UTILIZAÇÃO DE SEGMENTOS TRANSICIONAIS HOMORGÂNICOS EM SÍNTESE DE FALA CONCATENATIVA

SANDRA G. KAFKA, FERNANDO S. PACHECO, IZABEL C. SEARA, SIMONE KLEIN E RUI SEARA

LINSE: Laboratório de Circuitos e Processamento de Sinais
Departamento de Engenharia Elétrica
Universidade Federal de Santa Catarina
88040-900 - Florianópolis - SC – Brasil
{kafka, fernando, izabels, klein, seara}@linse.ufsc.br

Resumo - Neste trabalho, propomos a utilização de um único segmento transicional homorgânico representando os demais segmentos de uma mesma classe de homorgânicos. O objetivo dessa estratégia é obter uma economia no inventário de unidades necessárias à síntese de fala concatenativa. Resultados experimentais mostraram a eficiência e robustez da abordagem proposta, quando comparada ao desempenho acústico-perceptual da síntese com o banco de unidades completo, não havendo qualquer prejuízo à inteligibilidade e/ou à naturalidade da fala sintetizada.

Abstract - This paper proposes the use of a single homorganic transitional segment for representing the other segments of the same homorganic class. The goal of this strategy is to obtain a reduction in the inventory of units, which are needed in concatenative speech synthesis. Experimental results have shown the effectiveness and robustness of the proposed approach as compared with the whole inventory for acoustic-perceptual performance, without any significant loss in both intelligibility and naturalness from the synthesized speech.

Keywords - demisyllable, homorganic transitional segment, inventory of units, concatenative speech synthesis, text-to-speech.

1. Introdução

Atualmente, a abordagem concatenativa em sistemas texto-fala (TTS) é a que tem proporcionado melhores resultados. Nessa técnica, um conjunto de unidades da fala é previamente gravado e armazenado, por exemplo: difones, demissílabas, trifones, etc. A fala é então sintetizada pela simples concatenação dessas unidades. No entanto, um problema ainda não totalmente resolvido em síntese concatenativa é o concernente aos fenômenos co-articulatórios entre segmentos adjacentes. Em geral, cuidados adicionais devem ser tomados a fim de que mudanças espectrais abruptas entre segmentos adjacentes concatenados não causem perda de naturalidade na fala sintetizada. Barbosa (1999) menciona que, mesmo com polifones, não é possível fazer constar no sinal de fala pré-armazenado fenômenos co-articulatórios muito extensos, pois a combinatória levaria a um aumento contra-producente no número de unidades concatenantes.

Normalmente, para a formação do repertório de unidades em sistemas de texto-fala, busca-se diferenciar a natureza dos vários tipos de segmentos. No entanto, Bhaskararao (1999) observou que um único segmento transicional para uma vogal seguida de qualquer consoante oclusiva apresentando um mesmo ponto de articulação poderia ser usado para sintetizar diferentes segmentos dentro dessa mesma classe de homorgânicos[1]. Esse autor notou também que a utilização desses segmentos homorgânicos em limites co-articulatórios funcionavam relativamente bem, proporcionando uma redução significativa no número de unidades necessárias ao banco de um sintetizador concatenativo de fala sem nenhum prejuízo à inteligibilidade ou à naturalidade da fala.

Assim, norteados por este mesmo princípio, neste artigo, apresentaremos os fundamentos acústico-perceptuais que nos levaram a utilizar uma única consoante homorgânica como representante de toda uma classe de homorgânicas em segmentos VC[2] para o português brasileiro. Resultados experimentais, mostrando que não houve perda relevante de inteligibilidade e/ou naturalidade na fala sintetizada decorrente dessa estratégia, são apresentados e discutidos.

2. Classificação dos Segmentos Consonantais do Português Brasileiro

Os segmentos consonantais são sons de fala produzidos com algum tipo de obstrução nas cavidades supraglotais, podendo haver obstrução total/parcial da corrente de ar ou fricção. Na caracterização desses sons no português brasileiro, devem ser consideradas as seguintes situações:

- vibração ou não das cordas vocais;
- som nasal ou oral;
- articuladores envolvidos na produção do segmento;
- obstrução ou não da corrente de ar.

