Investigando a robustez de uma metodologia para determinação do valor de base da frequência fundamental

*Probing the robustness of a methodology to determine the base value of the fundamental frequency*

**Resumo:** Este trabalho investiga a robustez de uma metodologia proposta pelos foneticistas suecos Traunmüller e Eriksson para determinar o valor de base, um estimador estatístico do valor típico da frequência fundamental (F0) de um falante a partir da média e do desvio-padrão da F0. A metodologia consiste em estimar uma constante, *k*, que indica quantos desvios-padrão abaixo da média de F0 do falante o valor de base está. O método para estimar a constante foi criado e testado a partir de amostras de fala atuada. Verificamos nesse trabalho se a aplicação da mesma técnica a amostras de fala não atuada produz resultados comparáveis aos reportados por Traunmüller e Eriksson. A investigação usou amostras de fala produzidas por dez falantes nativos de alemão, estoniano, francês, inglês britânico, italiano, português brasileiro e sueco em três estilos de elocução: entrevista, leitura de frases e leitura de palavras. Os resultados indicam que a variabilidade causada pelos estilos de enunciação na F0 permite a aplicação da metodologia a amostras de fala não atuada. Os valores da constante derivados dos dados não atuados são próximos aos reportados pelos autores suecos, indicando que ela é robusta tanto do ponto de vista dos falantes quanto das línguas.

**Palavras-chave:** Entoação; Valor de base; Frequência fundamental.

**Abstract:** This paper probes the robustness of Traunmüller and Eriksson’s methodology to determine the base value of the speaking fundamental frequency, an estimator of a speaker typical F0 value. The methodology entails the estimation of a constant, *k*, indicating how many F0 standard deviations below the F0 mean the base value for a speaker lies. The methodology was originally developed from acted speech samples. Here we test if *k* values can be successfully obtained from non-acted samples and how they compare to the ones reported by Traunmüller and Eriksson. A speech corpus of speech samples differing in speaking styles (spontaneous interview, sentence reading, word list reading) from seven languages (English, Estonian, French, German, Italian, Brazilian Portuguese, Swedish) was used. Results show that *k* values estimated from non-acted speech are roughly the same as those reported in Traunmüller and Eriksson original paper. We speculate that deviations can be explained by the fact that some speakers make extensive use of non-modal register.

**Keywords:** Intonation; Base value; Fundamental frequency.

**1 Introdução**

Ao longo de um enunciado, a frequência fundamental da voz (F0) varia em função de fatores de natureza diferente: fatores linguísticos de escopo amplo, como a modalidade, ou locais, como a composição fonética dos segmentos que compõem o enunciado – cf. a distinção entre micro e macromelodia em Hirst

(2005); fatores paralinguísticos, como o estado emocional do falante no momento da enunciação e, ainda, fatores orgânicos, como o sexo e as idiossincrasias do trato vocal do falante – principalmente massa e comprimento das pregas vocais

(TITZE, 1994). Essa multiplicidade de fatores dificulta a estimativa do valor médio ou típico da F0 que um falante emprega em suas produções faladas.

Em alguns cenários, é interessante que a estimativa de valor típico da F0 reflita fatores orgânicos mais do que fatores linguísticos. Duas situações deste tipo são, por exemplo, a comparação de vozes com finalidade forense

(JESSEN, 2008) e o uso da voz como meio para autenticação da identidade de um usuário em aplicações de segurança

(SCHULTZ, 2007). Em aplicações como essas, a influência que o conteúdo linguístico de enunciados específicos possa vir a exercer sobre os contornos de F0 produzidos por um indivíduo não está no centro das atenções. O que se busca, ao contrário, é minimizar essas influências de forma a fazer os fatores orgânicos/biológicos ressaltarem no estimador estatístico de valor típico.

Pode-se pensar em uma situação em que o inverso seja verdadeiro, isto é, em que o interesse se volta para as os efeitos de um contraste linguístico sobre o comportamento da F0, independente dos falantes que expressam esse contraste. Pode-se estar interessado, por exemplo, em estabelecer qual é o efeito da modalidade interrogativa sobre o contorno de F0. Não interessam, nesse caso, diferenças mais ou menos esperadas entre falantes, como o fato da F0 de homens ser em geral menor do que a de mulheres. O importante é tentar neutralizar essas características idiossincráticas e colocar em deixar em relevo o modo pelo qual a variação de F0 expressa o contraste linguístico em questão. Procedimentos de normalização da curva de F0 podem ser usados para esta finalidade e são geralmente empregados em cenários nos quais diferentes falantes produzem repetições de enunciados em que existe algum contraste linguístico sob investigação. Estes procedimentos, de forma geral, requerem o uso de uma estimativa do valor típico da F0 (em geral a média aritmética) de cada um dos falantes que contribuíram com enunciados para um determinado corpus

(JASSEM, 1975; MAIDMENT; LECUMBERRI, 1996; ROSE, 1991).

Ambos os cenários discutidos acima deixam clara a importância e a utilidade de estudar as características estatísticas das curvas de F0, em especial a adequação das diferentes maneiras de obter uma estimativa do valor típico ou tendência central destas amostras. No contexto da estatística descritiva há diversos procedimentos para determinar o valor mais representativa de uma amostra de dados, cada qual com vantagens e limitações próprias

(KENNEY; KEEPING, 1962). A média e mediana são estimadores de localização versáteis, no sentido de que podem ser aplicados a amostras de qualquer natureza, desde que a variável observada possa ser medida em uma escala intervalar ou proporcional

(STEVENS, 1946). A média aritmética é o estimador de tendência central de F0 cujo uso é mais prevalente na literatura, apesar de sua sensibilidade à presença de assimetria na amostra, que é bastante comum em dados de F0

(JASSEM, 1975). A mediana, mais robusta à presença de assimetrias e valores extremos é uma alternativa à media – cf. a proposta de

(LOOZE; HIRST, 2014) para o uso da mediana como valor de referência para um procedimento de normalização de contornos de F0.

