Um Conjunto de Visemas para uma Cabeça Falante do Português do Brasil

José Mario De Martino, Léo Pini Magalhães

Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas, Campinas, SP, Brasil
martino@dca.fee.unicamp.br, leopini@dca.fee.unicamp.br

Resumo

A leitura orofacial, também denominada leitura labial, pode ser descrita como a técnica que procura facilitar a compreensão da fala às pessoas com capacidade auditiva diminuída, complementando a informação auditiva deficiente com a informação visual, associada à articulação das palavras, disponível na face do falante. A leitura orofacial pode ser um importante recurso de comunicação, principalmente para os deficientes auditivos parciais que já tenham adquirido a fala. Este trabalho, apresenta resultados parciais do desenvolvimento de uma cabeça falante para o Português do Brasil. Pretende-se que esta cabeça falante possa ser utilizada para o treinamento, aprendizado, avaliação e aperfeiçoamento do ensino da técnica de leitura orofacial. No artigo é apresentada a metodologia adotada para a definição de um conjunto de visemas para o Português do Brasil. Os visemas são representações visuais das realizações sonoras da fala, que, aos moldes dos fonemas, têm função distintiva e identificadora. O conjunto de visemas proposto contempla o fenômeno de coarticulação adjacente antecipatória e perseveratória e será utilizado como referência para a animação da cabeça falante.

1. Introdução

A comunicação através da fala é efetuada principalmente, quando não exclusivamente, através da percepção e interpretação da informação acústica produzida e modulada pela movimentação articulatória do trato vocal. Não obstante, as pistas visuais induzidas na face do locutor por esta movimentação articulatória podem contribuir para a percepção da fala. A constatação de que a inteligibilidade da fala em situação de degradação do sinal acústico por ruído pode ser melhorada através da observação da face do locutor [1], atesta a participação das pistas visuais no processo de percepção. Em geral, é assumido que as indicações da face não participam significativamente na percepção da fala até que o canal acústico falhe, e nestas situações de degradação do sinal sonoro, a falha é compensada pela informação visual. Assim, pode ser entendido que a informação fornecida pelas pistas visuais da face é redundante, pelo menos parcialmente, à informação transportada pelo sinal acústico.

Extrair das pistas visuais da face informações lingüísticas é denominado de leitura orofacial, ou ainda leitura labial. A leitura orofacial é praticada, em maior ou menor grau, por todos, mesmo que de forma inconsciente [2]. Por outro lado, a ruptura da coerência visual-auditiva da fala pode levar à percepção distorcida e incorreta da mensagem transmitida, como revelado pelo denominado efeito McGurk [3]. No experimento realizado por McGurk, imagens de vídeo de um locutor articulando os fonemas /gɑ/ foi dublada com o áudio de /bɑ/. O resultado foi percebido pela maioria da população de teste como /dɑ/. O efeito McGurk indica a importância de contemplar com fidedignidade os aspectos articulatórios da produção da fala que dão origem às pistas visuais apresentadas na face quando da implementação da cabeça virtual falante, que inerentemente tem a intenção de explorar o canal visual para a comunicação.

Os sons da fala são produzidos pela modificação controlada do fluxo de ar pulmonar. Estas modificações são efetuadas principalmente pelo posicionamento das pregas vocais, do véu palatino, da língua, da mandíbula e dos lábios. Grande parte destes movimentos articulatórios ocorre no interior da cavidade oral sem que seja possível a sua visualização. Conseqüentemente, o contraste visual dos segmentos é reduzido a um conjunto mais restrito de parâmetros do que o conjunto das possibilidades articulatórias, tornando a percepção visual menos capacitada para a discriminação entre segmentos do que a percepção auditiva. Assim, na percepção visual da fala, os padrões de movimentação articulatória visualmente contrastáveis acabam por ser associados a mais de um segmento sonoro da língua. Notadamente, o vozeamento e a nasalidade apresentam efeito acústico marcante, porém não permitem o contraste visual [4]. Segmentos sonoros que não são possíveis de serem diferenciados visualmente são denominados de homofemas (homo + (morf)ema). Um padrão visual de movimentação articulatória representante de um grupo homofemas é denominado visema. Uma das questões principais deste trabalho é, além da identificação de um grupo de homofemas para o português do Brasil, o estabelecimento de visemas para os grupos de homofemas identificados.

