DESENVOLVIMENTO E ANÁLISE DE MODELOS DE DETECÇÃO AUTOMÁTICA DE FRONTEIRAS PROSÓDICAS NA FALA ESPONTÂNEA

Development and analysis of models for the automatic detection of prosodic boundaries in spontaneous speech

[1] TEIXEIRA, Bárbara[*]
[2] RASO, Tommaso
[3] BARBOSA, Plínio

[1] Universidade Federal de Minas Gerais
[2] Universidade Federal de Minas Gerais
[3] Universidade Estadual de Campinas

Resumo: *A fala é segmentada em unidades entonacionais marcadas por fronteiras prosódicas por motivos cognitivos e linguísticos. Este trabalho tem como objetivo: investigar os parâmetros fonético-acústicos que orientam a produção e percepção de fronteiras prosódicas; desenvolver uma ferramenta de detecção automática das fronteiras. Duas amostras de trechos de fala espontânea monológica masculina foram segmentadas em unidades entonacionais por dois grupos de segmentadores treinados. As fronteiras prosódicas percebidas pelos segmentadores foram anotadas como terminais (conclusivas) ou não-terminais (não-conclusivas). Um script do Praat foi utilizado para extrair automaticamente parâmetros fonético-acústicos ao longo do sinal sonoro. Desenvolveu-se um treinamento de modelos compostos pela combinação de múltiplos parâmetros destinados à identificação automática de fronteiras marcadas pelos segmentadores. Utilizou-se o algoritmo Linear Discriminant Analysis (LDA) e considerou-se como fronteira prosódica posições em que pelo menos 50% dos segmentadores indicaram uma fronteira do mesmo tipo. O modelo de detecção automática de fronteiras terminais apresenta uma convergência de 80% em relação às fronteiras terminais marcadas pelos segmentadores na amostra I. Para as fronteiras não-terminais, foram obtidos três modelos de classificação estatística. Juntos, os três modelos apresentam uma convergência de 98% em relação às fronteiras não-terminais indicadas pelos segmentadores na amostra I. Os modelos foram validados posteriormente na amostra II. Os resultados da validação indicam que o desempenho do modelo dedicado às fronteiras terminais é de 75% acerto na segunda base de dados. Os modelos para as fronteiras não-terminais identificam 88% das fronteiras não-terminais marcadas na amostra II.*
Palavras-chave: *Fronteiras prosódicas; Detecção automática; Fala espontânea; Segmentação da fala.*

Abstract: *Speech is segmented into intonational units marked by prosodic boundaries for cognitive and linguistic reasons. This work aims: to investigate the phonetic-acoustic parameters that guide the production and perception of prosodic boundaries; to develop an automatic tool for detect prosodic boundary. Two samples of male spontaneous speech were segmented into intonational units by two groups of trained annotators. The prosodic boundaries perceived by the annotators were noted as terminal (conclusive) or non-terminal (non-conclusive). A Praat script was used to automatically extract phonetic-acoustic parameters along speech signal. A training of models composed by the combination of multiple parameters destined to the automatic identification of boundaries marked by the annotators was developed. The Linear Discriminant Analysis (LDA) algorithm was used and positions at which at least 50% of the annotators indicated a boundary of the same type were considered as prosodic boundary. The terminal model shows a convergence of 80% in relation to the terminal boundaries marked by the annotators in sample I. For the non-terminal boundaries, three statistical classification models were obtained. Together, the three models show a 98% convergence in relation to the non-terminal boundaries indicated by annotators in sample I. The models were validated later in sample II. The results of the validation indicate that the performance of the model dedicated to the terminal boundaries is 75% correct in the second database. The models for non-terminal boundaries identify 88% of the non-terminal boundaries marked in sample II.*
Keywords: *Prosodic boundaries; Automatic detection; Spontaneous speech; Speech segmentation.*

1 Introdução

Diversos estudos linguísticos mostram que uma das funções da prosódia consiste em segmentar unidades entonacionais e estabelecer suas respectivas fronteiras prosódicas. Tais unidades indicam agrupamentos relevantes do ponto de vista comunicativo da linguagem e as fronteiras são marcadas por parâmetros fonético-acústicos usados naturalmente pelos falantes para organizar o discurso e promover a efetiva comunicação oral. Este trabalho visa desenvolver e analisar modelos de detecção automática de fronteiras prosódicas em fala espontânea. A ideia da ferramenta é funcionar a partir de dois critérios relacionados entre si. Os critérios são os parâmetros fonético-

