Análise de modelos de classificação estatística para a segmentação (semi)automática da fala

Bárbara Falcão [1], *Maryualê Malvessi Mittmann*[2]

[1]Universidade Federal de Minas Gerais, ² Universidade do Vale do Itajaí; Centro Univ. FACVEST

barbaraheloha@gmail.com; maryuale@gmail.com

Resumo

O fluxo da fala é segmentado em pequenos fragmentos determinados por fronteiras prosódicas, por motivos cognitivos e linguísticos. Este trabalho utiliza dados de *corpora* de fala espontânea para investigar os parâmetros acústicos associados à percepção de fronteiras prosódicas de valor conclusivo (terminal) e não-conclusivo (não-terminal). A amostra foi segmentada em unidades V-V e um conjunto de parâmetros acústicos extraído a cada unidade, junto da anotação humana sobre a presença de fronteira a cada ponto. Dois modelos de classificação estatística foram utilizados, RF e LDA (*Random Forest* e *Linear Discriminant Analysis*), para gerar modelos de combinações de parâmetros capazes de predizer a realização das quebras percebidas pelos falantes. Os resultados indicam sucesso relativo de ambos os modelos na identificação de fronteiras terminais e não-terminais. O modelo LDA apresentou maior índice de acerto na previsão de fronteiras terminais e não-terminais do que o RF, porém com uma taxa de falsos positivos maior. Verificou-se a validade de utilização de modelos de classificação estatística para a identificação de fronteiras prosódicas; as próximas fases da pesquisa enfocarão o refinamento do treinamento do modelo LDA.

Palavras chave: fronteiras prosódicas, segmentação da fala espontânea; *script* do Praat.

1. Introdução

Este trabalho investiga os correlatos acústicos associados às fronteiras prosódicas em dados de fala espontânea em português brasileiro (PB). Os resultados possibilitarão a criação de um *script* do Praat para detecção automática ou, pelo menos, semiautomática de fronteiras prosódicas. O *script* contribuirá para o melhor entendimento da segmentação da fala, pois utilizará como critério de referência para a detecção automática das fronteiras simultaneamente a percepção humana e parâmetros acústicos. O *script* também auxiliará na compilação de *corpora* de fala espontânea, porque poderá tornar o processo de segmentação da fala mais rápido, poupando-se tempo e esforços humanos, o que pode ser visto como uma contribuição para a linguística de *corpus* em geral.

2. Referencial Teórico

É comumente reconhecido que, durante a comunicação oral entre falantes, o fluxo da fala é segmentado em pequenos fragmentos, também chamados de unidades entonacionais ou prosódicas, marcados por fronteiras prosódicas [1]–[4]. Um dos motivos que possivelmente explicaria a segmentação da fala em pequenos fragmentos delimitados por fronteiras prosódicas é o limite da capacidade da memória de trabalho do ser humano. A segmentação da fala em unidades delimitadas por fronteiras prosódicas também pode ser justificada por razões linguísticas. Neste caso, a segmentação é importante para captação do real domínio das relações linguísticas da fala.

As unidades prosódicas podem ser analisadas segundo perspectivas teóricas diferentes: sintáticas [5], [6], pragmáticas [7], [8] e cognitivas [9], [10]. As fronteiras prosódicas, contudo, podem ser estudadas *per se*, independentemente da perspectiva teórica [11].

Há um acordo geral sobre os parâmetros acústicos que marcam as fronteiras prosódicas. Entre eles estão o *reset* da frequência fundamental (f0) e da intensidade, o alongamento pré-fronteiriço, a mudança na taxa de articulação e de variação da f0, a pausa, a laringalização. Entretanto, ainda não há acordo sobre o peso de cada parâmetro para o estabelecimento de uma fronteira [12] .

O estabelecimento de uma correlação entre fronteiras prosódicas e parâmetros acústicos é extremamente complexo. A complexidade emerge por pelo menos duas causas principais: a) há uma grande quantidade de parâmetros acústicos no fluxo da fala; b) as fronteiras não constituem uma decisão categórica [11], [13]. Enquanto algumas fronteiras prosódicas são muito salientes e são percebidas por (quase) todas as pessoas, outras têm bem menos acordo quanto à sua percepção. Nesses casos, muitos autores acabam tomando decisões com base em razões teóricas, o que gera um efeito de circularidade [14], [15]. Por isso, é importante estudar as fronteiras prosódicas independentemente da análise funcional das unidades que elas geram.

