Tentativa de Disposição de Registros Entoacionais num Eixo Horizontal Organizado pela Tensão Entoacional

Waldemar Ferreira Netto
wafnetto@usp.br
Universidade de São Paulo

Daniel Oliveira Peres
doliveiraperes@gmail.com
Universidade de São Paulo

Marcus Vinícius Moreira Martins
marcusmartins@usp.br
Universidade de São Paulo

Maressa de Freitas Vieira
maressafv@gmail.com
Instituto Federal de São Paulo – *Campus* de Avaré

Gradus
Revista Brasileira de Fonologia de Laboratório

Vol. 2, nº 1
Dezembro de 2017

https://gradusjournal.com

Bibtex: @article{wafnetto2017tent, author = {Waldemar Ferreira Netto and Daniel Oliveira Peres and Marcus Vinícius Moreira Martins and Maressa de Freitas Vieira}, issn = {2526-2718}, journal = {Gradus}, month = {dec}, number = {1}, pages = {14-29}, title = {Tentativa de Disposição de Registros Entoacionais num Eixo Horizontal Organizado pela Tensão Entoacional}, volume = {2}, year = {2017}}

Resumo

Este ensaio analisa a expressão da emoção na entoação do português brasileiro (PB) por diferentes gêneros (masculino e feminino). Os trechos do discurso emocional foram coletados na Internet e divididos em colérico, neutro, triste e SNI (simulacros de entoação neutra) para ambos os gêneros. Os parâmetros entoacionais foram analisados automaticamente pelo software ExProsodia, que se baseou em medidas de F0 (Hz) e duração (ms). Os resultados mostraram que o discurso emocional no PB pode ser caracterizado pelos graus de tensão na produção de fala. Desse ponto de vista, cada gênero e emoção relacionada pode ser localizado ao longo do eixo de tensão entoacional proposto: (i) triste>neutro>simulacro de entoação neutra>colérica, para locutores masculinos; (ii) simulacro de entoação neutra/neutra/triste>colérica, para locutores femininos. Foi possível verificar que diferença de gênero se relaciona às características fisiológicas expressas por F0 analisado automaticamente e que a tensão entoacional no discurso emocional em ambos os gêneros também pode ser uma abordagem confiável para a análise automática da expressão de emoção no PB.

Palavras-chave: fala emotiva; análise automática; fala espontânea; entoação

Abstract

This study aims to analyze the expression of emotion in Brazilian Portuguese (BP) intonation by different genres (masculine and feminine). The emotional speech excerpts were collected on the Internet and divided into angry, sad, neutral and simulacrum of neutral intonation for both genres. The intonational parameters were analyzed automatically by the software ExProsodia, which were based on measurements of F0 (Hz) and duration (ms). The results showed that emotion speech in BP could be explained by the degrees of tension in speech production. In this view, each emotion and gender type are located along an axis of intonational tension: (i) sad>neutral>simulacrum of neutral intonation>angry, for male speaker; (ii) simulacrum of neutral intonation/neutral/sad>angry, for female speaker. The difference of gender could be related to physiological features expressed by F0. The intonational tension in emotional speech in both genres seems to be a reliable approach in order to analyze the expression of emotion in BP.

Keywords: Emotion speech; Automatic analysis; Natural speech corpus; Intonation

Introdução

A análise da manifestação das emoções associadas à fala tem sido objeto de especulação científica desde o século XIX. Darwin já afirmara que a tonalidade da voz tem relação com certos sentimentos, exemplificando que uma pessoa delicadamente reclamando de maus-tratos, ou de um pequeno sofrimento, quase sempre fala com voz aguda.[1] Spencer[2] afirmara "That certain tones of voice and cadences having some likeness of nature are spontaneously used to express grief, others to express joy, others to express affection, and others to express triumph or martial ardour, is undeniable." Embora tais proposições se estendessem para além da preocupação com manifestação das emoções, o reconhecimento da entoação como um fato comunicativo, voluntário ou não, teve seu início no século XIX, mas não recebeu uma atenção aprofundada nos estudos referentes à comunicação humana.

1. Darwin, *A expressão das emoções nos homens e nos animais* (2001).

2. Spencer, "The origin of music" (1890).

Numa das primeiras pesquisas que procurou descrever de forma mais sistemática a relação entre a variação de frequência e a manifestação das emoções na fala, Skinner [3] verificou que a frequência média na fala, provocada pela alegria (*happiness*) era mais aguda do que a provocada pela tristeza (*sadness*). Sua pesquisa baseou-se na hipótese de que um estado emocional geral seria induzido pela audição prévia de músicas tristes ou alegres e, ainda, acompanhada da audição de textos igualmente tristes ou alegres. Esse estado emocional provocaria naturalmente as manifestações sonoras correspondentes na fala. Para tanto, ele gravou e analisou a expressão curta "ah" de cada um de seus sujeitos. A interpretação do valor médio da curva de F0 dessa expressão foi feita a partir da comparação com o harmônico mais grave do espectrum de um diapasão calibrado em 1000 Hz.

3. Skinner, "A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness" (1935).

Alguns anos depois, Fairbanks e Pronovost[4] procuraram estabelecer a relação entre as variações da entoação e a manifestação das emoções na fala, bem como o julgamento dessas emoções por sujeitos ouvintes. A partir da fala simulada com leituras feitas por atores, os autores analisaram comparativamente as variações de F0 para as manifestações emocionais de desprezo (*contempt*), raiva (*anger*), medo (*fear*), tristeza (*grief*) e indiferença (*indifference*). Para essa comparação, estabeleceram quatro parâmetros: o valor médio da frequência em que ocorrem as manifestações emocionais (*pitch level*) medida em Hz; a variação tonal média, medida em tons musicais (*wide mean inflectional range*); a extensão tonal em que ocorrem essas manifestações, medida em tons musicais (*wide total pitch range*); e, a taxa de variação tonal em que ocorrem essas manifestações emocionais (*pitch change*) (medida em tons por segundo). As comparações foram feitas baseadas nos valores máximos e mínimos encontrados para esses parâmetros. Os resultados obtidos mostraram que manifestações de raiva e de medo ocorrem com a frequência média mais aguda e que indiferença ocorre com a mais

4. Fairbanks e Pronovost, "Vocal pitch during simulated emotion" (1938); Fairbanks e Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion" (1939).