[*]Correspondência dirigida para: barbaraheloha@gmail.com

acústicos extraídos automaticamente em conjunto com a percepção de segmentadores treinados para identificar perceptualmente a presença de fronteiras prosódicas. O principal propósito deste trabalho é contribuir ao melhor entendimento teórico acerca da percepção humana das unidades entonacionais e suas fronteiras prosódicas. Do ponto de vista prático, este trabalho visa desenvolver uma ferramenta computacional que auxilie a compilação de corpora de fala espontânea de português do Brasil (PB).

2 Referencial teórico

A fala é realizada e percebida em unidades entonacionais delimitadas por fronteiras prosódicas (Schubiger, 1958; Chafe, 1980; Schuetze-Coburn, 1994; Ladd, 2008). As unidades podem ser analisadas funcionalmente segundo perspectivas teóricas diferentes: sintáticas (Cooper e Paccia Cooper, 1980; Selkirk, 2005), pragmáticas (Halliday, 1965; Cresti, 2000; Szczepek Reed, 2012) e cognitivas (Chafe, 1994; Bybee, 2010). No entanto, por se tratar de um fenômeno perceptual que desempenha um papel crucial na compreensão dos textos falados, as fronteiras podem ser estudadas *per se*, independente da perspectiva teórico-funcional (Barth-Weingarten, 2016). Perspectivas perceptuais argumentam que os tipos de fronteiras são associados à percepção de conclusão ou de continuação da unidade (Pike, 1945; Pierrehumbert, 1980; Schegloff, 1998; Swerts, 1994). Em geral, esses trabalhos têm uma visão categórica sobre o fenômeno e a nomenclatura adotada estabelece que as fronteiras que indicam perceptualmente a continuação da unidade são chamadas de fronteiras não-terminais, já as fronteiras que indicam a finalização da unidade são chamadas de fronteiras terminais.

Este trabalho se insere na perspectiva que visa compreender as unidades entonacionais e respectivas fronteiras prosódicas no plano perceptual. De fato, este trabalho visa compreender melhor as configurações de parâmetros fonético-acústicos associados às fronteiras perceptualmente relevantes, pois essas configurações ainda não são completamente compreendidas. Adotou-se o enunciado como unidade de segmentação tal como ele é definido na *Language into Act Theory* (Cresti, 2000; Moneglia; 2005). Esta escolha é motivada por uma teoria pragmática para estudo da fala espontânea. No entanto, a teoria pressupõe que a delimitação do enunciado e unidades internas a ele seja orientada pela percepção de fronteiras conclusivas ou não-conclusivas, conforme trabalhos de base perceptual dedicados ao tema. Assim, a identificação das fronteiras não é condicionada pela teoria e não gera um efeito de circularidade.

3 Metodologia

Foram usados trechos de fala monológica masculina com alta qualidade acústica. Os trechos foram extraídos dos *corpora* C-ORAL-BRASIL (Raso e Mello, 2012; Raso, Mello e Ferrari, em preparação), compreendem fala formal, informal e fala televisiva, são compostos por em média 190 palavras e são organizados em duas amostras (Amostra I; Amostra II). Os trechos foram segmentados autonomamente por dois grupos de segmentadores *experts*. Os grupos que segmentaram as amostras I e II são compostos respectivamente por 14 e 19 segmentadores. Os segmentadores foram treinados anteriormente e já apresentavam experiência em segmentação prosódica da fala. A tarefa dos segmentadores consistia em marcar na transcrição dos trechos posições em que fronteiras foram percebidas, inserindo uma barra simples (/) para fronteira não-terminal e uma barra dupla (//) para terminal. Os trechos foram anotados em cinco camadas de TextGrid do Praat (Boersma e Weenink, 2015). Foram adotadas as seguintes camadas: I) Segmentação em unidades Vogal-Vogal (unidades V-V) e etiquetagem em caracteres ASCII; II) Anotação das fronteiras não-terminais marcadas, informando o número de pessoas que marcaram a fronteira; III) Anotação das fronteiras terminais marcadas, informando o número de pessoas que marcaram a fronteira; IV) Anotação do intervalo referente a pausas silenciosas; V) Transcrição ortográfica.

