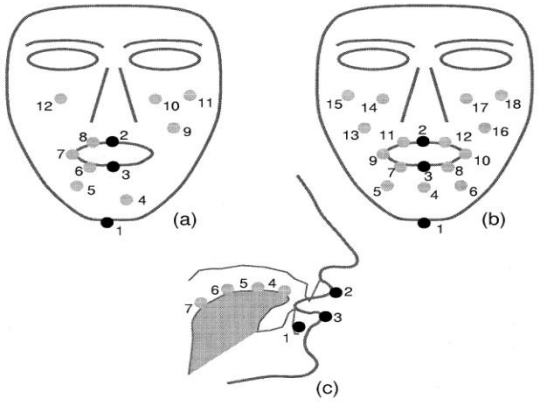

Figura 1. Posição dos marcadores OPTOTRAK. **[2]**

Mais recentemente, estudos de **[3]** constataram as diferenças existentes nas informações visuais quando indivíduos eram submetidos a diferentes níveis de ruído. O efeito natural de movimentar a face ao produzir a acústica da fala é enfatizado para locutores submetidos a ambientes com ruídos, o que altera a correlação existente entre os parâmetros visuais e acústicos da fala.

Uma ferramenta para fazer o estudo do movimento facial é o fluxo óptico, que é uma técnica utilizada para estimar movimento em sequências de imagens. Fluxo óptico é a distribuição da velocidade aparente do movimento dos padrões de intensidade em uma imagem. Em **[4]** foi feito uso deste algoritimo para observar quantitativamente a amplitude do movimento facial em função do ruído ao qual o testado estava submetido.

Em paralelo com o estudo do movimento facial, é feito o estudo da acústica da fala. O sinal acústico pode ser obtido através de um microfone e de um processo de conversão A/D (analógico-digital). Após a conversão o sinal é reamostrado a uma taxa apropriada para sua análise. Neste trabalho seguem os passos descritos por **[1]**, **[2]**, **[3]** e **[5]**, que o sinal acústico é transformado em parâmetros LSP, que são fortemente ligados a geometria do trato vocal.

Aguirre, **[6]** refere à identificação de sistemas como sendo uma área do conhecimento que estuda como se modelar e analisar sistemas a partir de observações, ou seja, dados. Um modelo matemático de um sistema é um análogo a tal sistema. Há várias formas em que as equações que descrevem o comportamento de um sistema podem ser escritas. Será chamada de representação a forma em que um modelo matemático é expresso **[6]**. Em algumas situações aproximações lineares são suficientes para aplicações práticas, entretanto, numa série de aplicações modelos lineares não são satisfatórios. A escolha de modelos não lineares, entretanto, acarreta um inevitável aumento na complexidade dos algoritmos a serem utilizados **[6]**.

Barbosa, **[3]** através de marcadores na região da face do testado consegue extrair o movimento facial durante a fala e relacioná-lo com os parâmetros acústicos utilizando modelos lineares e não lineares de predição.

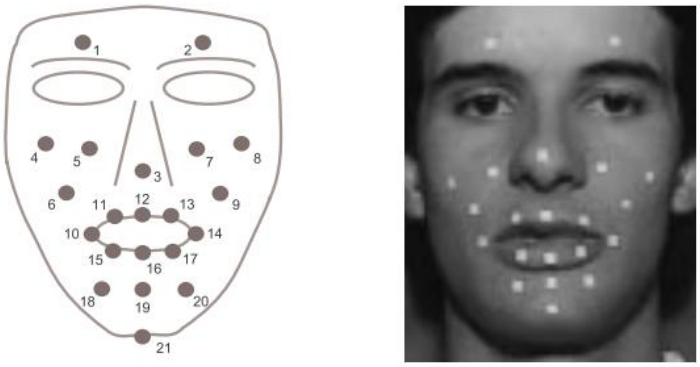

Figura 2. Marcadores pintados na face do locator utilizados para medição do movimento facial.

Neste trabalho, são obtidos dados do fluxo óptico relativos ao movimento da face de um voluntário que são comprimidos através da análise por componentes principais e, após são obtidos os parâmetros LSP relativos à fala. De posse destes dados é feita uma modelagem matemática de representação em tempo discreto, como ARX, ou no espaço de estados. para

que se estime os parâmetros LSP tendo em mãos somente dados modelados em componentes principais do fluxo óptico.

É importante salientar que, apesar das limitações, a correlação entre a parte visual da fala e a acústica da fala existe e é facilmente comprovada pelo fato de existirem pessoas que conseguem entender o que o outro fala somente pelo movimento de seus lábios (leitura labial). Em alguns momentos, a informação visual pode ser o único recurso para o reconhecimento da mensagem expressa, o que demonstra o interesse deste estudo, que pode contribuir no desenvolvimento de técnicas para a comunicação envolvendo pessoas com dificuldades na produção da fala ou na sua audição **[7]**.

1.1 Estimativa do Fluxo Óptico

Um problema fundamental no processamento de imagens em sequência é mensurar o fluxo óptico (ou velocidade da imagem). O objetivo é calcular uma aproximação para um campo 2-d de movimento, indicando a velocidade de cada ponto da superfície da imagem, a partir dos padrões espaço-temporais de intensidade da imagem **[8]**.

Os métodos para a computação do Fluxo Ótico podem ser classificados em três grupos principais: Técnicas diferenciais, Técnicas de correlação e Técnicas baseadas em Frequência de energia.

1.1.1 Técnicas Diferenciais, Método de Horn & Schunck

Técnicas diferenciais computam a velocidade através de derivadas espaço-temporais da intensidade da imagem ou versões filtradas das imagens (usando filtros passa-baixas ou passa-banda). Os primeiros casos utilizavam derivadas de primeira ordem e baseavam-se na translação das imagens **[8]**:

$$I(x,t) = I(x - vt, 0) \tag{1}$$

onde $v = (x,t)^T$. Expandindo a Equação 1 em serie de Taylor **[8]**:

$$\nabla \mathrm{I}(\mathrm{x},\mathrm{t}).\,\mathrm{v} + I_t(x,t) = 0 \tag{2}$$

onde $I_t(x,t)$ denota a derivada parcial em relação ao tempo de $I(x,t)$, $\nabla \mathrm{I}(\mathrm{x},\mathrm{t}) = (I_x(x,t), I_y(x,t))^T$, e $\nabla \mathrm{I}.\,\mathrm{v}$ denota o produto escalar. Com isso se obtem a componente

normal do movimento espacial dos contornos de intensidade constante, $v_n = sn$. A velocidade normal s e a direção normal n são dadas por **[8]**:

$$s(x,t) = \frac{-I_t(x,t)}{||\nabla I(x,t)||} \quad (3)$$

$$n(x,t) = \frac{\nabla I(x,t)}{||\nabla I(x,t)||} \quad (4)$$

Existem duas componentes desconhecidas de v na Equação 3, restringida por uma única equação linear. Outras restrições são, portanto, necessárias para encontrar os dois componentes de v.

1.1.1.1 Suposições e Restrições

Para evitar variações no brilho que ocorrem devido a efeitos de sombras, Horn & Shunck 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. Esta última condição assegura que o brilho da imagem é diferenciável **[9]**.

Horn & Schunck derivam a equação que retrata a mudança na iluminação em uma imagem a um ponto para o modelo de movimento da iluminação. Seja a iluminação de uma imagem no ponto (x, y), no plano da imagem no tempo t descrito por $E(x, y, t)$ **[9]**. Considerando-se o que acontece quando o modelo se move, a iluminação de um ponto particular no modelo é constante, então:

$$\frac{dE}{dt} = 0 \quad (5)$$

Expandindo a Equação 5 usando a regra da cadeia, e fazendo as substituições $u = \frac{dx}{dy}$ e $v = \frac{dy}{dt}$, chegamos à equação linear de variáveis u e v.

$$E_x u + E_y v + E_t = 0 \quad (6)$$

A Figura 1 ilustra a influência de uma fonte de iluminação na computação do Fluxo Óptico. Ao girar uma esfera com superfície de mesma intensidade mantendo sobre esta uma fonte de iluminação constante e fixa, o fluxo óptico será zero por não haver mudança de intensidade durante o giro. Porém, ao se mover a fonte de iluminação, a mudança de intensidade causada

por tal iluminação causará um campo de fluxo óptico aparente mesmo sem movimento da esfera.