O valor de base (*base value* ou *base line* em inglês) é um estimador estatístico de localização proposto por Traunmüller e Eriksson

([S.d.]) especialmente para amostras de F0 e leva em conta as especificidades típicas desse tipo de amostra. Uma dessas especificidades é que a variação de F0 em geral não é simétrica, como apontamos no parágrafo anterior. Quando os falantes fazem excursões entoacionais, o movimento, na grande maioria das vezes, é ascendente, fato que se revela nos histogramas de distribuições de F0 como uma assimetria positiva. Eriksson

(2011) sugere que o nível de F0 que pode ser considerado típico para um falante é aquele logo acima do mínimo necessário para manter a fonação. Movimentos abaixo deste nível seriam, segundo o autor, menos comuns porque poderiam resultar em vozeamento não-modal. Em situações que fazem a variabilidade da F0 aumentar, como por exemplo falar com maior envolvimento emocional, essa tendência à assimetria se mostra ainda mais claramente. O gráfico da figura 1 mostra o contorno de F0 normalizado temporalmente da mesma frase[[1]](#footnote-1) lida pela mesma pessoa, um ator, simulando três níveis de envolvimento emocional. Em verde, o contorno da elocução com um nível neutro ou típico de envolvimento, em vermelho baixo envolvimento e em azul alto grau de envolvimento. É bastante evidente no gráfico que quanto maior é o movimento envolvimento, maior a gama de valores explorados pelas excursões de F0. As excursões, no entanto, têm uma direção preferencial: as curvas, independentemente do nível de envolvimento, raramente descem abaixo de um ponto em torno de 100 Hz, que funciona como um piso a partir do qual o falante expande a gama tonal. Esse ponto seria o valor de base para esse falante.

Traunmüller e Eriksson ([S.d.]), os autores suecos desenvolvem uma metodologia para estimar o valor de base (Fb) de uma amostra de F0 e propõem a fórmula Fb = Fmédia - *kσ*, em que Fmédia e *σ* são, respectivamente, o valor da média aritmética e do desvio padrão de F0 de uma amostra de F0 e *k* é uma constante determinada empiricamente. Em um experimento com fala atuada emulando diferentes funções paralinguísticas, Traunmüller e Eriksson ([S.d.]) obtiveram o valor de 1,5 para a constante, mas indicaram que este valor não é fixo e pode apresentar uma variação entre 1,1 e 2 – valores obtidos em replicações subsequentes do experimento original.

Lindh e Eriksson

(2007), em um estudo posterior, revisaram o valor de *k* para 1,43 e sugeriram uma formulação alternativa para o cálculo do valor de base, a qual mostraram ser mais robusta do que a original. Nessa formulação, chamada por eles de *alternative base value*, assumindo uma distribuição normal para os dados de F0, o ponto 1,43·*σ* abaixo da média corresponde, aproximadamente, ao 7º percentil da distribuição empírica de F0.

No presente trabalho testamos a robustez da metodologia apresentada por Traunmüller e Eriksson ([S.d.]) para a determinação do valor de base da frequência fundamental da voz. Para isso, ela será aplicada a amostras de fala não atuada, produzidas por falantes de sete línguas, alemão, estoniano, francês, inglês britânico, italiano, português brasileiro e sueco, a fim de observar se os valores da constante *k* estimados a partir da fala não atuada são comparáveis aos obtidos pelos autores a partir da fala atuada e determinar a sensibilidade do valor de base a variações de *k*.



Figura 1: Contornos normalizados temporalmente da mesma frase interpretadas em três níveis de envolvimento emocional por um ator sueco.

**2. A metodologia de Traunmüller e Eriksson**

Utilizamos, neste trabalho, a metodologia descrita em Traunmüller e Eriksson ([S.d.]) para derivação da fórmula do valor de base. Os autores sugerem que o valor de base pode ser entendido intuitivamente como o valor de F0 que corresponderia à situação em que o falante produzisse fala sem nenhuma variação entoacional, isto é, com variabilidade de F0 nula, condição que corresponde ao conceito de carreador na teoria da modulação, proposta por Traunmüller

(1994). O desvio-padrão da F0 dessa situação idealizada seria zero e a média observada refletiria a F0 típica ou preferida daquele falante. A fórmula proposta pelos autores, mencionada na seção anterior, calcula o valor de base a partir de duas incógnitas, Fmédia e *σ*, que podem ser facilmente estimadas a partir de amostras de F0, e uma constante, *k*. A metodologia apresentada pelos autores no trabalho citado apresenta uma maneira empírica de chegar a um valor para *k* a partir de um *corpus* de gravações. Essa seção apresenta os princípios fundamentais dessa metodologia.

Uma vez que amostras de fala reais sempre apresentarão alguma variabilidade, é preciso estimar o valor de F0 que corresponderia a um contorno perfeitamente monotônico a partir de dados naturais. Essa estimativa é feita a partir da aplicação da técnica de regressão linear. Para tanto, é preciso dispor de uma série de pares de dados <média, desvio padrão>, extraídos de um *corpus* de fala natural. Traunmüller e Eriksson recorreram à fala atuada, em função da possibilidade oferecida por esse estilo de enunciação de eliciar o mesmo conteúdo linguístico sob diferentes condições paralinguísticas, que induzem a produção de variabilidade nos contornos de F0. A partir da distribuição dos dados no plano cartesiano formado pelas dimensões média e desvio-padrão a técnica de regressão linear permite estimar o valor que a média de F0 teria se o desvio-padrão fosse nulo, o qual corresponderá ao valor de base para aquele falante.

