Sobre a percepção de pausa na fala espontânea

Bárbara Helohá Falcão Teixeira
Programa de Pós-Graduação em Estudos Linguísticos
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
barbaraheloha@gmail.com

Tommaso Raso
Faculdade de Letras
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
tommaso.raso@gmail.com

Plínio Almeida Barbosa
Instituto de Estudos da Linguagem
Universidade Estadual de Campinas
Campinas, Brasil
pabarbosa.unicampbr@gmail.com

Resumo: O trabalho investiga a percepção de pausa dentro de um projeto de detecção automática de fronteiras prosódicas percebidas por segmentadores treinados na segmentação da fala espontânea. O longo do projeto ficou clara a necessidade de construir modelos separados para identificar fronteira com e sem pausa. Por consequência se tornou fundamental definir a pausa. O objetivo deste paper é chegar a uma medida de duração mínima necessária para que um silêncio seja relevante perceptualmente para ouvintes (ou seja, possa ser identificado como pausa), de modo a minimizar os erros em uma detecção automática de pausa, o que necessariamente gera uma fronteira prosódica. Todos os dados são extraídos das seções monológicas do corpus C-ORAL-BRASIL. Os dados utilizados para este trabalho foram segmentados em unidades entonacionais por dois grupos, compostos respectivamente por 14 e 19 segmentadores diferentes. As fronteiras percebidas pelos segmentadores foram marcadas como terminais, não-terminais ou disfluências. Os segmentadores com melhor concordância quanto à percepção das fronteiras anotaram as pausas percebidas nas posições de fronteira. Foi calculado o coeficiente Kappa relativo às marcações de pausas percebidas pelos anotadores. Uma série de variáveis fonético-acústicas foram anotadas com o objetivo de compreender o que contribui na percepção da pausa. As variáveis anotadas são a duração do silêncio, a taxa de articulação e o contexto fonético do silêncio, incluído fatores segmentais e prosódicos. Foi usado o Modelo Linear Generalizado para investigar as variáveis significativas para a percepção de pausa em posição de fronteira terminal, não- terminal e disfluência. Os resultados mostram que alguns fenômenos são importantes apenas para a percepção de pausa em posição de um tipo específico de fronteira. Outros, no entanto, se mostram importantes para a percepção de pausa em todos os três tipos de fronteira. Os resultados também sugerem que a duração ideal para o silêncio ser relevante perceptualmente é 100 ms.

Pausa; percepção; fronteiras prosódicas.

I. INTRODUÇÃO

A duração da pausa silenciosa necessária para a percepção de interrupção de fluxo de fala (pausa) é objeto de discussão na literatura da área. As propostas de duração mínima necessária são diversificadas: 190 ms [1], 95 ms [2], 100 ms [3], 120 ms [4], dentre outras. Atualmente, não há muito conhecimento sobre quais elementos, além da duração do silêncio, podem influenciar a percepção de pausa. Este trabalho analisa a percepção de pausa dentro de um projeto de detecção automática de fronteiras prosódicas percebidas na fala espontânea [5]. Neste projeto, chegamos à conclusão que é necessário criar modelos separados para fronteiras acompanhadas por pausas e para fronteiras sem pausa. Portanto, se torna necessário quantificar a duração mínima ideal do silêncio que possa ser percebido como pausa, de modo a fornecer essa medida para o programa e minimizar os erros na segmentação automática. De fato, a percepção de pausa é um fenômeno, influenciado por outros fatores além da duração do silêncio. O trabalho utiliza a segmentação de referência fruto do projeto detecção automática de fronteira prosódica. Por segmentação de referência entende-se uma segmentação fruto do acordo entre os segmentadores (um grupo de 14 anotadores e outro de 19). A partir desta segmentação foi pedido

a 14 deles que anotassem, sem mudar a segmentação, se percebiam ou não pausa nas posições marcadas como fronteiras. Utilizamos o termo "silêncio" para nos referirmos ao fenômeno físico mensurável e o termo "pausa" para o fenômeno perceptual. As variáveis investigadas são a duração do silêncio, a quantidade de anotadores que perceberam a pausa, o tipo de fronteira prosódica (se terminal, não-terminal ou disfluência) associada ao silêncio, a taxa de articulação, o contexto fonético em volta do silêncio e a presença de *reset* ou *shift* de f0.