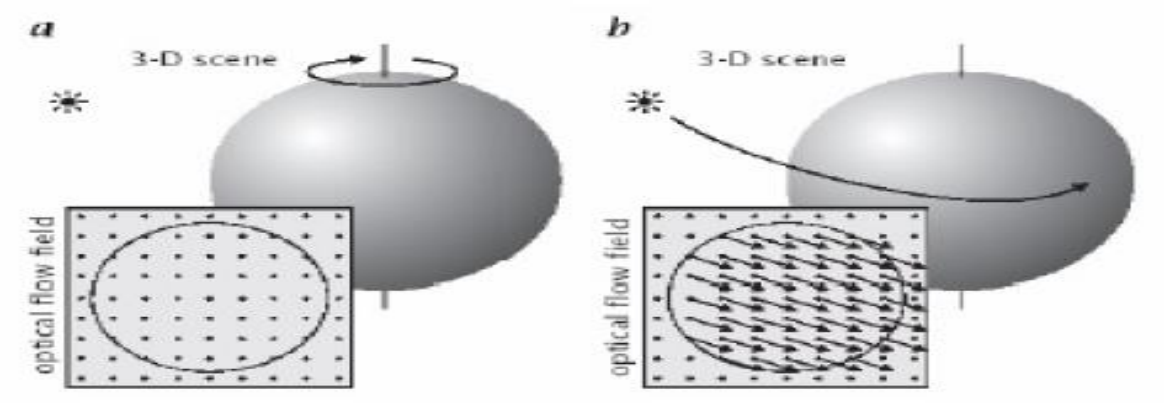

Figura 3 - Limitação Horn & Schunck: a) Um giro da esfera com a iluminação fixa determina um fluxo óptico=0. b) Um movimento da fonte de iluminação causa um campo de fluxo óptico aparente sem movimento da esfera **[1]**.

Outra restrição colocada na solução para Fluxo Ótico é a restrição de suavização, na qual pontos vizinhos de um objeto em movimento possuem velocidades similares, consequentemente o padrão de iluminação na imagem varia suavemente em quase toda parte. Um algoritmo baseado na restrição de suavização é provável que tenha dificuldades em estimações onde um objeto oculta o outro **[9]**.

$$\nabla^2 u = (\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 \; e \; \nabla^2 v = (\frac{\partial v}{\partial x})^2 + (\frac{\partial v}{\partial y})^2 \tag{7}$$

1.1.1.2 Estimação das Derivadas Parciais e Estimação do Laplaciano do Fluxo de Velocidade

É preciso estimar as derivadas da iluminação do conjunto discreto de imagens medidas disponíveis. É importante 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 **[9]**.

Para realizar esta estimação utiliza-se um ponto no centro de um cubo formado por oito medições cuja relação 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 do cubo.

$$E_x \approx \frac{1}{4}\left[E_{ii+1,jj+1,kk} - E_{ii+1,jj,kk} + E_{ii,jj+1,kk} - E_{ii,jj,kk} \ldots + E_{ii+1,jj+1,kk+1} - E_{ii+1,jj,kk+1} + E_{ii,,jj+1,kk+1} - E_{ii,jj,kk+1}\right] \quad (8)$$

$$E_y \approx \frac{1}{4}\left[E_{ii,jj,kk} - E_{ii+1,jj,kk} + E_{ii,jj+1,kk} - E_{ii,jj,kk} \ldots + E_{ii+1,jj+1,kk+1} - E_{ii+1,jj,kk+1} + E_{ii,,jj+1,kk+1} - E_{ii,jj,kk+1}\right] \quad (9)$$

$$E_z \approx \frac{1}{4}\left[E_{ii+1,jj,kk+1} - E_{ii+1,jj,kk} + E_{ii,jj+1,kk} - E_{ii,jj,kk\ldots.} + \quad E_{ii+1,jj+1,kk+1} - E_{ii+1,jj+1,kk} + E_{ii,,jj+1,kk+1} - E_{ii,jj+1,kk}\right] \quad (10)$$

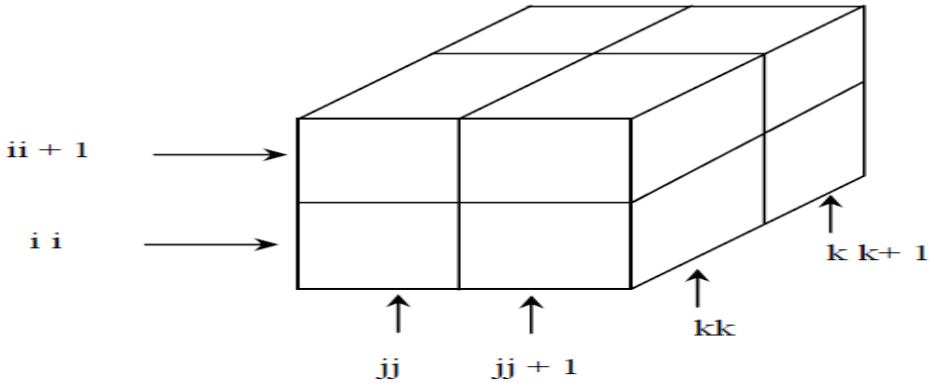

Figura 4 - As três derivadas parciais da iluminação da imagem ao centro do cubo são estimadas pela média das primeiras diferenças nas quatro bordas paralelas do cubo. A coluna índice j correspondente a direção x na imagem e a coluna índice i a direção y. Enquanto k representa a direção do tempo **[9]**.

É necessário aproximar o Laplaciano de u e v. Uma convincente aproximação é dada pela fórmula **[9]**.

$$\nabla^2 u \approx k\left(\bar{u}_{i,j,k} - u_{i,j,k}\right) e\ \nabla^2 v \approx k\left(\bar{v}_{i,j,k} - v_{i,j,k}\right) \quad (11)$$

Onde ū e v̄ são médias locais dos vetores de velocidade. Eles são estimados pela subtração do valor em um ponto a uma média ponderada dos valores vizinhos. Assim, as equações de $\bar{u}$ e $\bar{v}$ se tornam **[9]**.

$$\bar{u}_{i,j,k} \approx \frac{1}{6}\left[u_{ii-1,jj} + u_{ii,jj+1} + u_{ii+1,jj} + u_{ii,jj-1}\right] \ldots + \frac{1}{12}\left[u_{ii-1,jj-1} + u_{ii-1,jj+1} + u_{ii+1,jj+1} + u_{ii+1,jj-1}\right] \quad (12)$$

$$\bar{v}_{i,j,k} \approx \frac{1}{6}\left[v_{ii-1,jj} + v_{ii,jj+1} + v_{ii+1,jj} + v_{ii,jj-1}\right] \ldots + \frac{1}{12}\left[v_{ii-1,jj-1} + v_{ii-1,jj+1} + v_{ii+1,jj+1} + v_{ii+1,jj-1}\right] \quad (13)$$

1.1.1.3 Minimização dos Erros

Depois de feitas as estimativas e aproximações, surge o problema de minimizar a soma dos erros nas equações para a taxa de mudança da iluminação da imagem **[9]**.

$$\varepsilon_b = E_x u + E_y v + Et \tag{14}$$

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

$$\varepsilon_c^2 = (\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2 + (\frac{\partial v}{\partial x})^2 + (\frac{\partial v}{\partial y})^2 \tag{15}$$

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 **[9]**.

A minimização a ser alcançada achando valores satisfatórios para a velocidade do fluxo ótico (u, v).Usando o cálculo de variação será obtido **[9]**:

$${E_x}^2 + E_x E_y = \alpha^2 \nabla^2 u - E_x E_t \tag{16}$$

$${E_y}^2 v + E_x E_y u = \alpha^2 \nabla^2 v - E_y E_t \tag{17}$$

Usando a aproximação do Laplaciano da Equação 11, realizando algumas substituições, e resolvendo as equações para u e v encontra-se:

$$\left({E_x}^2 + \alpha^2\right)u + E_x E_y \bar{v} = \alpha^2 \bar{u} - E_x E_t \tag{18}$$

$$\left({E_y}^2 + \alpha^2\right)v + E_x E_y \bar{u} = \alpha^2 \bar{v} - E_y E_t \tag{19}$$

O determinante da matriz de coeficiente é igual a:

$$\alpha^2\left({E_x}^2 + {E_y}^2 + \alpha^2\right) \tag{20}$$

Resolvendo u e v encontra-se:

$$\left({E_x}^2 + {E_y}^2 + \alpha^2\right)u = \left(\alpha^2 + {E_y}^2\right)\bar{u} - E_x E_y \bar{v} - E_x E0_t \tag{21}$$

$$\left({E_x}^2 + {E_y}^2 + \alpha^2\right)v = \left(\alpha^2 + {E_x}^2\right)\bar{v} - E_x E_y \bar{u} - E_x E_t \tag{22}$$

1.1.1.4 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 **[9]**. A solução interativa pode ser expressa por:

$$u^{n+1} = u^{-n} - E_x \left(\frac{E_x u^{-n} + E_y v^{-n} + E_t}{{E_x}^2 + {E_y}^2 + \alpha^2} \right) \tag{23}$$

$$v^{n+1} = v^{-n} - E_x \left(\frac{E_x u^{-n} + E_y v^{-n} + E_t}{{E_x}^2 + {E_y}^2 + \alpha^2} \right) \tag{24}$$

Onde n denota o número da iteração.

