RELATÓRIO DE PESQUISA

Efeito do estilo de elocução e do falante sobre o tamanho mínimo de amostra para estimativa da taxa de produção da fala

Pablo ARANTES
Universidade Federal de São Carlos (UFSCar)

RESUMO

Investigamos o papel do estilo de elocução e do falante sobre a estimativa do tamanho mínimo de amostra necessário para estimar de forma estável a taxa de produção da fala. Os estilos de elocução comparados são a entrevista semiespontânea e a leitura de frases. São analisadas 20 amostras de fala, 10 em cada estilo, de 5 falantes do sexo masculino e 5 do feminino. Os tempos de estabilização são definidos como o momento ao longo da série temporal da taxa de produção calculada de forma cumulativa no qual há uma redução da variabilidade na série de dados. Dois critérios para definir estabilidade são apresentados e comparados, um baseado em análise estatística e outro em um limiar perceptual. Testamos também o efeito do aumento progressivo (começando com 30 s e chegando a 300 s) da amostra submetida aos procedimentos de detecção de estabilidade. Os resultados mostram que os tempos médios de estabilização dependem do critério usado para sua detecção, mas são em geral mais longos para o estilo semiespontâneo, variando entre 60 e 70 s para a leitura e 80 e 110 s para a fala semiespontânea. Os tempos de estabilização tendem a ser mais longos quando a duração da amostra é maior. O sexo do falante não tem impacto relevante sobre o tempo de estabilização. As estimativas de tempo de estabilização variam entre diferentes falantes quase tanto quanto a variabilidade intrafalante. Os resultados são relevantes para a área da fonética forense porque sugerem com base em uma metodologia explícita e reprodutível qual é a duração mínima que uma amostra de fala precisa ter para que

OPEN ACCESS

EDITADO POR
- Luma Miranda (ELTE)
- Manuella Carnaval (UFRJ)
- Carolina Gomes da Silva (UFPB)

AVALIADO POR
- Saulo Mendes Santos (UFMG)
- Ubiratã Kickhofel Alves (UFRGS)

DATAS
- Recebido: 30/10/2023
- Aceito: 09/01/2024
- Publicado: 27/05/2024

COMO CITAR
Arantes, P. (2024). Efeito do estilo de elocução e do falante sobre o tamanho mínimo de amostra para estimativa da taxa de produção da fala. *Revista da Abralin*, v. 23, n. 2, p. 125-159, 2024.

se estime a partir dela a taxa de produção da fala para fins de comparação de locutor.

ABSTRACT

We investigated the role of speaking style and individual speakers on estimating the minimum sample size required for stable estimation of speaking rate. The compared speaking styles are semi-spontaneous interviews and sentence reading. We analyzed 20 speech samples, 10 in each style, from 5 male and 5 female speakers. Stabilization times are the point along the time series defined by successive values of cumulative speaking rate where variability is reduced. Two criteria for defining stability are presented and compared, one based on the change point statistical analysis and one on a perceptual threshold. We also tested the effect of progressively increasing the sample size submitted to stability analysis (starting with 30 seconds and reaching up to 300 seconds). The results show that average stabilization times depend on the criteria used for detection, but are generally longer for the semi-spontaneous style, ranging from 60 to 70 seconds for reading and 80 to 110 seconds for semi-spontaneous speech. Stabilization times tend to be longer as the sample duration increases. Speaker sex has no significant impact on stabilization times. Estimates of stabilization time vary among different speakers almost as much as intra-speaker variability. The results are relevant to forensic phonetics applications because they suggest, based on an explicit and reproducible methodology, what is the minimum duration a speech sample needs to have in order to estimate from it the speech production rate for speaker comparison purposes.

PALAVRAS-CHAVE

Prosódia. Taxa de produção da fala. Taxa de elocução. Taxa de articulação. Fonética forense.

KEYWORDS

Prosody. Speaking rate. Speech rate. Articulation rate. Forensic Phonetics.

Introdução

A prosódia é o subnível de análise no âmbito da fonética e da fonologia que considera os fenômenos cujo escopo são domínios mais extensos do que segmentos individuais (LEHISTE, 1970). Uma dimensão da organização prosódica no nível temporal é a taxa de produção de unidades linguísticas produzidas por unidade de tempo, responsável pela percepção do que é comumente referido no senso comum como "velocidade de fala", parâmetro que varia em um contínuo que vai da fala lenta à rápida. Do ponto de vista da pesquisa linguística, a taxa de produção da fala é um parâmetro bastante estudado em razão de sofrer o efeito de diferentes variáveis linguísticas e paralinguísticas relevantes, como variedade dialetal, estilo de elocução, estado emocional e sexo do falante, entre outras (BARBOSA, 2019, p. 54).

Propomos a adoção da expressão "taxa de produção da fala" como um termo guarda-chuva, uma vez que a mensuração da taxa de produção de unidades linguísticas durante a enunciação pode se basear em critérios diferentes e receber designações específicas em função disso (BARBOSA 2019, p. 53-55). O tratamento dado às pausas, por exemplo, dá origem a dois tipos de taxa, que refletem aspectos da fala ligeiramente diferentes. Quando as pausas que ocorrem entre duas unidades linguísticas são contadas como parte da duração de uma delas, a medida resultante é chamada de "taxa de elocução". Quando as pausas não entram no cômputo da duração, ela é chamada de "taxa de articulação". A unidade linguística tomada como referência pode variar, indo da palavra ao fone individual, gerando medidas com diferentes graus de refinamento. A unidade temporal também pode ser diferente, sendo o minuto e o segundo duas escolhas típicas. Nesses dois casos, no entanto, não há nomes específicos associados às diferentes possibilidades. O escopo de medição da taxa é um terceiro parâmetro que pode variar. Nesse sentido, fala-se de "taxa global" ou "taxa local". Usa-se taxa global quando o escopo da medida é mais largo, isto é, inclui um número elevado de unidades linguísticas, como um enunciado completo (PFITZINGER, 1996) ou todo o conteúdo de fala entre duas pausas (KENDALL, 2013). Fala-se em taxa local quando ela é determinada a partir de janelas de análise de escopo mais restrito, que abrangem um número menor de unidades linguísticas, que podem ser tanto sílabas quanto fones individuais dependendo do algoritmo específico adotado. Taxas locais tendem a refletir alongamentos e compressões na duração segmental de caráter mais local, como o fenômeno conhecido como alongamento final (EDWARDS; BECKMAN; FLETCHER, 1991; WIGHTMAN et al., 1992), que podem estar correlacionados com a sinalização de proeminências e fronteiras prosódicas ao longo do eixo sintagmático da fala. Barbosa (2006), por exemplo, propõe uma medida para a taxa local de produção da fala, aplicável a diversos estilos de elocução, que identifica modulações temporais relativas ao longo de um enunciado e pode ser usada para identificar, por exemplo, pontos nos quais acontecem alongamentos significativos do ponto de vista da produção e da percepção, que podem ser atribuídos à presença de proeminências rítmicas.