II. METODOLOGIA

A. Dados

Os dados compreendem trechos de fala monológica masculina segmentados em unidades entonacionais por dois grupos de segmentadores treinados. Os grupos são compostos por 14 e 19 segmentadores diferentes. As fronteiras percebidas foram marcadas como terminais (TB), não-terminais (NTB) ou disfluências. Os áudios foram anotados em seis camadas do objeto de anotação TextGrid do Praat [6]. Para os fins deste trabalho, a camada relevante é apenas a camada de anotação dos intervalos referentes aos silêncios. Todos os dados são extraídos do corpus C-ORAL-BRASIL I (informal)
[7] e do C-ORAL-BRASIL II, sub-corpora de mídia e formal em contexto natural [8] por um total de 2686 palavras e 18 minutos.

Um *script* [9] foi usado para reconhecer, de acordo com a percepção dos segmentadores, as posições de TB, NTB e disfluência. O *script* considera como fronteira posições em que pelo menos 50% dos segmentadores indicaram uma fronteira do mesmo tipo. Para os casos em que os segmentadores não alcançaram um acordo mínimo de 50% devido aos empates, um segmentador *expert* decidiu se a posição compreende ou não uma fronteira e a natureza desta fronteira. O segmentador *expert* decidiu 19 posições sobre 2514. A Tabela 1 apresenta todas as fronteiras analisadas neste trabalho.

Tabela 1: Posições analisadas

Posições	Frequência
TB com silêncio	115
TB sem silêncio	26
NTB com silêncio	259
NTB sem silêncio	312
Disfluência com silêncio	35
Disfluência sem silêncio	86

B. Anotadores

Os anotadores da tarefa de percepção de pausa foram escolhidos com base no cálculo do coeficiente Kappa de Fleiss [10]. O Kappa observado comparou a anotação individual do segmentador com a anotação fruto do acordo de pelo menos 50% entre os segmentadores na tarefa de percepção das fronteiras. Foram avaliados 19 segmentadores de um dos grupos. Dentre os 19 segmentadores, 15 obtiveram um Kappa acima de 0,8. A seleção final para a tarefa deste trabalho considerou 11 segmentadores desse grupo de 15, mais três anotadores que apresentaram alto acordo entre eles após o período de formação [8]. Portanto, a tarefa de percepção de pausas foi realizada por 14 anotadores.

C. Tarefa de percepção de pausas

Os anotadores receberam os áudios, a transcrição e a segmentação ideal fruto do acordo de pelo menos 50% dos anotadores na tarefa de percepção de fronteira. Na transcrição, as fronteiras TB, NTB e disfluência já estavam marcadas.

A tarefa consistia em ouvir os áudios as vezes que fosse necessário e anotar a eventual percepção de pausa em correspondência das fronteiras anotadas. Foi calculado o coeficiente Kappa [8] referente às marcações de pausas percebidas pelos anotadores. O coeficiente Kappa foi calculado de forma geral considerando todos os 14 anotadores e todas as posições de fronteira.

D. Fatores analisados na tarefa de percepção de pausa

O nível de saliência perceptual das pausas foi medido com base no número de anotadores que marcaram a percepção de pausa em uma mesma posição. Foi estabelecido que as pausas com alto nível de saliência perceptual seriam aquelas percebidas por pelo menos 8 dos 14 anotadores. As pausas percebidas por 7 ou menos anotadores foram consideradas pausas de baixa saliência perceptual. Nosso objetivo era estabelecer a duração mínima de um silêncio perceptualmente relevante, de modo a fornecer essa instrução ao programa de segmentação e gerar a menor quantidade de erro, ao considerar modelos separados com e sem silêncio na fronteira percebida. Todos os silêncios, independentemente da duração e da saliência perceptual, foram analisados. Os fatores considerados como potenciais responsáveis pela percepção de pausa são os seguintes:

1. duração do silêncio;

2. tipo de fronteira (TB, NTB, Disfluência);

3. taxa de articulação dos áudios, calculada dividindo o número total de unidades V-Vs do áudio pela duração dos trechos de fala, isto é, desconsiderando os silêncios;

4. informações sobre o contexto fonético dos silêncios: (a) duração bruta da vogal antes do silêncio; (b) duração bruta da coda imediatamente anterior ao silêncio (se houver); (c) duração bruta da rima imediatamente anterior ao silêncio; (d) duração bruta da consoante imediatamente posterior ao silêncio (se houver); (e) presença ou ausência de coda [w], /R/ ou /S/ antes do silêncio; (f) fone imediatamente posterior ao silêncio.

