Efeitos do vozeamento não modal na descrição estatística de amostras de frequência fundamental

Isabela de Jesus Silveira[1]*; Pablo Arantes*[2]

[1] Universidade Federal de São Carlos, Departamento de Letras
[2] Universidade Federal de São Carlos, Departamento de Letras

isabelajsilv@gmail.com; pabloarantes@gmail.com

Resumo

Neste trabalho testamos o efeito da presença de vozeamento não modal em amostras de fala sobre uma metodologia desenvolvida previamente para a determinação do tamanho mínimo da amostra de fala necessário para o cálculo de três diferentes estimadores estatísticos de tendência central ou valor típico da frequência fundamental (F0): a média, a mediana e o valor de base. No *corpus* usado para testar a metodologia, composto por amostras de fala de 70 falantes e sete línguas diferentes, quase 80% dos falantes apresentam uma porcentagem não nula de valores de F0 que estão abaixo da faixa do registro modal de fonação. Essa porcentagem pode ser de até 30% do total de valores de F0. A presença de um número significativo de valores de F0 no registro não modal aumenta a variabilidade da amostra e pode introduzir vieses tanto nos valores dos estimadores estatísticos quanto no tempo necessário para sua estabilização. Isso torna importante o estudo do impacto do vozeamento não modal na metodologia desenvolvida anteriormente.

Para realizar esse objetivo foram necessários processos que replicavam a metodologia utilizada em trabalhos anteriores [1], [2] como a extração dos contornos de F0 de todos os falantes em todos os estilos presentes no *corpus* (totalizando 210 amostras), a geração de histogramas desses contornos de F0 a fim de determinar visualmente os limites do registro modal e do não modal em organizações bimodais dos contornos, e algumas estatísticas que apresentaram a forte prevalência do registro não modal no número total das amostras de fala das sete línguas analisadas.

Além disso, também foram realizados testes estatísticos para a determinação da influência do registro não modal nos valores dos estimadores estatísticos e em seus respectivos tempos de estabilização. Os resultados mostraram que a exclusão da parte não modal dos contornos de F0 não provocou mudanças significativas nos tempos de estabilização. Os resultados são os mesmos para os três estimadores testados, média, mediana e valor de base.

1. Introdução

A frequência fundamental (F0), o correlato acústico da vibração das pregas vocais, é um parâmetro cuja variação ao longo da produção da fala pode ser matéria de estudos sob diferentes perspectivas. O fato de haver aspectos idiossincráticos na variação de F0 e de ser possível, com algumas restrições, falar a respeito da F0 típica de um falante, possibilitou a investigação de duas questões muito relevantes: a primeira é a duração mínima que uma amostra de fala deve ter para que a estimativa de valor central ou típico de F0 obtida a partir dela possa ser considerada representativa do falante que a produziu; a segunda diz respeito à escolha da medida estatística que melhor caracteriza a tipicidade de um valor de F0, dada uma determinada amostra de valores de F0. Ambas as questões já foram objeto de pesquisa de trabalhos anteriores [1], [2] e auxiliaram o presente projeto. Os trabalhos citados contribuíram de maneira original para a investigação das duas questões referidas anteriormente por conta da introdução de um método estatístico para determinação do tamanho mínimo de amostra: o método *changepoint analysis*, que será melhor comentado na sessão "Materiais e Métodos".

Com relação ao estudo do tempo de estabilização realizado pelos trabalhos acima, foram feitas comparações dos tempos de estabilização determinados pela técnica *changepoint analysis* para três estimadores estatísticos (média aritmética, mediana e valor de base). Os resultados obtidos mostraram que os estimadores testados atingem um patamar baixo de variabilidade entre 5 e 10 segundos em um *corpus* composto por leituras da passagem "O vento norte e o sol", lida por falantes de 26 línguas (um falante por língua). O valor de base apresenta os tempos de estabilização menores e menos variáveis e a média aritmética e a mediana os maiores e mais variáveis.