A aplicação da regressão linear estima a inclinação da reta que melhor descreve a relação linear entre os pontos presentes no plano. Se usarmos a equação *y = a·x + b* para descrever essa reta, então a inclinação corresponde ao parâmetro *a*, *y* corresponde aos valores de desvio padrão e *x* aos valores da média de F0. O valor de base corresponderia ao valor médio de F0 para o qual o desvio padrão seria nulo, o que corresponderia, linguisticamente, à F0 que um falante produziria numa fala hipotética perfeitamente monotônica, não influenciada pelos diversos fatores que produzem variação em seu valor.

Dada a reta estimada pela análise de regressão linear, o valor de base (Fb), isto é, o ponto em que a linha de regressão cruza a linha horizontal *y =* 0 pode ser obtido através da expressão Fb = −*b/a*.

Pode-se reescrever a equação da reta de maneira a evidenciar a dependência de Fb em relação à média e ao desvio-padrão de F0 substituindo-se as variáveis *x* e *y* por Fmédia e *σ*. Substituindo Fb, Fmédia e *σ* na fórmula Fb = Fmédia - *kσ* pelos valores estimados na amostra analisada, obtemos um valor para a constante *k*. Esse valor de *k* pode ser usado na fórmula Fb = Fmédia - *kσ* para determinar o valor de base de qualquer amostra de F0 Mesmo sendo derivado a partir de dados de apenas um falante, os autores afirmam que o valor da constante *k* obtido dessa maneira deve, em princípio, funcionar bem para encontrar o valor de base em amostras de fala de qualquer falante em qualquer língua.

A figura 2 é uma representação esquemática das informações da regressão linear relevantes para a aplicação da metodologia de Traunmüller e Eriksson. Na figura, os pontos azuis são hipotéticos pares de valores <média, desvio-padrão> coletados em um *corpus*, o quadrado vermelho está localizado no ponto que corresponde à média das médias e à média dos desvios-padrão. A linha azul é a linha de regressão linear estimada a partir dos pontos, *a* indica o coeficiente de inclinação da reta, *b* o ponto em que a reta intercepta o eixo *y* e Fb é a localização do valor de base, isto é, o ponto no eixo *x* (média de F0) quando o desvio-padrão (eixo *y*) tem valor 0.



Figura 2: Representação esquemática das informações da regressão linear relevantes para a aplicação da metodologia de Traunmüller e Eriksson.

**3. Materiais e métodos**

**3.1 Materiais de fala**

O material de fala usado no experimento vem do *corpus* coletado no âmbito 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 seis línguas: alemão, estoniano, francês, inglês britânico, italiano, português brasileiro e sueco. As amostras das línguas individuais foram coletadas por pesquisadores integrantes do projeto em países em que cada uma das línguas é falada. Em função da uniformidade dos procedimentos de coleta, o *corpus* permite a comparação interlinguística do fenômeno de interesse em línguas com características diversas. São contempladas seis línguas da família indo-europeia (três do ramo românico e três do ramo germânico) e uma da família urálica (estoniano). 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. A literatura mostra que a variação no estilo de elocução é um dos fatores que causam variabilidade em medidas de longo de termo de F0, como a média e o desvio-padrão

(ESKÉNAZI, 1993; HOLLIEN; HOLLIEN; JONG, 1997; LLISTERRI, 1992). Essa variabilidade é importante no contexto do presente trabalho porque permite a aplicação da regressão linear como método para estimar como o valor médio de F0 varia em função do desvio-padrão, um dos fundamentos da metodologia de Traunmüller e Eriksson ([S.d.]), descrita na seção 2. 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) fez 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 da etapa anterior e na sua apresentação ao participante na forma de uma lista a ser lida. Foram analisadas amostras de fala de dez falantes de cada língua, cinco do sexo masculino e cinco do feminino, uma amostra de cada estilo, totalizando 210 amostras de fala (= 7 línguas ×10 falantes × 3 estilos).

**3.2 Extração dos dados**

A primeira parte da análise consistiu na extração dos valores de F0 de cada uma das 210 amostras de fala do *corpus*. A extração aconteceu em duas etapas. Na primeira etapa, o contorno de F0 foi extraído por meio do uso de um *script* do programa Praat escrito pelo primeiro autor, que otimiza a escolha dos parâmetros *floor* e *ceiling* do algoritmo de extração de F0 do Praat. Essa otimização tem como objetivo diminuir os erros de estimação de F0 mais comuns produzidos pela função *To Pitch*, baseada na técnica da autocorrelação. Na segunda etapa, os arquivos *Pitch* gerados na fase anterior foram corrigidos manualmente. Nessa fase, um segundo *script* foi usado para auxiliar a identificação dos erros não eliminados na etapa anterior. O *script* identifica duas amostras de F0 sucessivas, separadas por 80 milissegundos ou menos, em que o primeiro valor é 1,5 vezes maior ou menor do que o segundo. Nos pontos do contorno de F0 apontados pelo *script* como suspeitos de conter erro de extração, o trecho do oscilograma correspondente foi examinado visualmente para que fosse possível decidir se os valores de F0 estimados pelo Praat naquele trecho correspondiam à periodicidade identificada visualmente na forma de onda.

**3.3 Verificação da congruência entre variação na média e no desvio-padrão**

Na etapa de análise seguinte, um outro *script* do Praat processou os contornos corrigidos de F0 das 210 amostras, para extrair os valores de média e desvio-padrão de cada um. Esses valores foram usados para verificar um pressuposto da metodologia dos autores suecos. Para que a técnica de regressão linear possa ter sucesso na estimativa do valor da constante *k*, é necessário que haja variação nas médias e nos desvios-padrão dos três estilos de fala, e que a variação no desvio-padrão seja diretamente proporcional à variação na média, isto é, o estilo com maior valor de média deve apresentar também o maior valor de desvio-padrão e vice-versa. Caso isso não ocorra, a reta estimada pela regressão pode ter um coeficiente de inclinação nulo ou negativo, o que resulta em um valor negativo para a constante *k*. Valores negativos para *k* não fazem sentido, pois resultariam em valores de base localizados acima da média de F0, contrariando a intuição que fundamenta a proposição do valor de base.