Neste trabalho, nos preocupamos com a questão da duração mínima que uma amostra de fala deve ter para que seja possível inferir de maneira estável o valor de diferentes modos de determinar a taxa de produção da fala. Do ponto de vista teórico, é possível pensar que para a fala ser eficiente como modalidade de externalização da língua, os diferentes sons linguísticos precisam ser distinguíveis.

Portanto, a fala é uma atividade que, por necessidade, implica produzir padrões acústico-articulatórios variáveis. Essa variação se impõe porque uma língua deve dispor de um repertório variado de unidades distintivas e também pelo fato de uma mesma entidade em sentido abstrato poder ser realizada de maneiras diferentes dependendo do contexto. Essas considerações nos fazem pensar sobre quanto material de fala é necessário para que uma amostra de fala contenha dados suficientes a respeito de um determinado parâmetro fonético para representar de modo adequado a gama do repertório de unidades distintivas e as possibilidades de variação contextual dessas unidades em uma determinada variante linguística. Transpondo a discussão para o domínio prosódico, a questão se torna: quanto material de fala (sílabas, palavras ou outra unidade) é necessário para que a taxa de produção estimada a partir dele se torne representativa do comportamento de longo prazo do falante?

Do ponto de vista metodológico, podemos imaginar, a título de exemplo, um pesquisador interessado em desenhar um experimento no qual falantes de diferentes regiões dialetais do país serão gravados para determinar se o valor médio da frequência fundamental dos falantes varia conforme a região. Uma decisão importante, nesse caso, é definir quanto tempo de fala de cada participante é necessário gravar para que a medida a ser extraída tenha confiabilidade. Se a amostra for muito breve, a medida pode não refletir o padrão típico de cada região. Se, por outro lado, a amostra precisar ser muito longa, isso poderá implicar muito trabalho manual para fazer a gravação, dar tratamento e fazer a medição de cada amostra coletada. É importante, portanto, que uma decisão a esse respeito seja tomada com base em um critério relevante e objetivo. Kendall (2013) pode ser citado como um exemplo de trabalho que investiga essa questão no âmbito da pesquisa sociolinguística baseada em corpora de fala de grande escala. O autor investiga o impacto do tamanho da amostra de fala de cada falante sobre a variabilidade da medida da taxa de elocução. Os resultados de seu trabalho serão discutidos mais adiante.

1. Uso da taxa de produção da fala na fonética forense e as implicações da determinação da duração mínima de amostra para sua estimação

Nesta seção discutimos o uso aplicado que a área da fonética forense faz do parâmetro taxa de produção da fala e, a seguir, explicamos porque a questão da determinação da duração mínima que uma amostra de fala deveria ter para sua estimação tem grande interesse nesse contexto.

Do ponto de vista da fonética forense, Eriksson (2011, p. 52–54) justifica o interesse pelo estudo da taxa de produção da fala a partir do seguinte raciocínio: "sabe-se que padrões motores muito automatizados variam de indivíduo para indivíduo, mas tendem a ser estáveis e invariantes uma vez que tenham sido estabelecidos firmemente"[1]. Essa estabilidade já foi demonstrada para atividades motoras

[1] No original: "highly automatic motor patterns have been shown to vary from individual to individual but tend to be stable and invariant once they have been firmly established" (Tradução nossa).

não relacionadas à fala, como a marcha (*gait*) e a datilografia (*typing*); a articulação da fala é uma atividade motora e, portanto, pode apresentar essa mesma estabilidade, muito embora essa afirmação ainda precise de maior grau de corroboração. A literatura especializada já demonstrou que a taxa de produção da fala, em especial a taxa de articulação, tem um razoável poder discriminador em tarefas de comparação de locutores (KÜNZEL, 1997; OLIVEIRA, 2021), isto é, ela varia entre diferentes falantes mais do que entre diferentes amostras de fala de um mesmo falante. Levantamentos feitos entre especialistas forenses de todo o mundo a respeito das práticas que adotam em exames de comparação de vozes (GOLD; FRENCH, 2011, 2019) indicam que 93% deles regularmente investigam variáveis temporais e, desses, 73% analisam algum tipo de taxa de produção da fala, seja a de elocução ou de articulação. Apesar do uso bastante difundido da taxa de produção da fala na prática pericial, pouco se investigou a respeito do tamanho mínimo que uma amostra de fala deve ter para que o valor da taxa de produção extraído das amostras analisadas em exames periciais seja representativo do comportamento de longo termo do falante.