A literatura da área argumenta que há dois tipos de fronteiras, que veiculam a percepção de conclusão ou de continuação do enunciado [16]–[19]. Em um primeiro momento, optou-se por limitar a análise distinguindo apenas entre esses dois tipos de fronteira. O primeiro tipo de fronteira será chamado de fronteira terminal, o segundo de fronteira não terminal.

3. Objetivos

O presente trabalho tem como objetivo correlacionar as fronteiras prosódicas aos parâmetros acústicos. O estabelecimento dessa correspondência tem objetivos teóricos e práticos, dentre eles:

- Contribuir ao melhor entendimento teórico sobre a segmentação da fala, pois as fronteiras prosódicas sempre são marcadas por fenômenos fonéticos;

- Investigar os parâmetros acústicos que orientam a produção e a percepção dos dois tipos de fronteiras prosódicas;
- Possibilitar a criação de um *script* do Praat para segmentação, pelo menos parcial, do fluxo da fala.

4. Metodologia

Nesta seção, apresentamos os dados, o tratamento dos dados e a análise estatística.

4.1. Dados

O trabalho utiliza onze trechos de fala espontânea monológica com excelente qualidade acústica. Os trechos foram extraídos dos *corpora* de Português Brasileiro C-ORAL-BRASIL I e II [20], [21] e são compostos por aproximadamente 200 palavras. Os trechos são distribuídos em três gêneros de *corpora* de fala espontânea da seguinte forma:

a) Três monólogos formais em contexto natural (dois masculinos e um feminino) extraídos do *corpus* C-ORAL-BRASIL II;
b) Quatro monólogos de mídia televisiva (três masculinos e um feminino) extraídos do *corpus* C-ORAL-BRASIL II;
c) Quatro monólogos informais em contexto natural (dois masculinos e dois femininos) extraídos do *corpus* C-ORAL-BRASIL I.

4.2. Tratamento dos dados

Cada trecho foi segmentado autonomamente por quatorze segmentadores *experts*, membros da equipe do Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL) da Universidade Federal de Minas Gerais. Os segmentadores eram apresentados às gravações e aos textos transcritos, sem nenhuma marcação de fronteiras. A tarefa dos segmentadores consistia em marcar os pontos em que as fronteiras prosódicas eram percebidas, inserindo uma barra simples (/) para fronteira não terminal e uma barra dupla (//) para fronteira terminal.

Todos os segmentadores haviam sido treinados e já apresentavam, mesmo que em diferentes medidas, experiência em segmentação prosódica da fala. O treinamento, de duração média de três meses, havia consistido em: a) introdução ao assunto; b) marcação de fronteiras prosódicas em textos enviados pelos membros mais antigos; c) discussão semanal sobre a marcação de fronteiras.

Os trechos foram segmentados em unidades V-V usando o *software* Praat [22]. As unidades V-V são calculadas tomando o tempo que vai do *onset* de uma vogal ao *onset* da vogal sucessiva. A adoção de unidades V-V é justificada pela importância delas para a estruturação rítmica dos enunciados [23].

Foram adotadas as seguintes camadas de anotação do Praat: a) transcrição fonética larga; b) anotação das fronteiras prosódicas marcadas pelo grupo de segmentadores como não-terminais, informando o número de pessoas que marcaram a fronteira; c) anotação das fronteiras prosódicas marcadas pelo grupo de segmentadores como terminais, informando o número de pessoas que marcaram a fronteira; d) anotação do intervalo referente a pausas silenciosas; e) transcrição ortográfica do texto.