grave. No entanto, no teste de avaliação dessas emoções, foram consideradas como desprezo, tristeza e indiferença todas as leituras cuja frequência média fosse a mais grave. Quanto à extensão tonal, as manifestações de desprezo e de raiva foram as que apresentaram valores mais altos e a manifestação de indiferença apresentou a mais baixa. A manifestação de tristeza teve a menor variação tonal e a manifestação de raiva, a maior. Quanto à taxa de variação tonal por segundo, a mais rápida foi a manifestação de raiva e a mais lenta foi a de medo. Fairbanks e Hoaglin[5] analisaram a taxa de duração das mesmas emoções e verificaram que as manifestações de tristeza e de indiferença apresentaram as menores taxas de duração, atribuindo esse fato aos prolongamentos das fonações e às pausas. A partir dos anos 60, essa preocupação foi retomada com diversos autores.[6]

Em investigação semelhante à de Skinner [7], Bachorowski e Owren[8] analisaram um segmento vocálico de fala de sujeitos que eram submetidos a situações provocadoras de emoções positivas e de emoções negativas. Tomando medidas de F0, *jitter* e *shimmer*, os autores chegaram a resultados semelhantes: as situações em que emoções positivas eram estimuladas estabeleceram F0 mais agudo do que as que provocaram emoções negativas. Como os autores não trataram de nenhuma emoção específica, como nos trabalhos anteriores, é possível estabelecer que as manifestações de emoções negativas, que decorriam de um teste no qual o sujeito não conseguia alcançar os resultados previstos, eram mais propriamente relacionadas à frustração ou tristeza. A partir do ano 2000, o número de investigações que procuram descrever a relação entre a manifestações das emoções e as características acústicas da fala cresce vertiginosamente.[9]

Em trabalho mais recente, Bänzinger e Scherer,[10] num estudo quantitativo, verificaram que a variação global de F0 era afetada diretamente pelo estímulo emocional representado na fala e era a variação mais importante para a discriminação das categorias emocionais observadas. A partir de 1998,[11] com o trabalho de Slaney e McRoberts,[12] dados espontâneos de fala dirigida às crianças começaram a ser utilizados em estudos de análise automática da fala. Recentemente, um grande número de pesquisas tem utilizado a fala espontânea.[13]

Os estudos que tratam de fala emotiva em português brasileiro aparecem com maior frequência a partir da década de 1990. Colamarco e Moraes[14] analisaram 16 repetições de uma sentença padrão combinando emoções e tipos de sentença. O resultado apontou para uma independência entre a entoação com função gramatical e a entoação expressiva ligada à manifestação das emoções.

5. Fairbanks e Hoaglin, "An experimental study of the durational characteristics of the voice during the expression of emotion" (1941).

6. Markel, "The reliability of coding paralanguage: pitch, loudness, and tempo" (1965); Constanzo e Constanzo, "Voice quality profile and perceived emotion" (1969); Williams e Stevens, "Emotions and speech: some acoustical correlates" (1972); Scherer, "Vocal affect expression: a review and a model for future research" (1986). Para uma revisão de trabalhos desse período, cf. Scherer, 1986.

7. Skinner, "A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness" (1935).

8. Bachorowski e Owren, "Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context" (1995).

9. Ang et al., "Prosody-based automatic detection of annoyance and frustration in human-computer dialog" (2002); Fujisawa et al., "On the role of pitch intervals in the perception of emotional speech" (2003); Toivanen et al., "Automatic discrimination of emotion from spoken finnish" (2004); Vogt e André, "Comparing features sets for acted and spontaneous speech in view of automatic emotion recognition" (2005); Cook et al., "Evaluation of the affective valence of speech using pitch substructure" (2006); Vidrascu e Devillers, "Five emotions classes detection in real-world call center data: the use of various types of paralinguistics features" (2007); Rong et al., "Acoustic Features Extraction for Emotion Recognition" (2007); Neiberg e Elenius, "Automatic recognition of anger in spontaneous speech" (2008); Busso et al., "Analysis of emotionally salient aspects of fundamental frequency for emotions detection" (2009); Yang e Lugger, "Emotion recognition from speech signals using new harmony features" (2010); Laukka et al., "Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation" (2011).

10. Bänzinger e Scherer, "The role of intonation in emotional expressions" (2005).

11. Bartliner et al., "The automatic recognition of Emotions in Speech" (2011).

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

13. Bartliner et al., "The automatic recognition of Emotions in Speech" (2011).

14. Colamarco e Moraes, "Emotion expression in speech acts in Brazilian Portuguese: production and perception" (2008).

O estudo de Vassoler e Martins[15] analisou trechos de fala atuada lidos por três atrizes profissionais, subdivididos em raiva e neutro. Como resultado das análises, os trechos de fala com raiva obtiveram maiores valores de F0, ou seja, foram produzidos num registro mais alto que os trechos de fala neutra. Os autores forneceram duas explicações, uma de ordem fisiológica e outra linguística. Na primeira, os músculos e as cartilagens ligados à produção da fala recebem maior tensão, provocando o aumento da pressão subglotal e, consequentemente, causando a elevação dos valores de F0.[16] Do ponto de vista linguístico, os padrões entoacionais entre os dois tipos de fala analisados permaneceu estável, sendo a implementação fonética – sujeita às condições de produção internas e internas ao sujeito – a principal fonte de diferença entre a fala neutra e a com raiva.

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

16. Titze et al., "Phonation threshold pressure in a physical model of the vocal fold mucosa" (1995).

Peres[17] analisou a emoção na fala por meio de análise de produção e percepção. A análise de produção foi baseada em parâmetros acústicos entoacionais e de qualidade vocal. Para a análise, 32 excertos de fala espontânea do português brasileiro foram selecionados e divididos igualmente entre raiva, medo, alegria e tristeza. O teste de percepção foi feito por ingleses e brasileiros. Como esperado, o grau de concordância entre os brasileiros foi mais alto do que entre os ingleses. Os participantes, ingleses e brasileiros, quando equivocados no julgamento, tenderam a associar raiva com alegria, e tristeza com medo. O alto número de respostas corretas dadas pelos participantes brasileiros pode ser explicado pelo papel do léxico e pelo conhecimento pragmático da língua, já o desempenho dos ingleses pode ser explicado pela falta deles.

