Primeiros experimentos com dados articulatórios e sua relação com a segmentação acústica[1]

Glauco F. G. Yared
Jaqueline V. Gonçalves
Plínio A. Barbosa
Luís G. P. Meloni
UNICAMP

Abstract

The main purpose of this work is to analyse the projections of maximum and minimum points of two articulatory trajectories during speech production over the acoustic signal, namely, the mouth aperture and the jaw opening. In order to do so, we have chosen a native French speaker who repeated 5 times the following sentence: "*C'est pas ububuz, c'est pas ybybyz, c'est pas iziziz, c'est pas azhazhaz*". It seems reasonable from the analyses shown to establish a correspondence between articulatory and acoustic events, which happens to be useful to help determine acoustic boundaries. The articulatory boundaries coincide with the middle of the vowels and the mouth aperture seems to be a more consistent parameter than jaw opening for establishing these regularities.

1. INTRODUÇÃO

Caracterizar unidades fonéticas em função de parâmetros físicos mensuráveis tem sido objeto de estudos de muitos trabalhos em fonética lingüística (FANT, 1973; HALLE & STEVENS, 1979; JAKOBSON, FANT & HALLE, 1969; LADEFOGED, 1971). De fato, estruturas fonéticas são caracterizadas por padrões articulatórios que estão em movimento contínuo, ao invés de configurações estáticas (BROWMAN & GOLDSTEIN, 1986; BROWMAN & GOLDSTEIN, 1988).

A utilização de informações articulatórias, conjuntamente com informações acústicas no processo de identificação de fronteiras fonéticas, pode ser de grande importância na segmentação de fones. Além disso, em aplicações que necessitem de segmentação automática de sentenças faladas como, por exemplo, em reconhecimento de fala, é esperado que o sistema se torne mais robusto com a utilização da informação articulatória (BENOÎT, 1996) na recuperação da informação fonética, em situações nas quais o sinal acústico não fornece uma definição precisa para a determinação das fronteiras entre os segmentos.

O artigo apresenta o estudo da existência de possíveis correspondências entre eventos articulatórios, tais como picos (pontos de máximo e mínimo) nos sinais de movimentação do queixo ou de abertura da boca, e a estrutura fonética-acústica de uma sentença do francês nativo. A partir dos resultados é avaliado se as informações articulatórias podem facilitar a determinação das fronteiras entre as unidades fonéticas.

Primeiramente avaliamos aqui se existe correspondência temporal entre fones de mesma natureza ou de natureza distinta. Na

seqüência, analisa-se a ocorrência dos eventos articulatórios no intuito de verificar se tal informação pode ser utilizada para marcar as fronteiras acústicas.

2. BASE DE DADOS

Nos experimentos utiliza-se uma base de dados coletada no Advanced Telecommunications Institute International (ATR), em Kyoto, Japão (YEHIA, RUBIN & BATESON, 1998). Tal base é constituída de dados acústicos e articulatórios gravados para um locutor francês nativo (CB), do sexo masculino, contendo cinco repetições da sentença "C'est pas ububuz, c'est pas ybybyz, c'est pas iziziz, c'est pas azhazhaz", pronunciada em uma única sucessão, sendo que cada uma tem oito segundos de duração.

As trajetórias de pontos emitidos por diodos emissores (LEDs) de luz infravermelha (Ireds) posicionados na face (lábio superior, lábio inferior e queixo) do locutor foram rastreadas a partir do Optotrack. Tal equipamento consiste de um conjunto de três câmeras capazes de rastrear a posição dos LEDs, e possibilita a aquisição de posição com uma taxa de amostragem de sessenta Hz.

No experimento também foi coletada a atividade de músculos da face através de eletromiografia, e por esse motivo o sinal acústico também foi coletado a 5000 Hz.