Por que essa questão é relevante no contexto forense? Jessen (2009) e Gfroerer (2003) afirmam que a quantidade limitada de material de fala é um problema comum no contexto de tarefas de natureza forense. Gfroerer (2003) menciona 20 segundos como uma duração típica de amostra de fala encontrada em casos forenses reais. Jessen (2009) diz que não há um limiar estabelecido abaixo do qual a comparação de duas amostras seria impossível e afirma que "at least something like eight seconds of speech from the anonymous speaker and at least about double that time for the suspect is recommended" (JESSEN, 2009, p. 16), muito embora o autor não apresente nenhuma fundamentação técnica para justificar os valores mencionados. Falando especificamente sobre o uso da frequência fundamental no contexto da fonética forense, Eriksson (2011) diz que a média e o desvio-padrão da frequência fundamental são muito empregados. Em seguida, acrescenta que "to some extent means and standard deviations depend on the duration of the speech sample but there is no general agreement on what minimum duration is required to yield representative results" (ERIKSSON, 2011, p. 49). A preocupação é que amostras muito curtas não contemplem de forma representativa a gama de variação nos padrões sonoros tipicamente apresentada por um falante, limitação que é muito relevante no contexto de uma tarefa de comparação de locutores com finalidade forense.

Ainda no contexto da fonética forense, a questão do tamanho mínimo de uma amostra necessário para a estimação de um determinado parâmetro acústico da fala é relevante por uma razão adicional. Na doutrina criminalística atual (MORRISON, 2010, 2009), verifica-se um movimento em direção ao emprego de metodologias de avaliação do valor probatório de evidências produzidas em exames periciais que se baseiam na comparação entre as observações obtidas em amostras particulares (tanto a chamada amostra questionada quanto aquelas coletadas dos indivíduos suspeitos) e dados que reflitam a distribuição populacional dos parâmetros estudados no exame pericial. Essa metodologia, portanto, exige a compilação de bancos de dados de amostras de fala que permitam a estimação da distribuição populacional dos parâmetros acústicos mais relevantes para a tarefa de comparação de vozes. Existem estatísticas sobre a distribuição populacional do parâmetro acústico f0 para populações falantes de línguas como o sueco (LINDH, 2006), mandarim (CAO; LEI, 2017), tcheco (SKARNITZL; VAŇKOVÁ, 2017) e inglês britânico

(HUDSON et al., 2007). Para o parâmetro taxa de articulação, podemos citar os dados de Künzel (1997) e Jessen (2007), para a língua alemã e Cao e Wang (2011) para o mandarim. Na montagem de bancos de dados coletados com esse tipo de objetivo em mente, dois fatores muito relevantes são a definição da quantidade de falantes a serem gravados e a quantidade de fala a ser gravada de cada falante.

2. Pesquisas prévias

A pesquisa na literatura que investiga especificamente a questão do estabelecimento de métodos para a determinação da duração mínima de amostra de fala para a estimação da taxa de produção da fala mostra a existência de poucos trabalhos a respeito do tema. Há um conjunto restrito de estudos que serão comentados a seguir. Eles serão brevemente descritos primeiramente e a seguir discutiremos alguns aspectos em que as metodologias propostas por eles são divergentes.

A referência cronologicamente mais antiga é Kendall (2013), um livro que reúne e sintetiza diversas pesquisas do autor na área da sociofonética, nas quais a principal variável fonética analisada é a taxa de produção da fala. Em um dos capítulos de caráter mais metodológico da obra, o autor se põe basicamente a questão da quantidade mínima de dados necessária para obter uma estimativa estável da taxa de produção da fala. No estudo reportado em Kendall (2013), o autor investiga a questão em um corpus de fala de grande proporção, contendo material lido por 80 falantes de inglês norte-americano de diferentes regiões do país. O autor apresenta uma metodologia para investigar a questão, que consiste em calcular a taxa de produção (em particular, a taxa de articulação, calculada em sílabas por segundo) de maneira cumulativa ao longo de um enunciado e tratar o resultado como uma série temporal e posteriormente identificar o ponto no tempo em que se pode dizer que a série de algum modo se estabiliza. Para o cálculo da taxa de articulação, o autor soma a duração de todas as sílabas presentes em um enunciado fonético (o termo original empregado é *phonetic utterance*, que o autor define como todo o material linguístico contido em um trecho de fala entre pausas) e dividindo-se, então, esse resultado pelo número de sílabas presentes no intervalo. O autor propõe como critério para definição de estabilização da série temporal o uso de um limiar derivado experimentalmente para a percepção de diferenças na taxa de produção de fala. Usando esse critério, o autor conclui que o tamanho mínimo da amostra de fala necessário para a estabilização da estimativa da taxa é algo entre 80 e 200 enunciados fonéticos. Como no corpus usado pelo autor a extensão média do enunciado é 12 sílabas e a taxa de elocução média no mesmo corpus é de 4,43 sílabas/s, podemos obter a duração típica da sílaba, calculando o recíproco desse valor, que dá aproximadamente 0,226 segundos. Multiplicando esse valor por 12, sabemos que o enunciado fonético típico dura 2,7 segundos. Assim, um conjunto de 80 enunciados fonéticos dura em média 3,6 minutos e um conjunto de 200 dura em torno de 9 minutos.

Arantes (2015) e Arantes e Lima (2017) realizaram estudos iniciais nos quais analisaram séries temporais constituídas pelos valores das taxas de articulação e elocução, calculadas de forma cumulativa ao longo de amostras de fala lida, com o auxílio da técnica estatística conhecida como *change point analysis* (KILLICK; ECKLEY, 2014). O objetivo foi encontrar o ponto no tempo (em termos da duração

da amostra de fala analisada a partir de seu início) em que o valor da taxa atinge um patamar de variabilidade que pode ser considerado estável. A mesma técnica estatística havia sido aplicada com sucesso para a determinação do ponto de estabilização em séries temporais definidas pelo cálculo cumulativo de estimadores estatísticos de tendência central da frequência fundamental (ARANTES, 2014; ARANTES; ERIKSSON, 2014; ARANTES; ERIKSSON; GUTZEIT, 2017).