Esses resultados foram posteriormente aprofundados em um projeto de iniciação científica, cujos resultados são reportados em [3], [4] e, de forma geral, podemos dizer que mostraram que não existe uma diferença significativa na variabilidade dos pontos de estabilização entre os falantes

(foram investigados 10 falantes por língua, igualmente divididos entre os sexos feminino e masculino) nem entre línguas (foram investigadas 6 línguas). No caso de algumas línguas, houve diferenças entre os estilos (foram estudados três estilos, entrevista semi-espontânea, leitura de frases e leitura de listas de palavras): o estilo entrevista apresenta tempos de estabilização mais variáveis do que os demais.

Os trabalhos [3], [4] permitiram mostrar, ainda que esse não fosse um dos objetivos principais do trabalho, que há grande variabilidade no modo como os falantes fazem uso de qualidades de voz que podem ser compreendidas como não modais, especialmente a chamada pela literatura de voz laringalizada ou crepitante, em português, e *creaky voice*, em inglês. Com isso, surgiu a necessidade da observação e do estudo mais profundo a respeito dos efeitos do registro não modal sobre o tempo de estabilização das medidas estatísticas de F0. Além disso, a investigação a respeito do uso do registro não modal enseja uma busca pela caracterização das idiossincrasias nos contornos de F0, que podem ser capturadas por medidas globais da conformação das distribuições e seu possível potencial na comparação de locutores, embora esse aspecto não vá ser explorado no presente trabalho.

2. Objetivos

No presente trabalho, os objetivos que nos propusemos foram, primeiramente, a extensão do estudo prévio reportado em [3] e [4] devido aos resultados indiretos obtidos na pesquisa, que evidenciaram a presença do registro não modal em muitas amostras de fala presentes no *corpus* por eles estudado. Para isso, replicamos o experimento realizado lá, mudando o foco, que antes estava na análise do contorno de F0 completo, isto é, contendo todos os valores, mas agora privilegiando e investigando o possível papel do registro não modal nos resultados dos tempos de estabilização de F0. Isso foi realizado por meio da comparação dos tempos de estabilização obtidos a partir da análise de contornos de F0 integrais com aqueles obtidos a partir de contornos em que os valores de F0 que podem ser considerados parte do registro não modal são removidos.

3. Materiais e Métodos

O *corpus* comum a esses trabalhos citados e à presente pesquisa insere-se no contexto do projeto internacional "A typology for word stress and speech rhythm based on acoustic and perceptual considerations", coordenado pelo professor Anders Eriksson da Universidade de Estocolmo, Suécia. O *corpus* compreende dados de sete línguas: português brasileiro, sueco, alemão, inglês britânico, francês, italiano e estoniano (a última não foi analisada nos trabalhos anteriores). As amostras das línguas individuais foram coletadas por pesquisadores integrantes do projeto em países em que cada uma das línguas é falada. Além da variedade de línguas, outra razão para a escolha desse *corpus* para uso no projeto é o fato das amostras de fala variarem em termos do estilo de elocução. Três estilos são coletados: entrevista, leitura de frases e leitura de palavras. No estilo entrevista, um entrevistador (em geral um membro da equipe do projeto) faz perguntas ao participante sobre assuntos como trabalho, estudos e outros interesses do entrevistado, visando obter respostas não planejadas e de extensão variável. Para o estilo leitura de frases, um membro da equipe do projeto selecionou frases ditas pelo participante na entrevista, transcreveu-as ortograficamente e pediu que o participante as lesse em voz alta. No estilo leitura de palavras, o procedimento consistiu na escolha de uma palavra de cada frase presente na etapa anterior e na sua apresentação ao participante na forma de uma lista a ser lida.