2.1 Compensação do movimento da cabeça

O rastreamento das trajetórias de pontos localizados na face durante a produção da fala, realizada através do Optotrack, fornece uma medida da superposição do movimento da cabeça e o movimento de tais pontos. No entanto, deseja-se observar apenas o movimento dos pontos da face, e dessa forma deve-se subtrair do movimento total, a movimentação da cabeça. Nesse sentido, determina-se um novo sistema de coordenadas cartesianas, localizado sobre a cabeça do locutor (FIG. 1), de tal forma que o movimento

relativo entre a cabeça do locutor e o sistema de coordenadas seja nulo, ou seja, as trajetórias rastreadas em relação ao novo sistema de coordenadas corresponderão apenas ao movimento dos pontos em questão. Os eixos $\hat{X}$ e $\hat{Y}$ pertencem ao plano frontal, e o eixo $\hat{Z}$ é perpendicular ao plano formado por $\hat{X}$ e $\hat{Y}$.

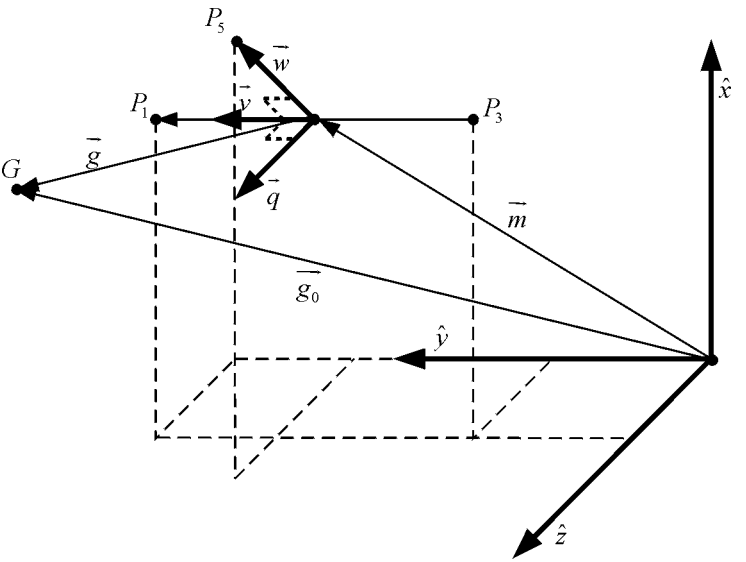

FIGURA 1 - Criação de um sistema de coordenadas localizado na cabeça do locutor.

Inicialmente têm-se os pontos P_1, P_3 e P_5 localizados sobre a cabeça do locutor. Pode-se determinar o ponto médio do segmento P_1P_3, que corresponderá ao vértice do novo sistema de coordenadas.

$$\vec{m} = \frac{\vec{p}_1 + \vec{p}_3}{2} \tag{1}$$

Na seqüência pode-se determinar os vetores na direção do segmento que liga o ponto médio ao ponto P_1 e o vetor na direção do segmento que liga o ponto médio ao ponto P_5, respectivamente.

$$\vec{v} = \vec{p}_1 - \vec{m} = \frac{\vec{p}_1 - \vec{p}_3}{2} \tag{2}$$

e

$$\vec{w} = \vec{p}_5 - \vec{m}. \tag{3}$$

Por último calcula-se o vetor ortogonal ao plano formado pelos vetores obtidos anteriormente através de um produto vetorial:

$$\vec{q} = \vec{w} \times \vec{v}. \tag{4}$$

As trajetórias dos pontos representadas no novo sistema de coordenadas podem ser obtidas pelas projeções do vetor $\vec{g}$ sobre os novos eixos definidos pelos vetores $\vec{w}$, $\vec{v}$ e $\vec{q}$, sendo $\vec{g}$ definido como:

$$\vec{g} = \vec{g}_0 - \vec{m}. \tag{5}$$

Os marcadores (LEDs) e o novo sistema de coordenadas estão ilustrados na FIG. 2.