1.2 Análise por Componentes Principais - PCA

A análise em componentes principais permite que o movimento facial seja representado por um conjunto muito menor de parâmetros **[3]**, **[5]** através da eliminação da redundância dos dados. O movimento facial é representado por uma matriz N X M que representa em componentes cartesianas x (horizontal) e y (vertical) o valor módulo do fluxo óptico em cada ponto em questão. O fluxo óptico é obtido entre quadros durante uma elocução.

Uma vez que os movimentos representados pelo Fluxo óptico são fortemente correlacionados, é possível trabalhar em uma base ortogonal em que os movimentos faciais são adequadamente caracterizados a partir de um conjunto de K componentes principais descorrelacionadas.

O cálculo das componentes principais inicia-se com a remoção da média de cada uma das linhas da matriz de dados M.

$$\mu_j = \frac{1}{m} \sum_{i-1}^{m} x_j^{(i)} \tag{25}$$

Sendo $x^1, x^2 \ldots, x^m$ os dados da matriz M, então troca-se cada $x_j^{(i)}$ por $x_j^{(i)} - \mu_j$.Após, calcula-se a matriz de covariância:

$$cov = \frac{1}{m} \sum_{i-1}^{m} (x^{(i)})(x^{(i)})^T \tag{26}$$

Em seguida, são computados os autovetores normalizados, e uma nova matriz elaborada composta por tais autovetores ordenados conforme seus autovalores correspondentes em ordem decrescente ao longo da diagonal principal, processo conhecido como transformada de Hotelling **[10], [11]**.

O valor da soma dos K primeiros autovalores dividida pela soma de todos os autovalores representa a variância correspondente às K primeiras componentes principais.

1.3 Técnicas de Aquisição dos Parâmetros da Acústica da Fala – Parâmetros LSP (*Line Spectrum Pairs*)

Para a representação da fala utiliza-se de coeficientes LSP (*Line Spectrum Pairs*) **[10]**, **[12]** que são eficientes por estarem ligados às frequências de ressonância do trato vocal dos 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 **[2]**. Os parâmetros LSP são fundamentados no princípio de conservação da envoltória do espectro da fala que, segundo **[13]**, é suficiente para assegurar a inteligibilidade do sinal **[7]**.

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 **[5]**, **[12]**, **[13]**. 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 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 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 pólos **[5]**, **[12]**, **[13].**

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 polos **[5]**. Assim, a função de transferência, no domínio z, do fluxo de volume glotal e da radiação labial pode ser representada aproximadamente como:

$$G(z)R(z) = \frac{K_1 K_2 (1-z_{-1})}{(1-z_a z_{-1})(1-z_b z_{-1})} \quad (27)$$

Onde $G(z)R(z)$ é a 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 **[7]**.

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 **[7]**:

$$\hat{s}(j) = -\sum_{i=1}^{p} \alpha_i s(j-i), \quad (28)$$

Em que $\hat{s}(j)$ é o valor do sinal predito, $s(j-i)$ são valores passados observados e α_i, $i = 1, \ldots, p$, 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 **[10]**, **[11]**:

$$H(z) = \frac{1}{1+\sum_{i=1}^{p} \alpha_i z^{-i}} \quad (29)$$

A análise por predição linear (LPC) objetiva uma boa estimação das propriedades espectrais do sinal. Para isto os coeficientes de predição $(\alpha_1, \ldots, \alpha_p)$ são obtidos em cada fala correspondente aos quadro, em que p é a ordem do filtro usada na análise. Para uma taxa de amostragem de 8 kHz, há aproximadamente 4 frequências de ressonância até a frequência de Nyquist (4kHz), 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 $p = 10$ **[13]**, **[14]**.

Equivalente aos parâmetros LPC, no domínio da frequência, um novo conjunto de parâmetros chamados LSP (Line Spectrum Pairs) é definido $(\omega_1, \theta_1, \ldots, \omega_{\frac{p}{2}}, \theta_{\frac{p}{2}})$. Estes parâmetros LSP (ω_i, θ_i) são obtidos a partir de um filtro de pólos estável **[7]**:

$$A_p(z^{-1}) = 1 + \sum_{i=1}^{p} \alpha_i z^{-i} \quad (30)$$

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:

$$P(z^{-1}) = A_p(z^{-1}) - \left(z^{-(p-1)}\right)A_p(z) = 1 + \left(a_1 - a_p\right)z^{-1} + \cdots + \left(a_p - a_1\right)z^{-(p)} - z^{-(p-1)} \quad (31)$$

$$Q(z^{-1}) = A_p(z^{-1}) - \left(z^{-(p-1)}\right)A_p(z) = 1 + \left(a_1 - a_p\right)z^{-1} + \cdots + \left(a_p - a_1\right)z^{-(p)} + z^{-(p-1)} \quad (32)$$

Reconstruindo:

$$A_p(z^{-1}) = \frac{1}{2}[P(z^{-1}) + Q(z^{-1})] \quad (33)$$

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 produtório, obtém-se:

Para p par:

$$P(z^{-1}) = (1 - z^{-1})\prod_{i=1}^{p/2}(1 - 2cos\omega_i z^{-1} + z^{-2}) \quad (34)$$

$$Q(z^{-1}) = (1 + z^{-1})\prod_{i=1}^{p/2}(1 - 2cos\omega_i z^{-1} + z^{-2}) \quad (35)$$

Para p ímpar:

$$P(z^{-1}) = (1 - z^{-1})\prod_{i=1}^{(p-1)/2}(1 - 2cos\omega_i z^{-1} + z^{-2}) \quad (36)$$

$$Q(z^{-1}) = \prod_{i=1}^{(p-1)/2}(1 - 2cos\omega_i z^{-1} + z^{-2}) \quad (37)$$

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

$$f_i = \omega_i(2\pi T) \quad (38)$$

$$g_i = \theta_i(2\pi T) \quad (39)$$

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

$$f = f_1, g_1, f_2, g_2, \ldots, f_{\frac{p}{2}}, g_{\frac{p}{2}} \quad (40)$$

1.4 Representações Lineares em Tempo Discreto e no Espaço de Estados

1.4.1 Modelo ARX (*Autoregressive with Exogenous Inputs*)

Existem algumas representações matemáticas que são especialmente adequadas à identificação de sistemas, usando-se algoritmos conhecidos para estimação de parâmetros **[6]**.

Dada a Equação 41:

$$A(q)y(k) = \frac{B(q)}{F(q)}u(k) + \frac{C(q)}{D(q)}v(k) \tag{41}$$

sendo q^{-1} o operador de atraso, de forma que *y(k)* q^{-1}= *y(k-1)*; *v(k)* ruído branco; n_y o máximo atraso entre os regressores de saída; n_u, n_v, n_d e n_f o máximo de atraso entre os regressores de entrada; *A(q), B(q), C(q), D(q)* e *F(q)* os polinômios definidos a seguir **[6]**:

$$\begin{aligned} A(q) &= 1 - a_1 q^{-1} \text{ -...- } a_{n_y} q^{-n_y} \\ B(q) &= b_1 q^{-1} \text{+...+ } b_{n_u} q^{-n_u} \\ C(q) &= 1 + c_1 q^{-1} \text{+...+ } c_{n_v} q^{-n_v} \\ D(q) &= 1 + d_1 q^{-1} \text{+...+ } d_{n_d} q^{-n_d} \\ F(q) &= 1 + f_1 q^{-1} \text{+...+ } f_{n_f} q^{-n_f} \end{aligned} \tag{42}$$

O modelo ARX pode ser obtido utilizando a Equação 41, tomando-se *C(q) = D(q) = F(q)* = 1 **[6]**, sendo A(q) e B(q) polinômios arbitrários, resultando em:

$$A(q)y(k) = B(q)u(k) + v(k) \tag{43}$$

Uma vez que o ruído v(k) aparece diretamente na equação, o modelo ARX é normalmente classificado como pertencente à classe de modelos de erro na equação **[6]**.

1.4.2 Espaço de Estados

Uma representação que pode ser usada para modelar relações entre variáveis internas ao sistema é a representação em espaço de estados **[6]**. Esse tipo de representação descreve o sistema no domínio do tempo e, segundo **[6]** é mais conveniente para poder se representar sistemas não-lineares e multivariáveis do que função ou matriz de transferência.