[1] Homorgânicos são dois ou mais fonemas que têm um mesmo ponto de articulação, diferenciando-se por outros traços (Dubois, 1973). Por exemplo, [t] e [d] são homorgânicos.

[2] O segmento VC é um dos constituintes das demissílabas e é constituído da parte estável da vogal até a transição para a consoante adjacente. O segmento CV é constituído da parte inicial da consoante até o início da vogal.

Tabela 1. Fonemas consonantais do português brasileiro, segundo Silva (1999)

	Bilabial		Labiodental		Dental-alveolar		Palatoalveolar		Palatal	Velar	
Oclusiva	p	b			t	d				k	g
Fricativa			f	v	s	z	ʃ	ʒ			
Nasal	m				n				ɲ		
Vibrante					r						
Flape					ɾ						
Lateral					l				ʎ		

Quanto à vibração das cordas vocais, temos os sons classificados como vozeados ou não-vozeados. Vozeados são aqueles em que as cordas vocais vibram durante a produção do som, ou seja, os músculos que formam a glote aproximam-se e, quando há a passagem do ar, eles vibram. Exemplos de sons vozeados: [b d v z ʒ g m n l ɾ r][3] (Maia, 1991 e Silva, 1999). Os não-vozeados são aqueles em que não há vibração das cordas vocais, ou seja, os músculos que formam a glote[4] encontram-se separados de maneira que o ar passa livremente (Silva, 1999). São não-vozeados os sons: [p t f s ʃ k][5] (Maia, 1991 e Silva, 1999).

Na caracterização de sons orais e nasais, é considerada a posição da úvula[6]. Se a úvula estiver levantada, fechando o acesso da passagem de ar para a cavidade nasal, temos os sons orais, pois não há ressonância nessas cavidades. Se, ao contrário, a úvula estiver abaixada, o ar passa pela cavidade nasal, produzindo, dessa forma, os sons nasais. São nasais no português brasileiro, as consoantes [m n ɲ][7] e as vogais nasais [ẽ, ẽ, ĩ, õ, ũ]; os demais sons são orais (Maia, 1991 e Silva, 1999).

Quanto à obstrução da passagem de ar nas cavidades supraglotais, ou seja, à maneira como a corrente de ar passa pelos canais supralaríngeos, podemos classificar os sons consonantais, segundo seu **modo de articulação**, como: oclusivos (oclusão total da passagem de ar) [p b t d k g]; nasais (corrente de ar interrompida na cavidade oral, passando pelas cavidades nasais) [m n ɲ]; fricativos (escape de ar feito como uma fricção) [f v s z ʃ ʒ]; vibrante (vibração intermitente da língua ou úvula) [r]; flape (vibração única e momentânea) [ɾ]; laterais (corrente de ar escapando somente pelos lados da cavidade bucal através da interposição da língua no centro da passagem) [l ʎ][8] (Maia, 1991 e Silva, 1999).

Quanto ao tipo de obstrução da cavidade bucal, ou seja, partindo da posição dos articuladores, podemos classificar os sons segundo seu **ponto de articulação** como bilabiais (articulados com os dois lábios) [p b m]; labiodentais (dentes e lábios inferiores) [f v]; dentais-alveolares (língua contra os dentes ou alvéolos [t d s z n r ɾ l]; palatoalveolares (língua ligeiramente mais recuada do que para o ponto alveolar) [ʃ ʒ]; palatais (parte frontal da língua contra o palato duro) [ʎ ɲ]; velares (dorso da língua contra o véu do palato) [k g] (Maia, 1991 e Silva, 1999). Essa classificação pode ser melhor visualizada na Tabela 1.