A identificação de homofemas está associada à identificação e contraste de diferentes padrões visíveis de movimentação articulatória. O grau de contraste

destes padrões depende de um conjunto de fatores, tais como, a habilidade articulatória do falante, as características físicas da face do locutor, a taxa de elocução, a capacidade do ouvinte/observador de perceber as pistas visuais, a iluminação e a distância e o ângulo de visão do ouvinte/observador [2]. Como indicado em [5] [6], a coarticulação é um outro fator lingüístico que fortemente influencia a percepção visual da fala.

A coarticulação se manifesta pela alteração do padrão articulatório de um dado segmento pela influência da articulação de outro adjacente ou, e em menor grau, próximo na cadeia da produção sonora. É possível distinguir a coarticulação perseveratória da coarticulação antecipatória. A coarticulação perseveratória tem lugar quando a articulação de um segmento é influenciada por segmento que o antecede na cadeia da produção sonora. Na coarticulação antecipatória, o segmento é influenciado por segmento que o sucede. Se o efeito da coarticulação não impacta significativamente a percepção acústica, permitindo que alofones gerados em contextos fonéticos diferentes sejam associados a um mesmo fonema, o mesmo não se aplica à percepção visual. A característica visual pode, devido à coarticulação, sofrer variações significativas a ponto de afetar a percepção dos grupos de homofemas.

2. Método e Instrumentação

Corpus

Para o estabelecimento de um conjunto de visemas para o português do Brasil foi efetuada a gravação em vídeo da fala de informante brasileiro nascido e criado na região da capital do Estado de São Paulo. O informante é estudante do curso de engenharia elétrica, não possuindo treinamento ou conhecimento específico em fonética e/ou fonoaudiologia. O informante foi instruído a produzir logatomas do tipo $'CV_1CV_2$, $'V_3V_2$; onde, utilizando o alfabeto fonético internacional, C = /p, t, k, (ɣ) ɾ, f, s, ʃ, l, ʎ/, V_1 = /i, a, u/, V_2 = /ɪ, ɐ, ʊ/ e V_3 = /i, e, ɛ, a, ɔ, o, u/. No logatoma, onde o segmento consonatal é o tepe alveolar /ɾ/, foi utilizado como primeiro segmento a velar fricativa /ɣ/, já que o tepe no português não ocorre no início de palavra. A ordem dos logotomas foi gerada aleatoriamente, e os mesmos foram apresentados para a leitura pelo informante na tela de um computador posicionado à sua frente. O locutor foi instruído para sempre iniciar e terminar a produção de um logatoma a partir de uma posição de repouso com a boca fechada.

Gravação

Durante a produção dos logatomas, o informante foi gravado em vídeo com o auxílio de duas câmeras JVC KY27C, sincronizadas e posicionadas em um ângulo de 90°. O áudio foi capturado com um microfone Shure SM58. As imagens e o áudio foram gravados em fitas S-VHS utilizando gravadores JVC BR-S622 DXU e JVC BR-S822 DXU.

Antes da gravação, pontos de interesse foram marcados com tinta branca na face do locutor. A figura 1 apresenta os pontos de interesse ao redor da boca e no queixo, cujas trajetórias foram medidas. Os visemas propostos neste trabalho resultam da análise da movimentação destes pontos.

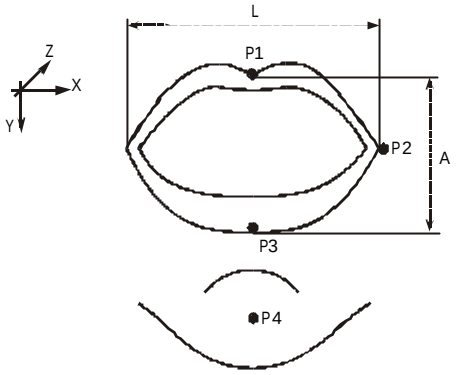

Fig. 1. Pontos de interesse.