Figura 3: Diferenças (em semitons) entre a média e o desvio-padrão dos estilos entrevista e leitura de frases e entrevista e leitura de palavras.

A figura 3 mostra um panorama da relação entre a variação nos parâmetros média e desvio-padrão nos três estilos de elocução nas setes línguas do *corpus*. A figura mostra as diferenças entre os valores da média e do desvio-padrão das amostras de fala tanto do estilo leitura de frases e quanto leitura de palavras em relação à média do estilo narrativa para os dez falantes de cada língua. A diferença entre os estilos foi calculada entre os valores de média e desvio-padrão expressos na escala de semitons. Esse procedimento foi adotado para que não houvesse grandes discrepâncias entre os dados dos falantes do sexo feminino e masculino. As falantes do sexo feminino são identificadas pela cor vermelha e os masculinos pela cor azul. Os cinco falantes de cada sexo são identificados por símbolos diferentes, conforme a legenda. Os pontos abaixo da linha horizontal pontilhada correspondem aos casos em que os valores para os estilos leitura de frases ou palavras (indicados por marcas no eixo horizontal) são menores do que os da narrativa para o falante em questão. Os pontos acima da linha correspondem a casos em que o valor do estilo narrativa é menor do que aquele ao qual ele é comparado.

A observação da figura 3 mostra que a influência dos estilos de elocução sobre a variabilidade da média e do desvio-padrão de F0 não é uniforme entre os falantes e entre as línguas. Os falantes m4 do português, m3 do sueco e m1 do inglês são exemplos em que diferenças no desvio-padrão entre os estilos vão na mesma direção das diferenças na média – no caso dos dois primeiros, frases > entrevista e palavras > entrevista e, no caso do último, frases < entrevista e palavras < entrevista. Essa configuração favorece a aplicação da metodologia de Traunmüller e Eriksson ([S.d.]). Há casos em que a diferença entre os estilos observada na média se dá em sentido oposto no desvio-padrão, como ilustram os falantes f2 do italiano e m4 do estoniano: frases > entrevista e palavras > entrevista nas médias, mas frases < entrevista e palavras < entrevista nos desvios-padrão.

No que diz respeito à influência da língua sobre os valores de média e desvio-padrão de F0, a figura 3 permite observar que no português o estilo entrevista tem médias e desvios-padrão maiores do que os outros dois estilos. No francês e no italiano, por outro lado, predominam casos em que o estilo entrevista tem as médias mais baixas, embora esse padrão não se reflita no desvio-padrão. A tabela 1 apresenta a porcentagem de falantes em cada língua cuja variação de desvio-padrão é diretamente proporcional à da média. Para esse cálculo, as comparações entrevista-frases e entrevista-palavras foram agrupadas. A inspeção da tabela confirma que o português foi a língua na qual a estratégia de manipulação do estilo de elocução foi mais bem-sucedida no sentido de produzir dados adequados á aplicação da metodologia a ser testada. Entre os estilos, a porcentagem é de 57% tanto nas comparações entrevista-frases quanto nas comparações entrevista-palavras. Entre os sexos, a porcentagem é de 51% para as mulheres e 63% para os homens. Os dados de todas as línguas foram agrupados para a realização do cálculo nas comparações entre estilos e sexos.

Tabela 1: Porcentagem de falantes cuja variação entre a média e o DP se dá no mesmo sentido

|  |  |
| --- | --- |
| **Língua** | **%** |
| Alemão | 60 |
| Estoniano | 55 |
| Francês | 55 |
| Inglês | 55 |
| Italiano | 55 |
| Português | 75 |
| Sueco | 60 |

 Uma das explicações para a incongruência entre o comportamento da média e do desvio-padrão em um número grande dos falantes da amostra pode ser o uso de um registro vocal não modal. A figura 4 mostra o histograma dos valores de F0 da amostra do estilo entrevista da falante f2 do italiano. Os valores de F0 estão expressos na escala OMe (*octave median*), proposta por De Looze e Hirst

(2014). Os valores de F0 em Hz (*fHz*) são transformados para a escala OMe (*fOMe*) por meio da fórmula *fOMe* = log2(*fHz*/*fmed*), onde *fmed* é o valor da mediana de F0 do falante, estimada a partir de todos os valores presentes no contorno a ser convertido[[2]](#footnote-2). O histograma indica que a amostra de F0 é bimodal. A parte da distribuição centrada em torno do valor -1 está uma oitava abaixo da mediana, que para esta falante é 207 Hz. A bimodalidade tem como efeito baixar a média da amostra (183 Hz para o contorno todo, 228 Hz excluindo da amostra de F0 os valores abaixo de -0.35 OMe), mas aumentar muito o desvio-padrão. No caso desta falante, quase 38% de todos os valores de F0 estão bastante abaixo do valor mediano, naquilo que se poderia chamar de uma região de vozeamento não modal.



Figura 4: Histograma da distribuição de F0 (na escala OMe) da falante italiana f2, estilo entrevista.