Desenvolveu-se uma versão estendida do *script ProsodyDescriptor* [24] para extrair uma série de parâmetros acústicos ao longo do sinal de fala. A versão estendida, denominada *BreakDescriptor,* extrai os parâmetros acústicos em todas as unidades V-V em uma janela centrada em toda fronteira de palavra fonológica, o que inclui as posições percebidas pelos segmentadores como fronteiras e também as não fronteiras. A janela inclui 10 sílabas fonéticas à esquerda e 10 sílabas fonéticas à direita de cada unidade V-V. O *BreakDescriptor* considera como fronteira prosódica as posições em que pelo menos 7 segmentadores perceberam uma fronteira, ou seja, pelo menos 50% de acordo. As demais posições, também localizadas em limites de palavras fonológicas, são consideradas pelo *script* como não fronteiras. No total, para cada unidade V-V, são calculadas 111 medidas acústicas globais e locais, possibilitando assim a análise de fenômenos prosódicos ao longo do sinal acústico. Os parâmetros extraídos são divididos em (a) medidas de taxa de elocução (*speech rate*) e ritmo; (b) medidas de duração normalizada do segmento; (c) medidas de frequência fundamental (f0); (d) medidas de intensidade; (e) medidas de pausa.

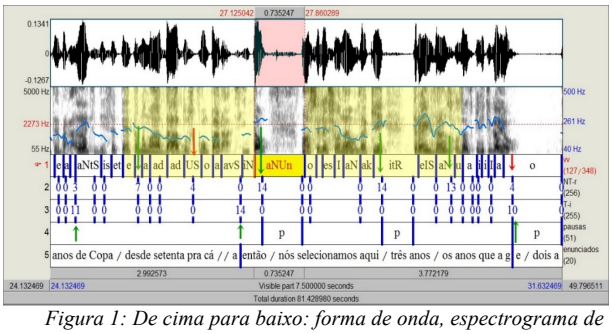

Figura 1: De cima para baixo: forma de onda, espectrograma de banda larga e camadas de anotação no Praat para trecho de fala espontânea. Destaca-se a janela usada para análise de correlato de fronteira do ponto central.

4.3. Análise estatística

Para o treinamento inicial do *script,* seis trechos correspondentes à fala masculina foram retidos.

As medidas extraídas dos trechos de fala espontânea monológica masculina foram submetidas a dois métodos de classificação estatística: Random Forest (RF) e Linear Discriminant Analysis (LDA). Estes modelos de classificação estatística foram utilizados para obtermos modelos de classificação hierárquica baseados na observação das variáveis de previsão, que, no nosso caso, são o conjunto de parâmetros acústicos em todas as fronteiras de palavras fonológicas e em torno delas, quer coincidente com fronteira ou não. Este processo possibilita identificar a combinação de medidas e pesos que melhor explicam a segmentação realizada perceptualmente pelos segmentadores.

Consideramos, para ambos os modelos de classificação, a presença de fronteira e a ausência de fronteira, tanto para as fronteiras terminais, quanto para as fronteiras não terminais. Consideramos também o poder de predição dos dois modelos de classificação. A predição compara a classificação prévia dos grupos com a classificação feita pelo modelo obtido, mostrando acertos e falsos alarmes para um conjunto de dados. Nesta etapa, utilizou-se 70% dos seis trechos de fala espontânea monológica masculina, independentemente se os 70% correspondiam às fronteiras ou às não fronteiras. Os 70% dos dados usados foram escolhidos aleatoriamente. O

treinamento inicial do RF e LDA utilizou uma parte dos 111 parâmetros acústicos extraídos pelo script.

Optou-se pelo treinamento total do modelo LDA, porque ele apresentou melhores resultados no treinamento inicial dos dois tipos de fronteira, reproduzindo melhor a segmentação de base perceptual do grupo.

O treinamento final do modelo LDA foi dividido em duas fases. Na primeira fase, as medidas extraídas pelo *BreakDescriptor* foram retiradas progressivamente do modelo de acordo com o peso hierárquico atribuído a cada uma delas. Desta forma, as medidas menos relevantes hierarquicamente foram eliminadas. Na segunda fase, as medidas extraídas pelo *BreakDescriptor* foram retiradas do modelo com base nos fenômenos fonéticos que elas representam. Assim, eliminou-se os fenômenos fonéticos menos relevantes para a marcação de fronteiras, conforme a literatura da área defende. Esse processo reduz o "ruído" no modelo, aumentando a proporção de acertos e reduzindo a proporção de falsos alarmes.