Para estabelecer uma referência fixa para as medidas das coordenadas espaciais dos pontos de interesse, tendo em vista movimentações involuntárias do tronco e pescoço, o informante utilizou um capacete especial fixo à cabeça.

Captura das trajetórias

As imagens e o áudio foram convertidos para formato digital utilizando o sistema Media 100 V.60. Após a digitalização o material foi segmentado nos diversos logatomas.

Para a medida da posição dos pontos quadra-a-quadro, foi utilizada uma ferramenta de *software* especialmente desenvolvida para este fim. Esta ferramenta, utilizando técnicas de processamento de imagem, efetua, após a identificação por um operador de pontos de referência no capacete e dos pontos de interesse da face na primeira imagem de uma seqüência, o cálculo das coordenadas dos pontos de interesse em todas as outras imagens do vídeo. O cálculo das posições tridimensionais dos pontos em cada imagem foi realizado com o auxílio de técnicas fotogramétricas (consulte, por exemplo, [1] para detalhes). A figura 2 apresenta a interface gráfica do usuário da ferramenta, onde é mostrado um par de imagens (câmera esquerda e câmera direita) do informante. Na figura também é possível observar o capacete que define a referência espacial para as medidas, assim como os pontos de interesse marcados na face do informante.

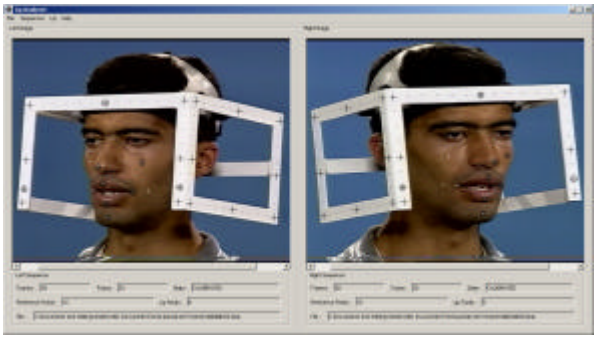

Fig. 2. Ferramenta de medida da movimentação articulatória.

Pré-processamento

Após a captura, as trajetórias tridimensionais dos pontos de interesse foram suavizadas por um filtro passa-baixa Butterworth de sexta ordem com freqüência de corte igual a 1/6 da freqüência de amostragem de 29,97 Hz - freqüência dos quadros do vídeo. Adicionalmente, as trajetórias foram deslocadas para indicarem posições relativas à posição inicial de repouso. O valor do deslocamento correspondente à posição de repouso foi calculado pela média das coordenadas dos 10 quadros iniciais e dos 10 finais da produção de todos os logatomas

O sinal de áudio digitalizado foi segmentado em fones através da inspeção do espectrograma.

Dentro de cada intervalo de produção sonora de cada fone, foi procurado o primeiro ponto crítico (derivada primeira, ou seja velocidade, igual a zero ou derivada segunda, ou seja aceleração, igual a zero) das trajetórias. Para o cálculo do ponto crítico foi considerada a coordenada de maior excursão da trajetória. Para os pontos P1, P3 e P4 (fig. 1), foi considerada a coordenada y. Para o ponto P2, foi considerada a coordenada z. Os casos em que não foi possível encontrar automaticamente o ponto crítico dentro do intervalo da produção sonora foram inspecionados manualmente. Nos casos, como o da produção do segmento /p/ comentado na seção 3, em que o ponto crítico ocorre antes da produção sonora foi considerado o ponto crítico antes da produção do fone. Em poucos casos em que não foi possível identificar um ponto crítico, foi considerado o meio da produção fone.