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

Neste estudo, optou-se pelo uso de fala espontânea por ela ser portadora da expressão autêntica da emoção na fala. A maioria dos estudos que trataram da fala expressiva faziam uso de sentenças com fala teatral ou outros tipos de elicitação.[18]

18. Scherer et al., "Vocal cues in emotion encoding and decoding" (1991).

A utilização de fala atuada ou elicitada tem a seu favor o controle dos estímulos em sentenças idênticas, pronunciadas nas mais variadas emoções e demais tipos de variação entoacional. Sem dúvida, essa característica permite ao experimentador um maior controle das variáveis que podem influenciar na produção e percepção da fala emotiva. Como argumento a favor do uso de fala atuada, Scherer[19] atentou para os problemas encontrados em gravações de fala espontânea, sem intervenção direta do experimentador, afirmando que "[...] *naturally recorded emotions are by definition singular cases, both in terms of speaker identity, situation context, and verbal content of utterance*". Segundo o autor, com essas características da fala espontânea, ficaria difícil a separação de quais variáveis estão de fato agindo para configurar a fala expressiva, configurando um problema quanto à ortogonalidade do experimento.

19. Scherer, "Speech and emotional states" (1981).

O estudo de Roberts,[20] entretanto, demonstrou que a fala teatral pode ser fortemente impregnada de estereótipos, afirmando que esse tipo de estímulo "*may merely reflect stereotypical behaviors that*

20. Roberts, "Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality" (2011).

actors are trained to adopt". A utilização desse tipo de fala poderia causar não só diferenças na produção, mas, provavelmente, na percepção dos estímulos.

A despeito da variação que possa haver entre os trechos espontâneos de fala emotiva, este trabalho dá preferência para esse tipo de produção pela possibilidade de obter dados importantes referentes à manifestação da emoção na fala.

Pressupostos teóricos

Para a realização desta pesquisa, baseamo-nos nos pressupostos do programa de pesquisa ExProsodia. Este programa objetiva a análise automática da entoação no português do Brasil, entendendo que entoação é uma sequência de tons, iguais ou diferentes, produzidos pela voz durante a fala. O desenvolvimento desse programa parte da hipótese de Xu e Wang[21] de que alguns fatos prosódicos têm restrições mecânico-fisiológicas e outros decorrem das necessidades expressivas dos falantes.

21. Xu e Wang, "Component of intonation: what are linguistic, what are mechanical/physiological?" (1997).

De maneira geral, a proposta teórica desse programa de pesquisa parte do princípio de que as unidades de segunda articulação — fonemas — não têm existência independente do léxico, ou seja, são dependentes da seleção lexical feita pelo falante. Apesar dessas unidades de segunda articulação (fonológica) existirem somente no ambiente lexical, têm um algoritmo de realização que produz alteração na corrente sonora da fala, concorrendo de forma coordenada com as demais manifestações vocais na mesma onda sonora. Por serem resultados de algoritmos de produção, as unidades de segunda articulação ocorrem linearmente e de forma discreta. Cada fonema tem seu próprio algoritmo de realização fonética. Desse ponto de vista, as unidades de segunda articulação decorrem de um sistema vibratório forçado. Entretanto, na medida em que a produção de voz parte da vibração das pregas vocálicas que advém do processo de expiração contínua, entende-se que seja um sistema vibratório autoexcitado, provido de uma fonte contínua de alimentação de energia. Dessa maneira, os mecanismos de controle da produção das vibrações estão sob monitoramento contínuo durante a produção de voz.[22]

22. Henrique, *Acústica Musical* (2002).

A produção da fala decorre da integração entre os dois sistemas vibratórios diferentes descritos acima: o sistema vibratório forçado, próprio das unidades da segunda articulação, e o sistema vibratório autoexcitado, próprio da voz humana. A integração entre esses dois sistemas vibratórios, cujos resultados manifestam-se na mesma onda sonora da fala, decorre da simultaneidade de sua ocorrência. Na medida em que o vozeamento, um sistema vibratório autoexcitado, necessita de um monitoramento continuo, relaciona-se diretamente com as condições imediatas de produção da fala e, portanto, está sujeito às condições físicas do próprio falante que provê a energia do sistema. Na medida em que as unidades de segunda articulação de-

correm de algoritmos de realização fonética (motora) desencadeados pelas necessidades lexicais expressivas do falante, estão sujeitas às condições cognitivas do próprio falante para a recuperação lexical e o consequente acesso aos algoritmos de realização fonética de cada uma das unidades de segunda articulação que compõem o item lexical recuperado. As unidades de segunda articulação que compõem cada item lexical estão diretamente associadas à história de cada um desses itens lexicais e, portanto, são definidas por questões sociais que não precisam estar diretamente relacionadas com os hábitos de vozeamento próprios do grupo de fala. Dessa maneira a integração entre o vozeamento e as unidades de segunda articulação não pressupõe uma adaptação perfeita na mídia sonora resultante.

Ferreira Netto[23] propôs que a entoação da fala pode ser decomposta em componentes estruturadoras e semântico-funcionais: finalização (F) e sustentação (S), de um lado, foco/ênfase (E), de outro. Pelo programa de pesquisa ExProsodia,[24] a produção da fala exige esforço para sustentar a voz com uma frequência relativamente estável, definida aqui como tom médio ideal (TM) de F0, que se repete nos momentos Z(t) mensurados de F0. A supressão desse esforço desencadeia uma declinação pontual que exige a retomada da tensão inicial. A sustentação (S) é consequência do esforço que se acrescenta a cada um dos momentos da fala, incluindo-se o inicial, para compensar a declinação pontual de finalização (F). Ritmo Tonal é consequência da ação dessas tendências que atuam em sentidos opostos, possibilitando a produção da fala. A componente F associa-se ao fato de que se trata do tom alvo da declinação pontual, estabelecida por um intervalo ideal decrescente de 7 st do TM obtido até o momento Z(t). TM é a tendência central dos valores válidos de F0 calculada como a média aritmética acumulada no tempo. A figura 1 ilustra como a série temporal é decomposta.

23. Ferreira Netto, *Decomposição da entoação frasal em componentes estruturadoras e em componentes semântico-funcionais* (2008).

24. 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); Peres et al., "Decomposição da entoação frasal em componentes estruturais e semântico-funcionais: um teste com análise da variação de gênero" (2009); Peres et al., "A influência da cadeia segmental na percepção de variações tonais" (2011).