3. Pistas Transicionais para as Consoantes

Kuehn e Moll (1972), observando os efeitos dos movimentos co-articulatórios, elaborando experimentos perceptuais, verificaram a taxa de identificação correta de consoantes e vogais em sílabas $C_1V_1C_2$ e em seqüências do tipo $C_1V_1V_2C_2$, agrupadas segundo o ponto de articulação, modo e vozeamento. Nesses experimentos, consoantes finais (C_2) junto com a transição vogal para consoante foram parcialmente ou completamente apagadas em diferentes pontos dentro da transição. Seus resultados mostraram que segmentos acústicos precedendo consoantes (V_1 em $C_1V_1C_2$ e V_2 em $C_1V_1V_2C_2$) contêm pistas perceptuais relacionadas principalmente ao ponto de articulação. Pistas perceptuais identificadoras de modo também aparecem, mas não são tão fortes quanto as de ponto.

Delattre et al. (1955), usando padrões formânticos em um sintetizador de fala, verificaram que o ponto de articulação de consoantes com constrições orais, principalmente as oclusivas, fricativas e nasais, é fornecido pela freqüência do segundo formante nas transições entre vogal e consoante. Seus experimentos perceptuais mostraram que a consoante [g] mais evidente é produzida com um segundo formante (F2) em 3000 Hz, o [d] mais evidente com um F2 em 1800 Hz e o mais claro [b] com F2 em 720 Hz, correspondendo assim essas três regiões de freqüência aos *loci* acústicos dessas consoantes.

Segundo, ainda, Wright et al. (1997), tanto as transições no *onset* (início - CV) quando na *coda* (final - VC) de um período de constrição consonantal fornecem pistas do ponto de articulação para consoantes que estão entre vogais (ver Fig. 1). No

[3] Os símbolos [ʒ, r, ɾ] correspondem aos sons das letras "j" como em *jaca*; "rr" como em *carro* e "r" (intervocálico) como em *caro*, respectivamente.
[4] Espaço entre os músculos estriados - cordas vocais.
[5] O símbolo [ʃ] corresponde ao som do "ch" como em *chato*.
[6] Apêndice muscular situado no extremo mais interno do "céu da boca", vulgarmente chamada de campainha (Jota, 1976).
[7] O som [ɲ] corresponde ao som do "nh" como em *unha*.
[8] O som [ʎ] corresponde ao som do "lh" como em *telha*.

entanto, experimentos perceptuais realizados por Fujimura et al. (1978) (*apud* Wright et al. (1997)) mostraram que, quando uma transição CV (consoante para vogal) para um *locus* conflitante com uma transição VC (vogal para consoante) eram ouvidas em conjunto[9], os ouvintes identificavam a consoante como tendo o ponto de articulação correspondente às transições de C para V.

Ficou evidente, desta forma, que poderíamos utilizar uma mesma transição em um mesmo grupo de homorgânicas, por exemplo, para todas as consoantes bilabiais. Então a transição para a consoante [p] em sílabas VC poderia ser a mesma das consoantes [b] e [m], já que elas têm o mesmo ponto de articulação. No entanto, nas transições CV, as características de ponto, modo e vozeamento da consoante específica seriam preservadas, já que nelas as consoantes são, perceptualmente, mais salientes do que em transições VC. Em nossa síntese, por exemplo, a palavra *machado* seria concatenada, no caso de demissílabas, da seguinte forma: {_m ma at XA At do o_}[10]. Conforme observamos, a demissílaba {at} (sílaba VC) se une à demissílaba {XA} (sílaba CV) sem nenhum prejuízo perceptual, já que, sendo os fonemas [t] ({T}) e [ʃ] ({X}) homorgânicos, a pista transicional CV é mais saliente do que a pista transicional VC, eliminando assim a necessidade de uma demissílaba {aX} para se unir a {XA}, no exemplo citado.

4. Constituição de um Banco de Unidades para a Síntese de Fala Concatenativa

Como visto anteriormente, um dos grandes problemas encontrados para a formação de um banco de unidades para a síntese de fala concatenativa é o número de unidades desse banco. Independentemente da filosofia empregada para a geração desse inventário, teremos sempre um grande número de unidades a segmentar. Por exemplo, se forem utilizadas demissílabas, necessitaremos de aproximadamente 2000 unidades, se optarmos por trifones, precisaremos então de aproximadamente 30.000, considerando-se para o português brasileiro um total de 35 fonemas.

