Gradus
Revista Brasileira de Fonologia de Laboratório

Retomada do tom médio após intervalos de sonoridade

Marcus Vinicius Moreira Martins*
Universidade Federal de Minas Gerais

Waldemar Ferreira Netto
Universidade de São Paulo

Abstract

Mean pitch analysis of locutions has been a strategy for evaluating emotional prosody since the late nineteenth century. The purpose of this paper is to verify if Mean Pitch is a measure that can be taken as reference for other evaluations of $F0$. Mean Pitch and moving Mean Pitch were extracted from narrative recordings divided in two groups: oral narrative (only males) and youtuber (males and females). The recovered $F0$ values were compared with the Mean Pitch and moving Mean Pitch measurements when speech interruptions occurred. The expected ratio between these values was 1. Although there was some differentiation between the groups, the ratios obtained ranged from 0.98 to 1.04, with a standard deviation ranging from 0.04 to 0.05. These results point to the fact that after speech interruptions, the speakers tend to recover the Mean Pitch values that preceded this interruption. The results corroborate the hypothesis that Mean Pitch is a safe reference measure for the evaluation of Fo variations.

*** Corresponding author**

E-mail marcusvmmartins@gmail.com

Address Departamento de Letras Clássicas e Vernáculas
Faculdade de Filosofia, Letras e Ciências Humanas
Universidade de São Paulo
Av. Prof. Luciano Gualberto, 403 - Butantã
05508900 - São Paulo, SP – Brazil

Article history
Received 2019-06-14
Revised 2019-09-11
Accepted 2019-11-22
Published 2019-12-31

Keywords:
prosody
intonation
mean pitch

Resumo

A análise da média global das locuções é uma estratégia para a avaliação da prosódia emocional. Esta pesquisa tem o propósito de verificar se o Tom Médio pode ser referência para outras avaliações de $F0$. De gravações de narrativas, extraíram-se o Tom Médio e o Tom Médio móvel de $F0$ de sujeitos agrupados. Compararam-se os valores de recuperação de $F0$ em relação a essas medidas quando ocorriam interrupções de sonoridade na fala. A razão esperada seria 1. Ainda que houvesse alguma diferenciação entre os grupos, as razões oscilaram entre 0.98 e 1.04, com desvio-padrão entre 0.04 e 0.05. Esses resultados apontaram para o fato de que após interrupções de sonoridade, os locutores tendem à recuperação dos valores do Tom Médio que antecediam essa interrupção. Os resultados corroboram a hipótese de que o Tom Médio é uma medida de referência segura para a avaliação das variações de $F0$.

Palavras-chave: Prosódia; Entoação; Tom médio.

Introdução

A análise da média variação global da frequência fundamental da fala tem sido constantemente tratada como uma variável acústica adequada para a avaliação da prosódia. Nos séculos XVIII e XIX, Spencer[1] e C. Darwin[2] já formulavam a hipótese de que a variação global de $F0$ de uma locução nos permite reconhecer o estado emocional do falante. No entanto, o que se nota é que os métodos para se estabelecer essa variação global não têm se realizado de forma unânime entre os pesquisadores. Neste artigo, apresentaremos uma alternativa para essa questão, a partir do conceito de Tom Médio (TM), por nós desenvolvido. Para isso, analisaremos o *pitch reset,* isto é, a retomada de *pitch* após a perda de sonoridade, utilizando dois princípios estatísticos:

(i) o de média global; e

(ii) o de média móvel

Na primeira parte do artigo, discutiremos o tratamento da análise da variação feito na literatura, seguido do conceito de interrupção da sonoridade da fala; na segunda parte, apresentaremos o modelo ExProsodia, que orienta esta pesquisa, e, por fim, a discussão de métodos e resultados.

[1] Spencer, "The origin of music" (1890).

[2] Darwin, *A expressão das emoções no homem e nos animais* (2000).

Estudos preliminares - análise da variação de $F0$

O cálculo da variação global de $F0$ tem sido fonte de constante debate nos estudos fonéticos da entoação. O interesse se justifica, uma vez que a forma adotada para se calcular esse valor se torna também elemento chave na compreensão e explicação de fenômenos entoacionais. Por este motivo, fizemos um recorte específico, a fim de apresentar apenas os autores que se utilizaram do cálculo da média para estabelecer esse valor. Os primeiros autores a dar um tratamento científico para o tema foram Fairbanks e Pronovost,[3] na década de 30. Para eles, uma estimativa da variação global poderia ser alcançada pela média aritmética de seis medidas individuais da variação de *pitch* para cada emoção. Posteriormente, na década de 50, Hanley[4] entendeu que a variação global poderia ser interpretada a partir de medidas da frequência média individual tomadas a cada intervalo de 26s. Nos anos 60, Peterson e Lehist[5] optaram pela frequência fundamental média que estivesse associada com cada núcleo de sílaba tônica. William e Stevens,[6] dez anos depois, ao analisar a locução espontânea de pilotos em situação de estresse muito forte, valeram-se da análise da média de amostras de alguns segundos da fala antes e depois do acidente.

Numa das primeiras tentativas de análise da percepção de emoções pela síntese eletrônica de fala, Brown e seus colegas[7] manipularam arquivos sonoros de locuções completas para suas experiências, reestabelecendo parâmetros a cada 10ms. Apple e seus colegas,[8] procurando automatizar a análise das emoções, usaram do mesmo recurso. Nos anos 90, Daly e Zue[9] optaram por uma abordagem mais intuitiva, solicitando a alguns foneticistas que definissem os estímulos sonoros das locuções como *low*, *high* ou *uncertain*. Slaney e McRoberts[10] fizeram uma análise global da média de cada locução usando medidas de frequência em Hz extraídas automaticamente e posteriormente convertidas em escala logarítmica e, ainda, divididas em três segmentos: inicial, medial e final.

Já nos anos 2000, Kang e seus colegas[11] procuraram compreender a variação emocional na entoação a partir de experimento em que tomavam leituras de palavras com diferentes emoções e avaliavam, dentre outros critérios, a frequência média medida em Hz para cada palavra. Fujisawa e seus colegas[12] questionaram o uso de um valor médio das locuções para a avaliação da entoação da fala associada às emoções, propondo que as dissonâncias de intervalos maiores e menores na fala teriam um impacto mais efetivo na percepção. Paulmann e seus colegas[13] fizeram uma avaliação de $F0$ valendo-se da análise automática da média das frequências estabelecida pelo software PRAAT para cada locução.

Como se pode ver, há uma grande profusão de métodos para

[3] Fairbanks e Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion" (1939).

[4] Hanley, "An analysis of vocal frequency and duration characteristics of selected samples of speech from three American dialect regions" (1951).

[5] Lehiste e Peterson, "Some basic considerations in the analysis of intonation" (1961).

[6] Williams e Stevens, "Emotions and speech: Some acoustical correlates" (1972).

[7] Brown et al., "Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech" (1974).

[8] Apple et al., "Effects of pitch and speech rate on personal attributions." (1979).

[9] Daly e Zue, "Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues" (1990).

[10] Slaney e McRoberts, "Baby ears: a recognition system for affective vocalizations" (1998).

[11] Kang et al., "Speaker dependent emotion recognition using speech signals" (2000).

[12] Fujisawa et al., "On the role of pitch intervals in the perception of emotional speech" (2003).

[13] Paulmann et al., "How aging affects the recognition of emotional speech" (2008).

avaliar a variação global de $F0$ a partir da média. De modo geral, os estudos estão relacionados à análise da fala emotiva e buscam generalizações sobre a dinâmica da variação de frequência. Além disso, é notório que os resultados apresentados corroboram as hipóteses de Darwin e de Spencer, que foram re-enunciadas por Ohala na hipótese do *frequency code*: "$F0$ agudo significa (em geral) pequenez, a atitude não ameaçadora, desejo de a boa vontade do receptor, etc., e $F0$ grave transmite grandeza, ameaça, autoconfiança e autossuficiência." [14] Por sua vez, as análises estruturais tendem a não fazer generalizações sobre a variação de *pitch*, embora note-se uma tendência a adotar como referência para o cálculo de $F0$ unidades fonológicas, como a sílaba e a mora, o que sugere uma relação mais explícita entre Fonética e Fonologia.

[14] Ohala, "An ethological perspective on common cross-language utilization of F_0 of voice" (1984).

Em nosso entender, o desafio metodológico é integrar a dinâmica da variação de $F0$ a elementos que, em tese, estruturam e compõem a prosódia. Ao tratar de questão semelhante, Xu e Wang[15] propuseram a noção de *pitch target*, sugerindo que a implementação da entoação parte de uma intenção do falante em atingir uma variação tonal específica, ou seja, um alvo tonal. Os autores definiram o *pitch target* como a menor unidade que se possa operar articulatoriamente e que esteja funcionalmente associada a tons ou a acentos tonais. Segundo eles, essas unidades seriam atribuídas a unidades segmentais como mora, sílaba ou alguma unidade maior. Nesse caso, qualquer avaliação decorreria da análise dessas unidades previamente definidas.