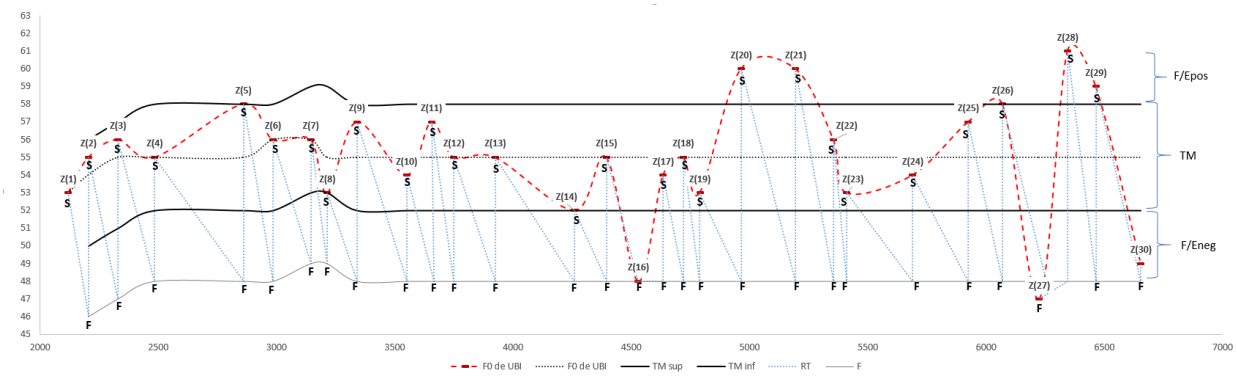

Figura 1: Representação do Ritmo Tonal da frase "Hoje, mais de 43 milhões de brasileiros já usam computador em casa.", marcado com linhas azuis pontilhadas, numa ocorrência efetiva de F0, marcada pela linha vermelha contínua. Os valores à esquerda estão em escala midi. As siglas Z indicam cada um dos momentos mensurados de F0 (UBIs); as siglas F, as finalizações supostas, sendo as que vão marcadas nos momentos Z(16), Z(27) e a última à direita (Z(30)) as que realmente se realizaram; as siglas S indicam os pontos de sustentação supostos, que estabelecem o Tom Médio. F/Epos refere-se aos valores que ocorreram acima do limite lateral superior do TM; F/Eneg refere-se aos valores que ocorreram abaixo do limite lateral inferior do TM.

Os valores válidos mensurados são os momentos de F0 que cumprem as restrições de altura, intensidade e duração. A série temporal se configura aditivamente como Z(t)=S(t)+F(t)+E(t). A seleção das unidades Z(t) — ora chamadas de UBI (*Unit of Base of Intonation*) — para análise é feita pelo aplicativo ExProsodia. O aplicativo faz a análise automática de porções da curva de frequência estabelecida por autocorrelação pelo software *Speech Filing System* (fig. 2).[25] Trata-se de um ambiente de computação livre que compreende diversas ferramentas para processamento de sinal, síntese e reconhecimento da fala.[26] Para a definição de uma unidade básica de entoação consideram-se três parâmetros: frequência maior do que 50 Hz e menor do que 700 Hz, intensidade suficiente para ser percebida e, garantidos os critérios anteriores, duração maior do que 20 ms. Por se tratar de porções portadoras de sonoridade, as UBIs tanto podem englobar sequências completas de segmentos soantes como em ['malɐ] 'mala' como apenas picos silábicos como em ['sapʊ]. Daí a necessidade de se tratar como uma unidade distinta das que tradicionalmente se consideram em análise fonética. Todos os parâmetros podem ser modificados pelo usuário. Ferreira Netto [27] faz uma descrição detalhada desse procedimento.

25. Huckvale et al., "The SPAR Speech Filing System" (1987).

26. Está disponível em http://www.phon.ucl.ac.uk/resource/LTF/ .

27. Ferreita Netto, "Análise automática de manifestações emocionais em PB: aplicações do programa ExProsodia" (2016).

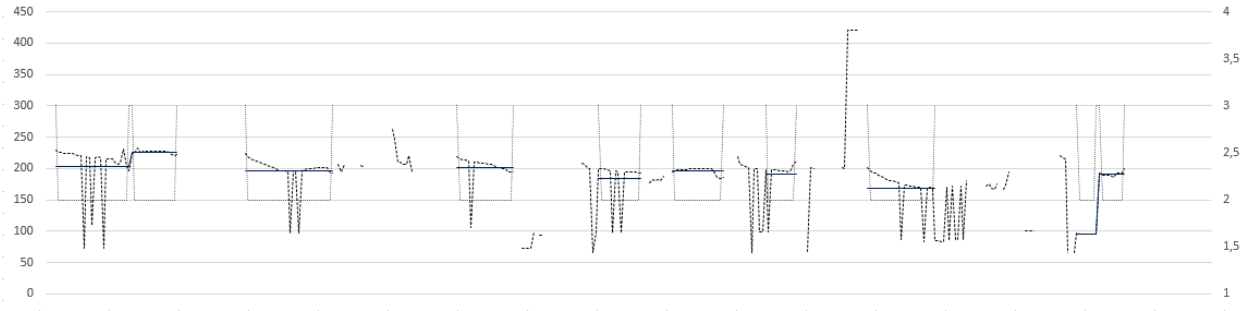

Figura 2: A linha tracejada representa a variação tonal de F0, as linhas pontilhadas verticais e horizontais mostram as porções de F0 que foram selecionadas a partir dos parâmetros estabelecidos (nesse caso, UBI≥40 ms, 50 Hz ≥ UBI ≤ 300, UBI > 0 rms) e as linhas contínuas horizontais, o valor médio das Unidades Básicas de Entoação, considerados para as análises posteriores. Os valores à esquerda estão em Hz e os valores à direita não têm significação; foram usados somente para composição gráfica da linha tracejada.