Dada a Equação 44:

$$\dot{\mathrm{x}}(t) = Ax(t) + Bu(t)$$

$$y(t) = Cx(t) + Du(t) \quad (44)$$

onde A, B, C e D são respectivamente, *n* x *n*, *n* x *p*, *q* x *n*, e *q* x *p* matrizes constantes. Em **[15]** é feita a descrição matemática de como são encontradas os estados iniciais x(0) e a entrada u(t), assim como proceder na discretização da Equação 44.

Neste trabalho a metodologia utilizada foi a estimação em espaço de estados utilizando subespaço, que em **[16]** descreve este método em 2 passos. O primeiro passo faz a projeção de certos subespaços gerados a partir dos dados, para encontrar uma estimativa da matriz de observabilidade prolongada e/ou uma estimativa de dados do sistema desconhecido. O segundo passo, recupera as matrizes do sistema a partir de qualquer presente estendido da matriz observabilidade ou os estados estimados **[16].**

1.5 RMSE – Erro Quadrático Médio

O índice RMSE (*Root Mean Square Deviation*), compara as predições do modelo com a média temporal do sinal. Ou seja, a média é usada como preditor trivial **[6].**

Este índice é calculado de acordo com a Equação 45:

$$RMSE = \frac{\sqrt{\sum_{k=1}^{N}(y(k)-\check{y}(k))^2}}{\sqrt{\sum_{k=1}^{N}(y(k)-\bar{y}(k))^2}} \quad (45)$$

onde $\check{y}(k)$ é a simulação livre do sinal e $\bar{y}(k)$ é o valor médio do sinal medido y(k), sendo que a média é calculada na janela de identificação **[6]**.

1.6 Objetivo Geral

O objetivo deste trabalho é analisar a relação existente entre movimento facial e acústica da fala, onde o movimento facial é extraído por meio do fluxo óptico e a acústica da fala analisada em parâmetros LPC e LSP. Tal relação será obtida através de estimadores lineares que serão desenvolvidos a partir de um banco de dados em que constam, os vetores de movimento facial e as variações em amplitude das faixas de frequências dos parâmetros LSP.

2 Materiais e Métodos

2.1 Aquisição de Dados

O primeiro passo para o estudo da acústica da fala e do movimento facial foi a aquisição de vídeos, onde locutores proferem sentenças. Como o objetivo era a estimação da acústica por meio do fluxo óptico, optou-se por trabalhar com um único indivíduo. Assim, as filmagens foram feitas utilizando este interlocutor, e a ele feitas 5 perguntas das quais não tinha conhecimento prévio de quais seriam. Os níveis de ruído que era exposto o interlocutor, através de um *headphone*, eram alterados e foram classificados como baixo, médio, alto, muito alto e sem ruído.

Para se evitar interferências, as perguntas foram escritas em folha de papel, e o interlocutor lia as perguntas para respondê-las. Estas perguntas eram repetidas conforme alterados os níveis de ruído. Além disso, a ordem dos níveis de ruído e das perguntas foi escolhida de maneira aleatória a fim de se evitar ao máximo alguma interferência por prévio conhecimento. As aquisições aconteceram com pequenos intervalos entre as perguntas. Como o ruído exposto não atingiu valores que danificam momentaneamente a audição, acredita-se que o tempo de recuperação não tenha afetado os experimentos.

Feita a aquisição dos dados o vídeo foi tratado como uma sequência de imagens conectadas, *frames*. Uma região retangular traçada sobre a região da boca do locutor foi traçada no primeiro frame e mantida ao longo de todo o vídeo.

2.2 Fluxo Óptico do Movimento Facial

Primeiramente, nos vídeos a serem analisados foi selecionada somente a região de interesse, em nosso caso a região que envolve os lábios bochechas e queixo.

Sob a região selecionada foi calculado o módulo do fluxo óptico ao longo do vídeo utilizando o método de Horn & Schunck.

O Fluxo Óptico foi extraído utilizando 30 iterações de acordo com a equação do método de Horn & Schunck as quais foram suficientes para que gerasse vetores de velocidade

condizentes com o que se espera do movimento facial. Também foi estimado o tamanho máximo deste vetor a fim de que fosse dado um limitador de busca para o algoritmo.

A Figura 3 ilustra alguns vetores obtidos através do uso do método de Horn & Schunck. Foram então selecionados alguns vetores, para fins de observação, correspondentes ao redor da boca do interlocutor através do uso de um filtro passa altas Butterworth de oitava ordem e frequência de corte de 15 Hz.

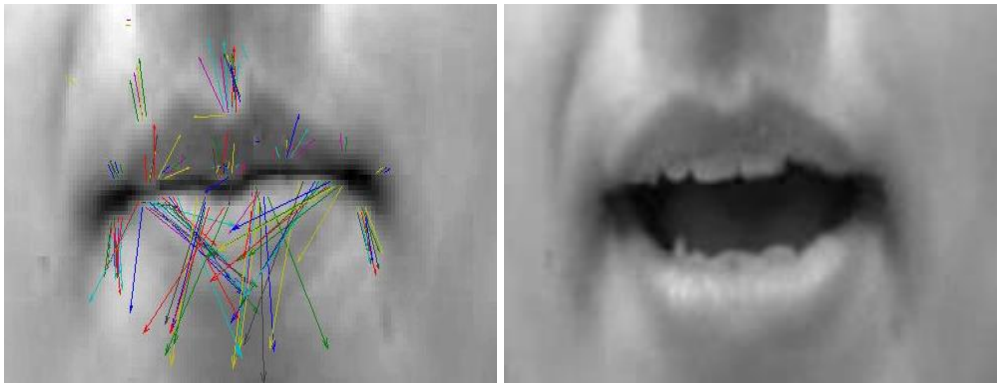

Figura 5 - Vetores com módulo e sentido da velocidade do movimento facial relativo entre dois frames.

2.3 Análise por Componentes Principais

A computação do fluxo óptico dos vídeos analisados gerou matrizes 100x120, região em torno da boca, as quais representavam, na forma retangular, os valores do movimento na direção vertical e horizontal de cada *pixel* ao longo de cada *frame*. Destas matrizes foram utilizados somente o módulo de cada valor computado.

A fim de reduzir o número de dados relativos ao módulo do fluxo óptico calculado, e que posteriormente será a variável de entrada para a determinação de um modelo matemático para representação do sistema desejado, as matrizes geradas pelo cálculo do fluxo óptico foram organizadas em componentes principais conforme descrito na seção 1.2.

Na estimação das componentes principais os autovetores indicam os coeficientes de cada componente principal e, os autovalores, representam a variância explicada por cada componente. A variância acumulada das 10 primeiras componentes principais do módulo do fluxo óptico calculado está apresentada na Figura 4.

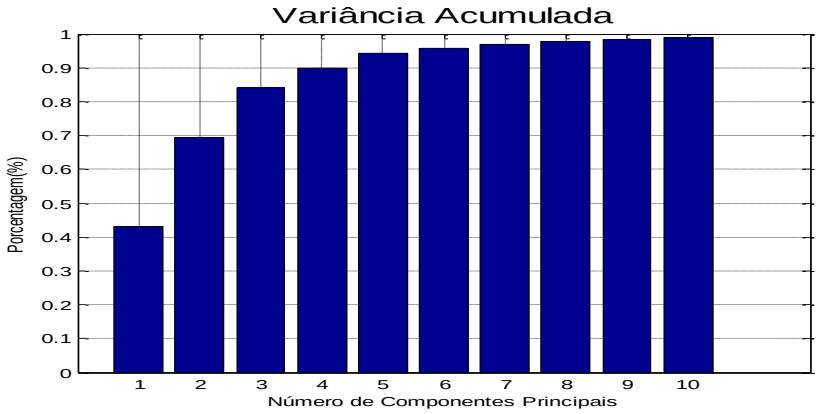

Figura 6 - Variância acumulada relativa às 10 primeiras componentes principais.

2.4 Análise dos Parâmetros Acústicos da Fala

Na análise do sinal acústico, o sinal de voz foi amostrado a uma taxa de 8040 amostras/s e dividido em 29 quadros/s. A cada quadro foi feita a análise LPC de ordem 10 e então os coeficientes LPC foram convertidos em LSP.

As Figuras 5, 6 e 7 referem-se aos parâmetros LSP relativos a data de nascimento enunciada pelo locutor sob condições normais, sob a inserção de um ruído caracterizado como médio e sob a inserção de um ruído caracterizado como alto respectivamente.