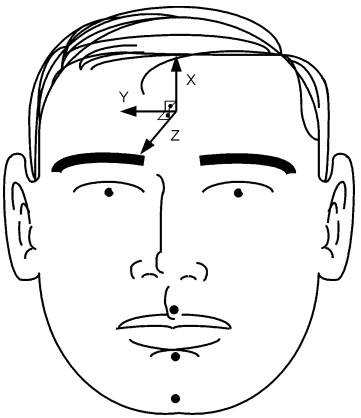

FIGURA 2 - Localização dos LEDs "•" (lábio superior, lábio inferior e queixo) e do novo sistema de coordenadas.

3. ANÁLISE DE FRONTEIRAS FONÉTICAS E ARTICULATÓRIAS

Os fenômenos acústicos observados durante o processo de produção da fala ocorrem como conseqüência da configuração do trato vocal e, dessa forma, estão ligados a eventos articulatórios. Deseja-se avaliar então quais eventos articulatórios se correlacionam com as unidades acústicas do sinal de fala.

Nesse sentido, observam-se os instantes de ocorrência dos pontos de máximo e mínimo do sinal de trajetória do LED localizado no queixo, a fim de analisar os instantes correspondentes no sinal acústico. O mesmo procedimento é adotado utilizando-se o sinal de abertura da boca, gerado a partir da diferença entre a trajetória do LED localizado no lábio superior e a do LED localizado no lábio inferior.

Procura-se observar a existência de regularidade na ocorrência dos eventos articulatórios representados pelos pontos de máximo e mínimo das trajetórias medidas, e se o processo de produção dos fones ocorre de forma semelhante nas vogais (/a/, /ɛ/, /i/, /y/, /u/, /ə/) e nas consoantes (/b/ e /z/) analisadas no experimento. Dessa forma, pode-se inferir se a informação articulatória facilita a segmentação da sentença falada, nos casos em que a extração da informação acústica é mais delicada.

As frases são segmentadas manualmente em fones de acordo com os espectrogramas observados, e os instantes de ocorrência dos picos dos sinais de trajetória do queixo e de abertura da boca são projetados sobre tais sentenças. Assim, pode-se medir a posição relativa de ocorrência do pico dentro do segmento correspondente a um fone. Em algumas repetições da sentença, não se observam picos dentro do intervalo de produção de determinados fones, e em outras, ocorre mais de um pico dentro do segmento de um mesmo fone acústico. Neste último caso, observa-se a média da ocorrência do pico nas demais repetições do fone e seleciona-se o pico mais próximo da média. Além disso, adota-se o critério de que todo pico ocorrido em uma posição menor que 1% do intervalo de duração do fone pertence ao fone anterior, isto é, alonga-se o intervalo anterior até o instante de ocorrência de tal pico. Tal critério é utilizado devido à possibilidade de existência de inconsistências na determinação precisa das fronteiras acústicas entre os fones.

Os procedimentos de análise são divididos em duas partes: projeções dos instantes de ocorrência dos picos do sinal de trajetória do LED localizado no queixo e projeções dos instantes de ocorrência dos picos do sinal de abertura da boca.

4. RESULTADOS

Os procedimentos de análise fornecem resultados sobre a organização temporal da produção fonética das vogais e consoantes analisadas no trabalho, baseados na informação articulatória. Tais resultados são observados para a trajetória do LED localizado no queixo e para a abertura da boca, como descritos na seqüência.

4.1 Sinal de trajetória da abertura da boca

Medem-se os instantes dos picos de máximo e mínimo do sinal de abertura da boca, relativos ao intervalo de duração do fone associado aos mesmos em termos percentuais.