**3.4 Extração da constante *k***

 Para estimar o valor da constante *k* usando a metodologia de Traunmüller e Eriksson é preciso dispor de uma distribuição de valores de média e desvio-padrão de F0. Esses valores foram gerados a partir dos contornos de F0, cuja extração é descrita na seção 3.2, segundo o procedimento descrito a seguir. Os arquivos de som foram segmentados manualmente para identificar os trechos de fala e as marcações foram armazenadas em arquivos *TextGrid* do programa Praat. Nas amostras do estilo entrevista foram marcados os trechos de fala entre pausas maiores do que 300 ms e nos estilos leitura de frases e leitura de palavras foram marcadas as frases e palavras individuais. Um *script* do Praat foi desenvolvido para selecionar aleatoriamente trechos marcados no arquivo *TextGrid* até que a duração acumulada desses trechos atinga pelo menos 60 segundos. O contorno de F0 dos trechos individuais selecionados é concatenado e a média e o desvio-padrão do contorno resultante são calculados. A operação é repetida dez vezes para cada para cada estilo de fala, de modo que são obtidos para cada falante trinta pares <média, desvio-padrão>. O procedimento de regressão linear é aplicado aos trinta pontos da amostra e o valor da constante *k* é determinado a partir dos parâmetros relevantes, segundo explicado na seção 2.

 Dado o componente aleatório no procedimento descrito no parágrafo anterior, decidimos investigar se as estimativas de *k* para cada falante produzidas por sua aplicação é estável. Para tanto, o procedimento descrito no parágrafo anterior foi repetido dez vezes para cada falante, de modo que para cada um deles obtivemos dez estimativas para o valor de *k*.

**4. Resultados**

 A figura 5 mostra os gráficos de dispersão e a curva de regressão linear ajustada aos dados dos 10 conjuntos coletados para o falante f1 do português. É possível ver que, apesar de haver alguma variabilidade, a distribuição dos dados em cada gráfico de dispersão é bastante similar, indicando que a estimativa de *k* é estável para esse falante em particular. Os valores de *k* variam entre 0,73 e 0,84, com coeficiente de variação de 0,4%. Os valores altos do coeficiente de determinação (*r2*) da regressão linear – entre 0,88 e 0,95 – indicam um bom ajuste da reta em relação aos pontos.



Figura 5: Gráficos de dispersão com reta de regressão linear superposta de 10 amostras da falante brasileira f1.

Em contraste, a figura 6 ilustra o caso de uma falante, f2 do português, cujo padrão de variação da média e do desvio-padrão não é adequado à aplicação da metodologia de estimação da constante *k*. É possível observar que a reta de regressão ora tem inclinação positiva (repetições 1 e 6, por exemplo), ora inclinação negativa (repetições 3 e 7, p.e.) e em alguns casos aparenta ter inclinação nula (repetição 9). Conforme é possível observar na figura 3, a falante apresenta diferenças na média entre os estilos (entrevista maior do que leitura de frases e palavras), embora o desvio-padrão seja basicamente o mesmo para os três estilos. Essa característica não faz dessa falante uma boa candidata à aplicação da metodologia de estimativa de *k* por meio da análise de regressão. Podemos ver isso na imensa variabilidade dos valores de *k* que a técnica estima para esse falante: mínimo de -11,36 e máximo de 38,89, com coeficiente de variação de 140%. Os valores de *r2* são bastante baixos, variando entre 0,002 e 0,26.



Figura 6: Gráficos de dispersão com reta de regressão linear superposta de 10 amostras da falante brasileira f2.

 A figura 7 mostra a distribuição dos valores da constante *k* estimados para as sete línguas do *corpus*. Os falantes estão dispostos no eixo horizontal e os valores estimados para a constante *k* aparecem no eixo vertical. Cada ponto corresponde a uma estimativa do valor de *k*. Em todas as línguas há falantes, como f2 do português, para os quais a aplicação da metodologia resulta em valores de *k* negativos, que não fazem sentido e são omitidos. O número de falantes em que casos desse tipo aconteceram variou entre um no italiano e quatro no francês e no estoniano. O valor médio de *k* para a amostra total é 2,24 com intervalo de confiança de 95% em torno da média de ± 0,13. A tabela 2 lista a média, o intervalo de confiança em torno da média e o coeficiente de variação das estimativas do coeficiente k para cada língua.

Tabela 2: Média, intervalo de confiança de 95% e coeficiente de variação da constante *k* para cada língua

|  |  |  |  |
| --- | --- | --- | --- |
| **Língua** | **Média** | **Intervalo de confiança** | **CV (%)** |
| Alemão | 2,8 | ± 0,32 | 47 |
| Estoniano | 2,14 | ± 0,27 | 48 |
| Francês | 2,23 | ± 0,4 | 72 |
| Inglês | 1,68 | ± 0,36 | 82 |
| Italiano | 2,71 | ± 0,43 | 69 |
| Português | 1,54 | ± 0,17 | 48 |
| Sueco | 2,48 | ± 0,3 | 51 |

 Para estabelecer a significância da variável independente língua sobre o valor de *k*, recorremos à aplicação de um teste estatístico de hipótese. A amostra não cumpre o pressuposto da homogeneidade de variância, necessário para o uso de um teste paramétrico, conforme testado pelo teste Fligner-Killeen: [Χ2 (6) = 17,3 *p* < 0,01]. O teste não paramétrico Kruskal-Wallis foi usado no lugar da análise de variância e indica um efeito estatisticamente significativo do fator língua sobre o valor médio de *k* [Χ2 (6) = 77 *p* < 0,001]. Análise das comparações pareadas indica que o português e o inglês, as línguas com os menores valores médios de *k*, formam um grupo homogêneo. As demais línguas não se agrupam de nenhuma maneira particular. Os valores do português e do inglês são os que mais se aproximam dos valores pontuais 1,5, usado em Traunmüller e Eriksson ([S.d.]), e 1,47, sugerido por Lindh e Eriksson (2007). A maioria das médias concentra-se em uma faixa muito próxima à indicada por Traunmüller e Eriksson ([S.d.]), que vai de 1,1 a 2. Considerando o limite inferior dos intervalos de confiança em torno da média, alemão, italiano e sueco ficam acima do limiar de 2.