Em seu estudo, Arantes e Lima (2017) estudaram fala lida e variaram as taxas de articulação e elocução em três níveis: normal (o nível habitual do falante), rápida e lenta em relação ao habitual. Os resultados mostram que o tempo médio de estabilização da taxa de articulação é 8,7 segundos e o da taxa de elocução é de 9,2 s, não havendo diferença estatística significativa entre esses dois valores. Há, entretanto, diferença entre a média da taxa rápida (7,8 s) e a da taxa lenta (11,1 s). Os autores usaram a unidade VV[2] como unidade linguística para o agrupamento das durações dos segmentos e o segundo como unidade de tempo. O número médio de unidades VV contidas no intervalo entre o início das amostras e o ponto de estabilização é 46 no caso da taxa de elocução e 54 no da taxa de articulação e essa diferença é significativa do ponto de vista estatístico. Não se registrou diferença estatística entre os tempos de estabilização dos três níveis das duas taxas.

Estudos posteriores (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) ampliaram a pesquisa inicial e incluíram o cálculo das taxas de articulação e elocução usando mais unidades linguísticas de agrupamento além de unidades VV (fone, sílaba fonológica e palavra fonológica), dois critérios para a contagem das unidades linguísticas (fonético e fonológico) e dois diferentes critérios para o estabelecimento do ponto de estabilização: o adotado por Kendall (2013) e o critério estatístico, adotado por Arantes e Lima (2017), ambos discutidos em detalhe na seção 4 do presente trabalho. Os resultados indicam que as variáveis testadas nos novos experimentos têm efeitos sobre os tempos de estabilização, embora as diferenças, mesmo quando sejam estatisticamente significativas, são de baixa magnitude, de modo que do ponto de vista prático não são muito relevantes. As médias de tempo de estabilização das diversas combinações entre os critérios testados não passam de 20 segundos (sugere-se a consulta aos textos citados para conferir a totalidade dos sumários estatísticos gerados). De modo geral, são valores muito próximos aos relatados em Arantes e Lima (2017) e, crucialmente, muito mais baixos do que os sugeridos pela literatura anterior, em especial Kendall (2013) e Jaffe e Breskin (1970), trabalho citado por Kendall (2013).

As metodologias empregadas por Kendall (2013) e por Arantes e colegas (ARANTES, 2015; ARANTES; LIMA, 2017) diferem em aspectos relevantes, principalmente em relação ao escopo do cálculo da taxa de produção da fala e ao critério adotado para a definição do ponto de estabilização. Em trabalhos posteriores de Arantes e colegas (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018), houve a tentativa de replicar aspectos da metodologia de Kendall (2013) aos dados analisados em Arantes e Lima (2017) e estudar possíveis diferenças nos resultados dos tempos de estabilização causadas pelas

[2] Unidade de agrupamento dos segmentos da cadeia da fala que tem tamanho isomórfico ao da sílaba fonológica. Compreende os segmentos produzidos entre dois ataques vocálicos consecutivos. Ver Pettorino *et al.* (2013) e Barbosa (2019, p. 40 e 55-60) para uma justificativa do uso das unidades VV no estudo de fenômenos prosódicos.

diferenças metodológicas. Descreveremos em detalhe, na seção 4.1, os aspectos técnicos das duas metodologias e das suas diferenças.

3. Objetivos do presente trabalho

As pesquisas descritas na seção anterior, apesar das diferenças metodológicas apontadas, têm em comum o fato de usarem dados de fala lida. Como dissemos na Introdução, o estilo de elocução, em particular a fala lida em comparação com diferentes gêneros de fala espontânea, é uma variável que afeta a taxa de produção da fala. Os dados da literatura mostram, no geral, que a fala lida tende a apresentar taxas de produção mais elevadas do que diferentes modos de elocução mais espontânea (BÓNA, 2014; CRYSTAL; HOUSE, 1982; HIROSE; KAWANAMI, 2002; HOWELL; KADI-HANIFI, 1991). Tendo em vista a influência robusta dos estilos de elocução sobre a taxa de produção da fala, um dos objetivos do trabalho reportado aqui é investigar o efeito da variação no estilo de elocução sobre o tempo mínimo necessário para a estimativa de taxa de produção da fala. Comparamos os tempos de estabilização da taxa de produção da fala lida e da fala em estilo semiespontâneo, conforme descreveremos na seção 4.2. Além disso, investigamos também o papel que os falantes exercem sobre a estimativa do tempo de estabilização. A observação desses efeitos se desdobra no exame do possível efeito do sexo do falante sobre os tempos de estabilização e também na estimativa da variabilidade dos tempos de estabilização observada entre os diferentes falantes da amostra. Finalmente, investigamos também o possível efeito da duração total da amostra de fala sobre a estimativa dos tempos de estabilização e a variação individual dessas estimativas.

4. Metodologia

4.1. Critérios para análise da taxa de produção da fala

Nos trabalhos anteriores a respeito do tema da estimativa de taxa de produção da fala que apresentamos na seção 2, a taxa investigada é calculada de forma cumulativa ao longo de um enunciado. Caso a unidade linguística adotada seja a sílaba, por exemplo, o cálculo da taxa se inicia pela primeira sílaba, depois abrange a primeira e a segunda, em seguida abrange da primeira até a terceira, de modo que, ao chegar à última sílaba do trecho analisado, todas elas são incluídas no cômputo. A expressão (1) a seguir formaliza esse cálculo.

$$cSR_i = \frac{i}{\sum_{j=1}^{i} dur_j} \quad (1)$$

Na expressão (1), cSR_i indica o valor da taxa cumulativa (*cumulative speaking rate*) desde a primeira unidade linguística até a unidade de índice i em uma dada amostra de fala segmentada. A fórmula será calculada no intervalo [1, n], sendo n o número total de unidades linguísticas na amostra de fala considerada. Os valores consecutivos da taxa de produção cumulativa são interpretados como os elementos de uma série temporal a ser analisada para determinar o ponto em que essa série alcança estabilidade. Como dissemos na seção 2, há diferentes critérios usados para determinar o que é estabilidade nesse contexto, que explicaremos mais adiante nesta seção.