5. presença ou a ausência de *reset* ou *shift* de f0 entre a fronteira analisada e o início da próxima unidade entonacional.

E. Análise estatística

O coeficiente Kappa relativo à percepção de pausa em todas as posições de fronteira foi de 0,65. O coeficiente Kappa para cada categoria é exposto abaixo.

Tabela 2: Coeficiente Kappa relativo à percepção de pausa em posições de fronteira

Categoria	Kappa
TB com pausa	0,79
TB sem pausa	0,53
NTB com pausa	0,57
NTB sem pausa	0,62
Disfluência com pausa	0,57
Disfluência sem pausa	0,82

Foi utilizada a regressão logística através de um Modelo Linear Generalizado (GLM) para compreender quais elementos têm impacto na percepção de pausa. A variável dependente investigada compreende a proporção de anotadores que

perceberam pausa nas posições de fronteira com silêncio. As variáveis independentes são aquelas apresentadas na seção anterior. Foram desenvolvidos três modelos diferentes para investigar a percepção de pausa em posição de TB, NTB e disfluência com silêncio. Os modelos permitem investigar o efeito das variáveis diversas no nível de saliência perceptual das pausas.

Os fatores relevantes que condicionam a percepção de pausa em posição de TB com silêncio são apresentados na Tabela 3.

Tabela 3: Fatores relevantes para a percepção de pausa em posição de TB com silêncio

Fenômeno	**Coeficiente estimado**	**Valor de p (p-value)**
Taxa de articulação	0,3230	0,005114
Duração do silêncio	0,0004	< 0,0000000000000002
Duração da coda antes do silêncio	0,0007	0,021658
Duração da consoante após o silêncio	0,0089	0,0000679
Presença de coda antes do silêncio - R	-0,8064	0,507475
Presença de coda antes do silêncio - S	-1,1213	0,030538
Presença de coda antes do silêncio - W	-0,7959	0,007443
Reset de f0 - Sim	0,5244	0,000374

Os fatores relevantes que condicionam a percepção de pausa em posição de NTB com silêncio são apresentados na Tabela 4.

Tabela 4: Fatores relevantes para a percepção de pausa em posição de NTB com silêncio

Fenômeno	**Coeficiente estimado**	**Valor de p (p-value)**
Taxa de articulação	0,1224	0,02105
Duração do silêncio	0,0034	< 0,0000000000000002
Duração da vogal antes do silêncio	0,0012	0,00391
Fone imediatamente posterior ao silêncio – Consoante vozeada	0,2186	0,05479

Os fatores relevantes para explicar a percepção de pausa em posição de disfluência com silêncio são apresentados na Tabela 5.

Tabela 5: Fatores relevantes para a percepção de pausa em posição de disfluência com silêncio

Fenômeno	Coeficiente estimado	Valor de p (p-value)
Taxa de articulação	-0,2859	0,02350
Duração do silêncio	0,0016	0,0000402
Duração da vogal antes do silêncio	0,0022	0,02032
Duração da consoante após o silêncio	-0,0158	0,00219
Fone imediatamente posterior ao silêncio – Consoante vozeada	1,4693	0,00087

Abaixo, são apresentadas as pausas percebidas pelos segmentadores.

Tabela 6: Pausas percebidas pelos segmentadores

Posição	Frequência	Alta saliência perceptual	Baixa saliência perceptual
Pausas	409	377 (92%)	32 (8%)
Pausas com duração acima de 100 ms	395	366 (93%)	29 (7%)
Pausas com duração menor ou igual a 100 ms	14	11 (79%)	3 (21%)

A seguir, são apresentadas as pausas com duração entre 50 ms a 90 ms.