A análise automática vale-se de ambientes de computação, aliados a linguagens de programação geralmente disponíveis nesses mesmos ambientes (como é o caso do Excel, cuja linguagem, VBA, é a que utilizamos; mas também do PRAAT, do Speech Filing System, do MATLAB, dentre muitos outros), além das linguagens independentes (como JAVA, C, Pascal, dentre outras), atuam como laboratórios digitais que permitem a reprodução e a simulação de eventos sonoros digitalizados. O aplicativo ExProsodia, de maneira um pouco diferente de outros softwares de análise da fala – em que se atua diretamente sobre as imagens geradas a partir de dados digitalizados – faz seus procedimentos a partir dos dados numéricos dos eventos sonoros digitalizados. A extração desses valores, especialmente relativos a frequência, intensidade e ordem de ocorrência, é feita pelo software Speech Filing System, que permite extrair as respectivas matrizes unidimensionais em formato de texto. É a partir dessas matrizes que o aplicativo ExProsodia faz

a análise automática. Por se desenvolver em ambiente vinculado a uma planilha numérica, como várias macros entrelaçadas, o aplicativo gera uma matriz multidimensional em planilhas do Excel® a partir das quais se podem fazer todas as operações matemáticas, estatísticas e gráficas que se mostrarem necessárias. Como são macros vinculadas à planilha, é possível atualizarem-se com novos procedimentos a qualquer momento.

A seleção das unidades básicas de entoação, as UBIs, se faz diretamente na matriz de dados digitais, definindo-se padrões para comparação. Para esta análise, a UBI foi definida a partir de 50 Hz até o máximo de 700 Hz com duração mínima de 60 ms. Os valores máximos e mínimos da escala de frequência possibilitaram que não houvesse nenhuma restrição necessária quanto ao gênero dos sujeitos que produziram os registros entoacionais. A duração mínima de 60 ms está de acordo com os trabalhos de Schaeffer[28] e de Winckel,[29] que estabelecem como ”constante de integração” ou ”espessura do presente” a duração mínima de 50 ms, acrescentando uma margem de erro de 20%, tendo em vista a heterogeneidade da amostra.[30] Uma vez definida a duração mínima da UBI, a rotina busca nas colunas relativas à frequência e à intensidade valores numa sequência mínima de 12 valores, uma vez que a janela de correlação feita pelo Speech Filing System é de 5 milissegundos; assim, ocorrendo 12 valores subsequentes acima de 50 Hz, abaixo de 700 Hz e com intensidade maior do que zero, estabelece-se o início de uma UBI. Sua finalização pode decorrer tanto da mudança dos valores em relação aos parâmetros estabelecidos, quando de uma variação de frequência que provoque modalização acima ou abaixo de 3 semitons. Uma vez finalizada, define-se o valor médio das frequências e das intensidades, em rms, para cada UBI. O Tom Médio se forma a partir da média acumulada no tempo dessas UBIs.

28. Schaeffer, *Traité des objets musicaux: essai interdisciplines* (1966).

29. Winckel, *Music, sound and sensation. A modern exposition* (1967).

30. Boemio et al., “Hierarchical an asymmetric temporal sensitivity in human auditory cortices” (2005).

O intervalo máximo entre as UBIs, por sua vez, foi definido em 500ms. A partir dessa duração entende-se ter havido uma pausa. Serra[31] verificou que na fala espontânea do português do Brasil, a duração de 714 ms é o valor médio para a percepção da pausa em fronteiras de sintagmas mas que, na leitura, esse valor cai para 473 ms. Resultados semelhantes foram obtidos por Duez[32] em relação ao francês para a fala espontânea de discursos políticos, quanto a pausas de até 400 ms. Tendo em vista estarmos lidando com gravações feitas espontaneamente, optamos novamente por uma margem de erro maior, estabelecendo o valor de 500 ms como o valor mínimo de uma pausa e, portanto, o intervalo máximo entre duas UBIs consecutivas. Por se tratar de unidade que não corresponde necessariamente a uma interrupção da fala, mas tão somente a ausência de sonoridade concorde com os parâmetros estabelecidos, a opção pela duração de 500 ms segue a proposição de Marcuschi.[33] Ainda que essa margem de segurança seja especulativa, trata-se de se garantir que as pausas possam ser minimamente consideradas intencionais.

31. Serra, “Fraseamento prosódico e percepção no português do Brasil: para o estudo dos estilos de fala” (2010).

32. Duez, “Perception of silent pauses in continuous speech” (1988).

33. Marcuschi, *Análise da conversação* (1986).

Materiais e Métodos

Os dados de fala espontânea analisados foram coletados na internet em sites que disponibilizam *podcasts*: Webcombrasil,[34] A Voz do Brasil,[35] Podcast UNESP.[36] Também foram coletadas gravações de vídeos no site YouTube.[37]. Todos os arquivos sonoros foram extraídos com o software *Soundtap Streaming Audio Recorder® v2.11.*[38] Os arquivos sonoros foram segmentados com o programa *Adobe Audition 3.0.1.* [39] A edição realizada foi a filtragem das vozes de terceiros e de ruídos indesejáveis, mantendo exclusivamente a voz a ser considerada. Para isso foi aplicado o efeito *Dynamic EQ effect* com frequência zero sobre o trecho a ser eliminado. Assim, foi possível manter a duração original de todos os arquivos. A análise e conversão da curva de frequência fundamental e da curva de intensidade para arquivos de texto foi realizada pelo software *Speech Filing System Release 4.8/Windows Win SFSVersion 1.7.*[40] Todas as demais análises foram feitas pelo aplicativo ExProsodia®, registrado pela Universidade de São Paulo produzido por Waldemar Ferreira Netto.[41] As análises estatísticas foram realizadas pelo software KyPlot® version 2.0 beta 15 (32 bit) produzido e registrado por Kuichi Yoshida.

Para a análise das variáveis gênero e manifestação emocional, foram selecionados 75 arquivos sonoros, distribuídos em grupos por gênero e registro. Os grupos compreendiam: locução colérica feminina-LCF (n=10), locução colérica masculina-LCM (n=10), locução com simulacro de entoação neutra feminina-LSNI-F (n=6), locução com simulacro de entoação neutra masculina-LSNI-M (n=6), locução neutra feminina-LNF (n=11), locução neutra masculina-LNM (m=11), locução triste feminina-LTF (m=9) e locução triste masculina-LTM (n=11). A avaliação dos registros entoacionais como colérico, SNI, neutro e triste decorreu de interpretação semântica. Registros coléricos foram tomados de gravações de reclamações de serviços feitas por telefone; registros SNI foram definidos pela incoerência notada entre a semântica do texto e sua entoação – trata-se de fala monotônica produzida com prevalência de ritmo e melodia plana, geralmente usada por falantes com determinados distúrbios psíquicos ou em situação de estresse[42] –; registros tristes referem-se a descrições de mortes de parentes próximos ou de situações extremamente vexaminosas; registros neutros foram tomados de entrevistas nas quais o entrevistado, objeto da coleta, tratava de assuntos relativos à sua profissão. Embora não houvesse restrições quanto à qualidade da gravação, algumas gravações tiveram de ser eliminadas ora pela baixa intensidade ora pelo uso excessivo de filtros. Os parâmetros que foram utilizados estão na Tabela 1 abaixo. Na tabela, as siglas F0 refere-se aos valores em Hz extraídos da curva entoacional produzida pelas unidades básicas da entoação (UBI); TM refere-se à média acumulada no tempo das UBIs, considerando-se um limite lateral de 3 semitons acima e 4 semitons abaixo; F/Epos