[15] Xu e Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese" (2001).

A hipótese deste trabalho parte de princípio semelhante. Entendemos que o falante tende a estabelecer um valor tonal de referência, que chamamos de **Tom Médio** (TM), a partir do qual ele fará as variações melódicas necessárias para funções estruturais, como a marcação de foco e a finalização; ao mesmo tempo, o Tom Médio o orientaria na implementação de emoções e atitudes de forma global. Assim, por um lado, os valores do Tom Médio seriam uma manifestação primária do *frequency code* proposto por Ohala e, por outro, seriam uma referência para o estabelecimento de elementos estruturais da entoação.

Interrupções de Sonoridade: *pitch reset* e pausas

É importante destacar que o conceito de *pitch reset* não tem uma definição canônica na literatura linguística; desta forma, neste artigo assumimos que ele seja o resultado da diferença de *pitch* entre duas unidades tonais adjacentes. Hirschberg e Pierrehumbert[16] entendem que a variação sistemática do *pitch range* é um marcador acústico importante para a segmentação do discurso, pois, estabelecida uma nova referência tonal, o falante teria que operacionalizar a elocução a partir deste novo ponto. Neste cenário, o valor de *pitch reset* poderia ser compreendido como um marcador

[16] Hirschberg e Pierrehumbert, "The intonational structuring of discourse" (1986).

acústico relevante para análise linguística. Um exemplo tratado pelas autoras, ainda que sem essa terminologia, diz respeito à marcação de foco, o qual bloquearia o processo de declinação frasal da entoação, uma vez que o novo valor de $F0$ estabeleceria um novo *intermediary Phrase* (iP).

No estado atual de nosso modelo, propomos que as pausas, e não a marcação de foco, seja o elemento linguístico mais evidente para a análise do *pitch reset.* A nosso ver, as pausas representam um correlato acústico-articulatório mais preciso do fenômeno a ser estudado, dado que as interrupções de sonoridade começam no momento de inércia da glote até sua retomada de movimento. No entanto, é preciso que esse intervalo de duração seja perceptível ao falante. Para definir essa dimensão perceptual, utilizamos o conceito psicofísico de *just noticeable difference* (JND),[17] que é essencialmente a quantidade mínima de variação que uma grandeza física deve sofrer para que ocorra uma experiência perceptiva notável.

A discussão a respeito da menor duração perceptível dessas interrupções começa ainda na década de 70. Boomer e Dittmann,[18] em trabalho que procurava diferenciar tipos de pausas, verificaram que a partir de 100ms seus sujeitos já eram capazes de fazer essa discriminação. Optaram, entretanto, por definir um valor intermediário de 200ms. Na década de 80, Duez,[19] em seus experimentos com estímulos diversificados quanto à presença ou não da cadeia segmental, encontrou um limiar seguro entre 180ms e 250ms quanto ao reconhecimento das pausas silenciosas por seus sujeitos em falas com presença de cadeia segmental.

Fletcher[20] seguiu a proposição de Duez quanto à duração e assumiu como pausas somente intervalos com duração de 200ms ou mais, que não apresentassem ruídos periódicos ou aperiódicos, que tivessem intensidade maior do que os ruídos de fundo. Em relação a essa caracterização de pausa, Duez,[21] em outro experimento, verificou que os ouvintes poderiam tomar, como pausas, eventos linguísticos de natureza diversificada, como a duração vocálica. Embora o alongamento vocálico precedendo pausa tenha sido predominante, a interação entre a variação de frequência e de intensidade na vogal que precedia a pausa também teve um efeito significativo.

Friedman e O'Connel,[22] comparando a percepção de pausas em línguas diferentes, encontraram como durações significativas um valor médio de 348ms para o inglês e 462ms para o alemão. Lövgren e van Doorn[23] encontraram valores bastante baixos para a definição de um limiar de percepção das pausas, 98ms. No entanto, optaram por durações acima de 212ms. Oliveira,[24] analisando pausas em narrativas, optou por uma duração de 250ms. Silva[25] verificou que a partir de 300ms as pausas são reconhecidas, embora pausas mais longas sejam mais consistentemente tomadas como

[17] Roederer, *Introdução à Física e Psicofísica da Música* (1998).

[18] Boomer e Dittmann, "Hesitation pauses and juncture pauses in speech" (1962).

[19] Duez, "Perception of silent pauses in continuous speech" (1985).

[20] Fletcher, "Some micro and macro effects of tempo change on timing in French" (1987).

[21] Duez, "Acoustic correlates of subjective pauses" (1993).

[22] Friedman e O'Connell, "Pause reports for spontaneous dialogic speech" (1991).

[23] Lövgren e van Doorn, "Influence of manipulation of short silent pause duration on speech fluency" (2005).

[24] Oliveira, "The Role of Pause Occurrence and Pause Duration in the Signaling of Narrative Structure" (2002).

[25] Silva, "A relação entre produção e percepção de pistas prosódicas na segmentação de narrativas espontâneas" (2017).

fronteiras de unidades de informação.

Como se pode ver, as durações perceptíveis de pausa variam, em geral, entre 200ms e 450ms. Para este trabalho, considerando que o objeto em foco não é a definição da duração das pausas silenciosas, mas a variação de frequência que ocorre nas unidades tonais antes e depois delas, entendemos que uma duração mínima de 300ms é suficiente. Em testes-piloto com os dados desta pesquisa, notamos que valores abaixo desse limiar tendem a ocorrer com pouca frequência; além disso, parecem violar de forma consistente o princípio adotado das JNDs, uma vez que podem não ser entendidos como interrupções de sonoridade pelos ouvintes.

A partir destes princípios, estabelecemos que a interrupção de sonoridade ocorre no momento t, em que $F0 < 50$Hz (considerando a inércia da própria glote) e intensidade $I = 0$ rms.[26] Dessa forma, as pausas foram definidas como quaisquer eventos, em que $t_n \geq 300$ms e $F0 < 50$Hz e $I = 0$ rms. Na seção a seguir, apresentamos o modelo que embasa a análise aqui apresentada e que foi utilizada para o desenvolvimento do programa ExProsodia.

[26] *Root mean square*, ou raiz do valor quadrático médio.

O programa ExProsodia

F0, Tom Médio e Tom Médio móvel

Em trabalho realizado anteriormente,[27] descrevemos a hipótese de que a entoação poderia ser tratada como uma série temporal. Séries temporais são um tipo de processo estocástico, cujas variáveis aleatórias $X : \Omega \rightarrow R$ são indexadas por elementos t pertencentes a um intervalo temporal T. [28] Entender um sistema de acumulação como série temporal é vantajoso, pois ao se tomar qualquer variável t como referência, pode se analisar seus estados passados e projetar, por modelamento, sua variação esperada no futuro. A partir dessa hipótese, propomos que os movimentos de longo prazo se caracterizam por uma linha de tendência ascendente ou descendente, estabelecendo uma certa analogia com o movimento linear descendente para a entoação frasal proposto por 't Hart *et al.*,[29] no modelo IPO. O Tom Médio corresponderia a essa tendência global e o Tom Médio móvel corresponderia aos movimentos tonais com tendência limitada a alguns momentos, estabelecidos em t.

[27] Ferreira Netto, "Variação de frequência e constituição da prosódia da língua portuguesa" (2006).

[28] Morettin, *Ondas e Ondaletas: da Análise de Fourier à Análise de Ondaletas de Séries Temporais* (1999).

[29] 't Hart et al., *A perceptual study of intonation: an experimental-phonetic approach to speech melody* (1990).

A partir desse pressuposto, o programa de pesquisa ExProsodia propõe que a avaliação da entoação da fala tome o Tom Médio como referência básica para todas as variações da frequência fundamental. Matematicamente, essa medida é a média acu-

mulada no tempo de todas as porções de fala que estão de acordo com critérios de frequência, intensidade e duração previamente definidos. Para os propósitos desta pesquisa, utilizou-se a equação 1:

$$\mathrm{TM}(t) = \frac{(t-1) * \mathrm{TM}_{(t-1)} + z(t)}{t}$$

Equação 1: Cálculo do Tom Médio. TM é o Tom Médio, z é cada uma das unidades usadas para calcular o Tom Médio e t é a posição do valor de z obtido na série temporal em que ele se encontra.