Agrupamento por similaridade

Os valores da posição (x, y, z) e da velocidade (dx/dt, dy/dt, dz/dt) no ponto crítico foram utilizados como parâmetros em análise de agrupamento baseada no algoritmo k-means [8]. Para tanto, cada fone foi representado por um vetor no espaço $\Re^{24}$: 6 valores, (x, y, z) e (dx/dt, dy/dt, dz/dt), para quatro pontos. Para cada fone foi efetuada análise de agrupamento considerando os contextos fonéticos apresentados na tabela 1. Na tabela, o fone sob análise, listado na coluna à esquerda, é apresentado em negrito nos contextos fonéticos da coluna à direita.

Fone	Contextos
/i, a, u/	$^{\prime}\mathbf{C}\mathbf{V}CV_2$, $^{\prime}\mathbf{V}V_2$
/ɪ, ɐ, ʊ/	$^{\prime}CV_1\mathbf{C}\mathbf{V}$, $^{\prime}V_3\mathbf{V}$
/e, ɛ, ɔ, o/	$^{\prime}\mathbf{V}V_2$
/p, t, k, (ɣ) ɾ, f, s, ʃ, l, ʎ/	$^{\prime}\mathbf{C}V_1CV_2$, $^{\prime}CV_1\mathbf{C}V_2$

Tabela 1. Contextos Fonéticos

A análise de agrupamento foi realizada variando-se de forma sistemática o número de grupos procurados, indo de 2 até um tamanho máximo igual ao número de contextos fonéticos. Foram efetuadas 40 repetições para cada procedimento de agrupamento, cada uma a partir de condições inicias diferentes geradas aleatoriamente. O critério de escolha da solução final foi o de menor coeficiente Davies-Bouldin [8]. Soluções finais com coeficiente Davies-Bouldin maior do que 1.0 foram descartadas e, para estes casos, foi considerado um único grupo.

3. Resultados

As figuras 3 a 6 apresentam as trajetórias dos pontos P1, P2, P3 e P4 durante a produção de /ˈpapɐ/. O sistema de referência para a correta interpretação dos gráficos é apresentado no canto superior esquerdo da figura 1. As linhas verticais tracejadas marcam os intervalos da produção dos fones estabelecidos pela análise do espectrograma do sinal de áudio. É interessante observar que o fechamento dos lábios para a produção da plosiva bilabial /p/ ocorre antes da produção sonora do respectivo segmento. Tal fato é notado, por exemplo, na figura 5, onde ocorre um mínimo local da coordenada y antes do início da produção sonora no quadro 38. Uma vez que as coordenadas apresentadas nos gráficos representam deslocamento a partir da posição de repouso, valores para x, y, e z próximos de zero indicam uma configuração próxima da posição de repouso, quando a boca encontra-se fechada. Também é interessante observar que o informante, antes de iniciar a produção do logatoma /ˈpapɐ/ abriu ligeiramente a boca, fechando-a para a produção do primeiro /p/.

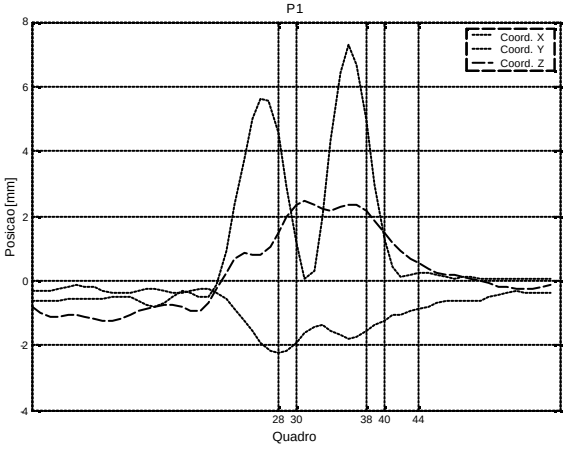

Fig. 3. Trajetória do ponto P1 durante produção de /ˈpapɐ/.