Conforme dissemos na Introdução, quando todas as unidades linguísticas que compõem um determinado trecho de fala a ser analisado são incluídas no cálculo da taxa de produção, dizemos que essa é uma taxa global, que podemos abreviar por r_g. Usamos esse valor como referência para calcular o erro de estimação da taxa quando calculada de forma parcial, isto é, levando em conta um subconjunto do total de unidades do enunciado ou trecho de fala completo. No caso das pesquisas resenhadas na seção 2, interessa também determinar o erro de estimação que temos quando se compara o valor da taxa de produção no ponto de estabilização, que podemos abreviar por r_{st}, com o valor da taxa calculada de forma global. Podemos definir o erro de estimação, abreviado por e, como a diferença entre o valor da taxa de produção de fala no ponto de estabilização (r_{st}) e o valor global da taxa (r_g). Para facilitar a interpretação do resultado, a diferença é transformada numa porcentagem relativa ao valor da taxa global, seguindo a expressão (2), mostrada a seguir.

$$e = \frac{r_{st} - r_g}{r_g} \cdot 100 \qquad (2)$$

A determinação do erro de estimativa da taxa de produção no ponto de estabilização é relevante porque esse erro é o critério sugerido por Kendall (2013) para a determinação do ponto de estabilização da taxa de produção cumulativa. Mesmo no caso em que esse não é o critério usado para determinar o ponto de estabilização, é importante saber quanto de erro se pode esperar quando se toma um subconjunto da amostra de fala em vez de sua totalidade para o cálculo da taxa de produção.

Os trabalhos de Kendall (2013) e de Arantes e colegas (ARANTES, 2015; ARANTES; LIMA, 2017; ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) diferem em dois grandes aspectos em relação aos procedimentos adotados para a análise da taxa de produção da fala. O primeiro deles é o escopo ou domínio do cálculo da taxa de produção, seja ela a de articulação ou elocução. Kendall (2013) propõe determinar a taxa de produção (no caso de seu trabalho, apenas a taxa de articulação é estudada) somando-se a duração de todas as unidades linguísticas (a unidade linguística usada por ele é sempre a sílaba) presentes em um enunciado fonético (o termo original empregado por ele é *phonetic utterance*) e dividindo-se esse resultado pelo número de sílabas presentes no intervalo. A definição de enunciado fonético que é possível depreender da leitura de Kendall (2013) é que se trata do material fonético produzido entre pausas silenciosas consecutivas, dado um certo limiar para que o que for considerado pausa exclua eventos como o silêncio de uma plosiva, por exemplo. Nos estudos de Arantes e colegas, o cálculo é feito da maneira mais granular possível, isto é, computando o valor da taxa uma unidade linguística por vez, seja ela qual for: fone, sílaba, unidade VV ou palavra. Nos trabalhos iniciais

(ARANTES, 2015; ARANTES; LIMA, 2017) usou-se a unidade VV e nos trabalhos posteriores (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) as outras unidades foram incluídas.

A segunda diferença metodológica notável diz respeito ao critério para definição do ponto de estabilização da série temporal da taxa de produção cumulativa. Kendall (2013) propõe o uso de um procedimento baseado em um limiar de indistinção perceptual da taxa de produção da fala derivado por Quené (2007). Segundo Kendall (2013, p. 20):

> "Quené (2007) estudou a literatura a respeito de limiares de percepção e notou a escassez de estudos relativos à fala (a maioria tematiza a taxa de produção na música) e realizou um estudo experimental para examinar o limiar de indistinção experimental para a taxa de elocução. Seus experimentos chegam a um limiar de indistinção de 5% relativamente à taxa de produção de um enunciado. Variações na taxa que excedam esse limiar provavelmente são percebidas e podem ser relevantes na comunicação falada."[3]

Partindo desse limiar de 5% em relação a um valor de referência, Kendall (2013) define como ponto de estabilização da taxa de produção o momento no tempo em que a diferença entre a estimativa da taxa de produção da fala cumulativa e taxa global daquela amostra entra no intervalo ± 5% e não sai mais dele pelo restante da amostra.

O critério para determinação do ponto de estabilidade da taxa de produção adotado por Arantes e colegas em seus trabalhos baseia-se na aplicação da técnica estatística *change point analysis* (KILLICK e ECKLEY, 2014). Essa técnica toma como entrada uma série temporal e encontra o ponto no tempo a partir do qual é possível identificar uma mudança significativa na variância subjacente à série de valores. No caso das séries de taxa de produção da fala, a variância nos valores da taxa cumulativa tende a diminuir conforme aumenta o intervalo abrangido pelo cálculo da taxa, isto é, conforme se avança no tempo em um enunciado e a técnica *change point analysis* indica o ponto a partir do qual se pode considerar que essa variabilidade mudou de forma significativa.

Ilustramos, por meio da Figura 1, a diferença entre os dois critérios de definição do ponto de estabilização (mencionados no item 1 acima) e exploramos os resultados produzidos por eles em um dos dados que foram analisados na fase anterior da pesquisa, publicada em Arantes, Eriksson e Lima (2018). Trata-se da série temporal produzida por um falante masculino, taxa em nível lento. O tipo de taxa mostrada é a de elocução, calculado em fones por segundo. A figura mostra a evolução do valor cumulativo da taxa de elocução ao longo da duração da leitura do texto. Ela foi calculada, nesse caso, fone a fone, desde o primeiro até o último que compõe aquela amostra.

---

[3] No original: "Quené (2007) reviewed the literature on JND (Just Noticeable Difference) and noted the paucity of studies relevant to speech communication (most have been about tempo in music) and conducted an experimental study to examine the JND for speech rate. His experiments 'provide an estimated JND of 5% of the base tempo of a speech utterance. Tempo variations exceeding this [difference limen] are likely to be noticeable, and relevant in speech communication".