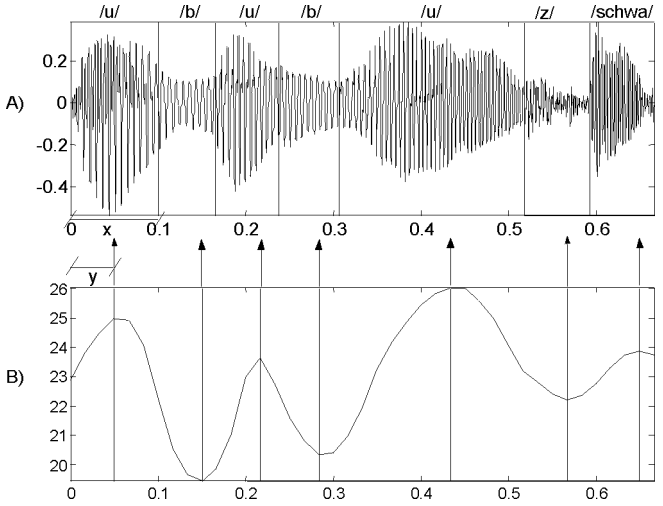

FIGURA 3 - Projeção dos eventos articulatórios sobre os segmentos acústicos. A) Segmentação acústica em fones do trecho "ububuz"; B) Picos no sinal de abertura da boca.

Dessa forma, os instantes de ocorrência dos picos (t_{picos}) são definidos em relação ao início do segmento acústico no qual se inserem, medido a partir do início do intervalo *x*, de acordo com a relação abaixo:

$$t_{pi\cos} = \frac{y}{x} \qquad (6)$$

onde *x* corresponde ao intervalo do fone obtido pela segmentação manual (FIG. 3.A) e *y* ao intervalo de ocorrência do pico (FIG. 3.B).

Em seguida, realiza-se o teste estatístico ANOVA sobre tais valores obtidos para as vogais (/a/, /ɛ/, /i/, /y/, /u/, /ə/) e para as consoantes (/b/ e /z/). Tal teste realiza uma comparação entre as médias dos conjuntos de dados de forma indireta, através da análise de variâncias, assumindo-se como hipótese nula que todos os dados provêm da mesma população.

Os resultados do teste estatístico encontram-se nas TAB. 1, 2, 3 e 4.

TABELA 1
Teste ANOVA para as vogais analisadas (/a/, /ɛ/, /i/, /y/, /u/, /ə/).

	/a/	/ə/	/ɛ/	/i/	/y/	/u/
p	0,128	0,572	0,981	0,203	0,242	0,109
F	2,45	0,610	0,02	2,04	1,67	2,86

TABELA 2
Teste ANOVA para comparação entre vogais (/u/×/y/ e /a/×/y/×/u/).

	/u/×/y/	/a/×/y/×/u/
p	0,092	$5{,}06 \times 10^{-6}$
F	2,26	9,20

TABELA 3
Teste ANOVA para as consoantes analisadas (/b/ e /z/).

	/b/	/z/
p	$7{,}24 \times 10^{-4}$	0,656
F	15,89	0,56

TABELA 4
Teste ANOVA para comparação entre vogal e consoante
(/i/ × /z/, /ə/ × /z/, /u/ × /z/).

	/i/ × /z/	/ə/ × /z/	/u/ × /z/
p	0,421	0,448	0,432
F	1,08	1,03	1,04

A TAB. 1 mostra que as vogais presentes na sentença são produzidas aproximadamente da mesma forma, quando se analisa cada uma separadamente, uma vez que o teste ANOVA fornece valores maiores que 0,05 para a variável estatística p. Porém, a TAB. 2 mostra que vogais distintas são produzidas de forma diferente quando comparadas entre si. Interessante notar que quando o formato labial na produção de fones é semelhante, como é o caso do /u/ e do /y/, o teste indica uma regularidade na produção dos mesmos, o que não ocorre para produção de fones de movimento articulatório muito diferentes como mostrado para os fones /a/, /y / e /u/.

No entanto, a TAB. 3 mostra que o fone /b/ não é produzido de forma independente da posição em que tal fone se insere na sentença, pois o valor de p está abaixo de 0,05. Dessa forma, pode-se interpretar que as vogais analisadas não se mostram sensíveis às consoantes vizinhas, enquanto que a consoante /b/ é afetada pelas vogais vizinhas, do ponto de vista articulatório. No caso do fone /z/, pode-se observar que a informação articulatória extraída a partir do movimento de abertura da boca não é suficiente para evidenciar a influência das vogais vizinhas sobre tal fone. A TAB. 4 mostra que o fone /z/ é produzido da mesma forma que as vogais vizinhas, comprovando tal fato. A informação articulatória provavelmente mais apropriada para a realização dos testes estatísticos, neste caso, deve ser a do movimento da língua.