Na medida em que o Tom Médio é somente uma medida de referência, pressupõe-se que haja uma variação vertical, em semitons (st), que os ouvintes percebam como sendo o mesmo estímulo, também calculada pelo princípio das JNDs. Para tanto, foram definidos experimentalmente um limite superior para o Tom Médio, apresentado na equação 2 como TM_{sup}, e um limite inferior, estabelecido em 3 como TM_{inf}. TM_{sup} e TM_{inf} estão, respectivamente, 3 st acima e 4 st abaixo do Tom Médio.[30]

[30] Consoni, "Aspectos da percepção da proeminência tonal em português brasileiro" (2011); Martins, "Aspectos da percepção e do controle entoacional do Português Brasileiro" (2012); Peres et al., "A influência da cadeia segmental na percepção de variações tonais" (2011); Consoni e Ferreira Netto, "A percepção de variação em semitons ascendentes em palavras isoladas no Português Brasileiro" (2016).

$$\mathrm{TM}_{sup}(t) = \mathrm{TM}(t-1) * 2^{0,25}$$

Equação 2: Limite superior do tom médio, 3st acima do Tom Médio.

$$\mathrm{TM}_{inf}(t) = \mathrm{TM}(t-1) * 2^{-0,33}$$

Equação 3: Limite inferior do tom médio, 4st abaixo do Tom Médio.

A definição das unidades z usadas para a extração do Tom Médio faz-se automaticamente por meio da amostragem de *frames* de 5ms de $F0$ definido pelo SFS.[31] A duração mínima definida para cada unidade z, doravante chamada de unidade básica de entoação (UBI) é de 20ms (4 *frames*). Considera-se que a sequência de *frames* deve estar de acordo com critérios pré-definidos:

[31] *Speech Filing System* v.4.7/Windows SFSWin (2008) https://www.phon.ucl.ac.uk/resource/sfs/. Huckvale et al., "The SPAR speech filing system" (1987).

(i) frequência fundamental $F0$ mínima de 50Hz e máxima de 700Hz;

(ii) intensidade I maior do que 0 rms; e

(iii) oscilações de frequência φ que não ultrapassem 3st.

Uma vez definido um conjunto de 4 ou mais *frames* subsequentes que estejam de acordo com esses critérios, assume-se tratar-se de uma UBI, que será considerada por sua posição na série temporal, por sua duração e pelo valor da média aritmética dos valores de frequência e de intensidade que a compõem. Assim, uma UBI mínima tem como parâmetros: UBI $= t \geq 20\,\mathrm{ms} - 4\,\textit{frames}$; $50 > F0 > 700$Hz; $I >$ rms; $\varphi < 3\,\mathrm{st}$, sendo limitada pela violação de qualquer um dos parâmetros.

A UBI aqui definida é unidade básica da análise automática processada pela rotina ExProsodia. O princípio linguístico adotado para sua definição é de que a variação de $F0$ não se dá exclusivamente dentro dos limites da uma unidade fonológica, como o segmento ou a sílaba, e pode ter dimensões variadas. Deste modo, uma sequência como [da.dʊ] teria somente uma UBI, definida pela sonoridade dos elementos que a compõem. Por sua vez, a palavra [da.tɐ] teria, em tese, duas UBIS: da/-a. Na fig. 1, pode-se ver uma análise processada a partir desse princípio superposta ao espectrograma de um fragmento de fala, em que se ouve "que eu acho que é importante é você entrar nesses lugares e se tornar parte [pausa] da equipe". As 19 UBIs estão indicadas pelos marcadores vermelhos unidos pela linha vermelha pontilhada, que representa a variação de $F0$ da elocução.

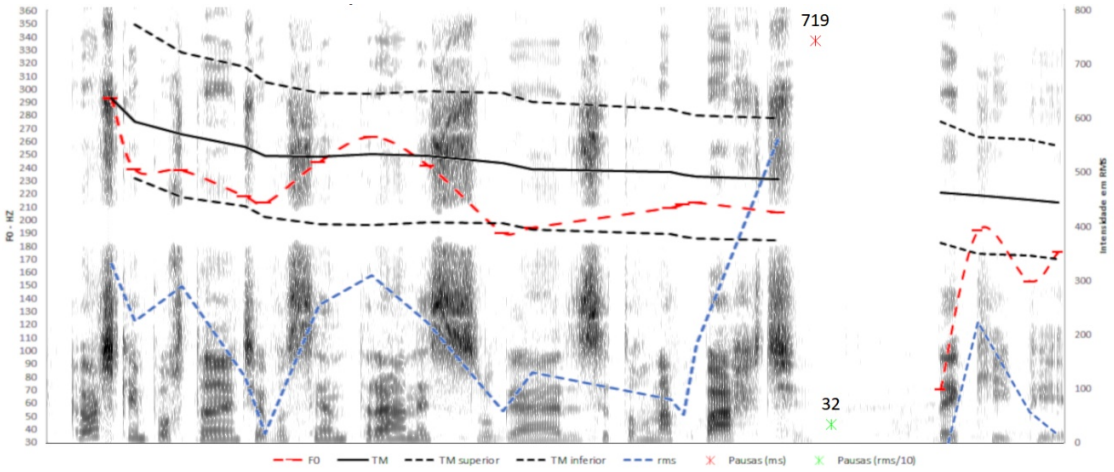

Figura 1: Superposição do espectrograma realizado pelo SFS e pela representação gráfica feita pelo ExProsodia. A linha preta contínua representa o Tom Médio, as linhas pretas tracejadas acima e abaixo do Tom Médio representam F/E_{sup} e F/E_{inf} respectivamente. A linha azul representa as variações de intensidade. O asterisco vermelho mostra a duração da pausa e o marcador verde mostra o valor médio de intensidade multiplicado por 10^{-1} rms.

Embora haja alguma imprecisão na superposição das imagens, extraídas de *softwares* diferentes, na fig. 1 é possível verificar que as UBIs não correspondem exatamente às porções do espectrograma em que ocorrem sons periódicos com muito baixa intensidade ou ruídos decorrentes seja de consoantes ou de fonte externa, de modo que se nota não haver relação unívoca entre UBIs e unidades fonológicas.

Por sua vez, o Tom Médio móvel corresponderia aos movimentos tonais com tendência limitada a alguns momentos t, i.e. em pontos específicos da elocução. Na decomposição da série temporal, a componente Tom Médio móvel corresponde à média móvel, tal como é descrita por Spiegel.[32]

[32] Spiegel, *Estatística* (1985).

$$\mathrm{TM}_{móvel}(t) = \frac{(n/10 - 1) * \mathrm{TM}_{(t-1)} + z(t)}{t}$$

Equação 4: Valores do Tom Médio móvel $\mathrm{TM}_{móvel}$ para cada momento t. n representa o total de UBI definidas até o momento t, TM é o Tom Médio, z é cada uma das unidades usadas para calcular o Tom Médio e t é a posição do valor de z obtido na série temporal em que ele se encontra.

Para textos curtos, o Tom Médio móvel não apresenta variações significativas em relação ao Tom Médio. Mas isso não é verdadeiro

para textos longos. A fig. 2 apresenta o resultado da análise entoacional de uma narrativa analisada neste trabalho com duração de 56s, aproximadamente. Nela podem-se ver, separadamente, as componentes que formam a série temporal da entoação.

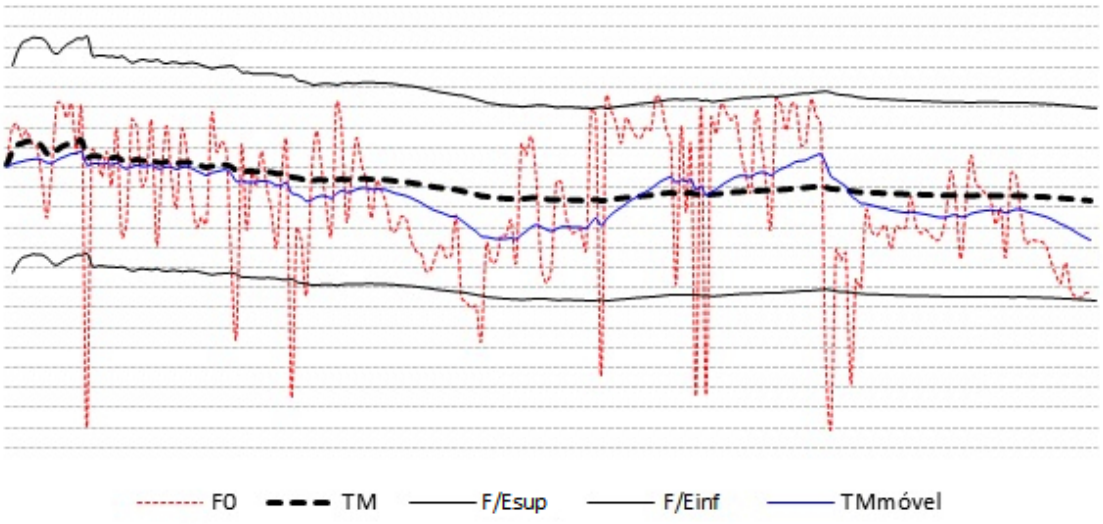

Figura 2: Gráfico feito pelo aplicativo ExProsodia. A linha vermelha representa $F0$, a linha preta tracejada representa o Tom Médio, as linhas pretas contínuas acima e abaixo do Tom Médio representam F/E_{sup} e F/E_{inf}, respectivamente, e a linha azul representa o Tom Médio móvel.