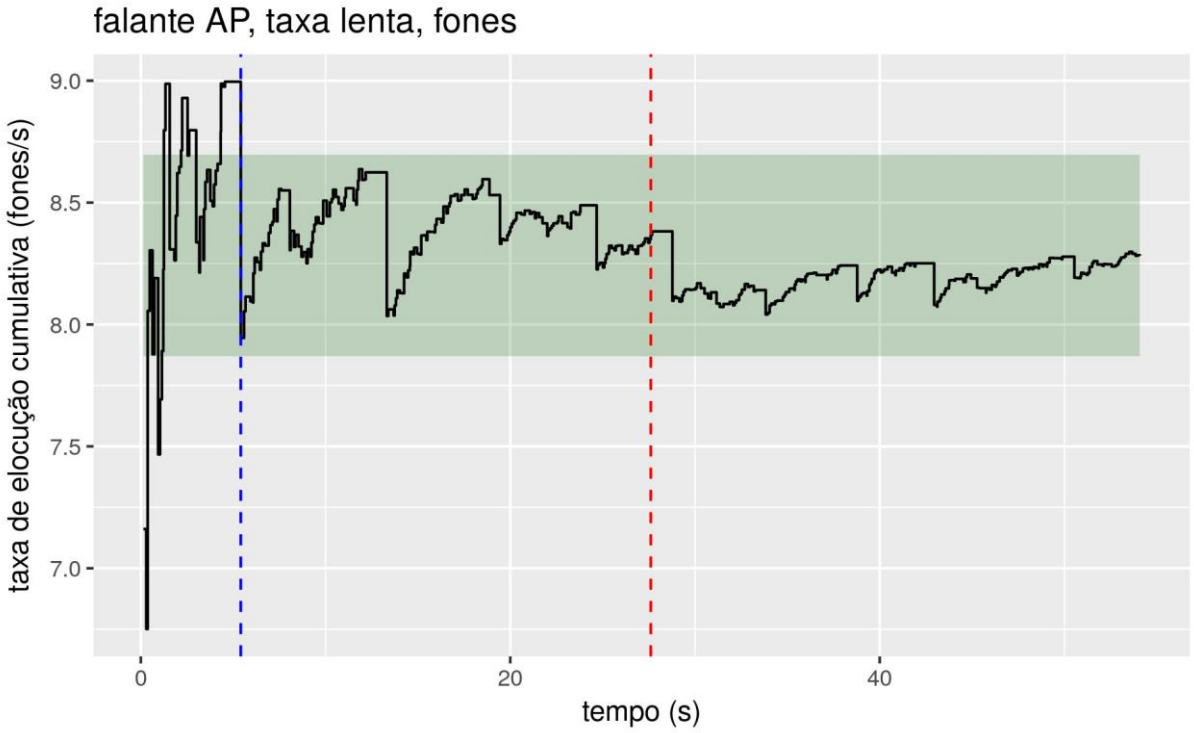

FIGURA 1 – Ilustração dos dois procedimentos para determinação do ponto de estabilização da taxa cumulativa de produção da fala.
Fonte: elaborada pelo autor

Na figura 1, a linha vertical tracejada vermelha indica o ponto de estabilização identificado pela técnica estatística *change point analysis* (27, 6 s). A aplicação do critério baseado no limiar de indiferença perceptual proposto por Kendall (2013) dá como resultado o ponto indicado pela linha vertical tracejada azul, localizado em 5,4 s a partir do início da leitura. As linhas horizontais superior e inferior do retângulo verde indicam, respectivamente, os limites de 5% acima e 5% abaixo do valor da estimativa da taxa de elocução global, isto é, no ponto final da gravação, quando todos os fones são levados em conta no cálculo da taxa (nesse caso, um valor ligeiramente acima de 8,25 fones/s). A partir de 5,4 s, todos os valores da taxa cumulativa ficam dentro do limiar. No exemplo, o critério perceptual estima um tempo de estabilização mais curto do que o critério estatístico.

Pensando no erro de estimação da taxa de produção que aceitamos ao definir o ponto de estabilização, no exemplo da figura 1 o erro é de aproximadamente -4% no caso do critério perceptual[4]. O uso do critério estatístico gera, no exemplo da figura, um erro de 1%, menor do que o gerado pelo segundo perceptual. A variância da série temporal é 13 vezes menor depois do ponto de estabilização no caso do critério perceptual e 16 vezes menor no caso do critério estatístico.

[4] A aplicação do critério de estabilização baseado no limiar de indiferença perceptual sempre gerará, por força da própria definição da metodologia, erros de estimação que serão, em valores absolutos, menores do que 5%. O uso do critério estatístico não limita de antemão o erro de estimativa a um intervalo definido.

Vamos explorar agora as possíveis diferenças que a escolha metodológica a respeito do escopo do cálculo da taxa de produção da fala cumulativa pode causar quando aplicado ao mesmo dado. A figura 2 mostra a mesma série temporal presente na figura 1, a progressão da taxa de elocução cumulativa, calculada fone a fone, mas agora também os pontos na leitura em que ocorrem pausas silenciosas, indicadas pelos retângulos verdes. A base dos retângulos indica a duração de cada pausa.