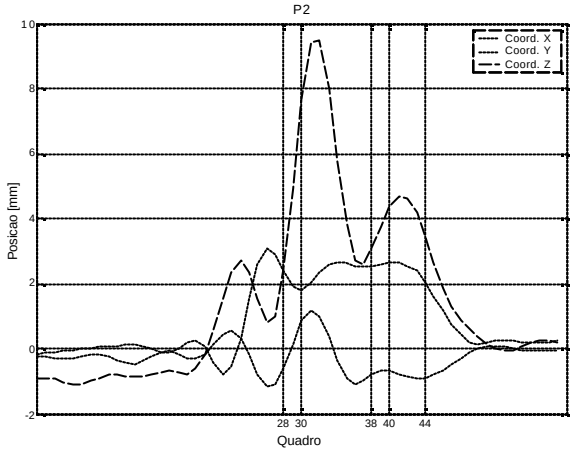

Fig. 4. Trajetória do ponto P2 durante produção de /ˈpapɐ/.

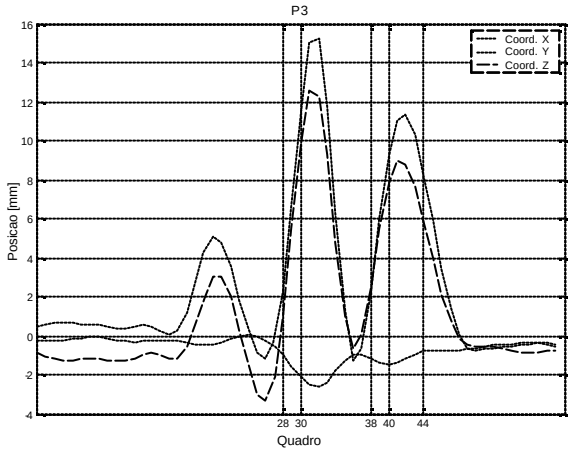

Fig. 5. Trajetória do ponto P3 durante produção de /ˈpapɐ/.

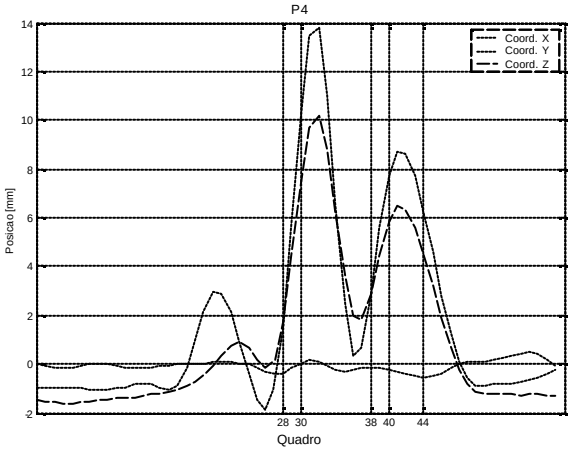

Fig. 6. Trajetória do ponto P4 durante produção de /ˈpapɐ/.

A tabela 2 apresenta os resultados do processo de agrupamento. Na tabela, as seguintes convenções foram empregadas V_1 e V_2 como na seção 2, C^1 = /p, t, k, ɾ, f, s, ʃ, l, ʎ/, C^2 = /p, k, (ɣ) ɾ, f, s, l, ʎ/, V^1 = /i, a/, V^2 = /ɪ, ɐ/, V^3 = /i, u/ e V^4 = /i, ʊ/.

A partir das coordenadas tridimensionais dos centróides dos grupos encontrados foram estimadas a abertura A e a largura L da boca (cf. figura 1), assim como a protrusão do lábio superior e inferior. As figuras 7 e 8 apresentam gráficos envolvendo estes parâmetros para os grupos dos segmentos vocálicos.