Tabela 7: Pausas com duração entre 50 ms e 90 ms percebidas pelos segmentadores

Duração da pausa	Frequência	Alta saliência perceptual	Baixa saliência perceptual
Entre 50 ms e 60 ms	2	2 (100%)	0 (0%)
Entre 61 ms e 70 ms	6	3 (50%)	3 (50%)
Entre 71 ms e 80 ms	3	3 (100%)	0 (0%)
Entre 81 ms e 90 ms	3	3 (100%)	0 (0%)

III. DISCUSSÃO DOS RESULTADOS

A. Percepção de pausa em posição de TB

A percepção de pausa em posição de TB com silêncio é condicionada por oito fatores. As variáveis numéricas significativas são a duração do silêncio, a taxa de articulação, a duração da coda antes do silêncio e a duração da consoante após o silêncio. Os resultados indicam que quanto maior a taxa de articulação, maior é o número de segmentadores que marcam a percepção de pausa em posição de TB. Esta mesma relação ocorre com a duração do silêncio, com a duração da consoante após o silêncio e também com a duração de qualquer coda antes do silêncio. Além disso, variáveis relacionadas ao reset de f0 e à presença de coda são significativas. Quando comparada com a ausência de *reset* de f0, a presença de *reset* de f0 entre a fronteira com silêncio e o início da próxima unidade entonacional favorece a percepção de pausa, aumentando o número de segmentadores que marcam a percepção de pausa em posição de TB com silêncio. A

presença de coda [w], /R/ ou /S/ antes do silêncio é significativa, mas diminui o número de segmentadores que indicam a percepção de pausa.

A percepção de pausa em posição de TB é condicionada por duas variáveis relacionadas às codas. Com a variável duração da coda antes do silêncio, é avaliada a duração de qualquer coda antes do silêncio, ou seja, sem especificar de qual coda se trata. Com a variável presença de coda [w], /R/ ou /S/, é analisado o efeito da presença destas codas especificas antes do silêncio, mas a duração não é observada diretamente. De fato, a duração de qualquer coda antes do silêncio aumenta o número de segmentadores que marcam pausa, porém, a variável que codifica a coda [w], /R/ ou /S/ antes do silêncio diminui este número. Desta forma, os resultados sugerem que a duração de qualquer coda antes do silêncio aumenta o número de segmentadores que indicam a percepção de pausa em posição de TB com silêncio, porém, não é possível afirmar qual coda (se [w]/, /R/ ou /S/) é mais relevante.

B. Percepção de pausa em posição de NTB

A percepção de pausa em posição de NTB é condicionada por quatro fenômenos. As variáveis numéricas significativas são a taxa de articulação, a duração do silêncio e a duração da vogal antes do silêncio. Assim como em TB, quanto maior a duração do silêncio, maior é o número de segmentadores que marcam a percepção de pausa. A mesma relação de proporção é evidenciada com a taxa de articulação e a duração da vogal antes de silêncio.

O fone logo após o silêncio também é importante. Neste caso, quando é analisado o contraste entre a presença de consoante não vozeada e a presença de consoante vozeada ambas após o silêncio, a presença de consoante vozeada, independentemente do modo de articulação, é significativa e aumenta o nível de saliência perceptual da pausa.

C. Percepção de pausa em posição de disfluência

A percepção de pausa em posição de disfluência é condicionada por cinco fatores. A duração do silêncio e, depois, a taxa de articulação. Neste caso, quanto menor a taxa de articulação, maior é o número de segmentadores que marcam a percepção de pausa. O fato dos silêncios diante das disfluências serem percebidos com maior facilidade devido à menor taxa de articulação é uma diferença entre o modelo em discussão e os demais modelos, pois, nos modelos TB e NTB acontece o inverso.

A duração da vogal antes do silêncio é outro elemento significativo. Neste caso, quanto maior a duração da vogal, maior a saliência perceptual da pausa. A duração da consoante localizada após o silêncio é significativa: quanto menor a duração da consoante, maior é o nível de percepção da pausa. A presença de consoante vozeada após o silêncio, independentemente do modo de articulação, é significativa. Assim, para a percepção de pausa em posição de disfluência com silêncio, é importante que a sílaba após o silêncio se inicie com consoante vozeada.

D. Duração mínima necessária para a percepção de pausa

Foram analisadas 409 posições de silêncio. Os resultados mostram que 92% dos silêncios foram percebidos como pausa por pelo menos 8 anotadores. A quase totalidade dos silêncios com duração acima de 100 ms (93%) foi percebida como pausa por pelo menos 8 anotadores. Além disso, os resultados evidenciam que 79% dos silêncios com duração menor que 100 ms foram percebidos como pausa por pelo menos 8 anotadores. Todos os silêncios com duração entre 50 ms e 60 ms foram percebidos como pausa por pelo menos 8 anotadores. Aproximadamente 50% dos silêncios com duração

entre 61 ms e 70 ms foram percebidos como pausa por pelo menos 8 anotadores. Todos os silêncios com duração entre 71 e 100ms foram percebidos por pelos menos 8 anotadores. Desta forma, com exceção dos silêncios com duração entre 61 ms e 70 ms, todos os silêncios com duração entre 50 ms e 90 ms foram percebidos como pausa por pelo menos 8 anotadores.