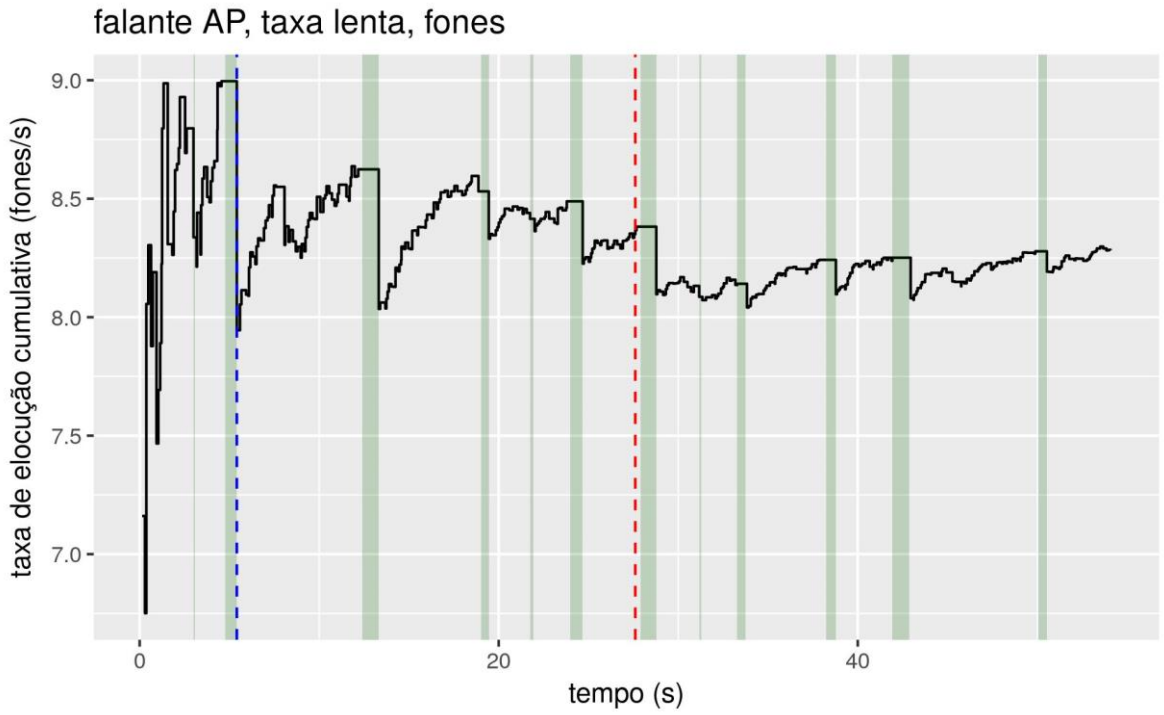

FIGURA 2 – Pausas que delimitam os enunciados fonéticos no enunciado. Pontos de estabilização são os mesmos mostrados na figura 1.
Fonte: elaborada pelo autor

Podem ser identificadas doze pausas na figura 2, que definem, portanto, treze enunciados fonéticos assim como definidos por Kendall (2013). É possível observar que dentro de um enunciado fonético a taxa cumulativa apresenta uma certa variação: em geral, o valor aumenta do início até o final de cada enunciado, fenômeno que poderia ser em parte explicado pelo fenômeno do alongamento final que mencionamos anteriormente. Essa dinâmica interna da taxa de produção ao longo de um enunciado, que é revelada quando a taxa é calculada mais granularmente, poderia ser uma justificativa para preferir o cálculo feito de maneira mais agregada. Quando se olha para os enunciados fonéticos mais à direita no gráfico, no entanto, vemos que a amplitude na variação da taxa ao longo do enunciado fonético reduz-se à medida que o tempo passa e mais dados são incorporados à medida cumulativa da taxa. A diminuição da volatilidade da taxa cumulativa ao longo do enunciado justifica a adoção de um cálculo mais granular, isto é, calculado usando unidades de menor extensão temporal, como o fone ou a sílaba, por exemplo.

Na figura 3, mostramos duas séries de dados. A série de pontos ligados por uma linha tracejada são os valores da taxa de elocução calculada de forma isolada nos doze primeiros enunciados fonéticos do exemplo explorado nas duas figuras anteriores. A série de pontos ligada pela linha contínua mostra o valor da taxa calculada de maneira cumulativa nesses mesmos doze enunciados, o que gera uma variação muito mais suave, com valores para a taxa de elocução oscilando em torno da taxa de 8 fones/s.

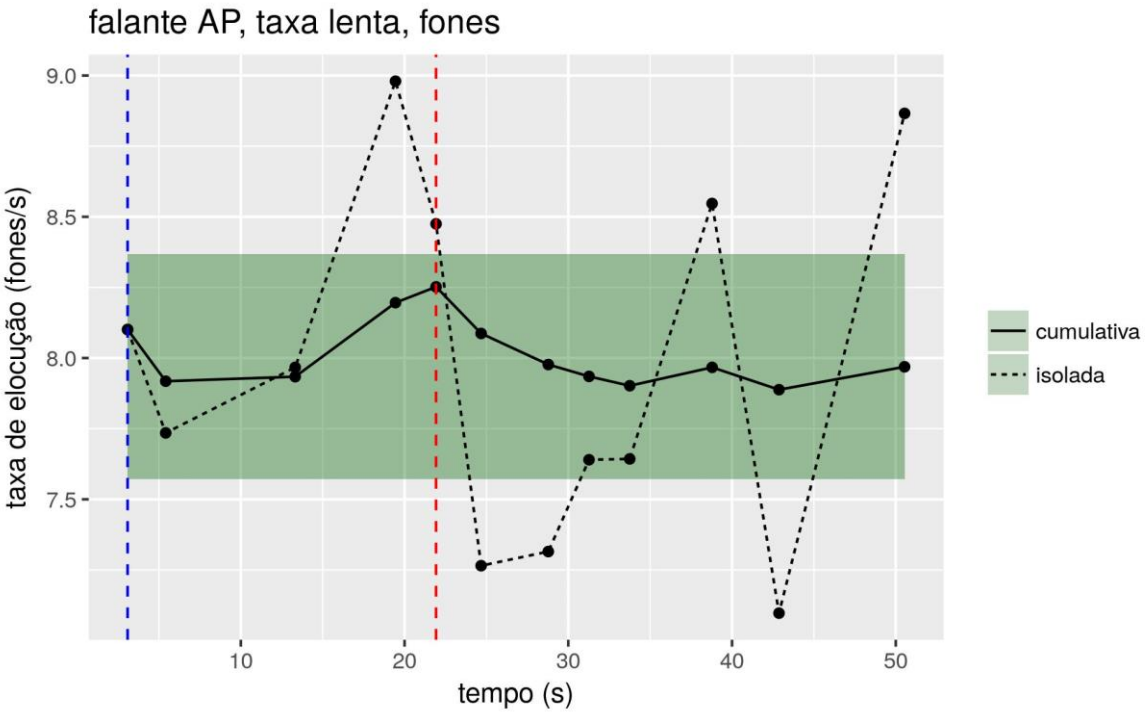

FIGURA 3 — Taxa de elocução calculada tomando como escopo os enunciados fonéticos. A linha tracejada liga os valores em cada enunciado fonético isolado e a linha contínua os valores calculados de maneira cumulativa. Pontos de estabilização são indicados pelas linhas verticais vermelha (critério da variância) e azul (critério do limiar perceptual).
Fonte: elaborada pelo autor.