Um script do Praat foi utilizado para extrair automaticamente diversos parâmetros fonético-acústicos em todas as unidades V-Vs em uma janela centrada em toda fronteira de palavra fonológica. A janela inclui 10 unidades anteriores e posteriores em relação à cada fronteira de palavra fonológica. Considerou-se como fronteira posições em que pelo menos 50% dos segmentadores indicaram uma fronteira da mesma natureza. Assim, nas amostras I e II, posições

de fronteira prosódica são respectivamente aquelas em que pelo menos 7 e 10 segmentadores marcaram uma fronteira do mesmo tipo. As demais posições são consideradas posições de nenhuma fronteira. No total, os dados compreendem 116 fronteiras terminais, 534 não-terminais e 1744 posições de nenhuma fronteira. As medidas extraídas compreendem medidas de: I) Taxa de elocução e ritmo (6 medidas); II) Duração normalizada e suavizada dos segmentos silábicos (34 medidas); III) Frequência fundamental (65 medidas); IV) Intensidade (4 medidas); V) Pausa (2 medidas).

Utilizou-se o algoritmo supervisionado *Linear Discriminant Analysis (LDA)* para realizar o treinamento de modelos de classificação automática das fronteiras. O processo de treinamento consistiu em construir modelos compostos por múltiplos parâmetros fonético-acústicos extraídos pelo script, foi realizado heuristicamente e todas as medidas extraídas foram avaliadas. No caso das fronteiras terminais, foram usadas posições de fronteira terminal (TB), não-terminal (NTB) e nenhuma fronteira (NB) com a categoria TB englobando apenas posições de TB e NO-TB englobando posições de ausência de TB, isto é, NTB em conjunto com NB. Devido a uma maior dificuldade do algoritmo para detectar fronteiras não-terminais, o treinamento destinado a essas fronteiras utilizou posições de fronteira NTB e NB. Em relação às fronteiras não-terminais, o treinamento consistiu em treinar o modelo heuristicamente, eliminar as fronteiras não-terminais identificadas pelo modelo e desenvolver outro processo de treinamento heurístico para detectar as fronteiras remanescentes não classificadas corretamente, quantas vezes fosse necessário refazer este processo. Os dados da amostra I foram diretamente submetidos ao treinamento do algoritmo, os dados referentes à amostra II foram utilizados na validação dos modelos desenvolvidos. Após a validação dos modelos, os dados foram combinados em uma única base de dados e os modelos obtidos foram reaplicados na base de dados completa. Adicionalmente, foram observados os tipos de acordo e desacordo entre segmentação do grupo e predição dos modelos.

4 Resultados

Foram obtidos quatro modelos de detecção automática de fronteiras prosódicas. Um deles é destinado às fronteiras marcadas como terminais pelos segmentadores, os demais se dedicam às fronteiras não-terminais. Abaixo, apresentamos o desempenho alcançado pelos modelos no desenvolvimento e validação deles.

Tabela 1: Modelos de detecção automática de fronteiras

Modelo	Principais parâmetros	Amostra	Etapa	Frequência de fronteiras identificadas	%
TB	Pausa e f0	I	Desenvolvimento	56	80
		II	Validação	34	74
NTB-1	Duração e pausa	I	Desenvolvimento	152	68
		II	Validação	193	66
NTB-2	Speech rate e f0	I	Desenvolvimento	57	25
		II	Validação	55	19
NTB-3	Duração e f0	I	Desenvolvimento	11	5
		II	Validação	10	3

A reaplicação dos modelos na base completa de dados apresentou os resultados dispostos na tabela (2).

Tabela 2: Situações de acordo e desacordo entre segmentadores e modelos aplicando-os na base completa de dados

Desempenho	Marcação do grupo	Predição do modelo	Modelo	Frequência	%

Acordo	TB	TB	TB	90	77,6
Acordo	NTB	NO-TB	TB	352	65,9
Acordo	NB	NO-TB	TB	1680	96,3
Acordo	NTB	NTB	NTB-1	351	65,7
			NTB-2	213	39,9
			NTB-3	121	22,7
Acordo	NB	NB	NTB-1	1289	73,9
			NTB-2	350	20
			NTB-3	86	4,9
Desacordo	TB	NO-TB	TB	25	21,6
Desacordo	NTB	TB	TB	162	30,3
Desacordo	NB	TB	TB	0	0
Desacordo	TB	NTB	NTB-1	101	87,1
			NTB-2	32	27,6
			NTB-3	28	24,1
Desacordo	TB	NB	NTB-1	15	12,9
			NTB-2	73	62,9
			NTB-3	88	75,9
Desacordo	NTB	NB	NTB-1	181	33,9
			NTB-2	289	54,1
			NTB-3	412	77,2
Desacordo	NB	NTB	NTB-1	447	25,6
			NTB-2	648	37,2
			NTB-3	287	16,5