Uma compactação no banco de unidades pode ser obtida através de estratégias capazes de reduzir o número de unidades necessárias à formação desse banco sem prejuízo na qualidade da fala sintetizada. Bhaskararao (1999) obteve sucesso sintetizando unidades através de um único segmento transicional de uma vogal seguida de qualquer consoante oclusiva membro de uma mesma classe homorgânica (mesmo ponto de articulação), incluindo também as fricativas homorgânicas.

Este autor utilizou-se de um segmento transicional em posição de coda (consoante final) na unidade VC, ou seja, o segmento [ak], para a geração das seguintes seqüências de vogais e consoantes sem nenhum prejuízo perceptual relevante: [ak], [akʰ], [ag], [agʰ], [ax] e [aχ][11]. Esse sucesso se deve ao fato de que, quando os articuladores se movem da posição de uma vogal para uma oclusiva de pontos de articulação comuns, os movimentos são similares, independentemente de serem vozeados ou não-vozeados.

A partir deste estudo, verificamos que também poderíamos aplicar esta técnica para a redução do inventário de unidades para o português brasileiro. Essa redução, no entanto, foi estendida para um número maior de segmentos.

A demissílaba consiste de dois tipos de segmentos: o primeiro, denotado por CV, constitui-se da parte inicial da consoante até a parte vocálica inicial, e o segundo, denotado por VC, constitui-se da parte estável da vogal até a transição para a consoante adjacente. Nossa estratégia foi então utilizar um mesmo segmento VC como representante de cada classe de homorgânicas. Assim, {Vp} foi representante da classe das bilabiais; {Vt} foi representante das dentais-alveolares e das palatoalveolares; {Vf}, representante das labiodentais e {Vk}, representante das velares (ver Tabela 2).

Esta estratégia só pôde ser utilizada para os segmentos do tipo VC, pois suas pistas perceptuais são bem menos evidentes do que as CV. Ou seja, cada sílaba sintetizada é composta de segmentos VC e CV, que juntos não conflitam quanto aos *loci*. No entanto, a consoante que vai diferenciar o modo de articulação em questão, caracterizando a consoante sintetizada, é a do segmento do tipo CV. Nesse caso, cada uma das consoantes em segmentos CV deve representar seu vozeamento, modo e ponto de articulação correspondentes.

O uso da estratégia de aplicação de um único segmento consonantal homorgânico representante de diferentes classes de segmentos nos limites articulatórios vogal/consoante, trouxe uma redução de aproximadamente 75% na quantidade de segmentos do tipo VC para a formação do banco de unidades concatenantes, conforme pode ser constatado na Tabela 3. Salientamos que, dentre as demissílabas, inserimos alguns polifones necessários à construção, por exemplo, de unidades com "s" ou "r" em coda silábica (final de sílaba), como em *arp* na palavra *carpa*.

[9] Por exemplo, na palavra "cata" ['kate], a vogal da sílaba "ca" [ka] traz em suas transições de "c" para "a" (consoante para vogal) e de "a" para "t" (vogal para consoante) informações sobre pontos de articulação conflitantes quanto ao *locus*, já que [k] é velar e [t]é alveolar.

[10] Neste exemplo e nos demais, _ corresponde ao silêncio em início e final de palavra; as letras maiúsculas correspondem ao contexto tônico e as minúsculas, ao contexto átono.

[11] O símbolo [ʰ] corresponde à aspiração das consontes antecedentes e [χ], a uma fricativa velar.