5. Resultados e discussão

Apresentamos os resultados obtidos durante o treinamento inicial a partir das amostras dos dados.

5.1. Treinamento inicial do modelo RF

Apresentamos abaixo as fronteiras prosódicas identificadas perceptualmente pelos segmentadores e o poder de predição alcançado pelo modelo de classificação RF. O Random Forest utilizou em seus cálculos as ocorrências de presença e ausência de fronteiras terminais e não-terminais, discriminadas na Tabela 1.

Tabela 1: *RF - treinamento inicial - Frequência bruta de identificação de fronteiras terminais e não terminais*

Fronteira	Terminal	Não-terminal
Presença	47	185
Ausência	785	646

O modelo RF identificou 28% das fronteiras terminais. Em outras palavras, o modelo RF apresentou uma convergência de 28% com as fronteiras marcadas como terminais pelos segmentadores. O modelo RF obteve também apenas 1% de falsos alarmes em relação à ausência de fronteiras terminais. Com isso, o modelo RF apresentou uma convergência de 99% com os pontos em que os segmentadores não perceberam fronteiras terminais.

Com relação às fronteiras não terminais, o modelo RF identificou 19% das fronteiras não terminais. Ou seja, o modelo RF apresentou uma convergência de 19% com as fronteiras marcadas pelos segmentadores como não-terminais. O modelo RF obteve 6% de falsos alarmes em relação à ausência de fronteiras não terminais. Assim, o modelo RF apresentou uma convergência de 94% com os pontos em que os segmentadores não perceberam fronteiras não terminais.

5.2. Treinamento inicial do modelo LDA

Apresentamos abaixo as fronteiras prosódicas identificadas perceptualmente pelos segmentadores e o poder de predição alcançado pelo modelo de classificação LDA. As ocorrências de presença e ausência de fronteiras terminais e não terminais identificadas pelo modelo LDA estão discriminadas na tabela 2.

Tabela 2: *LDA - treinamento inicial - frequência bruta de identificação de fronteiras terminais e não terminais -*

Fronteira	Terminal	Não-terminal
Presença	75	142
Ausência	1076	1010

O modelo LDA identificou 57% das fronteiras terminais, ou seja, o modelo LDA apresentou uma convergência de 57% com as fronteiras marcadas pelos segmentadores como terminais. O modelo LDA obteve 2% de falsos alarmes em relação à ausência de fronteiras terminais. Assim, o modelo LDA apresentou uma convergência de 98% com os pontos em que os segmentadores não perceberam fronteiras terminais.

Em relação às fronteiras não-terminais, o modelo LDA identificou 38% das fronteiras, isto é, o modelo LDA apresentou uma convergência de 38% com as fronteiras marcadas pelo grupo de segmentadores como não terminais. O modelo LDA obteve 5% de falsos alarmes em relação à ausência de fronteiras não terminais. Com isso, o modelo LDA apresentou uma convergência de 95% com os pontos em que os segmentadores não perceberam fronteiras não-terminais.

5.3. Treinamento final do modelo LDA

Apresentamos abaixo os resultados obtidos durante o treinamento final do modelo LDA. Iniciou-se a primeira fase do treinamento final do LDA com 111 medidas. Retirou-se progressivamente as medidas menos relevantes hierarquicamente para a classificação estatística até o número de medidas utilizadas no modelo ser reduzido a 21.

5.3.1. Primeira fase do treinamento final - LDA

As fronteiras prosódicas identificadas perceptualmente pelos segmentadores e o poder de predição alcançado pelo modelo de classificação LDA, utilizando 111 medidas está apresentado na Tabela 3.

Tabela 3: *LDA - treinamento final 111 medidas - frequência bruta de identificação de fronteiras terminais e não terminais*

Fronteira	Terminal	Não-terminal
Presença	38	179
Ausência	759	618

O modelo LDA identificou 76% das fronteiras terminais, ou seja, o modelo LDA apresentou uma convergência de 76% com as fronteiras marcadas pelos segmentadores como terminais. O modelo LDA obteve 2,6% de falsos alarmes em relação à ausência de fronteiras terminais. Deste modo, o modelo LDA apresentou uma convergência de 97,4% com os pontos em que os segmentadores não perceberam fronteiras terminais.