34. http://www.webcombrasil.com.br/ .
35. http://www.ebcservicos.ebc.com.br/programas/a-voz-do-brasil .
36. http://podcast.unesp.br/ .
37. http://www.youtube.com/ .
38. NCH Software, *Soundtap Streaming Audio Recorder v2.11* (2007).
39. Adobe Systems, *Adobe Audition 3.0.1 build 8347.0* (2012).
40. Huckvale et al., "The SPAR Speech Filing System" (1987).
41. Ferreira Netto, "ExProsodia" (2010).
42. Martins e Ferreira Netto, "Proposal of description for an intonation pattern: The simulacrum of neutral intonation" (2017).

refere-se aos valores em HZ que ocorreram acima do limite lateral superior do TM; F/Eneg refere-se aos valores em Hz que ocorreram abaixo do limite lateral inferior do TM; intraUBI refere-se à duração interna das UBIs, mensuradas em milissegundos; e interUBI refere-se aos valores, mensurados em milissegundos que ocorrem entre as UBIs.

Os parâmetros para os quais encontrou-se P (<0,01) significativo foram agrupados conforme sua origem (cf. Tabela 1).

F0	***TM***	**FEpos**	**FEneg**	***intraUBI***	***interUBI***
menor_F0_UBI	*TM-mUBI*	*maior_FEpos_UBI*	*maior_FEneg_UBI*	*maior_intraUBI*	*media_interUBI*
maior_F0_UBI	*menor_TM*	*media_FEpos_UBI*	*media_FEneg_UBI*	*media_intraUBI*	*dp_interUBI,*
extensao_F0	*maior_TM*	*dp_FEpos_UBI*	*dp_FEneg_UBI*	*dp_intraUBI*	*mediana_interUBI*
dp_F0_UBI,	*TM*	*skew_FEpos_UBI*	*skew_FEneg_UBI*	*skew_intraUBI*	*cv_interUBI*
skew_F0_UBI	*dp_TM*	*extensao_FEpos_UBI*	*extensao_FEneg_UBI*	*mediana_intraUBI*	
mediana_F0_UBI	*skew_TM*	*cv_FEpos_UBI*	*cv_FEneg_UBI*	*cv_intraUBI*	
cv_F0_UBI,	*mediana_TM*	*kurt_FEpos_UBI*	*kurt_FEneg_UBI*		
kurt_F0_UBI	*cv_TM*				
UBI_final					

Tabela 1: P significativo, conforme origem.

Análise de dados

A Tabela 2 abaixo parte de um conjunto de análises estatísticas cujo propósito é o de verificar se os parâmetros estabelecidos pelo programa ExProsodia podem servir como critérios para a definição de características particulares para as manifestações entoacionais da emoção na fala de língua portuguesa. As análises estatísticas foram realizadas com todos os parâmetros em relação a todos os registros entoacionais. Utilizou-se o teste z com nível de confiança α<=0,01.

	LCF	**LCM**	**LSNI-F**	**LSNI-M**	**LNF**	**LNM**	**LTF**	**LTM**
F0	**1**	0	0	**-1**	0	**-1**	0	**-1**
TM	**1**	0	0	**-1**	0	**-1**	0	**-1**
FEpos	**1**	0	0	0	0	0	0	**-1**
FEeneg	**1**	0	0	0	0	**-1**	0	**-1**
IntraUBI	**-1**	**-1**	0	**1**	0	0	0	0
InterUBI	0	0	0	0	0	0	0	0

(1): parâmetros sempre acima da média para ztest, com P>0,99
(-1): parâmetros sempre abaixo da média para ztest, com P<0,01
(0): parâmetros dentro da média.

Tabela 2: Resultados de análises estatísticas.

Em seguida, a partir da moda de valores positivos — significativamente acima da média (P<0,01) — ou negativos — significativamente abaixo da média (P<0,01) — maior do que n/2 em cada categoria, foram definidos para cada grupo a tendência geral.

Os resultados apontaram os seguintes fatos:

a) os registros LCM, LSNI-F, LNF e LTF apresentam tendência a estar de acordo com os valores médios em relação a todos os parâmetros;

b) o registro LTM tende a estar abaixo da média em relação aos parâmetros diretamente relacionados à entoação (F0, TM FEpos e FEneg);

c) o registro LCF tende a estar acima da média em relação aos parâmetros diretamente relacionados à entoação (F0, TM FEpos e FEneg;

d) os registros LSNI-M e LNM a estar abaixo da média quanto aos parâmetros relacionados a F0 e a TM.

De maneira geral, pode-se pensar que, em relação aos parâmetros que consideram variações de frequência, LCF e LTM encontram-se nos extremos opostos, e LCM, LSNI-F, LNF e LTF encontram-se em posição central (fig. 3).

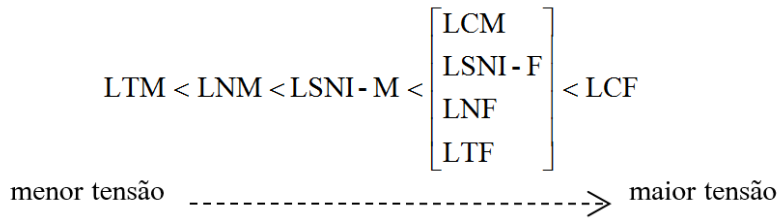

Figura 3: Registros masculinos e femininos organizados por ordem de tensão.

Em relação aos parâmetros que consideraram a duração das UBIs (IntraUBI) e entre os segmentos (InterUBI), pode-se notar que LCF e LCM tendem a ter duração abaixo da média e apenas LSNI-M, acima. Todos os demais tendem a estar de acordo com valores médios. Em relação à duração entre UBIs, nenhum registro apresentou variação, tendendo a estar de acordo com valores médios.