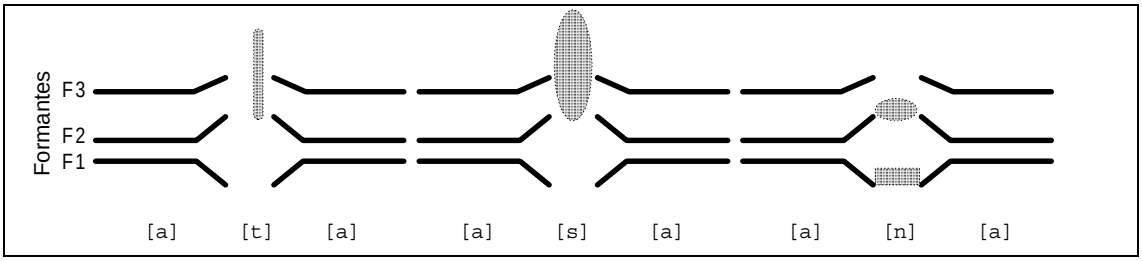

Figura 1. Ilustração esquemática das pistas do ponto de articulação em seqüências VCV, nas quais a vogal é um [a] e as consoantes são uma oclusiva alveolar não-vozeada [t], uma fricativa alveolar não-vozeada [s] e uma nasal alveolar [n], conforme Wright et al (1997).

Tabela 2. Correspondência de segmentos consonantais homorgânicos na síntese concatenativa

Segmentos sintetizados	Correspondência fônica
Vogal + [p]	→ oclusivas bilabiais [p b] → nasal [m]
Vogal + [t]	→ oclusivas alveolares [t d] → vibrante [r] → flape [ɾ] → lateral [l] → nasal [n] →fric. dental-alveolares [s z] → fric. palatoalveolares [ʃ ʒ]
Vogal + [f]	→ fric. labiodentais [f v]
Vogal + [k]	→ oclusivas velares [k g]

Tabela 3. Número de unidades necessárias à sintese concatenativa dos segmentos VC com e sem homorgânicas

Segmentos sintetizados	Sem homorgânicas	Com homorgânicas
[p b m]	3×12 vogais=36	1×12 vogais=12
[t d], [r] [ɾ], [l] [n] [s z] [ʃ ʒ]	10×12 vogais=120	1×12 vogais=12
[f v]	2×12 vogais=24	1×12 vogais=12
[k g]	2×12 vogais=24	1×12 vogais=12
Total de segmentos VC	204	48

5. Resultados Experimentais

Após a concepção das unidades VC, usando um único representante de uma classe de homorgânicas, fizemos testes de escuta informais, apresentando aos ouvintes exemplos de diferentes contextos sintetizados com uma mesma unidade VC.

Os testes de escuta em geral não mostraram nenhuma perda de inteligibilidade e/ou naturalidade causadas por tais artifícios. Esses testes foram divididos em duas seções. A primeira serviu para a verificação da inteligibilidade da fala sintetizada e a segunda, para a verificação da naturalidade dos sons sintetizados com um único segmento transicional para cada classe de homorgânicas.

Os testes constituíram-se de avaliações subjetivas efetuadas por ouvintes que eram falantes nativos do português brasileiro. Esses testes foram efetuados através da escuta de uma lista de palavras sintetizadas, contendo pares mínimos do tipo: {*paca* e *paga*}, {*pala*, *pata*, *passa*}, {*cama*, *capa* e *cabo*}, com uma mesma unidade VC em cada grupo de homorgânicas. Solicitava-se ao ouvinte que escrevesse as palavras escutadas para avaliarmos se houve problemas quanto à inteligibilidade dos sons sintetizados.

Uma segunda lista de palavras contendo os sons em análise inseridos em palavras em fala natural e em fala sintetizada foi apresentada aos ouvintes, pedindo que fosse observado se havia diferenças quanto à naturalidade entre as duas palavras em relação ao som que se estava avaliando. Nos testes, foram utilizadas palavras sintetizadas por bancos de unidades obtidos de dois locutores: um feminino e outro masculino.