Conforme se pode ver na fig. 2, a comparação das sequências de valores definidos para as séries temporais do Tom Médio e do Tom Médio móvel evidencia os momentos em que houve variações locais no conjunto entoacional da locução. De maneira geral, o Tom Médio mostra uma tendência global e o Tom Médio móvel, tendências momentâneas na entoação. Essa comparação também permite observar os pontos em que houve uma tendência continuada de foco ou ênfase F/E, seja acima, seja abaixo dos limites do Tom Médio. No esquema representado na fig. 3, podemos observar detalhadamente a relação entre Tom Médio, Tom Médio Móvel e $F0$, diante de uma pausa.

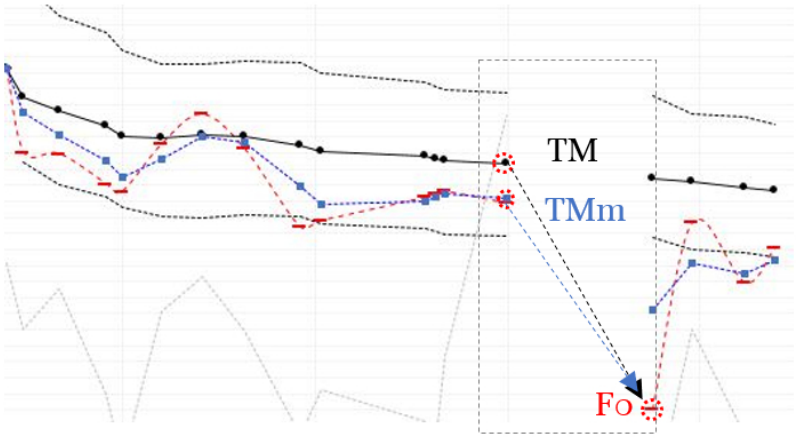

Figura 3: Variações de $F0$ antes e depois de pausa. TM=Tom Médio e TMm=Tom Médio Móvel. A linha contínua preta com marcadores circulares representa o Tom Médio (TM); a linha pontilhada azul com marcadores quadrados representa o Tom Médio móvel (TMm); a linha tracejada vermelha com marcadores retangulares representa $F0$ e a caixa em destaque, uma pausa. O gráfico foi feito a partir da Rotina ExProsodia.

Na fig. 3, nota-se que o Tom Médio tem uma tendência à estabilidade em torno de um valor central, enquanto o Tom Médio

móvel se mostra mais suscetível às variações locais de $F0$. No entanto, é possível identificar que o Tom Médio móvel tende a se manter dentro dos limites laterais marcados pelas linhas pretas tracejadas mesmo naqueles pontos em que $F0$ viola esses limites. Essa tendência à manutenção do Tom Médio móvel no espaço perceptual do Tom Médio indica, em tese, uma correlação entre aquilo que o falante estabelece como parâmetro para elocução como um todo e aquilo que é estabelecido em cada sintagma entoacional, ou seja, o falante tende a manter um Tom Médio global, a despeito das interrupções de sonoridade, retomando a articulação dos tons, a partir de pontos fixados perceptualmente.

A partir disso, para verificar esta hipótese de manutenção do Tom Médio após as interrupções de sonoridade, optou-se por analisar as frequências que ocorrem antes e depois das pausas em duas situações distintas:

a) considerando o último valor do Tom Médio antes da interrupção comparativamente ao primeiro valor de $F0$ após a interrupção; e

b) considerando o último valor do Tom Médio móvel comparativamente ao primeiro valor de $F0$ após a interrupção.

Espera-se que o valor médio da razão entre eles seja 1, na medida em que, para se entender como manutenção do Tom Médio, as unidades devem apresentar uma tendência à igualdade. Nas seções a seguir, apresentamos os materiais e métodos adotados.

Descrição dos dados

Um dos pressupostos que orientou a coleta de dados foi a diversidade de gêneros textuais e dialetos, a fim de se testar a robustez do cálculo proposto. Para esta pesquisa preliminar, coletaram-se dados no site YouTube,[33] partindo das palavras-chave "fui assaltada" e "fui assaltado" como recurso de busca. A escolha destas palavras não foi aleatória e, sim, orientada pela natureza dos relatos encontrados após uma série de testes. A busca resultou em um grande número de arquivos de vídeo, gravados na sua maioria por *youtubers*. Desse conjunto, foram selecionados 15 em que os locutores eram mulheres e 15 em que eram homens. A seleção se baseou na qualidade da gravação, especialmente quanto a ausência de ruídos de fundo e músicas. A extração do som dos vídeos foi feita com o aplicativo MediaHuman Audio Converter.[34] Além dos vídeos, também foram utilizadas 15 gravações de narrativas do ciclo de Lampião, obtidas de pesquisa realizada anteriormente[35] e não balanceadas por sexo/gênero, sendo 11 homens e 4 mulheres.

Todos os arquivos sonoros foram processados usando o Adobe Audition 3.[36] Alguns arquivos passaram por um filtro Dy-

33 https://youtube.com/.

34 *MediaHuman Audio Converter* versão 1.9.6.5 (2011) https://www.mediahuman.com/pt-br/audio-converter/.

35 Conceição et al., "Análise da ênfase prosódica em narrativas orais do ciclo de Lampião" (2016).

36 *Adobe Audition* 3.0.1 build 8347.0 (2012) https://www.adobe.com/br/products/audition.html.

NAMIC EQ, que cobre os sons indesejados sem perder a duração e a frequência fundamental. Também, em alguns casos, foi necessário eliminar vinhetas iniciais e finais, ou intercalação de músicas ou de efeitos sonoros. A duração dos áudios processados e analisados oscilou entre 1 min19 s e 18 min47 s. O cálculo de $F0$ e da curva de intensidade foi realizado pelo *software* SFS por autocorrelação. Posteriormente, extraíram-se as matrizes de frequência e de intensidade. As matrizes foram analisadas pelo aplicativo EXPROSODIA.[37]

[37] FERREIRA NETTO, "ExProsodia" (2010).

Resultados

Dados globais

A partir dos dados apresentados na seção anterior, foram verificadas as diferenças médias, em Hz, obtidas pela razão ρ entre o valor de $F0$ da última UBI antes da supressão de sonoridade (Fof) e o valor Tom Médio da primeira UBI após o intervalo (TM_i), no primeiro procedimento (eq. 5), e entre o Tom Médio móvel (TMm_i), no segundo (eq. 6), formalizados abaixo. Os cálculos estatísticos e os gráficos foram feitos com os *softwares* EXCEL[38] e KYPLOT.[39]

$$\rho_1 = \frac{F0_f}{\mathrm{TM}_i}$$

$$\rho_2 = \frac{F0_f}{\mathrm{TM}m_i}$$

A distribuição que se obteve com a primeira avaliação com todos os dados é representada, na fig. 4, pela superposição das curvas de distribuição dos valores de ρ obtidas nos procedimentos (a) e (b). As curvas foram obtidas a partir do cálculo da média de $\rho 1$ (linha azul) e $\rho 2$ (linha vermelha) para todas as elocuções, tendo em vista que todas as gravações tiveram pelo menos mais de uma interrupção da sonoridade.

[38] *Microsoft Excel 2016 MSO* (16.0.8827.2131) 32 bits (2016) https://products.office.com/pt-br/excel.

[39] *KyPlot* versão 2.0 beta 15 (32 bit) (2001) http://www.kyenslab.com/en/index.html.

Equação 5: Procedimento (a). Razão ρ entre o valor de $F0$ da última UBI antes da supressão de sonoridade ($F0_f$) e o valor Tom Médio da primeira UBI após o intervalo (TM_i).

Equação 6: Procedimento (b). Razão ρ entre o valor de $F0$ da última UBI antes da supressão de sonoridade ($F0_f$) e o valor Tom Médio móvel da primeira UBI após o intervalo ($\mathrm{TM}m_i$).

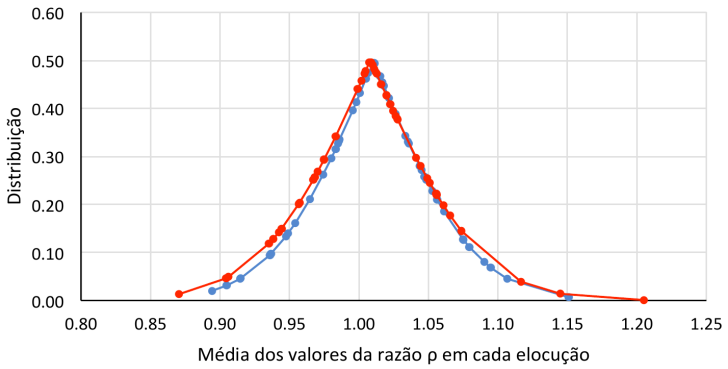

Figura 4: Gráfico de distribuição dos valores da razão ρ_1 e ρ_2 para os todos os dados $n = 45$.