Grupos	Contextos
i_1	/i/V_2, C^2/i/C^2
i_2	/tit/, /ʃiʃ/
a	/a/V_2, C/a/C
u	/u/V_2, C/u/C
I	V_3/ɪ/, C^1/ɪ/
A	V_3/ɐ/, C^1/ɐ/
U	V_3/ʊ/, C^1/ʊ/
e	/e/V_2
eh	/ɛ/V_2
oh	/ɔ/V_2
o	/o/V_2
p_1	/p/V^1, V^1/p/V_2, /upɐ/
p_2	/pu/, /up/V^4
t_1	/atɐ/, /ta/
t_2	/atʊ/, /itɪ/, /ti/
t_3	V^3/tʊ/, /utɪ/, /tu/
t_4	V^3/tɐ/, /atɪ/
k_1	/akɐ/, /ka/
k_2	/ik/V', /itɪ/, /ti/
k_3	V_1/kʊ/, /ku/
r_1	V^1/ɾ/V^2, /ɣ/V^1
r_2	V_1/ɾʊ/, /uɣ/V^2, /ɣu/
f_1	/af/V^2, /if/V_2, /f/V^1
f_2	/uf/V_2, /afʊ/, /fʊ/
s_1	V^1/s/V^2, /s/V^1
s_2	V_1/sʊ/, /su/
s_3	/us/V^2
sh_1	V_1/ʃ/V^2, /ʃ/V^1
sh_2	V^1/ʃʊ/
sh_3	/uʃʊ/, /ʃu/
l_1	/al/V^2, /ilɐ/, /la/
l_2	V^3/li/, /alʊ/, /li/
l_3	/ulɐ/, /lu/
l_4	V^3/lʊ/
lh_1	V^1/ʎ/V^2, /ʃ/V^1
lh_2	V_1/ʎʊ/
lh_3	/uʎ/V^2, /ʎʊ/

Tabela 2. Grupos encontrados.

4. Considerações finais

O conjunto de consoantes adotado baseia-se na premissa de que as pistas visuais aparentes na face de um falante não espelham a nasalidade, ou seja a movimentação do véu palatino, nem a vibração das pregas vocais. Assim, considera-se que os segmentos consononantais se organizam nos seguintes grupos de homofemas: /p, b, m/, /t, d, n/, /k, g/, /ɾ, ɣ/, /f, v/, /s, z/, /ʃ, ʒ/ e /ʎ, ɲ/.

Já para os segmentos vocálicos, é considerado que os segmentos /i, ɪ, ɐ, a, ʊ, u/ e / representam um conjunto mínimo de padrões articulatórios diferenciáveis, sendo que os segmentos /e, ɛ, ɔ, o/ são visualmente similares a um daqueles segmentos. Considerando a menor distância euclidiana dos centróides dos grupos encontrados foi possível identificar os seguintes grupos de homofemas: /e, ɐ/, /ɛ, a/, e /ɔ, o, u/. As figuras 7 e 8 apresentam a distribuição dos centroídes dos segmentos vocálicos identificados considerando os parâmetros protrusão do lábio superior, abertura e largura da boca.

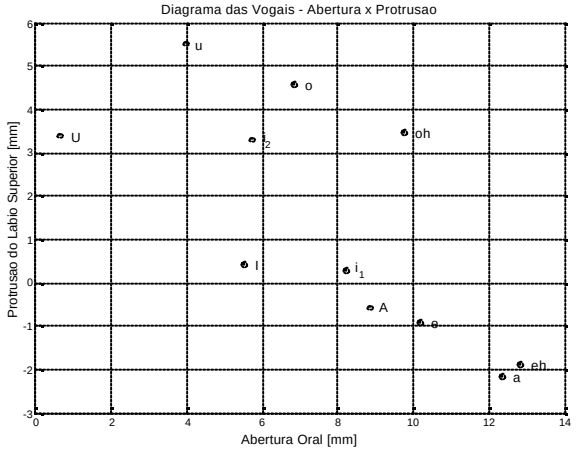

Fig. 7. Diagrama das Vogais - Abertura x Protrusão.