Quanto ao item inteligibilidade, em 100% das palavras apresentadas, houve reconhecimento dos pares mínimos avaliados, tanto com a voz feminina quanto com a masculina. Já, no quesito naturalidade, as avaliações subjetivas não obtiveram muito sucesso para um caso específico. Os ouvintes estranharam algumas das palavras que possuíam as consoantes “s” ou “r” em coda silábica (final da sílaba), achando as sintetizadas diferentes das naturais, principalmente quando a fala sintetizada era referente ao banco de unidades do locutor feminino. Isso ocorreu em situações particulares, como em segmentos fricativos não-vozeados em coda silábica, pois o vozeamento da consoante seguinte à coda silábica é, algumas vezes, prejudicado pela transição para uma consoante não-vozeada presente no segmento VC precedente. Um exemplo disso aparece na palavra *vesgo*. Como, nesse caso, teríamos o vozeamento assimilatório da consoante [s] em coda silábica, passando então para [z], essa palavra seria melhor sintetizada se tivéssemos ao invés das unidades: {_VE ESk go o_}, as unidades {_VE ESg go o_}. Nos casos em que segue ao segmento VC uma consoante não-vozeada como em *fisco*, esse problema não ocorre. Também naquelas em que não se tem sílaba travada (tipo CVC, por exemplo) não ocorre tal “estranhamento”,

parecendo a palavra sintetizada tão natural quanto a original (não sintetizada).

Os espectrogramas das palavras sintetizadas, utilizando a estratégia de redução de unidades aqui apresentada, já indicavam que teríamos um bom desempenho na síntese, pois não havia nenhum descasamento aparente entre os diferentes formantes. Na Fig. 2, podemos observar os espectrogramas mostrando a concatenação das unidades para formação das palavras *machado* [ma'ʃado] e *bala* ['bale]. Nas Figs. 2(a) e 2(c), temos os espectrogramas dessas palavras em fala natural e, nas Figs. 2(b) e 2(d), em fala sintetizada. A palavra *machado* foi concatenada através das unidades {_ma at XA At do o_} e a palavra *bala*, concatenada com {_BA At la a_}.

A partir destes exemplos, podemos constatar a redução na quantidade de segmentos, pois, ao invés de termos *machado* concatenado com as demissílabas {_ma ax XA Ad do o_}, reduzimos a necessidade de segmentos do tipo {Ad}, substituindo-os por um único representante dessa classe, a dos dentais e palatoalveolares, o segmento {At}. Em *bala*, teríamos sem nossa estratégia as seguintes unidades {_BA Al la a_}. Com nosso artifício, utilizamos no lugar da unidade {Al}, a unidade {At}, representante das dentais e palatoalveolares, a mesma unidade já utilizada na palavra *machado* {_ma at XA At do o_}.

6. Conclusões

Nossos resultados mostraram que a pista do ponto de articulação é robusta o suficiente para suportar a transição para consoantes com diferentes modos de articulação. Como exemplo, temos o segmento consonantal [t] que serviu de ponto de articulação para as dentais-alveolares e as palatoalveolares, em demissílabas tipo VC, sendo C o segmento consonantal da sílaba seguinte.

Com esta estratégia, reduzimos para aproximadamente 25% o número de unidades do banco, passando dos 204 segmentos VC necessários à formação do inventário de demissílabas a apenas 48 unidades. Os testes de escuta mostraram que não houve nenhum prejuízo quanto à naturalidade e/ou inteligibilidade de tais segmentos, o que ratifica a utilização deste artifício para formação de bancos de unidades em sistemas de síntese de fala. Estudos futuros devem ser feitos a fim de verificarmos a validade de tal estratégia aplicada a outros tipos de unidades concatenantes.

Agradecimentos

Agradecemos à Empresa de Telecomunicações DÍGITRO Tecnologia Ltda. pelo financiamento dado a esta pesquisa.