A primeira etapa, que consumiria bastante tempo e esforço, seria a extração do contorno de frequência fundamental de cada amostra de fala. No entanto, essa parte já havia sido realizada pelo projeto anterior [3], [4]. Uma vez que já possuíamos um contorno de F0 para cada amostra de fala, usamos um *script* do programa Praat [5] que selecionou dez subamostras de 30 segundos para cada estilo de fala, para cada falante. Esse valor foi escolhido em função de resultados anteriores, descritos em [1], que sugeriam que esse intervalo é suficiente para a estabilização da média, mediana e valor de base cumulativos em um conjunto de 26 línguas. Em seguida, outro *script* foi utilizado para obter as medidas estatísticas de localização calculadas de maneira cumulativa. As medidas foram calculadas de forma cumulativa em incrementos fixos de 200 ms. As medidas ou estimadores estatísticos investigados foram as mesmas sugeridas por [1]: a média aritmética, a mediana e o valor de base.

Sendo assim, com os contornos de F0 prontos, geramos os histogramas correspondentes as 210 amostras de fala presentes no *corpus*. O exame visual dos histogramas foi feito com o objetivo de identificar os casos em que o histograma mostra uma destacada multimodalidade (mais do que um pico proeminente). Para cada contorno de F0 extraído a partir dos arquivos sonoros do *corpus*, convertemos os valores em Hertz para a escala OMe (octave median), proposta por [6]. A conversão dos valores foi feita através da seguinte fórmula:

$$f_{Ome} = \log_2\left(\frac{f_{Hz}}{f_{med}}\right)$$

Na fórmula acima, f_{Ome} é o valor convertido para a escala OMe, f_{Hz} é o valor de F0 em Hertz e f_{med} é o valor da mediana de F0 do falante (estimada a partir de todos os valores presentes no contorno a ser convertido). A utilidade dessa escala consiste no fato de que ela usa um valor considerado típico para o falante – a mediana – como fator de normalização para todos os valores de um determinado contorno e expressa a variabilidade dos valores de F0 em torno do valor de referência em termos de oitavas. Essa operação permite identificar facilmente os valores que estão muito acima ou abaixo do valor da mediana nos histogramas.

Na observação dos histogramas, a conversão dos valores mostrados no eixo vertical para uma escala não-linear foi necessária nos casos em que o pico no histograma que corresponde a uma região de registro vocal não modal não era muito aparente. Em alguns casos, como o do contorno cujo histograma de uma amostra de fala do sueco (falante feminina, estilo entrevista) é mostrado na Figura 1, a bimodalidade no histograma é evidente qualquer que seja a escala usada para mostrar a frequência de ocorrência. Em contrapartida, em outros casos, para que a bimodalidade seja mais visível foi preciso recorrer à conversão, como é o caso da Figura 2. Testamos a transformação dos valores de frequência de ocorrência brutos em $\log_{10}$ e em sua raiz quadrada, com resultados parecidos. Na Figura 2, por exemplo, os valores de

frequência de ocorrência foram transformados em sua raiz quadrada. Trata-se de uma amostra produzida por uma falante feminina do português, no estilo leitura de palavras.

Para a identificação visual dessa tarefa, separamos os histogramas em duas regiões, uma em que o falante faz uso de um registro não modal e outra em que predomina o registro modal de vozeamento. Em diversos casos, os histogramas apresentam dois picos claramente identificáveis visualmente como no exame da Figura 1, que permite postular que o contorno de F0 representado ali consiste da sobreposição de duas distribuições de F0 com valores centrais diferentes.

Em histogramas como o mostrado na Figura 1, o limiar foi posicionado no ponto mais baixo do vale formado entre os picos que indicam a localização central dos registros não modal e modal. No caso da Figura 1, o limiar foi posicionado no ponto -0.4 OMe, que corresponde, na escala Hertz, a aproximadamente 146 Hz. Em casos como o mostrado na Figura 2, em que não há um pico tão bem definido correspondente ao registro não modal, procurou-se identificar uma região em que houvesse alguma descontinuidade entre os registros. No caso mostrado em 2, o limiar foi estabelecido no ponto -0.5 OMe, correspondente a 153 Hz.

A audição da amostra de fala e o exame do oscilograma e espectrograma em trechos cujo contorno de F0 é composto por valores próximos ao pico inferior do histograma mostram que nesses trechos os falantes fazem uso de vozeamento não modal, principalmente uma qualidade de voz laringalizada ou crepitante.