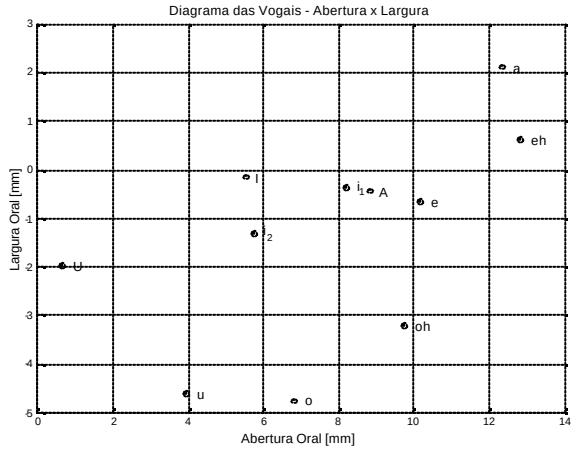

Fig. 8. Diagrama das Vogais - Abertura x Largura.

Como resultado final tem-se o conjunto de visemas marcados com fundo cinza na tabela 2. Cada um dos visemas é caracterizado por um contexto fonético e pelos parâmetros posição e velocidade dos pontos P1, P2, P3 e P4, definidos pelos centróides dos grupos identificados. Estes parâmetros serão utilizados como referência para o controle da cabeça falante integrada a um conversor texto-fala [9] em desenvolvimento. Para a animação da cabeça falante, os visemas são aproximados por um polinomio de terceiro grau. A figura 9 apresenta os valores medidos e aproximados pela curva de Hermite da coordenada y do ponto P3 durante a produção de /fafɐ/.

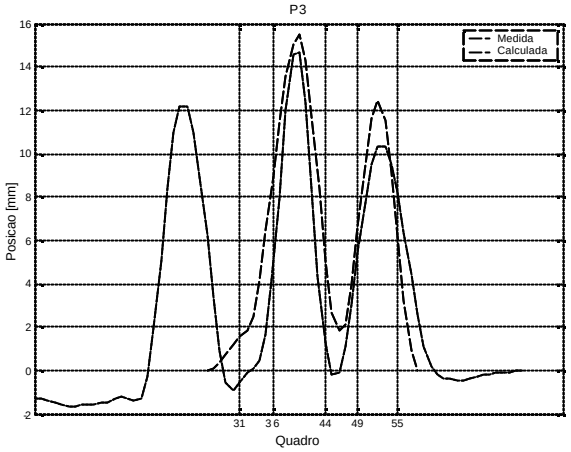

Fig. 9. Coordenada y do ponto P3 na produção de /fafɐ/

Referências

[1] Norman P. Erber. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*. Vol. 40, 1975, 481-492.

[2] Janet Jeffers, Margaret Barley. *Speechreading (Lipreading).* Charles C. Thomas Publisher. 1971.

[3] H. McGurk, J. MacDonald. Hearing lips and seeing voices. *Nature.* Vol. 264, 1976, 746-748.

[4] Norman P. Erber. Auditory, Visual, and Auditory-visual Recognition of Consonants by Children with Normal and Impaired Hearing. *Journal of Speech and Hearing Research*. Vol. 15, 1972, 413-422.

[5] André-Pierre Benguerel, Margaret Kathleen Pichora-Fuller. Coarticulation Effects in Lipreading. *Journal of Speech and Hearing Research*. Vol. 25, December 1982, 600-607.

[6] Elmer Owens, Barbara Blazek. Visemes Observed by Hearing-Impaired and Normal Adult Viewers. *Journal of Speech and Hearing Research*. Vol. 28, 1985, 381-393.

[7] N. Ayache. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception.* P. Sander tradutor. MIT Press.

[8] Anil K. Jain, richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[9] José Mario De Martino, Fábio Violaro. Um talking head par o português do Brasil que objetiva suportar a leitura labial. Contribuiciones Tecnológicas para a Discapacidad – J. García, Tamón Ceres, Luis Azevedo e Teodiano Freire eds. I Jornadas CYTED sobre Tecnologías de Apoyo a la Discapacidad. Maio 2003.,123-127.

Agradecimentos

Os autores agradecem as contribuições do Prof. Fábio Violaro no desenvolvimento deste trabalho.