O modelo LDA identificou 39% das fronteiras não terminais, ou seja, o modelo LDA apresentou uma convergência de 39% com as fronteiras marcadas pelos segmentadores como não terminais. O modelo obteve 5,1% de falsos alarmes em relação à ausência de fronteiras não terminais. Assim, o modelo LDA apresentou uma convergência de 94,9% com os pontos em que os segmentadores não perceberam fronteiras não terminais.

A seguir apresentamos as fronteiras identificadas perceptualmente pelos segmentadores e o poder de predição alcançado pelo modelo de classificação LDA, utilizando 21 medidas (Tabela 4).

Tabela 4: *LDA - treinamento final 21 medidas - frequência bruta de identificação de fronteiras terminais e não terminais*

Fronteira	Terminal	Não-terminal
Presença	48	174
Ausência	800	636

O modelo LDA identificou 60,4% das fronteiras terminais, ou seja, o modelo LDA apresentou uma convergência de 60,4% com as fronteiras marcadas pelos segmentadores como terminais. O modelo LDA obteve 4,5% de falsos alarmes em relação à ausência de fronteiras terminais. Deste modo, o modelo LDA apresentou uma convergência de 95,5% com os pontos em que os segmentadores não perceberam fronteiras terminais.

O modelo LDA acertou 25,2% das fronteiras não terminais, ou seja, o modelo LDA apresentou uma convergência de 25,2% com as fronteiras marcadas pelos segmentadores como não terminais. O modelo LDA obteve 3,9% de falsos alarmes em relação à ausência de fronteiras não terminais. Com isso, o modelo LDA apresentou uma convergência de 96,1% com os pontos em que os segmentadores não perceberam fronteiras não terminais.

5.3.2. Segunda fase do treinamento final – LDA

A segunda fase do treinamento final do LDA utilizou parte das medidas acústicas, escolhendo-as com base nos fenômenos fonéticos do sinal acústico. Verificou-se a necessidade de considerar medidas diferentes para cada tipo de fronteira a ser analisada, assim, as fronteiras terminais foram identificadas com base em 21 medidas, e as não-terminais com base em 52 medidas (Tabela 5).

Tabela 5: *LDA – 2ª fase do treinamento final - frequência bruta de identificação de fronteiras terminais e não terminais*

Fronteira	Terminal 22 medidas	Não-terminal 52 medidas
Presença	47	180
Ausência	776	595

O modelo LDA identificou 76,5% das fronteiras terminais, ou seja, o modelo LDA apresentou uma convergência de 76,5% com as fronteiras marcadas pelos segmentadores como terminais. O modelo LDA obteve 4,2% de falsos alarmes em relação à ausência de fronteiras terminais. Deste modo, o modelo apresentou uma convergência de 95,8% com os pontos em que os segmentadores não perceberam fronteiras terminais.

O modelo LDA identificou 41,6% das fronteiras não terminais, ou seja, o modelo LDA apresentou uma convergência de 41,6% com as fronteiras marcadas pelos segmentadores como não-terminais. O modelo LDA obteve 5% de falsos alarmes em relação à ausência de fronteiras não terminais. Assim, o modelo LDA apresentou uma convergência de 95% com os pontos em que os segmentadores não perceberam fronteiras não-terminais.

6. Conclusões

Os resultados do treinamento inicial indicam que o modelo *Linear Discriminant Analysis* é melhor para identificar tanto as fronteiras não terminais, quanto as fronteiras terminais, porque o índice de acerto para os dois tipos de fronteira é bem maior, se comparado ao índice de acerto do modelo *Random Forest*.

A segunda fase do treinamento final do LDA sugere a adequação desse método de classificação estatística para a modelagem de um *script* para detecção (semi)automática de fronteiras prosódicas, pois, após a eliminação dos fenômenos fonéticos menos relevantes para a marcação de fronteiras prosódicas, de acordo com a literatura, observa-se que o modelo aproximou-se mais às decisões tomadas pela maior parte dos segmentadores.