Os valores dos limiares obtidos pela identificação visual da bolsista e do orientador foram armazenados em tabelas, as quais relacionavam o falante e o estilo de fala, e auxiliaram a extração das primeiras informações estatísticas, que serão mais bem descritas a seguir.

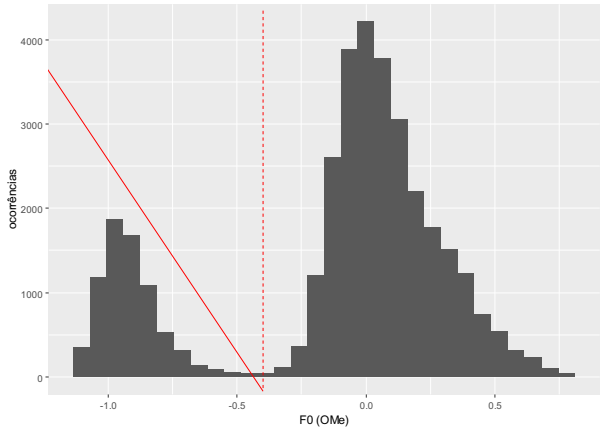


Figura 1: *Histograma da amostra de fala do estilo entrevista de uma falante do sueco. A linha vermelha tracejada indica o limiar identificado entre o registro modal e o não modal.*

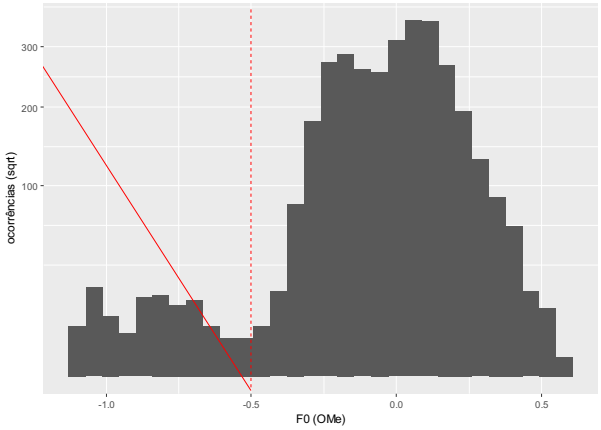


Figura 2: *Histograma da amostra de fala do estilo leitura de palavras de uma falante do português brasileiro. Valores do eixo vertical convertidos em sua raiz quadrada. A linha vermelha tracejada indica o limiar identificado entre o registro modal e o não modal.*

Com essa etapa realizada, surgiram as primeiras estatísticas a respeito do uso de vozeamento não modal, as quais são apresentadas a seguir. Considerando todas as línguas (7), todos os estilos (3) e todos os falantes (10), há 91 casos em que a análise visual indicou a presença de multimodalidade no histograma da distribuição de F0. Isso indica que 43,3% dos contornos são afetados pela presença de vozeamento não modal. Fazendo a análise por estilo de elocução temos as seguintes porcentagens de uso de vozeamento não modal (considerando todas as línguas): 62% no estilo entrevista, 50% na leitura de frases e 40% na leitura de palavras. Analisando as línguas separadamente temos os resultados mostrados a seguir na Tabela 1.

	F 1	**F 2**	**F 3**	**F 4**	**F 5**	**M 1**	**M 2**	**M 3**	**M 4**	**M 5**
Alemão	2	2	1	1	1	1	2	1	2	0
Estoniano	0	3	1	0	0	0	0	0	0	0
Francês	1	1	2	1	2	2	1	2	1	0
Inglês	2	2	2	2	3	2	2	0	2	2
Italiano	1	2	1	1	2	0	0	1	2	2
Português	2	2	2	2	2	3	1	2	1	1
Sueco	1	2	3	2	1	1	2	0	0	0