TABELA 5
Média e desvio padrão dos instantes de ocorrência dos picos nas vogais (/a/, /ə/, /ɛ/, /i/, /y/, /u/).

	/a/	/ə/	/ɛ/	/i/	/y/	/u/
μ_{picos}	0,545	0,552	0,648	0,610	0,729	0,697
σ_{picos}	0.115	0.227	0.319	0.212	0.108	0.09

TABELA 6
Média e desvio padrão dos instantes de ocorrência dos picos nas consoantes (/b/ e /z/).

	/b/	/z/
μ_{picos}	0,739	0,854
σ_{picos}	0.135	0.116

Na TAB. 5 encontram-se as médias dos instantes de ocorrência dos picos nas vogais em análise. Pode-se observar que as médias estão compreendidas no intervalo medido em segundos $0{,}5 \leq \mu_{pi\cos} \leq 0{,}80$. Tais valores sugerem que existe uma tendência da existência de eventos articulatórios em torno de 65% do intervalo de duração das vogais. Assim, pode ser interessante utilizar essa informação para a segmentação em difones, considerando-se neste caso que o início do difone não se encontra exatamente na metade do fone anterior, mas tem o início em torno de 65% do intervalo de duração do fone. Tal afirmação pode ser baseada no fato de que os eventos articulatórios são bastante relacionados com a estrutura acústica da sentença, pois a produção do sinal acústico ocorre a partir da configuração do trato vocal. A mesma conclusão pode ser estendida para as consoantes em análise, conforme mostra a TAB. 6.

4.2 Sinal de trajetória do ponto no queixo

O mesmo procedimento é adotado na obtenção dos resultados utilizando-se o sinal de trajetória do LED localizado no queixo. Os valores obtidos através do teste estatístico encontram-se nas TAB. 7, 8, 9 e 10.

TABELA 7
Teste ANOVA para as vogais analisadas (/a/, /ɛ/, /i/, /y/, /u/, /ə/).

	/a/	/ə/	/ɛ/	/i/	/y/	/u/
P	0,304	0,390	0,100	0,837	0,712	0,446
F	1,32	1,31	5,45	0,05	0,15	0,66

TABELA 8
Teste ANOVA para comparação entre vogais (/u/×/y/ e /a/×/y/×/u/).

	(/u/×/y/)	/a/×/y/×/u/
p	0,195	0,010
F	1,83	3,79

TABELA 9
Teste ANOVA para as consoantes analisadas (/b/ e /z/).

	/b/	/z/
p	$2,35 \times 10^{-6}$	0,214
F	222,61	2,13

TABELA 10
Teste ANOVA para comparação entre vogal e consoante (/i/ x /z/, /ə/ x /z/, /u/ x /z/).

	/i/ × /z/	/ə/ × /z/	/u/ × /z/
p	0,557	0,077	0,226
F	0,87	2,74	1,91

Os resultados obtidos com o sinal de abertura da boca são confirmados pelos resultados obtidos com a trajetória do marcador localizado no queixo. Assim, as conclusões dos testes descritos no item anterior podem ser estendidas para os testes com o LED localizado no queixo, observando-se também as TAB. 11 e 12.

TABELA 11
Média e desvio padrão dos instantes de ocorrência dos picos nas vogais (/a/, /ə/, /ɛ/, /i/, /y/, /u/).

	/a/	/ə/	/ɛ/	/i/	/y/	/u/
μ_{picos}	0,535	0,522	0,804	0,628	0,755	0,579
σ_{picos}	0.136	0.154	0.219	0.32	0.284	0.223

TABELA 12
Média e desvio padrão dos instantes de ocorrência dos picos nas consoantes (/b/ e /z/).