5 Discussão dos resultados

O resultado alcançado pelo modelo TB nas amostras I e II é parcialmente compatível. Com a amostra II, observa-se que há uma queda de 6% de desempenho em relação ao desempenho alcançado com a amostra I. Esse declínio pode ser justificado pelo fato da amostra originalmente utilizada no desenvolvimento do modelo (amostra I) possuir uma grande quantidade fronteiras terminais marcadas com pausa silenciosa, o que não é verdade para a amostra II. Na amostra II, apenas 35% das fronteiras terminais são seguidas por pausa. Em relação à categoria NO-TB, os resultados obtidos pelo modelo TB indicam que aproximadamente 66% das fronteiras não-terminais e 96% das posições de nenhuma fronteira foram corretamente marcadas como NO-TB pelo modelo TB. Em relação às não-terminais, os resultados indicam que o modelo NTB-1, composto principalmente por medidas de pausa e duração normalizada, é o mais explicativo e generalizável. O modelo NTB-2 explica aproximadamente 22% das fronteiras não-terminais, utilizando principalmente medidas de taxa de elocução e articulação. O modelo NTB-3 explica um número de casos bem menor, suas principais medidas são a taxa de picos de saliência duracional das unidades V-Vs em torno das fronteiras. O processo de treinamento buscou investigar subgrupos de fronteiras não-terminais marcados por diferentes configurações de parâmetros e utilizou as posições de NTB e NB, o procedimento consistiu em desenvolver gradativamente modelos NTB à medida que algumas fronteiras não fossem identificadas. A reaplicação dos modelos NTB na base completa de dados usando todas as posições (TB, NTB e NB), não eliminando NTB corretamente identificadas para aplicar o modelo subsequente indica que algumas NTB são reconhecidas pelos três modelos, dois modelos ou apenas por um modelo. Outras, todavia, não foram identificadas por nenhum dos três modelos NTB. Isso deve ao fato de que, a partir do momento em que certas fronteiras foram identificadas pelo modelo NTB-1, durante o treinamento dos modelos, eliminou-se a possibilidade destas fronteiras serem reconhecidas pelos

demais modelos NTB, já que elas foram retiradas dos dados. Em outras palavras, devido à metodologia de empregada durante a fase de desenvolvimento dos modelos, fronteiras NTB corretamente reconhecidas por NTB-1 sequer tiveram a "oportunidade" de serem reconhecidas pelos modelos NTB-2 e NTB-3. Assim como fronteiras identificadas pelo modelo NTB-2 não tiveram a "oportunidade" de serem reconhecidas pelo modelo NTB-3. Quanto às demais situações de acordo entre segmentadores e modelos NTB, dentre os três modelos NTB obtidos, NTB-1 é o modelo cuja composição é melhor para identificar as posições de nenhuma fronteira, identificando corretamente 73,9% das posições de NB.