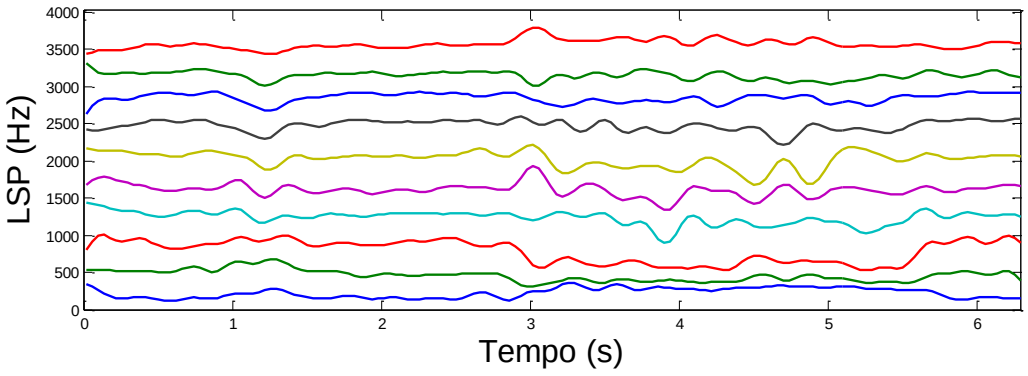

Figura 7 - Parâmetros LSP. Locutor submetido a condições sem ruído.

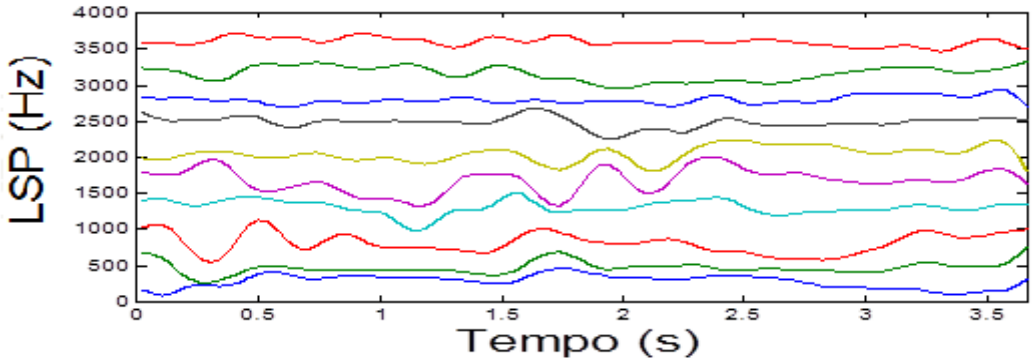

Figura 8 - Parâmetros LSP. Locutor submetido a um ruído médio.

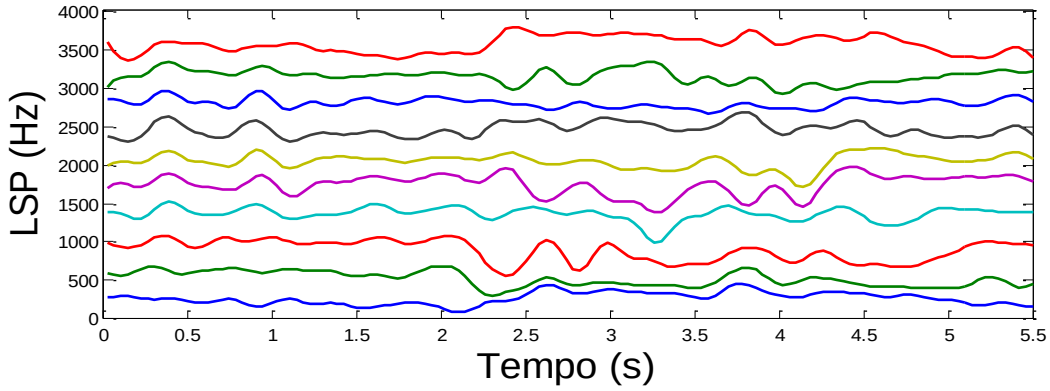

Figura 9 - Parâmetros LSP. Locutor submetido a um ruído alto.

Para criação dos modelos matemáticos é necessário que tanto os valores de entrada quanto os valores de saída a serem modelados possuam o mesmo número de valores de dados. As componentes principais referentes ao fluxo óptico possuíam aproximadamente doze mil amostras enquanto que, devido à taxa de amostragem, os parâmetros LSP possuíam em torno de duzentas amostras.

Para resolver tal problema os dados LSP foram reamostrados de forma a possuírem o mesmo número de amostras que os dados das PCA's sem perder as informações relevantes nos pontos que serão de interesse para a modelagem. As Figuras 8, 9 e 10 ilustram os parâmetros LSP após a amostragem.

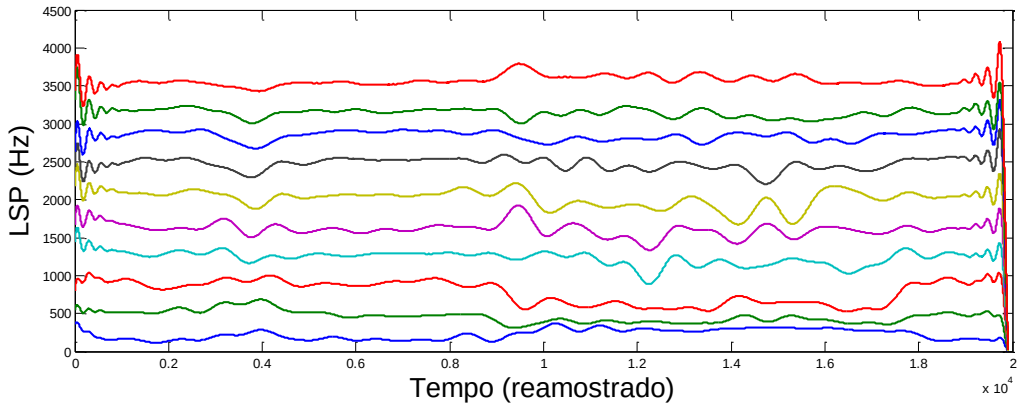

Figura 10 - Parâmetros LSP reamostrados. Locutor submetido a nenhum ruído.

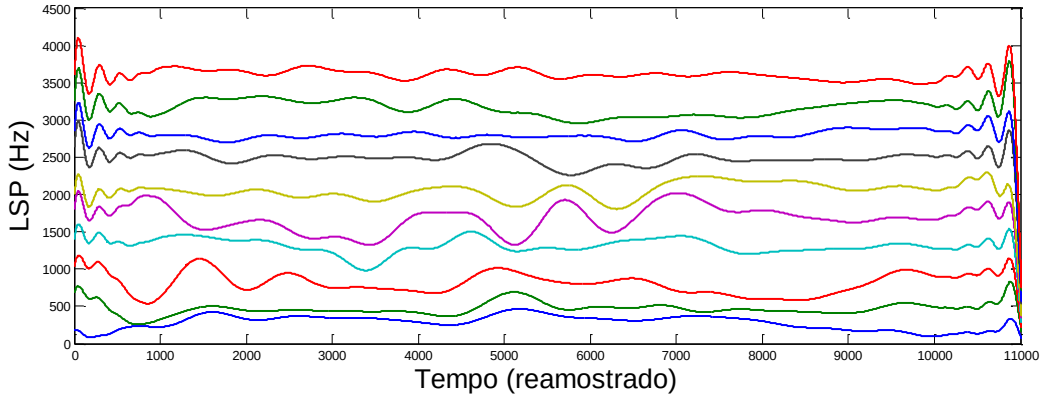

Figura 11 - Parâmetros LSP reamostrados. Locutor submetido a um ruído médio.

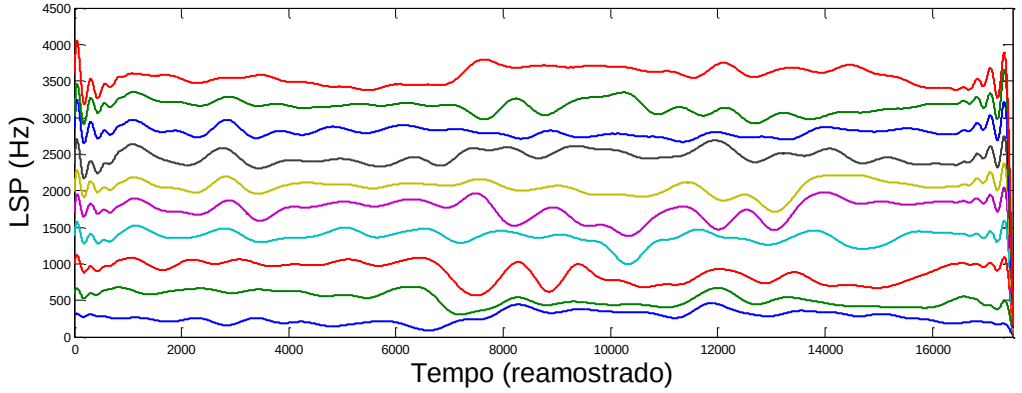

Figura 12 - Parâmetros LSP reamostrados. Locutor submetido a um ruído alto.