Tabela 1: *Número de estilos (máximo de 3) em que há presença de vozeamento não modal em função dos falantes e das línguas. Na primeira linha da tabela, F ou M fazem referência ao sexo do falante (F para falantes do sexo feminino e M para falantes do sexo masculino) e o número que segue é o identificador do falante individual dentro de cada língua.*

	Entrevista	**Leitura de**	**Leitura de**

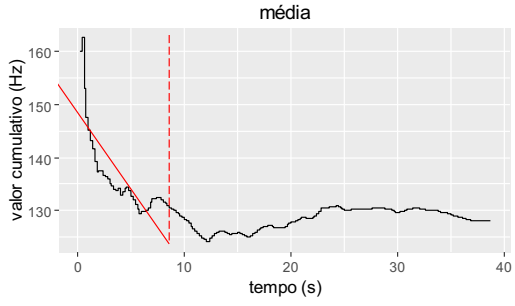
		palavras	frases
Alemão	40	50	40
Estoniano	10	10	20
Francês	80	20	30
Inglês	60	90	40
Italiano	60	20	40
Português	70	70	40
Sueco	50	40	30

Tabela 2: *Porcentagem de falantes que fazem uso de vozeamento não modal em função do estilo de elocução e da língua.*

Após a análise dos histogramas, iniciou-se o momento de processar todos os dados e estatísticas provenientes do *corpus* de forma a replicar a metodologia utilizada em [3] e [4], que calculava o valor de estabilização da F0 de três estimadores estatísticos diferentes: a média, a mediana e o valor de base. A metodologia para determinação dos pontos de estabilização dos três estimadores foi reaplicada, mas aqui apenas a parte dos contornos de F0 que correspondem a vozeamento modal foram consideradas.

Para todos os contornos em que detectamos bimodalidade no histograma, calculamos o valor da média e do valor de base no contorno integral e no contorno filtrado e tiramos a diferença entre os dois. A diferença média na média é de 1,6 Hz (desvio-padrão de 5 Hz e amplitude entre 0 e 36 Hz) e no valor de base é 5,4 Hz (desvio-padrão de 5,4 Hz e amplitude entre 0 e 91 Hz). Apesar das diferenças médias serem relativamente pequenas, a variabilidade e amplitude dessa diferença são altas, justificando seu estudo mais detalhado.

Para identificar o instante de tempo em que a variância dos estimadores diminui ou se estabiliza, momento que chamamos de ponto de estabilização, aplicamos a análise estatística *change point analysis* [7] em cada uma das três curvas de valor cumulativo. A figura 3 a seguir ilustra a aplicação da técnica. Essa técnica produz uma estimativa do ponto em que a variância da série temporal analisada sofre mudança e testa a hipótese de que os valores antes e aqueles depois do ponto de estabilização sejam estatisticamente diferentes. Como utilizamos 10 subamostras de cada estilo para cada falante, foi possível verificar a variabilidade da localização dos pontos de estabilização tanto do ponto de vista intrafalante quanto interfalante.



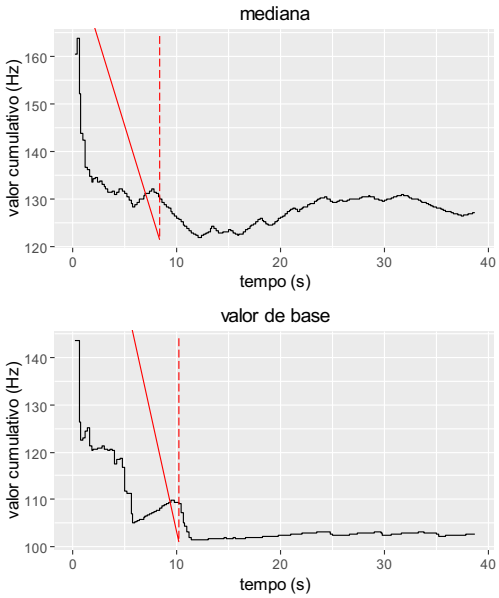


Figura 3: *Valores dos três estimadores estatísticos calculados de forma cumulativa ao longo de uma amostra de F0. A linha vermelha tracejada indica o ponto de estabilização determinado pela aplicação da técnica* change point analysis.