Quanto aos desacordos entre segmentadores e modelo TB, aproximadamente 21,6% das fronteiras terminais (presença de fronteira terminal) foi incorretamente marcado como NO-TB (ausência de fronteira terminal). Observou-se que isso ocorreu principalmente em posições de TB sem pausa. Observou-se também que cerca de 30% das NTB foram marcadas como TB pelo modelo TB. Neste caso, as fronteiras NTB foram imediatamente seguidas por pausa. O desacordo NB para os segmentadores versus TB para o modelo TB não ocorreu. A frequência de ocorrência do desacordo TB para os segmentadores versus NTB para os modelos NTB-1, NTB-2 e NTB-3 é 161. Com isso, algumas TB foram inadequadamente marcadas como NTB por mais de um dos modelos dedicados às NTB, porque os dados compreendem 116 fronteiras TB. Especificamente, o modelo com maior incidência deste erro é o NTB-1. Os resultados obtidos sugerem, então, que os modelos NTB-2 e NTB-3 são melhores para distinguir TB de NTB. A frequência de ocorrência do desacordo TB para os segmentadores versus NB para os modelos NTB totaliza 176. Neste tipo de desacordo, também há interseções e algumas TB foram marcadas como NB por mais de um dos modelos NTB. Como o modelo NTB-1 obteve a menor frequência de ocorrência do desacordo (TB para o grupo versus NB para o modelo), o modelo NTB-1 parece ter uma melhor capacidade para distinguir TB de NB. Juntos, os três modelos NTB marcaram posições de NTB como NB em 882 posições. Certas posições de NTB foram marcadas como NB por mais de um dos modelos NTB, mas este erro foi mais recorrente nos modelos NTB-3 e NTB-2. Assim, os resultados sugerem que o modelo NTB-1 apresenta uma melhor capacidade para distinguir NTB de NB. Os três modelos NTB marcaram posições de NB como NTB em 1282 posições. Como nos demais desacordos entre segmentadores e modelos NTB, há interseções e algumas posições de NB, de acordo com os segmentadores, foram marcadas como NTB por mais de um dos modelos voltados para as NTB. Apesar do modelo NTB-1 reconhecer a maior parte de NTB, os modelos NTB-2 e NTB-1 parecem ser menos eficientes para identificar posições de NB, pois a maioria das posições de NB marcadas equivocadamente como NTB foram etiquetadas justamente pelos modelos NTB-2 e NTB-1.

REFERÊNCIAS

BARTH-WEINGARTEN, D. *Intonation Units Revised: Cesuras in talk-in-interaction*. Philadelphia: John Benjamins Publishing Company, 2016.

BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer*, 2015.

BYBEE, J. *Language, Usage and Cognition*. Cambridge: CUP, 2010.

CHAFE, W. Discourse, consciousness anda time: The Flow and dsiplacement of Conscious Experience in Speaking and writing. Chicago: University of Chicago Press, 1994.

CHAFE, W. The Deployment of Consciousness in the production of a Narrative. In: _____. (Org.). *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood: Ablex, 1980. p. 9-50.

COOPER, W., PACCIA COOPER, J. *Syntax and Speech*. Cambridge: Harvard Universty Press, 1980.

CRESTI, E. *Corpus di Italiano parlato*. v. 1. Firenze: Accademia della Crusca, 2000.

HALLIDAY, M.A.K. *Speech and Situation*. Londres: University College, 1965.

LADD, R. *Intonational phonology*. 2ed. Cambridge: CUP, 2008.

MONEGLIA, M. The C-ORAL-ROM resource. In: CRESTI, Emanuela; MONEGLIA, Massimo (Org.). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamis. 2005. p. 1-69.

PIERREHUMBERT, J. *Phonetics and phonology of English intonation*. Dissertação. Massachusetts Institute of Technology. 1980.

PIKE, L. *The intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

RASO, T.; MELLO, H. (Org.). *C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informa*l. 1ed. Belo Horizonte: UFMG, 2012.

RASO, T.; MELLO, H.; FERRARI, L. (Org.). *C-ORAL-BRASIL II: corpus de referência do português brasileiro*. (em preparação)

SCHEGLOFF, E. Reflections on Studying Prosody in Talk-in-interaction. *Language and Speech* 41 (3-4): 235-263, 1998

SCHUBIGER, M. *English Intonation: Its Form and Function*. Tübingen: Niemeyer, 1958.

SCHUETZE-COBURN, S. *Prosody, syntax, and discourse pragmatics: Assessing information flow in German conversation*. Ph.D. Univ. da California, Los Angeles. 1994.

SELKIRK, E. Comments on Intonational Phrasing in English. In: FROTA,S.; VIGÁRIO, M. & FREITAS, M.J. (eds.) *Prosodies*, Berlim: Mouton de Gruyter, p. 11-58, 2005.

SWERTS, M. Prosodic Features of Discourse Units. 1994. Thesis (PhD) – Technische Universiteit Eindhoven, 1994.

SZCZEPEK REED, B. Prosody, Syntax and Action Formation: Intonation Phrases and Action Components. In: BERGMANN, P. et al. (eds), *Prosody and Embodiment in Interactional Grammar*, Berlim: Mouton de Gruyter, 2012, p. 142-169.