Fazendo uma comparação das Figuras 8, 9 e 10, com as Figuras 5, 6 e 7, nota-se um alto grau de semelhança havendo, para os parâmetros reamostrados, uma distorção no início e no final das amostras.

Para este trabalho, tal distorção não influenciará na criação dos modelos matemáticos, pois a análise será feita a palavras isoladas e seus parâmetros acústicos estão fora desta região de distorção.

2.5 Representações Lineares em Tempo Discreto e no Espaço de Estados

Para cada palavra foram feitos modelos preditores: ARX e em espaço de estado que, a princípio, foram os que apresentaram melhores resultados. Os modelos ARX encontrados para todos os casos foram de ordem nove e sem nenhum atraso entre os regressores de entrada e de saída, e os modelos em equação de espaço de estado foram de ordem 10 e sem nenhum atraso entre os regressores de entrada e de saída. A escolha da ordem dos modelos, bem como os atrasos referente aos regressores de entradas e saídas foram escolhidos conforme eram obtidos os melhores resultados.

3 Resultados e Discussões

Com a finalidade de analisar quantitativamente e qualitativamente os parâmetros LSP obtidos, foram utilizados indicadores matemáticos de medidas de dispersão sobre a frase enunciada pelo locutor para cada condição de ruído a que este estava submetido.

A princípio, de cada faixa de valores LSP foi feito o cálculo da variância e do desvio padrão. As Tabelas 1 e 2 contêm os valores de variância e desvio padrão respectivamente, para os casos em que o locutor não estava submetido a nenhum ruído (normal), a um ruído caracterizado como médio (médio) e a um ruído caracterizado como alto (alto).

Tabela 1 - Variância dos parâmetros LSP.

	Faixa LSP					
Nível de Ruído	**1**	**2**	**3**	**4**	**5**	
Normal	4690	6126	23642	8184	7537	
Médio	8900	7330	20354	9454	27853	
Alto	6489	9314	19108	8099	17822	
	Faixa LSP					
Nível de Ruído	**6**	**7**	**8**	**9**	**10**	**Soma**
Normal	12396	6671	4068	3134	4080	80.528
Médio	11052	5918	2711	10255	3448	107.275
Alto	9062	8084	3665	8038	12482	102.163

Tabela 2 - Desvio padrão dos parâmetros LSP.

	Faixa LSP					
Nível de Ruído	**1**	**2**	**3**	**4**	**5**	
Normal	68,48	78,27	153,73	90,47	86,81	
Médio	94,34	85,62	142,67	97,23	166,89	
Alto	80,55	96,50	138,23	89,99	133,50	
	Faixa LSP					
Nível de Ruído	**6**	**7**	**8**	**9**	**10**	**Soma**
Normal	111,34	74,64	63,78	55,98	63,86	847,36
Médio	105,13	76,93	52,07	101,27	58,64	980,79
Alto	95,12	89,91	60,54	89,65	111,72	985,71

Nota-se que a variância em torno de um valor médio, ao fazer um somatório de todas as faixas de frequência LSP, é maior para os casos em que o testado é submetido a um ruído

externo, assim como o desvio padrão. Isso é explicado pelo efeito Lombard, onde pessoas submetidas a um ruído externo tendem a movimentar mais os lábios e a face durante a fala e também tendem aumentar a intensidade do áudio (falar mais alto) durante uma conversação nessas condições. O receptor, por sua vez, tende a focalizar mais na região da boca com maior ou menor grau de resolução.

O próximo passo deste trabalho foi encontrar um modelo matemático que representasse as 10 faixas de valores dos parâmetros LSP em função das 5 primeiras componentes principais referentes ao fluxo óptico da região da face selecionada. As 5 primeiras componentes foram escolhidas por representarem cerca de 90% da variabilidade do movimento obtido por meio do fluxo óptico, como pode ser visto na Figura 4.

As falas enunciadas pelo locutor foram a sua data de nascimento e o nome de sua mãe. A partir deste ponto, para a criação dos modelos matemáticos foram selecionados trechos em que o testado enunciava apenas uma ou poucas palavras.

O primeiro trecho da fala selecionado para teste foram as palavras “noventa e um”. Foram selecionados os instantes em que a fala nesse trecho era iniciado e finalizado para todos os ruídos submetidos. Em seguida foram retiradas as componentes principais e os parâmetros LSP relativos a estes instantes encontrados. Após, os modelos foram obtidos. Desta forma a distorção referida anteriormente devido ao processo de reamostragem dos parâmetros LSP não têm relevância nos trechos selecionados para modelagem.

Em se tratando que havia três conjuntos de dados, com ruído alto, médio e nenhum ruído, modelos foram criados e validados utilizando cada um destes dados. Os mesmos dados que foram usados para modelagem do sistema também foram utilizados para a validação a fim de se encontrar um valor ótimo que servisse para comparação da validação dos demais modelos. Os dados de predição eram comparados entre si pelo valor RMSE atingido pela predição de cada modelo.

As Tabelas 3 e 4 representam os erros RMSE de predição dos modelos ARX e em espaço de estado respectivamente. Alguns modelos ARX apresentaram valores de predição acima de 10^{12} vezes maiores do que o resultado esperado, e para estes modelos não foi calculado os valores RMSE, sendo apenas representados por um * no lugar do valor.

Tabela 3 - Valores RMSE dos modelos ARX. A primeira coluna é relativa aos dados utilizados para modelagem e a segunda coluna relativa aos dados utilizados para validação do modelo. As demais representam o valor RMSE para cada faixa de valores LSP.

Dados Modelo	**Dados Validação**	**RMSE LSP 1**	**RMSE LSP 2**	**RMSE LSP 3**	**RMSE LSP 4**	**RMSE LSP 5**	
Alto	Alto	0,004	0,003	0,003	0,009	0,007	
Alto	Médio	42,097	81,696	30,836	57,457	39,504	
Alto	Normal	*	*	*	*	*	
Médio	Alto	3,738	4,434	2,735	3,081	4,108	
Médio	Médio	0,013	0,016	0,005	0,007	0,004	
Médio	Normal	*	*	*	*	*	
Normal	Alto	*	*	*	*	*	
Normal	Médio	*	*	*	*	*	
Normal	Normal	0,004	0,003	0,003	0,004	0,007	
Dados Modelo	**Dados Validação**	**RMSE LSP 6**	**RMSE LSP 7**	**RMSE LSP 8**	**RMSE LSP 9**	**RMSE LSP 10**	**Somatório RMSE**
Alto	Alto	0,005	0,009	0,019	0,021	0,030	0,110
Alto	Médio	49,080	187,617	319,843	460,057	832,917	2101,103
Alto	Normal	*	*	*	*	*	*
Médio	Alto	3,520	6,538	14,793	16,057	23,960	82,965
Médio	Médio	0,005	0,013	0,022	0,030	0,052	0,168
Médio	Normal	*	*	*	*	*	*
Normal	Alto	*	*	*	*	*	*
Normal	Médio	*	*	*	*	*	*
Normal	Normal	0,003	0,013	0,013	0,019	0,019	0,088

Tabela 4 - Valores RMSE dos modelos em espaço de estado. A primeira coluna é relativa aos dados utilizados para modelagem e a segunda coluna relativa aos dados utilizados para validação do modelo. As demais representam o valor RMSE para cada faixa de valores LSP.