4. Resultados e discussão

Para testar se existe uma diferença significativa do ponto de vista estatístico entre os tempos de estabilização dos contornos integrais e dos contornos filtrados (apenas parte modal) recorremos ao teste-*t* de amostras pareadas. Se o teste não atingir significância podemos dizer que a presença de valores de F0 muito baixos nos contornos de alguns dos falantes, associados a episódios de vozeamento não modal, não provocam um efeito importante no tempo de estabilização. Os testes-*t* pareados foram realizados separadamente para cada um dos estimadores. Reportamos também o resultado do teste de Fligner-Killeen de homogeneidade de variância, usado para comparar a variância das duas amostras, que permitirá dizer se uma das condições (contornos completos ou apenas registro modal) é mais variável do que a outra. Adotamos 5% como nível para rejeição da hipótese nula para todos os testes.

• **Média**: teste de Fligner-Killeen [X^2 (154) = 194,13 p < 0,05], teste-*t* [*t*(2099) = 0,25 ns].

• **Mediana**: teste de Fligner-Killeen [X^2 (152) = 141,46 ns], teste-*t* [*t*(2099) = 0,07 ns].

• **Valor de base**: teste de Fligner-Killeen [X^2 (140) = 138,22 ns], teste-*t* [*t*(2099) = 0,58 ns].

Interpretamos os resultados estatísticos como indicação da robustez do método empregado para a estimação dos pontos de estabilização. Um dos efeitos da presença de uma grande quantidade de valores muito baixos de F0 no contorno por conta do uso mais extenso do vozeamento não modal é o aumento da variabilidade do contorno. Essa maior variabilidade de F0 poderia implicar, em princípio, tempos de estabilização um pouco mais longos, uma vez que a absorção de valores de F0 muito diferentes no cálculo cumulativo do estimador provocaria variabilidade na estimativa. Os resultados que obtivemos, no entanto, parecem indicar que este não é o caso, ou que a possível vantagem dos contornos

filtrados é de magnitude tão baixa que não é captada pelos testes estatísticos. Embora não tenhamos replicado todos os testes estatísticos reportados por [2], a falta de efeito significativo nos testes pareados reportados acima são uma indicação forte de que obteríamos resultados idênticos ou muito parecidos aos reportados lá.

Referências

[1] Arantes P, Eriksson A. Temporal stability of long-term measures of fundamental frequency. In: Campbell N, Gibbon D, Hirst D, editors. Proceedings of the 7th International Conference on Speech Prosody. Dublin: ISCA; 2014. p. 1149–1152.

[2] Arantes P, Eriksson A, Gutzeit S. Effect of Language, Speaking Style and Speaker on Long-term F0 Estimation. Interspeech 2017.

[3] Gutzeit S, Arantes P. Estabilidade de medidas estatísticas de frequência fundamental: comparação da variabilidade intra e interlinguística. Anais do 22o Congresso de Iniciação Científica. São Carlos: Universidade Federal de São Carlos; 2014.

[4] Gutzeit S, Arantes P. Estabilidade de medidas estatísticas de frequência fundamental: comparação da variabilidade intra e interlinguística. Resumos da 67a Reunião Anual da SBPC. São Carlos: Sociedade Brasileira para o Progresso da Ciência; 2015.

[5] Boersma P. Praat, a system for doing phonetics by computer. Glot International. 2001;5(9/10):341–345.

[6] De Looze C, Hirst DJ. The OMe (Octave-Median) scale: A natural scale for speech melody. In: N.Campbell, Gibbon D, Hirst D, editors. Proceedings of the 7th International Conference on Speech Prosody. Dublin; 2014. p. 910–914.

[7] Killick R, Eckley IA. changepoint: An R Package for Changepoint Analysis. Journal of Statistical Software [Internet]. 2014;58(3):1–19. Available from: http://www.jstatsoft.org/v58/i03/