Neste gráfico, pode-se ver que, ainda que haja uma tendência à normalização dos dados em torno da média global obtida, os valores ficaram ligeiramente acima do que era esperado: $\bar{x}_0(1.01) > \bar{x}_e(1)$ ($n = 45$, $\sigma = 0.06$). Os mesmos valores foram encontrados em ambos os procedimentos. Considerando-se que as variações esperadas deveriam ser no máximo 1.19 e mínimo 0.8 — equivalentes a 3 st acima e abaixo do Tom Médio — é possível aceitar que, embora haja uma assimetria negativa (0.47 para o primeiro procedimento e -0.37 para o segundo, com mediana de 1.03 para ambos), as recuperações do Tom Médio tendem a se manter dentro dos limites previstos pela proposta do Tom Médio e seus intervalos laterais, como apontado na seção anterior.

A assimetria negativa aponta para uma pequena concentração dos dados acima da média, como já foi caracterizado, o que indica uma tendência à retomada do Tom Médio ligeiramente acima da tendência central estabelecida globalmente na elocução, o que também pode ser explicado pela mediana. Woods e colegas[40] afirmam que a mediana é um indicador robusto para especificar um valor "típico" que não sofre influências de valores extremos e que pode ser usado como um descritor para um grupo inteiro (p. 29 e 32). Pode-se pensar, portanto, que há uma tendência global para retomar o valor do Tom Médio e do Tom Médio móvel com uma variação ascendente em relação ao que seria esperado, mas ainda dentro dos limites do Tom Médio, como estabelecidos nas equações 2 e 3.

[40] Woods et al., *Statistics in language studies* (1986).

Dados separados por grupos de locutores

Embora seja possível caracterizar uma variação ascendente típica acima do valor esperado, ao se isolarem os grupos de falantes, a relação que se estabelece não é a mesma. Procedendo de maneira semelhante ao que se fez anteriormente para os dados globais, fez-se, para cada grupo de locutores, um gráfico com a mesma superposição de duas curvas: uma com a razão média entre o valor de $F0$ da última UBI antes da interrupção e o valor do Tom Médio UBI imediatamente após a interrupção ρ_1, e outra com Tom Médio móvel imediatamente após a interrupção ρ_2. Neste caso, a divisão de falantes se deu entre

A) *youtubers* do gênero feminino;

B) *youtubers* do gênero masculino; e

C) narrativas do ciclo de Lampião, sem distinção de gênero.

A Fig. 5 representa a média das razões ρ_1 e ρ_2 para cada uma destas condições, com $n = 15$ para cada uma delas.

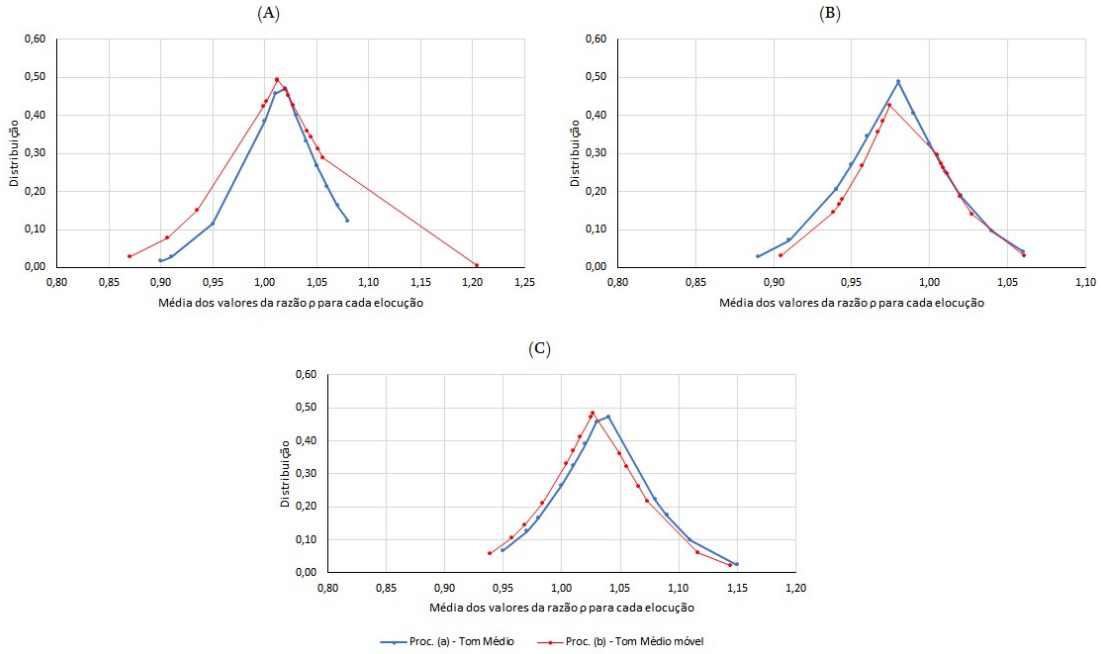

Figura 5: Distribuição das razões entre valores médios de $F0$ antes de interrupção e os valores do Tom Médio e do Tom Médio móvel após interrupção, com $n = 15$ para todos os casos. (A) *youtubers* do gênero feminino; (B) *youtubers* do gênero masculino; (C) narrativas do ciclo de Lampião, sem distinção de gênero.

Como se pode ver, na Figura. 5, em (A), a distribuição dos valores da razão entre Tom Médio móvel e $F0$ é semelhante à que se encontrou para os dados globais, $\bar{x}_0(1.01) > \bar{x}_e(1)$ — com $n = 15$ e $\sigma = 0.06$. A razão entre Tom Médio e $F0$, por sua vez, ficou um pouco acima desse valor, $\bar{x}_0(1.02) > \bar{x}_e(1)$ $\bar{\text{x}}\text{o}(1{,}02) > \bar{\text{x}}\text{e}(1)$ — com $n = 15$ e $\sigma = 0.06$. A mediana do grupo de *youtubers* femininos que se encontrou para essas duas medições foi 1.07, com assimetria negativa de -1.36 para o Tom Médio em relação a $F0$; e 1.06 com assimetria negativa de -1.08 para Tom Médio móvel em relação a $F0$.

Na mesma fig. 5, em (B), a distribuição dos valores da razão entre Tom Médio móvel e $F0$ é semelhante à que se encontrou para os dados globais, $\bar{x}_0(0.98) < \bar{x}_e(1)$ — com $n = 15$ e $\sigma = 0.04$. A razão entre Tom Médio e $F0$ foi semelhante, com $\bar{x}_0(0.98) < \bar{x}_e(1)$ — com $n = 15$ e $\sigma = 0.05$. A mediana do grupo de *youtubers* masculinos é 0.99 com assimetria positiva de 0.05 para Tom Médio móvel em relação a $F0$, e 0.99 para com assimetria positiva de 1.27 para Tom Médio móvel em relação a $F0$.

Ainda na fig. 5, em (C), que representa o grupo de locutores que fizeram narrativas do ciclo de Lampião, a distribuição dos valores da razão entre Tom Médio móvel e $F0$ ficou um pouco acima do que se encontrou para os dados globais, $\bar{x}_0(1.03) > \bar{x}_e(1)$ — com $n = 15$ e $\sigma = 0.06$. A razão entre Tom Médio e $F0$ manteve essa tendência e também ficou um pouco acima do valor esperado, $\bar{x}_0(1.04) > \bar{x}_e(1)$ — com $n = 15$ e $\sigma = 0.05$.

A análise multivariada por *cluster*, comparando todos os dados

de Tom Médio e de Tom Médio móvel entre os três grupos de narradores, apresentou um resultado significativo, que permite estabelecê-los como elementos semelhantes, ainda que os grupos se mantenham diferenciados internamente. O dendrograma da fig. 6 caracteriza bem essa relação:

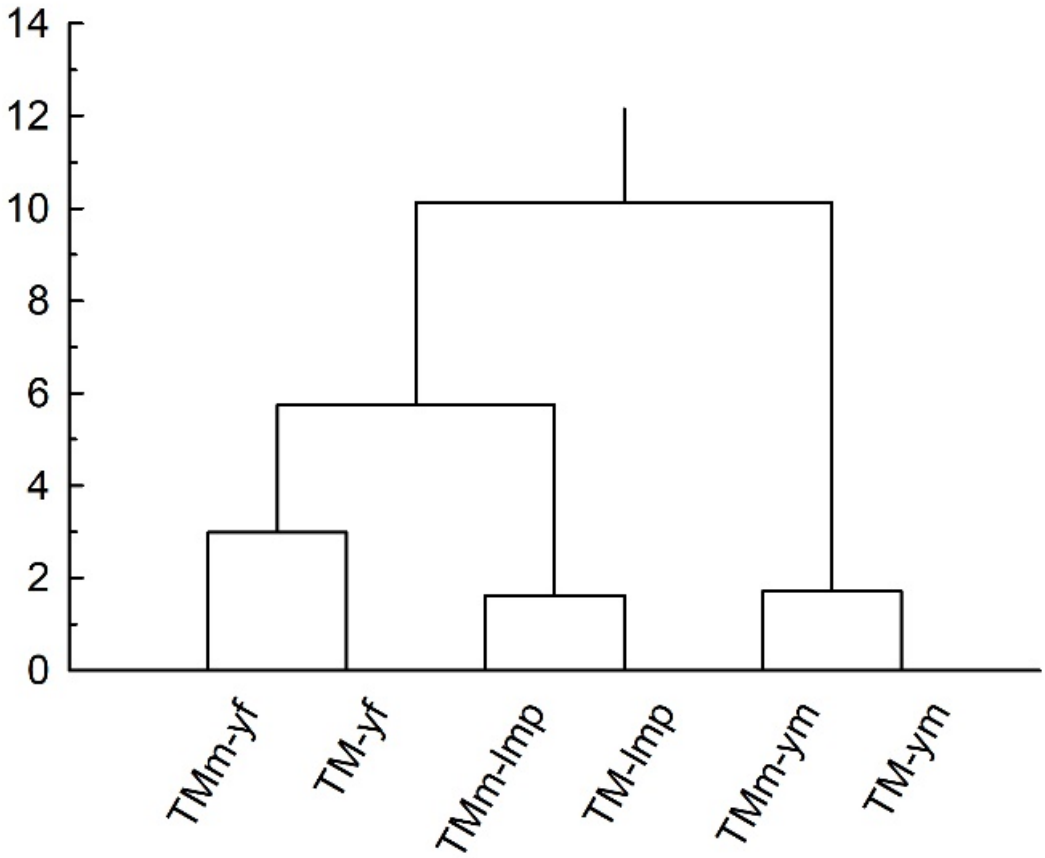

Figura 6: Dendrograma realizado por teste de cluster. TM – Tom Médio, TMm – Tom Médio móvel, yf – *youtubers* do gênero feminino, ym – *youtubers* do gênero masculino e lmp – narrativas do ciclo de Lampião.

O dendrograma realizado por teste de *cluster* mostra a tendência ao agrupamento das categorias. A extensão dos traços verticais mostra o maior ou menor distanciamento entre as categorias que se associam. Desta forma, como se pode ver na fig. 6, ocorre maior identidade entre as variáveis Tom Médio móvel e Tom Médio nos três grupos de falantes do que entre os grupos propriamente. A relação que se estabelece entre Tom Médio móvel e Tom Médio entre *youtubers* femininos — respectivamente TMm-yf e TM-yf — é menor do que a mesma relação dos demais grupos. Também se nota que locutores de narrativas do ciclo de Lampião — TMm-lmp para Tom Médio móvel do ciclo de Lampião e TM-lmp para Tom Médio do ciclo de Lampião — e *youtubers* femininos aproximam-se mais entre si, isolando a categoria dos *youtubers* masculinos — TMm-ym para Tom Médio móvel masculino e TM-ym para Tom Médio masculino.

Com a intenção de obter uma avaliação global dos resultados obtidos na análise de *cluster*, procurou-se verificar quais variáveis promovem a aproximação e o isolamento entre esses grupos,

por meio de da comparação entre os grupos pelo teste ANOVA, o qual teve como resultado a estatística $F0(2.64) > Fe(2.3)$ e $p < 0.05$, apontando para uma diferenciação significativa dos grupos analisados. Uma sequência de testes t e testes F, não pareados, comparando todas as amostras entre si, apresentou os seguintes resultados:

	$t(1.7)$ de $\bar{x}$	P	$F(2.5)$ de σ^2	P
TMm-yf ×TM-yf	0.10		1.89	
TMm-yf ×TMm-ym	1.39		3.29	0.02
TMm-yf ×TM-ym	1.52		2.60	0.04
TMm-yf ×TMm-lmp	0.64		1.77	
TMm-yf ×TM-lmp	0.91		1.73	
TM-yf ×TMm-ym	1.87	0.04	1.74	
TM-yf ×TM-ym	2.00	0.03	1.37	
TM-yf ×TMm-lmp	0.65		1.06	
TM-yf ×TM-lmp	0.97		1.09	
TMm-ym ×TM-ym	0.24		1.27	
TMm-ym ×TMm-lmp	2.57	0.01	1.85	
TMm-ym ×TM-lmp	2.91	0.00	1.91	
TM-ym ×TMm-lmp	2.66	0.01	1.46	
TM-ym ×TM-lmp	2.98	0.00	1.50	
TMm-lmp ×TM-lmp	0.32		1.03	

Quadro 1: Comparação entre os grupos e as médias pelas estatísticas t e F.

No quadro 1, é possível verificar que os grupos de *youtubers* diferenciaram-se um do outro de forma significativa especialmente pela variância, avaliada pelo teste F (condições: TMm-yf × TM-yf e TMm-yf × TM-ym). Nesse quadro, os valores entre parênteses na primeira linha representam os limites para a significância dos valores das estatísticas t e F. Esse fato evidencia que a variação decorre da dispersão dos dados em relação às médias.[41]

[41] Woods et al., *Statistics in language studies* (1986).

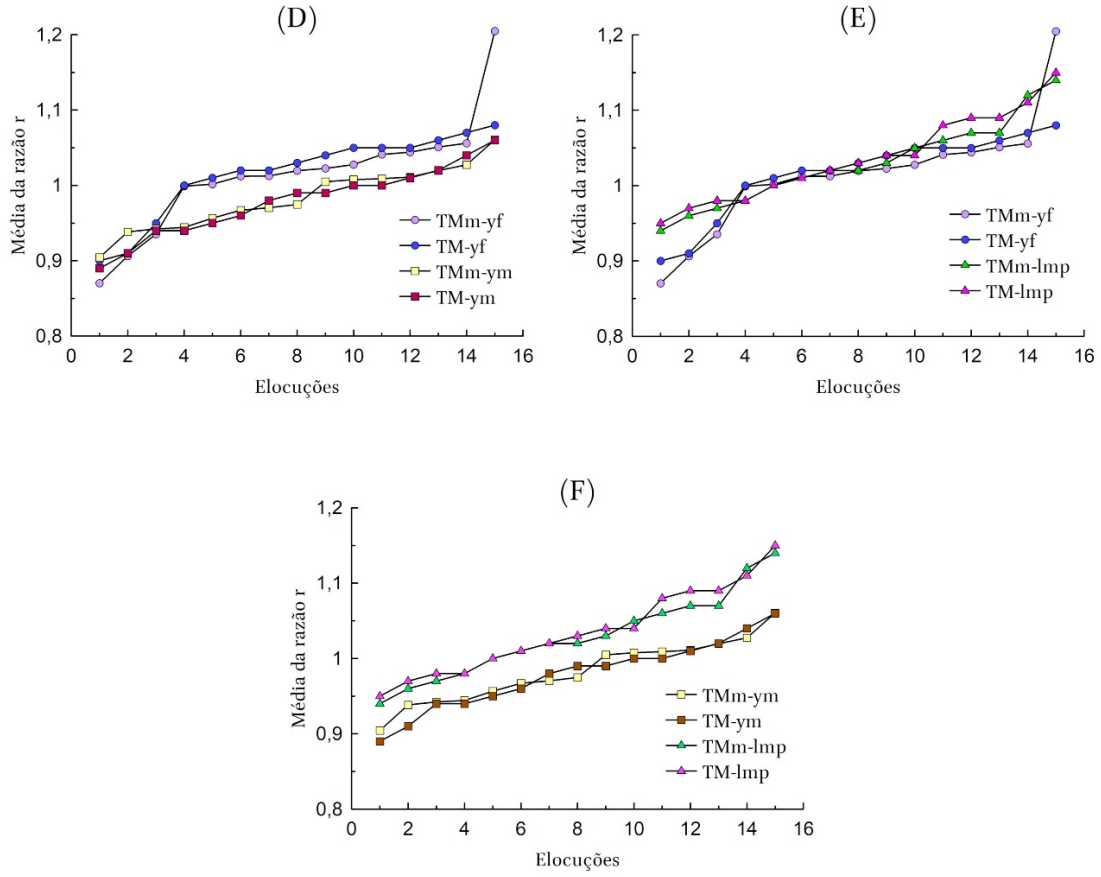

Figura 7: Gráficos comparando os dados organizados em ordem crescente das médias. TM – Tom Médio, TMm – Tom Médio móvel, yf – *youtubers* do gênero feminino, ym – *youtuber* do gênero masculino e lmp – narrativas do ciclo de Lampião.

Os gráficos da fig. 7 mostram a comparação dos dados organizados em ordem crescente das médias. Em (D) comparam-se o Tom Médio móvel e o Tom Médio dos grupos de *youtubers* femininos e masculinos; em (E), faz-se a mesma comparação entre os grupos de *youtubers* femininos e os locutores de narrativas do ciclo de Lampião; em (F), comparam-se os mesmos dados de tom médio entre os grupos de *youtubers* masculinos e os locutores de narrativas do ciclo de Lampião. Esses resultados corroboram os que se encontraram com a análise de *cluster*,[42] que mostram uma maior identidade entre as categorias de *youtubers* femininos e locutores das narrativas do ciclo de Lampião.