Dados Modelo	**Dados Validação**	**RMSE LSP 1**	**RMSE LSP 2**	**RMSE LSP 3**	**RMSE LSP 4**	**RMSE LSP 5**	
Alto	Alto	0,042	0,035	0,044	0,062	0,040	
Alto	Médio	15,407	18,531	5,248	6,254	3,635	
Alto	Normal	2,336	1,791	2,034	0,693	0,840	
Médio	Alto	2,176	1,287	1,831	4,153	1,103	
Médio	Médio	0,082	0,095	0,028	0,033	0,021	
Médio	Normal	1,774	1,389	2,058	2,514	2,458	
Normal	Alto	1,659	0,958	1,105	1,223	0,893	
Normal	Medio	1,161	1,998	1,240	1,970	1,419	
Normal	Normal	0,020	0,023	0,033	0,050	0,052	
Dados Modelo	**Dados Validação**	**RMSE LSP 6**	**RMSE LSP 7**	**RMSE LSP 8**	**RMSE LSP 9**	**RMSE LSP 10**	**Somatório RMSE**
Alto	Alto	0,026	0,032	0,068	0,062	0,100	0,512

Alto	Médio	3,752	8,585	13,088	16,583	24,830	115,913
Alto	Normal	1,236	1,773	1,490	1,546	1,975	15,714
Médio	Alto	1,268	1,046	2,257	1,679	2,517	19,317
Médio	Médio	0,021	0,050	0,076	0,111	0,158	0,675
Médio	Normal	0,689	1,209	2,014	1,464	1,379	16,948
Normal	Alto	1,271	1,397	1,019	2,097	1,592	13,213
Normal	Médio	0,477	1,076	2,036	3,221	2,016	16,615
Normal	Normal	0,018	0,047	0,046	0,059	0,060	0,408

Percebe-se que o valor ótimo para predição é equivalente a um RMSE em torno de 1 ou inferior. O modelo que mais próximo chegou a esse valor foi uma representação em espaço de estados utilizando os dados sem ruídos para modelar o sistema e os dados com ruído alto para validação. Para algumas faixas de valores LSP o melhor modelo obtido apresentou um valor de erro muito alto em relação ao que era esperado, porém foi possível verificar a relação existente entre a entrada e a saída predita. Vale salientar que foram usados apenas métodos lineares de predição.

Visto que as representações em espaço de estado obtiveram melhores resultados que as representações em ARX, foi selecionado outro vídeo no qual o falante enuncia a palavra: "Maria" e então feito o mesmo estudo que foi descrito anteriormente, desde a análise por componentes principais do fluxo óptico à representação por modelos, porém foram utilizados somente os modelos em representação de espaço de estado.

A Tabela 5 representa os erros RMSE de predição dos modelos em espaço de estado.

Tabela 5 - Valores RMSE dos modelos em espaço de estado. A primeira coluna é relativa aos dados utilizados para modelagem e a segunda coluna relativa aos dados utilizados para validação do modelo. As demais representam o valor RMSE para cada faixa de valores LSP.

Dados Modelo	**Dados Validação**	**RMSE LSP 1**	**RMSE LSP 2**	**RMSE LSP 3**	**RMSE LSP 4**	**RMSE LSP 5**	
Alto	Alto	0,015	0,021	0,014	0,010	0,009	
Alto	Médio	0,596	0,732	1,782	1,032	0,828	
Alto	Normal	1,246	2,648	1,138	2,550	2,669	
Médio	Alto	0,836	1,475	1,079	1,991	1,327	
Médio	Médio	0,101	0,086	0,151	0,089	0,065	
Médio	Normal	0,616	1,825	1,118	2,318	1,203	
Normal	Alto	0,635	2,911	1,410	2,394	1,958	
Normal	Medio	1,308	1,712	1,908	2,078	1,131	
Normal	Normal	0,112	0,094	0,092	0,131	0,092	

Dados Modelo	Dados Validação	RMSE LSP 6	RMSE LSP 7	RMSE LSP 8	RMSE LSP 9	RMSE LSP 10	Somatório RMSE
Alto	Alto	0,006	0,011	0,019	0,029	0,021	0,156
Alto	Médio	0,539	0,567	0,309	2,346	2,709	11,440
Alto	Normal	0,550	0,780	1,093	1,201	2,267	16,143
Médio	Alto	0,906	2,183	2,557	1,490	1,476	15,320
Médio	Médio	0,076	0,076	0,098	0,204	0,365	1,311
Médio	Normal	0,566	0,542	1,090	1,153	2,406	12,837
Normal	Alto	0,677	1,394	1,336	1,296	1,624	15,634
Normal	Médio	0,767	0,867	1,381	2,305	4,858	18,315
Normal	Normal	0,081	0,065	0,098	0,046	0,185	0,995

Novamente, percebe-se que o valor ótimo como resultados para predição é um RMSE em torno de 1. O modelo que mais próximo chegou a esse valor foi uma representação em espaço de estado utilizando os dados com ruído alto para modelar o sistema e os dados com ruído médio para validação. Deve-se notar que, para este melhor modelo, tanto nos dados utilizados para modelagem quanto os utilizados para validação, o locutor estava submetido a ruídos externos, mesmo que em intensidades diferentes. A presença deste ruído faz com que o locutor enuncie suas frases em um tom de voz mais elevado, assim como articular mais os lábios durante a fala, fato caracterizado pelo efeito Lombard. Essa maior articulação facial bem como a maior variância dos parâmetros LSP, mesmo que em pequena escala, em comparação com os dados obtidos sem ruído pode ter sido e grande importância no melhor resultado do modelo.

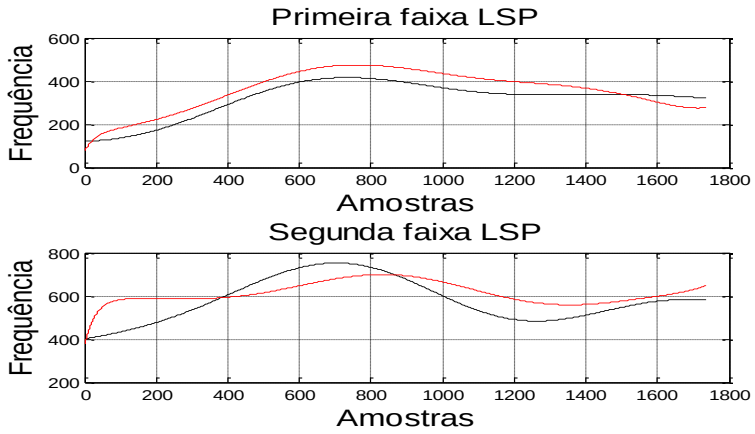

Figura 13 - Faixas 1 e 2 dos parâmetros LSP. Em vermelho valor simulado, em azul valor esperado.

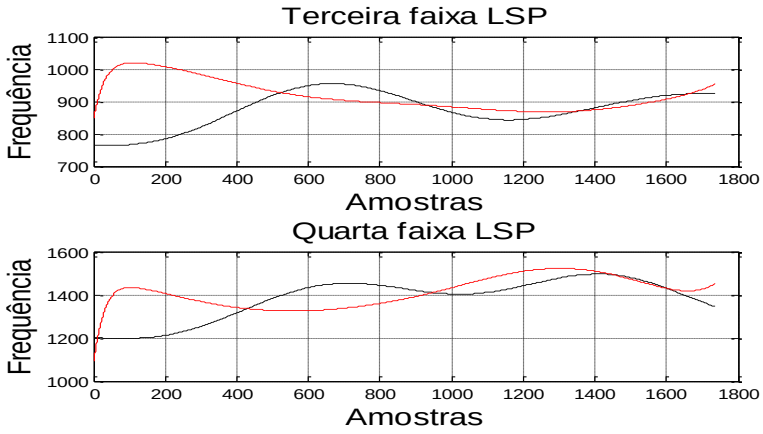

Figura 14 - Faixas 3 e 4 dos parâmetros LSP. Em vermelho valor simulado, em azul valor esperado.

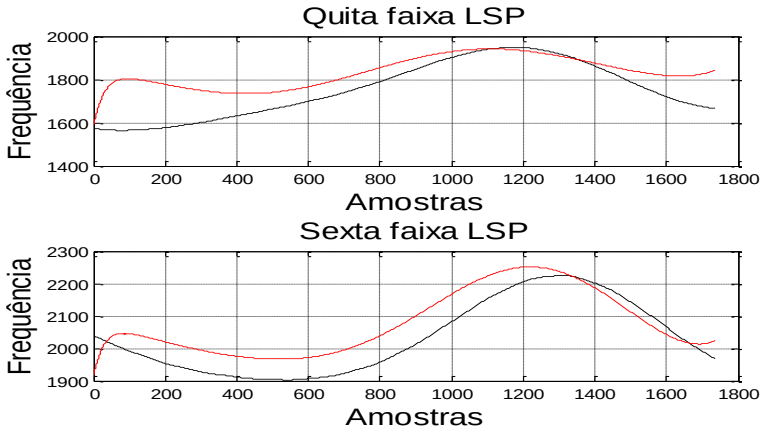

Figura 15 - Faixas 5 e 6 dos parâmetros LSP. Em vermelho valor simulado, em azul o valor esperado.