A diferença de tensão entre registros masculinos e femininos pode estar associada às variações naturais de F0 e de TM, na medida em que registros de gênero feminino tendem a ter valores maiores nos parâmetros diretamente relacionados a essas séries. Se fizermos uma distinção entre registros masculinos (fig. 4) e registros femininos (fig. 5), será possível verificar que há uma tendência nos registros femininos a estarem de acordo com os valores médios, mantendo acima da média especialmente o registro colérico (LCF). No caso dos registros masculinos, houve maior distribuição no eixo de tensão.

LTM < LNM < LSNI - M < LCM

Figura 4: Registros masculinos organizados por ordem de tensão.

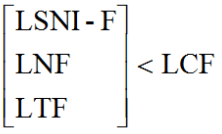

Figura 5: Registros femininos organizados por ordem de tensão.

Considerações finais

O estabelecimento de sequências de registros de manifestações emocionais conforme a tensão apresentada na entoação possibilita

a interpretação comparativa entre eles. Ainda que os resultados apresentados corroborem pesquisas anteriores, a comparação entre todos os registros, tal como vai na fig. 3, permite que se note que apenas a entoação colérica masculina não mostra características especiais em relação a valores médios, assemelhando-se aos registros femininos de entoação neutra, triste e simulacro de entoação neutra. Dessa maneira, o reconhecimento automático da entoação colérica exige o diferencial da duração das unidades portadoras da entoação. Nesse caso, a tendência à duração abaixo da média do registro colérico — masculino e feminino — complementa-se com os valores médios apresentados na entoação. Essa complementação entre duração e entoação já foi apontada em trabalho anterior.[43]

Na medida em que se pode descartar o registro colérico masculino dos registros com valores médios, é notável que os registros feminino neutro, triste e simulacro de entoação neutra apresentem-se de forma semelhante quanto aos critérios utilizados para a comparação. A constatação de que a fala feminina é mais estável quanto às variações de frequência já foi notada por Andrade.[44] A autora encontrou para as vozes masculinas uma variação média oscilando entre 110Hz e 146,7 Hz e, para as vozes femininas, uma concentração acentuada em torno de 203,5 Hz. Nesse caso, verifica-se a necessidade de investigações mais detalhadas dos parâmetros utilizados.

Entretanto, o registro SNI feminino, definido como simulacro de entoação neutra, apresenta-se na situação do registro neutro feminino, o que corrobora a hipótese da existência desse registro. Por ser definido pela sua incoerência entre semântica e entoação, ao contrário dos registros neutro e triste femininos, o registro SNI feminino exige, também, investigações mais detalhadas quanto aos parâmetros utilizados.

Fato semelhante não ocorre em relação aos registros masculinos, que estabelecem uma sequência bem definida de tensão entoacional. O registro simulacro de entoação neutra masculino interpõe-se entre o registro neutro e o colérico. Como se pode notar, o registro colérico masculino, à semelhança do feminino, tem a maior tensão entoacional. Já o registro neutro masculino tem menor tensão entoacional do que o simulacro de entoação neutra masculino. Apesar de permanecer entre os valores médios, o registro simulacro de entoação neutra masculino, desse ponto de vista difere do registro neutro, por apresentar, na sequência, um grau de tensão entoacional maior. Esse dado aponta para a existência do registro simulacro de entoação neutra masculino, a par do feminino, como um registro próprio da entoação na fala em língua portuguesa.

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

44. Andrade, "Determinação dos limiares de normalidade dos parâmetros acústicos da voz" (2003).

Referências

Adobe Systems (2012). *Adobe Audition 3.0.1 build 8347.0*. Software. San Jose.

Andrade, L. M. O. (2003). "Determinação dos limiares de normalidade dos parâmetros acústicos da voz". Dissertação de mestrado. São Paulo: Universidade de São Paulo.

Ang, J., R. Dhillon, E. Shriberg, A. Stolcke e A. Krupski (2002). "Prosody-based automatic detection of annoyance and frustration in human-computer dialog". In: *Proceedings of 7th International Conference on Spoken Language Processing*. Denver, pp. 2037–2040.

Bachorowski, J. A. e M. J. Owren (1995). "Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context". In: *Psychological Science*, pp. 219–224.

Bänzinger, T. e K. R. Scherer (2005). "The role of intonation in emotional expressions". In: *Speech Communication*, pp. 252–267.

Bartliner, A. et al. (2011). "The automatic recognition of Emotions in Speech". In: Cowie, T., C. Pelachaud e P. Petta. *Emotion-oriented Systems: The Humaine Book*. Berlin: Springer, pp. 71–99.

Boemio, A., S. Fromm, A. Braun e D. Poeppel (2005). "Hierarchical an asymmetric temporal sensitivity in human auditory cortices". In: *Nature Neuroscience*, pp. 389–395.

Busso, C., S. Lee e S. Narayanan (2009). "Analysis of emotionally salient aspects of fundamental frequency for emotions detection". In: *IEEE Transactions on Audio, Speech, and Language Processing 17(4)*, pp. 582–596.

Colamarco, M. A. e J. A. Moraes (2008). "Emotion expression in speech acts in Brazilian Portuguese: production and perception". In: *Proceedings of the Speech Prosody*. Campinas: Unicamp.

Constanzo, F. N. e P. Constanzo (1969). "Voice quality profile and perceived emotion". In: *Journal of Couseling Psychology 16(3)*, pp. 267–270.

Cook, N. D., T. Fujisawa e K. Takami (2006). "Evaluation of the affective valence of speech using pitch substructure". Inglês. In: *IEEE Transactions on Audio, Speech, and Language Processing 14(1)*, pp. 142–151.

Darwin, Charles (2001). *A expressão das emoções nos homens e nos animais*. São Paulo: Companhia das Letras.

Duez, D. (1988). "Perception of silent pauses in continuous speech". In: *Language and Speech 28(4)*, pp. 377–389.

Fairbanks, G. e L. W. Hoaglin (1941). "An experimental study of the durational characteristics of the voice during the expression of emotion". In: *Speech Monographs 6(1)*, pp. 85–90.

Fairbanks, G. e W. Pronovost (1938). "Vocal pitch during simulated emotion". In: *Science*, pp. 382–383.

Fairbanks, G. e W. Pronovost (1939). "An experimental study of the pitch characteristics of the voice during the expression of emotion". In: *Speech Monographs*, pp. 87–104.