 No Gráfico 8, apresentamos os valores de *r2*, estimados nas dez amostras de cada falante, agrupados por língua. Os falantes estão dispostos no eixo horizontal e os valores de *r2* no eixo vertical. Quanto mais próximo de 1 é o valor de *r2*, melhor é o ajuste da reta estimada a partir da técnica de regressão linear aos dados da amostra.

 Como é possível observar, o sueco e o italiano são as línguas que apresentam as maiores proporções de falantes com valores de *r2* acima de 0,75, que indica um bom ajuste da reta estimada a partir da técnica de regressão linear. A inspeção conjunta das figuras 7 e 8 sugere que os falantes que apresentam valores baixos de *r2* tendem a apresentar maior variabilidade nos valores de *k*. A correlação entre o valor médio de *r2* por falante e o desvio-padrão de *k* calculado por falante é de -0,74, indicando uma relação forte entre as duas variáveis. Uma análise de regressão simples foi usada para predizer os valores médios de *r2* a partir dos valores médios de desvio-padrão de *k*. Uma equação de regressão significativa foi encontrada [F(1, 47) = 58,7 *p* < 0,001], com coeficiente de determinação (*r2*) de 0,54. O aumento de uma unidade de desvio-padrão no valor de *k* implica na redução de aproximadamente 40% no *r2* da regressão linear.



Figura 7: Valores da constante *k* separados por falante (cinco femininos e cinco masculinos) e por língua.



Figura 8: Valores do coeficiente de determinação (*r2*) das análises de regressão linear agrupados por falante e por língua.

**5. Sensibilidade do valor de base em relação à constante *k***

O valor estimado para a constante *k* varia entre os falantes de uma mesma língua e entre línguas diferentes. Por conta disso, é importante ter uma ideia da variabilidade causada na estimativa do valor de base pelo uso de diferentes valores possíveis de *k*. Para isso, fizemos uma simulação em que o valor de *k* foi sistematicamente variado e o valor de base correspondente foi calculado. Utilizamos nessa simulação os contornos de F0 de todos os falantes e todos os estilos da amostra de dados do português. Para cada contorno, o valor de base da F0 foi calculado a partir da fórmula Fb = Fmédia - *kσ*, variando o valor de k entre 0,8 e 2,2, com passos intermediários em 1,27 e 1,73. Os valores mínimo e máximo estão próximos aos limites da faixa de variabilidade encontrada nas análises reportadas na seção 1.

A figura 9 mostra os resultados dessa variação, separados pelos estilos de fala. O sexo dos falantes é codificado pela cor e os diferentes falantes por símbolos diferentes. No eixo horizontal estão os quatro valores de *k* testados e, no eixo vertical, o valor de base para cada falante, em Hertz.



Figura 9: Variação do valor de base (Hz) em função do valor da constante *k* (formulação original).

Replicamos o teste com a formulação alternativa do valor de base

(LINDH; ERIKSSON, 2007), que estima aquele valor como um determinado quantil da amostra de F0.

A constante *k* pode ser entendida como a indicação de quantos desvios-padrão abaixo do valor da média o valor de base está localizado. Se assumirmos que os valores de F0 seguem uma distribuição normal centrada em zero e com desvio-padrão unitário, a função *pnorm(-k)* da linguagem de programação R retorna o valor cumulativo de probabilidade da distribuição normal compreendido no intervalo [-∞, -k]. Esse valor, que chamaremos de *q*, pode ser interpretado como o quantil que corresponde ao valor de base. Seguindo esse método, o valor de base foi estimado como sendo os quantis 0,01, 0,04, 0,1 e 0,21. Na formulação alternativa do valor de base, os autores sugerem o uso do quantil 0,074 para a determinação do valor de base.

Tabela 3: Quantis correspondentes ao valor de base e sua relação com os valores de *k*.

|  |  |
| --- | --- |
| Valor de *k* | Quantil correspondente ao Fb |
| 0,8 | 0,21 |
| 1,27 | 0,1 |
| 1,73 | 0,04 |
| 2,2 | 0,01 |

A figura 10 mostra os resultados da variação do valor de base segundo a formulação alternativa, separados pelos estilos de fala. O sexo dos falantes é codificado pela cor e os diferentes falantes por símbolos diferentes. No eixo vertical está o valor de base, em Hertz, para cada falante, e no eixo horizontal os valores do quantis que correspondem à localização do valor de base (ver tabela 3).



Figura 10: Variação do valor de base (Hz) em função do quantil (formulação alternativa).

Para os propósitos em que o uso do valor de base pode ser mais útil, robustez não significa que os valores retornados pela fórmula sejam estritamente invariantes para um mesmo falante, mas sim a preservação das diferenças entre os valores calculados pela fórmula para os diferentes falantes.

A tendência geral, dedutível a partir da fórmula, é que quanto maior *k*, menor será o valor de Fb. Observe-se o painel central da figura 9, que corresponde à leitura de frases. Ali, os valores de Fb obtidos quando *k* é igual a 0,8 estabelecem uma ordenação entre os falantes: f1 > f5 > f3 > f2 > f4 > m2 > m3 > m5 > m1 > m4. Apesar de diferenças nos valores absolutos do Fb para cada falante, a ordenação observada anteriormente permanece inalterada quando o valor de *k* sobe para 1,27; uma única alteração aparece quando *k* é igual a 1,73 (m3 = m5). Finalmente, quando *k* é igual a 2,2 há uma inversão (m5 > m3). Pelo menos no estilo leitura de frases, verificamos que o cálculo do valor de base segundo a formulação original é relativamente robusto em relação às possíveis variações de *k* no sentido que definimos anteriormente: a ordenação dos dez falantes em termos de seu valor de base fica quase inalterada não importa qual seja o valor definido para *k*.