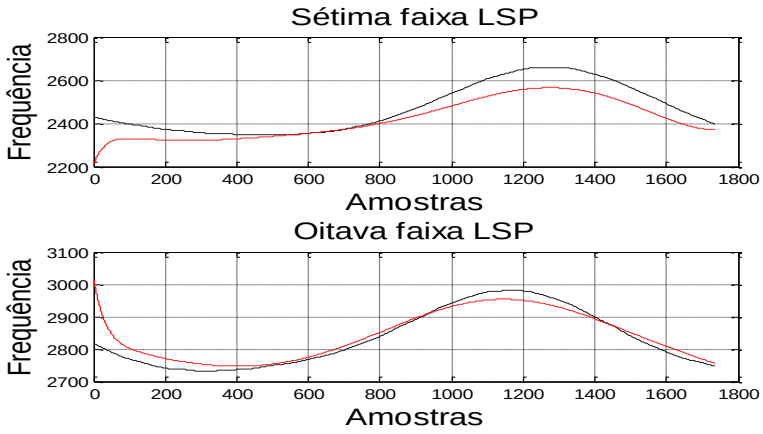

Figura 16 - Faixas 7 e 8 dos parâmetros LSP. Em vermelho valor simulado, em azul valor esperado.

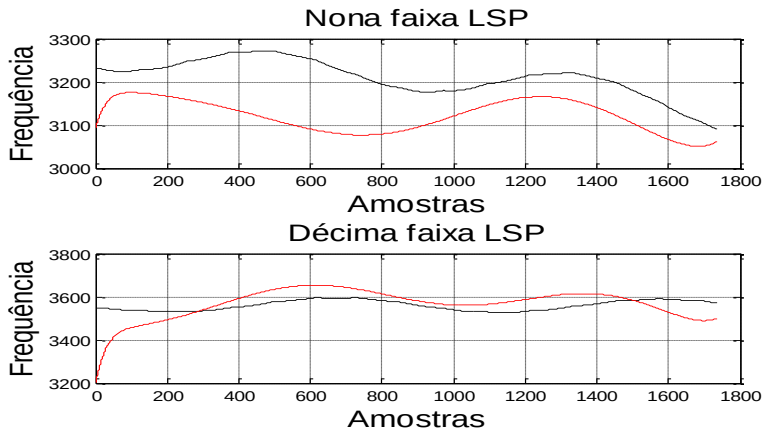

Figura 17 - Faixa 9 e 10 dos parâmetros LSP. Em vermelho valor simulado, em azul valor esperado.

As Figuras 11, 12, 13, 14 e 15 ilustram os valores preditos pelo modelo matemático encontrado e os valores esperados. Pela Tabela 5, pode-se notar que para a oitava faixa de valores LSP o erro RMSE de predição foi de apenas 0,309. Observando graficamente na Figura 14 a saída predita pelo modelo na oitava faixa, em torno de 2831 Hz, acompanhou satisfatoriamente a saída esperada.

Com o intuito de se medir a similaridade entre os valores preditos e os reais, fez-se uso da correlação cruzada entre os dois sinais. As equações do cálculo de tal correlação pode ser encontrado em **[17]**.

A Tabela 6 apresenta os valores dos coeficientes encontrados para cada faixa de valores LSP.

Tabela 6 - Coeficientes calculados para cada faxa de valores LSP preditos.

Faixa LSP	**1**	**2**	**3**	**4**	**5**
Coeficiente	0,941857	0,822617	0,585983	0,636792	0,912643
Faixa LSP	**6**	**7**	**8**	**9**	**10**
Coeficiente	0,977199	0,965699	0,963211	0,550631	0,600122

Os valores encontrados são referentes ao maior valor absoluto de cada sinal obtido através do calculo da correlação entre os sinais preditos e reais. Nota-se que o grau de similaridade chegou, para sexta faixa de valores LSP, a atingir 97,77%, melhor caso, e para a terceira faixa *atingir* somente 58,60%, pior caso.

4 Conclusões

Na produção da fala, a geometria do trato vocal determina suas frequências de ressonância e influencia no movimento da face. O trabalho desenvolvido demonstrou que padrões acústicos da fala e movimentos faciais estão relacionados e acoplados. Mostrou também que a relação entre o movimento de partes da face pode ser modelada por meio da análise em componentes principais do fluxo óptico e, que é possível, através de modelos matemáticos lineares, estimar os parâmetros LSP da acústica da fala fazendo uso desse gradiente de deslocamento (Fluxo Óptico - método de Horn & Schunck).

Neste estudo os modelos de equação em espaço de estado obtidos apresentaram, para todos os casos testados, resultados melhores que os modelos ARX. Algumas frequências dos parâmetros LSP apresentaram melhores resultados, como faixa em torno de 2033 Hz, que mostrou 97,71% de similaridade com os dados esperados, enquanto que a faixa em torno de 3210 Hz atingiu apenas 55,06%. Outro resultado verificado é que, em ambientes ruidosos é uma característica do ser humano aumentar seu tom de voz e articular mais o movimento de sua face em comparação com ambientes sem ruído.

Portanto, para trabalhos futuros, através do que foi exposto neste, é interessante verificar, por exemplo, se existem outros modelos preditores mais eficientes, lineares ou não, ou algum outro método de compressão, além da PCA, que possa estimar com maior acerto os parâmetros LSP através do fluxo óptico do movimento facial.

Referências Bibliográficas.

[1] Silva, Vinícius L. G. Análise do Efeito Lombard no movimento facial frpor meio de *Optical Flow*. Monografia – DEL – UFV – 2013.

[2] Yehia, H. C.; Rubin, P., Vatikiotis Bateson, E., October 1998. *Quantitative association of vocal-tract and facial behavior*. Speech Communication, v. 26, p. 23–43.

[3] Yehia, H. C.; Barbosa, A.V. (2001). *Measuring the relation between speech acoustics and 2D facial motion*, CEFALA - Center for Research on Speech, Acoustics, Language and Music.

[4] Vatikiotis-Bateson E.,I.-M. Eigsti, S. Yano, andK. G. Munhall (1998). *Eye movement of perceivers during audiovisual speech perception*, Perception & Psychophysics, vol. 60, no. 6, pp. 926–940.

[5] Barbosa, A. V. (2005). Um Estudo Sobre Relações Entre as Falas Audível e Visível (*a Study On The Relations Between Audible And Visible Speech*). Tese de Doutorado – CPDEE – UFMG.

[6] Aguirre, L. A. (2007). Introdução à Identificação de Sistemas: Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais. Editora UFMG, 3a ed. ISBN: 9788570415844

[7] Vatikiotis-Bateson E., Barbosa, A. V. C. C. Y. O. M. T. J., Yehia, H. C., August-September 2007. *Audiovisual lombard speech: Reconciling production and perception*. AVSP, p. 45–50.

[8] J. Barron, D. Fleet, and S.S. Beauchemin, *Performace of Optical Flow Techniques*, J. Comput. Vision, 12, 1994, 43-77

[9] Horn B. K. P. and Schunck B. G., *Determining optical flow.* AI 17, pp. 185-204, 1981.

[10] Barbosa, A. V. (2000). Codificação Audio-Visual Integrada da Fala. Dissertação Mestrado – CPDEE – UFMG.

[11] Moreira, K. S. Um estudo sobre as relações de padrões do movimento facial com a acústica da fala e com a identidade do locutor. Tese de doutorado – CPDEE – UFMG – 2008.

[12] Vatikiotis-Bateson E.; Yehia H. C. (2002). *Speaking mode variability in multimodal speech production*. IEEE Trans Neural Netw. 13(4):894-9. doi: 10.1109/TNN.2002.1021890

[13] Flanagan, J. L. *Speech analysis, synthesis, and perception*. Third edition. [S.l.]: Springer-Verlag, 1972

[14] 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

[15] Chen, Chi-Tsong. (1998). *Linear system theory and design*/ *by* Chi-Tsong Chen – 3rd ed.p. cm. – (The Oxford series in electrical and computer engineering). ISBN – 13978019511777-6 (cloth). ISBN – 019511777-8 (cloth)

[16] Favoreel, W., Moor, B. De, Overschee, P. V. (2000). *Subspace state space system identification for industrial process.* Jornal of Process Control 10, p. 149~155

[17] Huang, Xuedong (2001). *Spoken language processing: a guide to theory, algorithm, and system development*/ Xuedong Huang, Alex acero, Hsiao-Wuen Hon. ISBN 0130226165

[18] Sugamura, N., Itakura, F. (1986). *Speech analysis and syntesis methods develop 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.

[19] Gonzalez, Rafael C. 2005. "Processamento de imagens digitais". São Paulo:EdgardBlücher.

[20] Yehia, H. C., Itakura, F., October 1996. *A method to combine acoustic and morphological constraints in the speech production inverse problem*. Speech Communication, v. 18.2, p. 151–174

[21] Yehia, H. C., Itakura, F (1994). *Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal fourier analysis*, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. Vol. I, pp. 477-480.