[42] Cf. p. 24.

Na fig. 7 (D), é possível perceber que a dispersão dos dados em relação às médias ocorre entre os valores menores e os valores médios do grupo feminino de *youtubers*. A diferenciação do grupo masculino de *youtubers* em relação ao grupo feminino e ao grupo de locutores do ciclo de Lampião decorre da diferença dos valores geral do Tom Médio e Tom Médio móvel. Na comparação, como se pode ver na fig. 7 (E), o grupo masculino de *youtubers* manteve-se sempre com Tom Médio e Tom Médio móvel relativamente abaixo dos valores dos outros grupos. Conforme se pode notar na fig. 7 (F), não houve variação de dispersão entre o grupo masculino de *youtubers* e o de locutores de narrativas do ciclo de Lampião.

Considerações finais

A hipótese do programa ExProsodia de análise automática da entoação estabelece que, por se tratar de uma série temporal, $F0$ poderia ser decomposto em componentes.[43] Essa decomposição partiria de uma medida dinâmica de tendência central a que se chamou de Tom Médio. Definiu-se o Tom Médio pela média acumulada no tempo das frequências válidas subsequentes. Apesar de se tratar de uma medida dinâmica acumulada, adquire estabilidade e, por isso, pode ser tomada como referência a outras medidas. Nesta pesquisa foi possível verificar que, mesmo tomando sujeitos com origem e gênero diversos, os locutores tendem a manter $F0$ dentro dos limites de variação do Tom Médio quando ocorriam interrupções seguidas da retomada de sonoridade na fala.

Esta pesquisa partiu da análise de narrativas porque elas são textos que se mostram semanticamente coerentes e porque podem levar à estabilidade do Tom Médio. Pesquisas que envolvam dados extraídos de outros gêneros textuais poderão, talvez, apresentar resultados diferentes.

Ferreira Netto e seus colegas, em diferentes trabalhos,[44] utilizaram essa medida para o estabelecimento de parâmetros diversos que permitiram contrastar estados emocionais manifestos na fala. Peres[45] e Martins e colegas,[46] baseados no Tom Médio, verificaram que o limiar de dispersão de $F0$ perceptível por falantes do inglês era diferente daquele que se encontrou para o português do Brasil. A percepção dos falantes ingleses quanto a essa dispersão mostrou-se em torno de 2 st, estreitando a faixa do Tom Médio.

Nesta pesquisa, as diferenças que se notaram entre os grupos exigem ainda outras pesquisas com outros detalhamentos. Podem ser resultado da indiferenciação da origem dos sujeitos, como foi o caso do *youtubers*, ou da indiferenciação de gênero, como foi o caso dos locutores de narrativas do ciclo de Lampião, ou ainda devido à extensão das gravações. Em relação à origem dos sujeitos, ainda que de forma muito preliminar, os resultados obtidos corroboram os de Peres,[47] quando tratou da percepção diferenciada da entoação regional entre falantes do português do Brasil. Por se tratar de medidas dinâmicas, o Tom Médio e o Tom Médio móvel podem ser tomados em tempo real. Conceição e seus colegas,[48] com base na proposta de Wennerstrom,[49] encontraram variações no curso do tempo em $F0$ que corresponderiam a manifestações de expressividade semântica próprias dos textos analisados.

A questão apresentada por Fujisawa e outros,[50] e posteriormente também em Cook e outros,[51] de que haja um paradoxo em relação à percepção das variações de $F0$ que não pode ser explicada pelas análises fonológicas do contorno de $F0$, não levaram

43 Ferreira Netto, "Variação de frequência e constituição da prosódia da língua portuguesa" (2006); Ferreira Netto, "Decomposição da entoação frasal em componentes estruturadoras e em componentes semântico-funcionais" (2008); Ferreira Netto et al., "Análise automática de manifestações emocionais de tristeza e cólera em PB: abordagem pelo programa ExProsodia" (2013); Ferreira Netto, "Análise automática de manifestações emocionais em PB: aplicações do programa ExProsodia" (2016).

44 Ferreira Netto et al., "Análise automática de manifestações emocionais de tristeza e cólera em PB: abordagem pelo programa ExProsodia" (2013); Ferreira Netto et al., "Efeitos da entoação e da duração na análise automática das manifestações emocionais" (2014).

45 Peres, "The perception of emotion by native and non-native speakers" (2013).

46 Martins et al., *Diferença tonal mínima perceptível em português e inglês* (2017).

47 Peres, "A prosódia e o reconhecimento dialetal" (2016).

48 Conceição, "Avaliação do tom médio em manchetes telejornalísticas apresentadas por mulheres" (2016).

49 Wennerstrom, "Intonation and evaluation in oral narratives" (2001).

50 Fujisawa et al., "On the role of pitch intervals in the perception of emotional speech" (2003).

51 Cook et al., "Evaluation of the affective valence of speech using pitch substructure" (2005).

em conta as tendências centrais dessas variações de $F0$. Os autores retomaram a proposta de Rameau,[52] do início do século XVIII, que chamava a atenção para o fato de que a terça maior seria, naturalmente, animada e alegre, enquanto a terça menor, naturalmente, terna e triste. Os autores argumentaram que combinações de três ou mais tons têm o sentido universal dos modos maior ou menor e, portanto, a capacidade dos ouvintes normais de detectar o estado afetivo positivo ou negativo de um falante indica uma sensibilidade da orelha humana a informações contidas em $F0$. Peres e outros[53] verificaram que há diferenças significativas na percepção das variações de $F0$, tanto em sujeitos musicalmente treinados como não treinados, quando a cadeia segmental da fala está presente ou não. O limiar de percepção dessas variações de $F0$ cai para 1 st para os dois grupos de sujeitos quando não há cadeia segmental e 3 st quando há. O fenômeno parece ser linguisticamente condicionado, pois Peres[54] e Martins e seus colegas[55] encontraram resultados que estabelecem limiares diferentes para cada língua, especialmente quando há cadeia segmental presente no estímulo dos testes. Vassoler e Martins,[56] Ferreira Netto e seus colegas.[57] e Martins e Ferreira Netto[58] encontraram aspectos significativos na variação de $F0$ que permitiram a separação automática de manifestações emocionais. A comparação dos resultados das pesquisas mostradas acima mostra que se trata de um problema que ainda merece maior aprofundamento.

Os resultados desta pesquisa apontaram que a tendência de retomar o Tom Médio, respeitando seus limiares de variação de 3 st acima e abaixo, pode ser tomada como uma referência segura, pelo menos em relação ao português falado no Brasil.

52 Rameau, *Traité de l'harmonie reduite à ses principes naturels* (1722).

53 Peres et al., "A influência da cadeia segmental na percepção de variações tonais" (2011).

54 Peres, "The perception of emotion by native and non-native speakers" (2013).

55 Martins et al., *Diferença tonal mínima perceptível em português e inglês* (2017).

56 Vassoler e Martins, "A entoação em falas teatrais: uma análise da raiva e da fala neutra" (2013).

57 Ferreira Netto et al., "Efeitos da entoação e da duração na análise automática das manifestações emocionais" (2014).

58 Martins e Ferreira Netto, "Proposal of description for an intonation pattern: The simulacrum of neutral intonation" (2017).

Agradecimentos e critérios éticos

O material utilizado para a pesquisa e coletado no YouTube está de acordo com a lei de direitos autorais Lei Nº 9.610, de 19/02/1998, cap. IV, Art. 46, itens II e VIII. Em relação à ética na coleta e no uso dos dados, a coleta dos dados está conforme a Resolução do Plenário do Conselho Nacional de Saúde nº 510, 07/04/2016, Art 1º, parágrafo único itens V e VII. Além dos recursos da própria Universidade de São Paulo, esta pesquisa teve apoio do CNPq, processos 400145/2009-0; PQ 300235/2010-0; 421369/2018-3.

Referências

Apple, William, Lynn A. Streeter e Robert M. Krauss (1979). "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology* 37.5, pp. 715–727.

Boomer, Donald S. e Allen T. Dittmann (1962). "Hesitation pauses and juncture pauses in speech". *Language and speech* 5.4, pp. 215–220.

Brown, Bruce L., William J. Strong e Alvin C. Rencher (1974). "Fifty-four voices from two: the effects of simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental frequency on ratings of personality from speech". *The Journal of the Acoustical Society of America* 55.2, pp. 313–318.

Conceição, Gdalva da (2016). "Avaliação do tom médio em manchetes telejornalísticas apresentadas por mulheres". In: *ExProsodia. Resultados preliminares*. Organizador Waldemar Ferreira Netto. São Paulo: Paulistana, pp. 39–41.