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

Ferreira Netto, W. (2008). *Decomposição da entoação frasal em componentes estruturadoras e em componentes semântico-funcionais*. Apresentação. Niterói. URL: https://www.academia.edu/2272651/Decomposi%C3%A7%C3%A3o_da_entoa%C3%A7%C3%A3o_frasal_em_componentes_estruturadoras_e_sem%C3%A2ntico-funcionais (acesso em 11 de janeiro de 2018).

Ferreira Netto, W. (2010). "ExProsodia". In: *Revista da Propriedade Industrial - RPI 2038*, p. 167.

Ferreira Netto, W. M., V. M. Martins e M. F. Vieira (2014). "Efeitos da entoação e da duração na análise automática das manifestações emocionais". In: *Estudos Linguísticos*, pp. 22–32.

Ferreita Netto, W. (2016). "Análise automática de manifestações emocionais em PB: aplicações do programa ExProsodia". In: Ferreira Netto, Waldemar. *ExProsodia. Resultados Preliminares*. São Paulo: Paulistana, pp. 11–28.

Fujisawa, T., N. D. Cook e K. Takami (2003). "On the role of pitch intervals in the perception of emotional speech". In: *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo: Tokyo Institute of Technology.

Henrique, L. L. (2002). *Acústica Musical*. Lisboa: Calouste Gulbenkian.

Huckvale, M. A., D. T. Brookes, M. J. B. Johnson e L. Whitaker (1987). "The SPAR Speech Filing System". In: *European Conference On Speech Technology*. Edinburgh.

Laukka, P., D. Neiberg, M. Forsell, I. Karlsson e K. Elenius (2011). "Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation". In: *Computer Speech and Language*, pp. 84–104.

Marcuschi, L. A. (1986). *Análise da conversação*. São Paulo: Ática.

Markel, N. N. (1965). "The reliability of coding paralanguage: pitch, loudness, and tempo". In: *Journal of Verbal Learning and Verbal Behavior*, pp. 306–308.

Martins, M. V. e W. Ferreira Netto (2017). "Proposal of description for an intonation pattern: The simulacrum of neutral intonation". In: *The Journal of the Acoustical Society of America*, p. 3701.

NCH Software (2007). *Soundtap Streaming Audio Recorder v2.11*. Software. Greenwood Village.

Neiberg, D. e K. Elenius (2008). "Automatic recognition of anger in spontaneous speech". In: *Proceedings of the Interspeech 2008*. Brisbane.

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

Peres, D., F. Consoni e W. Ferreira Neto (2009). "Decomposição da entoação frasal em componentes estruturais e semântico-funcionais: um teste com análise da variação de gênero". In: *Osuchil - The Ohio State University Congress on Hispanic And Lusophone*. Ohio.

Peres, D., F. Consoni e W. Ferreira Netto (2011). "A influência da cadeia segmental na percepção de variações tonais". In: *LLJournal*. URL: http://www.academia.edu/1875927/A_influencia_da_cadeia_segmental_na_percepcao_das_variacoes_tonais (acesso em 11 de janeiro de 2018).

Roberts, L. (2011). "Acoustic effects of authentic and acted distress on fundamental frequency and vowel quality". In: *The 17th International Congress of Phonetic Sciences (ICPhS XVII)*. Hong Kong.

Rong, Jia, Yi-Ping Phoebe Chen, Morshed Chowdhury e Gang Li (janeiro de 2007). "Acoustic Features Extraction for Emotion Recognition". In: *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*. Melbourne, pp. 419–424.

Schaeffer, P. (1966). *Traité des objets musicaux: essai interdisciplines*. Paris: Seuil.

Scherer, K. R. (1981). "Speech and emotional states". In: Darby, J. *Speech evaluation in psychiatry*. New York: Grune & Stratton, pp. 189–220.

Scherer, K. R. (1986). "Vocal affect expression: a review and a model for future research". In: *Psychological Bulletin*, pp. 143–165.

Scherer, K. R., R. Banse, H. G. Wallbott e T. Goldbeck (1991). "Vocal cues in emotion encoding and decoding". In: *Motivation and Emotion*, pp. 123–148.

Serra, C. R. (2010). "Fraseamento prosódico e percepção no português do Brasil: para o estudo dos estilos de fala". In: *Scientibus*, pp. 33–58.

Skinner, E. R. (1935). "A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness". In: *Speech Monographs*, pp. 81–137.

Slaney, M. e G. McRoberts (1998). "Baby ears: a recognition system for affective vocalizations". In: *Proceedings of 1998 IEEE International Conference*. Seattle. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.5393&rep=rep1&type=pdf (acesso em 11 de janeiro de 2018).

Spencer, H. (1890). "The origin of music". In: pp. 449–468.

Titze, I. S., S. S. Schmidt e M. Titze (1995). "Phonation threshold pressure in a physical model of the vocal fold mucosa". In: *Journal of Acoustical Society of the America*, pp. 3080–3084.

Toivanen, J., E. Väyrynen e T. Sappänen (2004). "Automatic discrimination of emotion from spoken finnish". In: *Language and speec*, pp. 383–412.

Vassoler, A. M. O. e M. V. Martins (2013). "A entoação em falas teatrais: uma análise da raiva e da fala neutra". In: *Revista Estudos Linguísticos*, pp. 9–18.

Vidrascu, L. e L. Devillers (2007). "Five emotions classes detection in real-world call center data: the use of various types of paralinguistics features". In: *Online Proceedings International workshop on Paralinguistic Speech - between models and data*. Saarbrücken.

Vogt, T. e E. André (2005). "Comparing features sets for acted and spontaneous speech in view of automatic emotion recognition". In: *IEEE International Conference on Multimedia and Expo*. Amsterdam.

Williams, C. E. e K. N. Stevens (1972). "Emotions and speech: some acoustical correlates". In: *Journal of the Acoustical Society of America*, pp. 1238–1250.

Winckel, F. (1967). *Music, sound and sensation. A modern exposition*. New York: Dover Publications.

Xu, Y. E e Q. E. Wang (1997). "Component of intonation: what are linguistic, what are mechanical/physiological?" In: *International Conference on Voice Physiology and Biomechanics*. Evanston. URL: http://www.homepages.ucl.ac.uk/~uclyyix/voice.html (acesso em 11 de janeiro de 2018).

Yang, B. e M. Lugger (2010). "Emotion recognition from speech signals using new harmony features". In: *Signal Processing*, pp. 1415–1423.