Considere-se, agora, para o painel central da figura 10: o mesmo tipo de análise nos leva a observar que para parte dos falantes a ordenação tende a permanecer estável a despeito das mudanças no quantil que corresponde ao Fb, com exceção dos falantes f2, f5 e m3, que em algum momento apresentam mudança brusca na passagem de um valor de quantil a outro.

Para poder quantificar o grau de robustez das duas formulações do cálculo de Fb, a original e a alternativa, além dos diferentes estilos de fala, determinamos, para cada estilo e para cada valor de *k* ou *q*, a distância euclidiana entre os valores de Fb de todos os falantes, tomados em pares. O desvio-padrão das distâncias será então considerado um indicador de robustez, considerados os diversos agrupamentos de variáveis independentes (formulação do valor de base, estilo de fala e sexo dos falantes e os valores de *k* e *q*). Menores valores de desvio-padrão indicarão maior robustez.

Tabela 4: Desvio-padrão (Hz) das distâncias entre o valor de base dos falantes, agrupado pelos estilos de fala.

|  |  |  |
| --- | --- | --- |
| **Estilo de fala** | **Original** | **Alternativa** |
| Entrevista | 35,3 | 36,8 |
| Frases | 30,6 | 32,8 |
| Palavras | 33,5 | 33,5 |

Tabela 5: Desvio-padrão (Hz) das distâncias entre o valor de base dos falantes, agrupado pelo sexo dos falantes.

|  |  |  |
| --- | --- | --- |
| **Sexo do falante** | **Original** | **Alternativa** |
| Feminino | 21,2 | 30,6 |
| Masculino | 18,5 | 14,3 |

Tabela 6: Desvio-padrão (Hz) da distância entre o valor de base dos falantes, agrupado pelos valores de *k* e *q*.

|  |  |  |  |
| --- | --- | --- | --- |
| **Valores de *k*** | **DP** | **Valores de *q*** | **DP** |
| **0,8** | 30,4 | **0,21** | 32,2 |
| **1,27** | 30,7 | **0,1** | 32,5 |
| **1,73** | 31,3 | **0,04** | 33,8 |
| **2,2** | 33 | **0,01** | 33,2 |

Tabela 7: Desvio-padrão (Hz) da distância entre o valor de base dos falantes, agrupado pela interação entre estilos de fala e sexo dos falantes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Estilo de fala** | **Sexo do falante** | **Original** | **Alternativa** |
| Entrevista | Feminino | 24,8 | 37,2 |
| Masculino | 20,9 | 13,6 |
| Frases | Feminino | 13,1 | 27 |
| Masculino | 16 | 16 |
| Palavras | Feminino | 19 | 24,9 |
| Masculino | 14 | 11,8 |

As duas formulações parecem ter o mesmo grau de robustez quando se compara o fator estilo de fala, uma vez que o desvio-padrão das distâncias entre os falantes não varia muito em função desta variável. O sexo dos falantes apresentou uma relação de interação complexa: a formulação original parece ser mais robusta para as mulheres e a alternativa para os homens; além disso, de forma geral as duas formulações parecem mais robustas quando aplicadas aos dados dos falantes masculinos. A Tabela 6, que mostra a interação entre estilo de fala e sexo do falante mostra que a diferença de robustez mais pronunciada entre os sexos se dá na formulação alternativa, em especial no estilo entrevista.

A observação da figura 10 mostra que alguns falantes apresentam comportamento mais discrepante em relação aos demais em termos da mudança no valor de base em função da variação no valor do quantil associado a ele. No estilo entrevista, os falantes f3, f4, f5 e m3 têm mudanças mais abruptas, no estilo leitura de frases, os falantes f2, f5 e m3 devem ser destacados e no estilo leitura de palavras, os falantes f2 e f5. A observação dos histogramas dos contornos produzidos por estes falantes em cada estilo indica o uso sistemático do registro não modal de vozeamento, semelhante ao padrão mostrado na figura 4. Por conta disso, quando o valor do quantil que corresponde ao valor de base assume valores mais baixos, como 0,04 ou 0,01, o Fb estimado começa a estar localizado na região de registo não-modal.

**6. Conclusão**

O principal objetivo do presente trabalho é testar a robustez da metodologia desenvolvida e apresentada por Traunmüller e Eriksson ([S.d.]) para a determinação do valor de base da F0. O valor de base seria característico de cada falante, em tese invariante ou pelo menos bastante robusto a diversos fatores que afetam a F0 se determinado a partir de uma amostra suficientemente extensa. Na proposta dos autores, a fórmula para a determinação do valor de base depende do valor da média e desvio-padrão do falante, além de uma constante *k*, cujo valor é determinado empiricamente. No trabalho mencionado acima, os autores apresentam uma metodologia para a estimação da constante, baseado na aplicação de regressão linear a dados de média e desvio-padrão de F0. Nos experimentos descritos pelos autores, lança-se mão de fala produzida por atores, que simulam o efeito de fatores paralinguísticos, como por exemplo diferentes atitudes em relação aos enunciados produzidos (entusiasmo, tédio etc.). Esse recurso é usado para produzir enunciados idênticos do ponto de vista segmental, mas variáveis do ponto de vista da média e do desvio-padrão da F0. Uma característica fundamental que as amostras de F0 precisam exibir para que a metodologia seja aplicada é que a variabilidade nas médias e nos desvios-padrão sejam diretamente proporcionais, isto é, que as amostras com maior média sejam também as que apresentem os maiores desvios-padrão.