Entretanto, para os dois tipos de fronteiras, o LDA apresentou um índice de falsos alarmes relativos à ausência de fronteira superior ao índice do RF. Os resultados também indicam uma maior dificuldade para detectar automaticamente fronteiras não terminais. Por isso, na fase atual, os pontos de erro do modelo LDA estão sendo analisados, buscando tanto o aumento dos acertos para detectar fronteiras, quanto a redução dos falsos alarmes.

Agradecimentos

Agradecemos à equipe do LEEL pelo trabalho de segmentação manual das amostras de fala em unidades V-V e anotação da percepção das fronteiras prosódicas.

Referências

[1] M. Schubiger, “English Intonation: Its Form and Function,” p. 120, 1958.

[2] S. Schuetze-Coburn, M. Shapley, and E. G. Weber, “Units of intonation in discourse: a comparison of acoustic and auditory analyses.,” *Lang. Speech*, vol. 34, no. 3, pp. 207–234, 1991.

[3] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 1998.

[4] B. S. Reed, *Analyzing Conversation: An Introduction to Prosody*. London: Palgrave Macmillan, 2011.

[5] W. E. Cooper and J. Paccia-Cooper, *Syntax and speech*. Cambridge, MA, USA: Harvard University Press, 1980.

[6] E. O. Selkirk, “Sentence Prosody - Intonation, Stress and Phrasing,” in *The Handbook of Phonological Theory*, J. A. Goldsmith, Ed. London: Basil Blackwell, 1995, pp. 550–569.

[7] E. Cresti, *Corpus di Italiano parlato*, vol. 1. Firenze: Accademia della Crusca, 2000.

[8] B. S. Reed, “Prosody, syntax and action formation: Intonation phrases as ‘action components,’” in *Prosody and Embodiment in Interactional Grammar*, P. Bergmann, J. Brenning, M. Pfeiffer, and E. Reber, Eds. Berlin: Mouton de Gruyter, 2012, pp. 142–170.

[9] W. Croft, “Intonation units and grammatical structure,” *Linguistics*, vol. 33, no. 5, pp. 839–882, 1995.

[10] J. Bybee, *Language, usage and cognition*. Cambridge: Cambridge University Press, 2010.

[11] D. Barth-Weingarten, *Intonation Units Revisited: Cesuras in talk-in-interaction.* Amsterdam: John Benjamins, 2016.

[12] P. Auer, “Zum Segmentierungsproblem in der Gesprochenen Sprache,” *InLiSt - Interact. Linguist. Struct.*, vol. 49, no. November 2010, pp. 1–19, 2010.

[13] D. Bolinger, “Around the edges of language,” in *Intonation: Selected Readings*, D. Bolinger, Ed. Harmondsworth: Penguin, 1972, pp. 19–29.

[14] G. Brown, K. L. Currie, and J. Kenworthy, *Questions of intonation*. London: Croom Helm, 1985.

[15] B. Peters, K. J. Kohler, and T. Wesener, “Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache,” *Prosodic Struct. Ger. Spontaneous Speech*, no. 35a, pp. 143–184, 2005.

[16] L. Pike, “The intonation of American English,” University of Michigan, 1945.

[17] J. Pierrehumbert, “Phonetics and phonology of English intonation,” Massachusetts Institute of Technology, 1980.

[18] E. Schegloff, "Reflections on studying prosody in talk-in-interaction.," *Lang. Speech*, vol. 41, no. 3, pp. 235–263, 1998.

[19] B. S. Reed, "Turn-final intonation in English," in *Sound Patterns in Interaction*, E. Couper-Kuhlen and C. Ford, Eds. Amsterdam: John Benjamins, 2004, pp. 97–118.

[20] T. Raso and H. Mello, Eds., *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: UFMG, 2012.

[21] T. Raso and H. R. Mello, Eds., *C-ORAL-BRASIL II: Corpus de referência do português brasileiro falado formal*.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2015.

[23] P. A. Barbosa, "At least two macrorhythmic units are necessary for modeling brazilian portuguese duration," in *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustic*, 1996, pp. 85–88.

[24] P. A. Barbosa, "Semi-automatic and automatic tools for generating prosodic descriptors for prosody research," in *Proceedings of the Tools and Resources for the Analysis of Speech Prosody*, 2013, vol. 13, no. 2, pp. 86–89.