Conceição, Gdalva da, Amanda Lassak, Renata Rosa e Mayara de Sousa (2016). "Análise da ênfase prosódica em narrativas orais do ciclo de Lampião". In: *ExProsodia. Resultados preliminares*. Organizador Waldemar Ferreira Netto. São Paulo: Paulistana, pp. 64–66.

Consoni, Fernanda (2011). "Aspectos da percepção da proeminência tonal em português brasileiro". Tese de doutorado. São Paulo: Universidade de São Paulo.

Consoni, Fernanda e Waldemar Ferreira Netto (2016). "A percepção de variação em semitons ascendentes em palavras isoladas no Português Brasileiro". In: *ExProsodia. Resultados preliminares*. Organizador Waldemar Ferreira Netto. São Paulo: Paulistana, pp. 19–23.

Cook, Norman D., Takashi X. Fujisawa e Kazuaki Takami (2005). "Evaluation of the affective valence of speech using pitch substructure". *IEEE Transactions on Audio, Speech, and Language Processing* 14.1, pp. 142–151.

Daly, Nancy A. e Victor W. Zue (1990). "Acoustic, perceptual, and linguistic analyses of intonation contours in human/machine dialogues". *First International Conference on Spoken Language Processing*. Kobe, pp. 497–500.

Darwin, Charles (2000 [1872]). *A expressão das emoções no homem e nos animais*. São Paulo: Companhia das Letras.

Duez, Danielle (1985). "Perception of silent pauses in continuous speech". *Language and speech* 28.4, pp. 377–389.

Duez, Danielle (1993). "Acoustic correlates of subjective pauses". *Journal of Psycholinguistic Research* 22.1, pp. 21–39.

Fairbanks, Grant e Wilbert Pronovost (1939). "An experimental study of the pitch characteristics of the voice during the expression of emotion". *Speech Monographs* 6.1, pp. 87–104.

Ferreira Netto, Waldemar (2006). "Variação de frequência e constituição da prosódia da língua portuguesa". Tese de livre-docência. São Paulo: Universidade de São Paulo.

Ferreira Netto, Waldemar (2008). "Decomposição da entoação frasal em componentes estruturadoras e em componentes semântico-funcionais". *IV Congresso Internacional de Fonética e Fonologia*, pp. 26–27.

Ferreira Netto, Waldemar (2010). "ExProsodia". *Revista da Propriedade Industrial–RPI* 2038, p. 167.

Ferreira Netto, Waldemar (2016). "Análise automática de manifestações emocionais em PB: aplicações do programa ExProsodia". In: *ExProsodia. Resultados preliminares*. Organizador Waldemar Ferreira Netto. São Paulo: Paulistana, pp. 1–18.

Ferreira Netto, Waldemar, Organizador (2016). *ExProsodia. Resultados preliminares*. São Paulo: Paulistana.

Ferreira Netto, Waldemar, Marcus Vinícius Moreira Martins e Maressa Freitas de Vieira (2014). "Efeitos da entoação e da duração na análise automática das manifestações emocionais". *Estudos Linguísticos* 43.01, pp. 22–32.

Ferreira Netto, Waldemar, Daniel Oliveira Peres, Marcus Vinícius M. Martins, Renata Luzia Cezar Moraes de Rosa e Maressa Freitas de Vieira (2013). "Análise automática de manifestações emocionais de tristeza e cólera em PB: abordagem pelo programa ExProsodia". *Leitura* 52, pp. 43–65.

Fletcher, Janet (1987). "Some micro and macro effects of tempo change on timing in French". *Linguistics* 25.5, pp. 951–968.

Friedman, Lori A. e Daniel C. O'Connell (1991). "Pause reports for spontaneous dialogic speech". *Bulletin of the Psychonomic Society* 29.3, pp. 223–225.

Fujisawa, Takashi, Kazuaki Takami e Norman D. Cook (2003). "On the role of pitch intervals in the perception of emotional speech". *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokio.

Hanley, Theodore D. (1951). "An analysis of vocal frequency and duration characteristics of selected samples of speech from three American dialect regions". *Speech Monographs* 18.1, pp. 78–93.

't Hart, Johan, René Collier e Antonie Cohen (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press.

Hirschberg, Julia e Janet Pierrehumbert (1986). "The intonational structuring of discourse". *Proceedings of the 24th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 136–144.

Huckvale, Mark A. et al. (1987). "The SPAR speech filing system". *European Conference on Speech Technology* (Edinburgh, 1987). International Speech Communication Association, pp. 1305–1308.

Kang, Bong-Seok, Chul-Hee Han, Sang-Tae Lee, Dae-Hee Youn e Chungyong Lee (2000). "Speaker dependent emotion recognition using speech signals". *Sixth International Conference on Spoken Language Processing*. Beijing.

Lehiste, Ilse e Gordon E. Peterson (1961). "Some basic considerations in the analysis of intonation". *The Journal of the Acoustical Society of America* 33.4, pp. 419–425.

Lövgren, Tobias e Jan van Doorn (2005). "Influence of manipulation of short silent pause duration on speech fluency". *Disfluency in Spontaneous Speech Workshop 2005* (Aix-en-Provence, 2005). Editado por Estelle Campione e Jean Véronis. International Speech Communication Association, pp. 123–126.

Martins, Marcus Vinicius Moreira e Waldemar Ferreira Netto (2017). "Proposal of description for an intonation pattern: The simulacrum of neutral intonation". *The Journal of the Acoustical Society of America* 141.5, pp. 3701–3701.

Martins, Marcus Vinicius Moreira, Daniel Oliveira Peres e Waldemar Ferreira Netto (2017). *Diferença tonal mínima perceptível em português e inglês*. URL: https://www.researchgate.net/publication/322931630_Diferenca_tonal_minima_perceptivel_em_portugues_e_ingles.

Martins, Marcus Vinicíus Moreira (2012). "Aspectos da percepção e do controle entoacional do Português Brasileiro". Dissertação de mestrado. Universidade de São Paulo.

Morettin, Pedro Alberto (1999). *Ondas e Ondaletas: da Análise de Fourier à Análise de Ondaletas de Séries Temporais*. São Paulo: EDUSP.

Ohala, John J. (1984). "An ethological perspective on common cross-language utilization of F_0 of voice". *Phonetica* 41.1, pp. 1–16.

Oliveira, Miguel (2002). "The Role of Pause Occurrence and Pause Duration in the Signaling of Narrative Structure". *Advances in Natural Language Processing*. Editado por Elisabete Ranchhod e Nuno J. Mamede. Berlin: Springer, pp. 43–51.

Paulmann, Silke, Marc D. Pell e Sonja A. Kotz (2008). "How aging affects the recognition of emotional speech". *Brain and Language* 104.3, pp. 262–269.

Peres, Daniel Oliveira (2013). "The perception of emotion by native and non-native speakers". *First UCL Graduate Conference in Linguistics*, pp. 64–65.

Peres, Daniel Oliveira (2016). "A prosódia e o reconhecimento dialetal". In: *ExProsodia. Resultados preliminares*. Organizador Waldemar Ferreira Netto. São Paulo: Paulistana, pp. 91–103.

Peres, Daniel Oliveira, Fernanda Consoni e Waldemar Ferreira Netto (2011). "A influência da cadeia segmental na percepção de variações tonais". *LLJournal* 6.1.

Rameau, Jean-Philippe (1722). *Traité de l'harmonie reduite à ses principes naturels*. Paris: Ballard.

Roederer, Juan Gualterio (1998). *Introdução à Física e Psicofísica da Música*. Traduzido por Alberto Luis da Cunha. São Paulo: EDUSP.

Silva, Ebson Wilkerson Rocha da (2017). "A relação entre produção e percepção de pistas prosódicas na segmentação de narrativas espontâneas". Dissertação de mestrado.

Slaney, Malcolm e Gerald McRoberts (1998). "Baby ears: a recognition system for affective vocalizations". *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Volume 2. IEEE, pp. 985–988.

Spencer, Herbert (1890). "The origin of music". *Mind* 15.60, pp. 449–468.

Spiegel, Murray R. (1985). *Estatística*. São Paulo: McGraw-Hill.

Vassoler, Aline Mara de Oliveira e Marcus Vinícius Moreira Martins (2013). "A entoação em falas teatrais: uma análise da raiva e da fala neutra". *Estudos Linguísticos* 42.1, pp. 9–18.

Wennerstrom, Ann (2001). "Intonation and evaluation in oral narratives". *Journal of Pragmatics* 33.8, pp. 1183–1206.

Williams, Carl E. e Kenneth N. Stevens (1972). "Emotions and speech: Some acoustical correlates". *The Journal of the Acoustical Society of America* 52.4B, pp. 1238–1250.

Woods, Anthony, Paul Fletcher e Arthur Hughes (1986). *Statistics in language studies*. London: Cambridge University Press.

Xu, Yi e Q Emily Wang (2001). "Pitch targets and their realization: Evidence from Mandarin Chinese". *Speech communication* 33.4, pp. 319–337.