	/b/	/z/
μ_{picos}	0,66	0,60
σ_{picos}	0.409	0.243

5. CONCLUSÕES

Percebe-se que a movimentação articulatória no arranjo frasal proposto para este trabalho fornece resultados que indicam a existência de correspondência entre o domínio articulatório e o domínio acústico a partir da informação extraída dos sinais de movimentação do queixo e abertura da boca, ou seja, os pontos de máximo e mínimo observados nos sinais espaciais. Tal informação pode ser utilizada para identificar fronteiras de unidades acústicas conhecidas (difones e trifones, no caso de síntese acústica da fala). Além disso, os resultados obtidos com o sinal de abertura da boca apresentam uma variabilidade menor do que os obtidos com o sinal de movimentação do queixo.

A observação de regularidade na produção de vogais do ponto de vista articulatório, tanto pela informação extraída da abertura da boca, como pela informação extraída da movimentação do queixo, revela a possibilidade da utilização deste dado para a distinção entre

algumas vogais. Além disso, pode comprovar também que a consoante /b/ é sensível à presença das vogais vizinhas, enquanto que as vogais analisadas não se mostram influenciadas pelas consoantes mais próximas.

Sabe-se que a produção acústica depende diretamente da configuração do trato vocal e, portanto, está fortemente ligada aos eventos articulatórios estudados no trabalho (pontos de máximo e mínimo do sinal de abertura da boca e de movimentação do queixo). No caso ideal, todos os segmentos determinados por eventos articulatórios deveriam cair sobre a posição central do fone, ou sobre o onset e offset dos mesmos, o que permitiria encontrar facilmente as fronteiras entre unidades acústicas tais como difones e trifones. Os resultados mostram que os eventos articulatórios não ocorrem exatamente no centro das vogais, mas se aproximam do mesmo. Neste sentido, pretende-se realizar estudos futuros a fim de reavaliar a definição das fronteiras de tais unidades acústicas, no intuito de verificar se a modificação na localização das fronteiras entre as unidades acústicas pode trazer ganhos para os sistemas de síntese (maior naturalidade) e reconhecimento de fala (menor taxa de erro), levando em consideração a relação custo/benefício para a implementação do segmentador automático que também utiliza a informação articulatória. Além disso, o valor médio de ocorrência do pico em uma determinada posição dentro do fone pode ser utilizado como informação adicional em algoritmos de reconhecimento de fala.

Portanto, pode-se utilizar a informação de abertura da boca, por exemplo, de forma conjunta com a informação acústica durante o processo de determinação das fronteiras entre as unidades fonéticas, sempre que haja disponibilidade do material, a fim de se facilitar a segmentação.

NOTA

[1] Os autores agradecem ao suporte financeiro dado pela FAPESP a Jaqueline Vieira Gonçalves e a Sérgio Robertos Barros pela contribuição com o trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

BENOÎT, C. *Synthesis and automatic recognition of audio-visual speech*. London: The Institution of Electrical Engineers, 1996.

BROWMAN, C. P.; GOLDSTEIN, L. Towards an articulatory phonology. *Phonology Yearbook*, 3, p. 219-252, 1986.

_____. Some notes on syllable structure in articulatory phonology. *Phonetica*, 45, p. 140-155, 1988.

FANT, G. Distinctive features and phonetic dimensions. In: FANT, G. *Speech sounds and features*. Cambridge, MA, p. 171-191, 1973.

HALLE, M.; STEVENS, K. N. Some reflections on the theoretical bases of phonetics. In: LINDBLOM, B.; OHMAN, S. (Ed.) *Frontiers of speech communication research*. New York: Academic Press, 1979, p. 335-353.

JAKOBSON R.; FANT, C. G. M.; HALLE, M. *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge, Mass: MIT Press, 1969. (MIT Acoustics Laboratory Technical Report, 13.)

LADEFOGED, P. *Preliminaries to linguistic phonetics*. University of Chicago Press, 1971.

YEHIA, H.; RUBIN, P.; BATESON, E. V. Quantitative association of vocal-tract and facial behaviour. *Speech Communication*, 26 (1-2), p. 23-43, 1998.