No presente trabalho, testamos se o uso de diferentes estilos de elocução de fala não atuada é capaz de produzir o tipo de variabilidade na média e no desvio-padrão dos contornos de F0 necessário para a aplicação da metodologia para estimar o valor da constante *k*. Além desse fator, testamos ainda o papel de falantes e línguas como fonte de variabilidade na estimação de *k*. Para tanto, nossa investigação analisa dados produzidos por 70 falantes de sete línguas diferentes.

Os resultados reportados aqui indicam que a estratégia de usar diferentes estilos de elocução para conseguir variabilidade na média e no desvio-padrão dos contornos de F0 produz padrões que permitem a aplicação bem-sucedida da metodologia. O uso de registro não modal, extensivo no caso de alguns dos falantes do *corpus*, no entanto, é um fator que parece explicar os casos de insucesso. Em estudos posteriores pode ser interessante propor um critério objetivo para eliminar dos contornos os trechos de vozeamento não e verificar o impacto dessa eliminação nos resultados.

De modo geral, os valores de *k* estimados a partir das amostras de fala não atuada são bastante próximos àqueles que os autores suecos reportam em seu trabalho e que foram derivados de amostras de fala atuada. Portanto, pode-se dizer que a técnica é robusta ao uso de fala não atuada. Diferenças entre falantes existem, mas elas podem ser associadas em grande parte aos casos em que a regressão linear tem um valor de *r2* baixo e acontecem nos dados dos falantes com maior prevalência de uso do registro não modal. Verificamos a existência de diferenças entre as línguas, que, muito embora significativas do ponto de vista estatístico, não são de grande extensão. O valor médio de *k* de quatro das sete línguas está dentro do intervalo [1,1 2] relatado por Traunmüller e Eriksson ([S.d.]).

 Finalmente, os resultados da simulação apresentados na seção 5 mostram que o próprio valor de base é uma medida que é bastante robusta às variações no valor da constante *k*. Dado um grupo de falantes, sua ordenação baseada no valor de base é pouco alterada pelo valor de *k* que se escolha usar. Uma vez que um dos usos mais interessantes para o valor de base é como um estimador do valor típico ou característico de um falante, essa quase invariância nas distâncias entre o valor de base dos falantes é uma propriedade interessante.

**Agradecimentos**

Os autores agradecem ao professor Anders Eriksson, da Universidade de Estocolmo, pela cessão do *corpus* analisado no trabalho. A segunda autora agradece à FAPESP pela Bolsa de Iniciação Científica (processo 2014/21161-5).

**Referências Bibliográficas**

ERIKSSON, Anders. Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work. In: NEUSTEIN, A.; PATIL, H. A. (Org.). *Forensic Speaker Recognition: Law Enforcement and Counter-terrorism*. [S.l.]: Springer, 2011. p. 41-70.

ESKÉNAZI, Maxine. Trends in Speaking Styles Research. 1993, Berlin: ISCA, 1993. p. 501-509. Disponível em: <http://www.isca-speech.org/archive/eurospeech\_1993>.

HIRST, Daniel J. Prosodic aspects of speech and language. In: BROWN, K. (Org.). *Encyclopedia of Language and Linguistics*. [S.l.]: Elsevier Science, 2005. v. X. p. 167-178.

HOLLIEN, Harry; HOLLIEN, Patricia; JONG, Gea De. Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America*, v. 102, n. 5, p. 2984-2992, 1997.

JASSEM, Wiktor. Normalisation of F0 curves. In: FANT, GUNNAR; TATHAM, M. A. A. (Org.). *Auditory Analysis and Perception of Speech*. London: Academic Press, 1975. p. 523-530.

JESSEN, Michael. Forensic Phonetics. *Language and Linguistics Compass*, v. 2, n. 4, p. 671-711, 2008.

KENNEY, J. F.; KEEPING, E. S. Mathematics of Statistics. [S.l.]: Van Nostrand, 1962. p. 50-54.

LINDH, Jonas; ERIKSSON, Anders. Robustness of Long Time Measures of Fundamental Frequency. 2007, Antwerp, Belgium: [s.n.], 2007. p. 2025-2028.

LLISTERRI, Joaquim. Speaking styles in speech research. 1992, Dublin, Ireland: [s.n.], 1992.

LOOZE, Céline De; HIRST, Daniel. The OMe (Octave-Median) scale: A natural scale for speech melody. 2014, Dublin: [s.n.], 2014. p. 910-914.

MAIDMENT, J. A.; LECUMBERRI, M. L. Pitch analysis methods for cross-speaker comparison. 1996, Delaware: [s.n.], 1996.

ROSE, Philip. How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication*, v. 10, n. 3, p. 229-247, 1991.

SCHULTZ, Tanja. Speaker Characteristics. In: MÜLLER, CHRISTIAN (Org.). *Speaker Classification I: Fundamentals, Features, and Methods*. [S.l.]: Springer, 2007. p. 47-74.

STEVENS, S. S. On the theory of scales of measurement. *Science*, 1946.

TITZE, Ingo. *Principles of voice production*. Englewood Cliffs: Prentice Hall, 1994.

TRAUNMÜLLER, H. Conventional, biological and environmental factors in speech communication: a modulation theory. *Phonetica*, v. 51, p. 170-183, 1994.

TRAUNMÜLLER, Hartmut; ERIKSSON, Anders. *The frequency range of the voice fundamental in the speech of male and female adults*. [S.d.]. Disponível em: <http://www2.ling.su.se/staff/hartmut/f0\_m&f.pdf>.

1. A frase, em sueco, é “Nån av mammorna hann lämna honom” e uma tradução aproximada seria “Algumas das mães puderam deixá-lo”. As gravações foram cedidas por Anders Eriksson. [↑](#footnote-ref-1)
2. A utilidade dessa escala está 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 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. [↑](#footnote-ref-2)