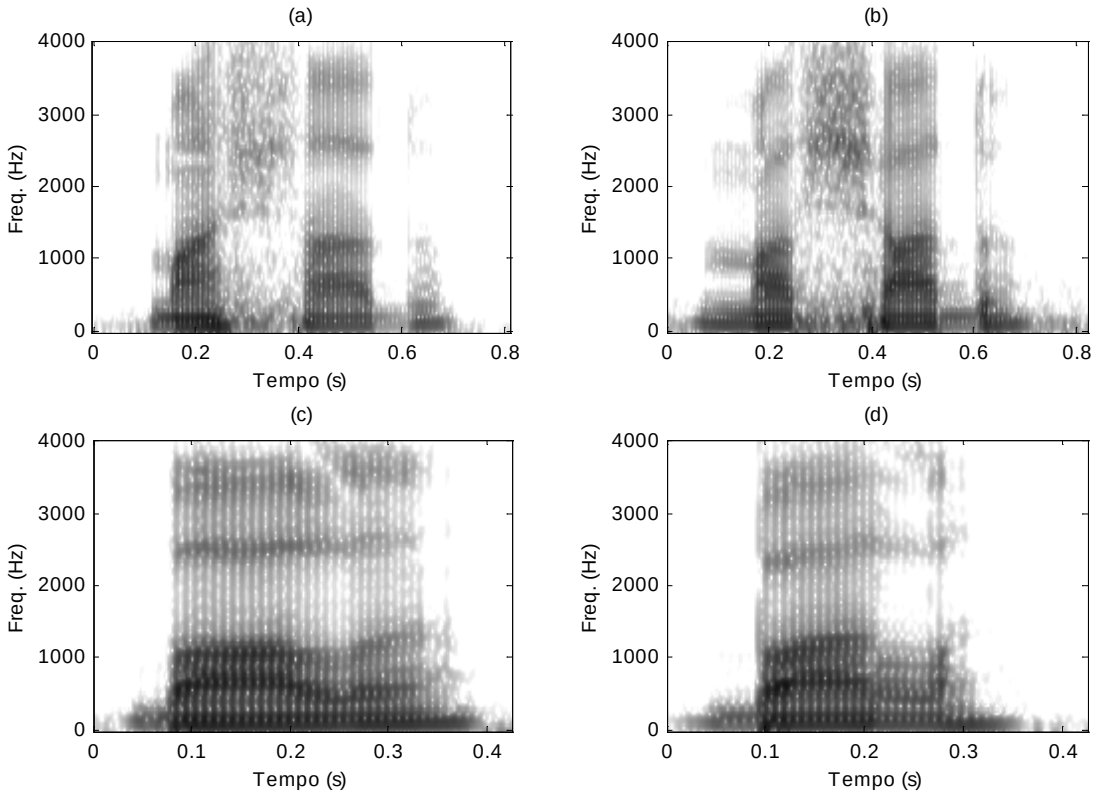

Figura 2. Concatenação, via demissílabas, das palavras *machado* (em (a) fala natural e em (b) fala sintetizada) e *bala* (em (c) fala natural e em (d) fala sintetizada).

Referências Bibliográficas

Barbosa, P. A. (1999). Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia da fala. In: *Estudos de Prosódia* (Scarpa, E. M.). pp. 21-52. Ed. da UNICAMP, Campinas.

Bhaskararao, P. (1999). Subphonemic segment inventories for concatenative speech synthesis. In: *Fundamentals of Speech Synthesis and Speech Recognition* (Keller, E.). Chap. 4, pp. 69-85. John Wiley & Sons, Chichester.

Delattre, P. C., Liberman, A. M. e Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**: 769-73.

Dubois, J. et al. (1973). *Dicionário de Lingüística*. Cultrix, São Paulo.

Fujimura, O., Macchi, M. J, Streeter, L. A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, **21**: 337-345.

Jota, Z. S. (1976). *Dicionário de Lingüística*. Presença, Rio de Janeiro.

Kuehn, D. P. e Moll, K. (1972). Perceptual effects of forward coarticulation. *Journal of Speech and Hearing Research*, **15** (3): 654-664.

Maia, E. M. (1991). *No Reino da Fala: A Linguagem e Seus Sons* 3 ed. Ática, São Paulo.

Silva T. C. (1999). *Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercício*. Contexto, São Paulo.

Wright, R; Frisch, S. e Pisoni, D. B. (1996-1997). *Speech Perception*. Progress Report n. 21 Indiana University. Disponível em: http://www.indiana.edu/~srlweb/publication/manuscript211.pdf.