A análise do contexto fonético das pausas de alta saliência perceptual com duração variando de 50 ms a 90 ms mostrou que 55% das pausas são seguidas por oclusiva não vozeada. Aproximadamente 18% das pausas de alta saliência perceptual com duração entre 50 ms e 90 ms são precedidas por um fone alongado. Cerca de 9% das pausas com alta saliência perceptual e duração entre 50 ms e 90 ms são seguidas por uma palavra perceptualmente proeminente. Aproximadamente 18% das pausas com alta saliência perceptual e duração entre 50 ms e 90 ms não são seguidas ou precedidas por palavras proeminentes, alongamentos ou oclusivas não vozeadas.

IV. CONCLUSÕES

Se analisarmos as pausas de alta saliência perceptual com duração entre 50 ms e 90 ms que são seguidas por oclusiva não vozeada, qualquer consideração sobre a duração mínima necessária para a pausa ser relevante perceptualmente é arbitrária, pois é impossível mensurar o quanto da percepção é atribuível ao silêncio e o quanto é atribuível à oclusão do fone após o silêncio. De fato, a definição da duração mínima ideal para ser relevante perceptualmente não é simples. Todos os silêncios com duração entre 75 ms e 90 ms foram percebidos por pelo menos 50% dos anotadores. Porém, são apenas três casos e não seria adequado estabelecer uma duração mínima ideal com base em poucos casos. Como 93% dos silêncios com duração acima de 100 ms foram percebidos pela maioria dos anotadores, a duração mínima ideal para a pausa proposta neste trabalho é 100 ms. Com esta proposta de duração, se evitam que sejam considerados pausas silêncios devidos à fase de oclusão das plosivas, cuja duração média se encontra na faixa de 100 ms (veja-se [11]). Isso, a nosso ver, gera uma perda mínima de pausas com duração menor, mas evita que muitos silêncios sejam considerados pausas indevidamente.

O trabalho mostra que a duração do silêncio não é o único elemento relevante para explicar a percepção de interrupção do fluxo de fala, que há diversos outros fatores que condicionam essa percepção, que esses fatores incluem a natureza da fronteira (outro argumento para diferenciar os tipos de fronteira), além da taxa de articulação e o contexto fonético. Nos dados, há até um silêncio de 54 ms que foi percebido como pausa por 11 anotadores. Este caso específico reforça a tese de que a duração do silêncio não é o único elemento relevante, pois este silêncio é o menor encontrado nos dados e, mesmo assim, possui um alto nível de saliência perceptual. Estabelecer uma medida mínima para percepção de pausa a ser fornecida para um segmentador automático é, portanto, uma questão de custo-benefício.

REFERÊNCIAS

[1] Agnello J. A study of intra- and inter-phrasal pauses and their relationship to the rate of speech. Ohio State University, 1963.

[2] Ruder K. Duration of silent interval as a perceptual cue of speech pauses. Perceptual and Motor Skills 1973; 36: 47-57.

[3] Henderson A, Goldman-eider F, Skarbek A. Sequential temporal patterns in spontaneous speech. Language and Speech 1966; 9: 207-216.

[4] Heldner M. Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. The Journal of the Acoustical Society of America 2011; 130: 508-513.

[5] Raso T, Teixeira B, Barbosa P. Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. J. of Speech Sci.;2020; 9(00):105-28.

[6] Boersma P, Weenink D. Praat: doing phonetics by computer: http://www.praat.org/; 2015.

[7] Raso T, Mello H. C-ORAL-BRASIL: *Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG; 2012.

[8] Raso T, Mello H, Ferrari L. *C-ORAL-BRASIL: Corpus de referência do português brasileiro falado formal*, no prelo.
[9] Barbosa P. BreakDescriptor: *Disponível com o autor*; 2016/2019.
[10] Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bulletin; 1971; 76:378–382.
[11] Barbosa P. *Incursões em torno do ritmo da fala*. Campinas: Pontes, 2006.