A comparação entre os valores máximo e mínimo nas figuras 2 e 3 indica que, apesar da diferença na maneira de calcular a taxa de produção, a amplitude de variação é muito parecida, com um mínimo pouco maior do que 7 e um máximo que não ultrapassa 9 fones/s. A aplicação dos dois critérios de identificação do ponto de estabilização à série reduzida de doze valores mostrada na figura 3 resulta em dois valores, cuja localização é dada pela posição dos traços verticais tracejados azul e vermelho. O azul indica o ponto de estabilização identificado pelo critério do limiar perceptual e o vermelho o ponto localizado pela técnica estatística *change point analysis*. A aplicação do critério do limiar perceptual indica que desde o primeiro enunciado fonético, que dura aproximadamente 3 s, a taxa já está dentro da banda de variação de ±5% relativamente ao valor da taxa global, indicada na figura 3 pelos limites horizontais do retângulo verde. O erro de estimação associado a esse ponto de estabilização é de 0,4% e não faz sentido pensar em grau de redução da variância uma vez que a série é considerada estável desde o primeiro ponto. O ponto de estabilização encontrado pela aplicação do critério estatístico é

de aproximadamente 22 s e acontece no quinto enunciado fonético. O erro de estimação é de aproximadamente 2,3% e a relação entre a variância antes e depois do ponto de estabilização é 0,65.

Os dois critérios para identificação da estabilidade da taxa cumulativa de produção da fala descritos nesta seção baseiam-se em princípios diferentes, embora tenham o mesmo propósito. Não estabelecemos aqui nenhum juízo de princípio a respeito da maior adequação de um dos dois para os propósitos do presente trabalho. Uma das vantagens da metodologia baseada na técnica estatística *change point analysis* é que ela pode ser aplicada, em princípio, a qualquer série temporal. Essa flexibilidade permite que a mesma ferramenta seja usada para estabelecer o tempo de estabilização de outros parâmetros acústicos medidos em escala intervalar ou proporcional. Como reportamos na seção 2, isso já foi feito para o parâmetro frequência fundamental e poderia ser estendido ainda a outros parâmetros relevantes para a pesquisa e para aplicações práticas. O critério baseado no limiar perceptual, em contraste, é específico para o parâmetro taxa de produção. Pode-se considerar essa uma vantagem, uma vez que incorpora conhecimento específico para o parâmetro fonético em análise. Os estudos de Arantes e colegas (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) sugerem que a diferença entre as duas metodologias não gerou efeitos importantes sobre os resultados de tempo de estabilização. Essa observação é interessante, pois sugere que os resultados são robustos e relativamente independentes do método usado. Os resultados do presente estudo poderão alinhar-se aos anteriores ou mostrar alguma diferença nas condições presentes no corpus de fala analisado aqui.

4.2. Material de fala

O material de fala que analisamos no presente trabalho são amostras de fala que foram coletadas no âmbito do projeto internacional "A typology for word stress and speech rhythm based on acoustic and perceptual considerations"[5]. O corpus é composto por três tipos de gravações: (i) uma *entrevista espontânea*, durante a qual um entrevistador faz perguntas abertas ao entrevistado, que pode falar livremente sobre o assunto com um mínimo de intervenção por parte do entrevistador; (ii) a entrevista foi transcrita por membros do projeto e dias depois o participante voltava ao laboratório para a *leitura de frases* retiradas da transcrição da sua própria entrevista; (ii) *leitura de palavras* isoladas, retiradas das frases selecionadas para a etapa anterior. O corpus completo é composto por gravações de sete línguas, sendo uma delas o português brasileiro. Dez falantes do PB foram gravados segundo esse protocolo e as amostras de fala desses falantes foram analisadas no presente trabalho. Os 10 participantes (5 do sexo feminino, 5 do sexo masculino) são falantes nativos da língua e representantes da variedade linguística típica do interior do estado de São Paulo, especialmente da região de Campinas, com idades variando entre 18 e 30 anos, todos com escolaridade universitária completa ou em andamento no momento da coleta dos dados. Uma das vantagens do desenho do corpus para nossos propósitos é que o mesmo material linguístico foi produzido

[5] Mais informações a respeito dos objetivos do projeto, procedimentos de coleta dos dados e informações sociodemográficas dos participantes podem ser vistas no endereço https://wordstress.ling.su.se/. O projeto é coordenado por Anders Eriksson, da Universidade de Estocolmo, Suécia.

nos três estilos, de modo que diferenças de tempo de estabilização podem ser atribuídas à diferença de estilo e não ao conteúdo segmental das amostras de fala. Não usamos as gravações de leitura de palavras isoladas neste estudo. A razão é que a duração das gravações nesse estilo tem em média 40 palavras e uma duração líquida (apenas fala, desprezando as pausas) típica de 30 segundos, que consideramos ser pouco material para a estimação da taxa de produção.

4.3. Variáveis dependentes e independentes

As variáveis independentes que controlamos são o estilo de elocução, com dois níveis (entrevista semiespontânea e leitura de frases), e o sexo do falante. A variável dependente a ser medida é o tempo de estabilização da taxa de articulação.

Limitaremos a medida de tempo de estabilização à taxa de articulação. A justificativa para essa limitação é a baixa magnitude da diferença dos tempos de estabilização entre taxa de articulação e elocução verificadas nos estudos anteriores (ARANTES, 2015; ARANTES; LIMA, 2017; ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018). A unidade linguística para o cálculo da taxa de articulação será apenas a unidade VV, também em função da pouca diferença nos tempos de estabilização entre as várias unidades testadas anteriormente.

A geração dos sumários estatísticos, dos gráficos e os testes estatísticos foi feita usando o ambiente de computação estatística R (R CORE TEAM, 2020).

4.4. Análise fonética

Os 20 arquivos de áudio, dois de cada estilo para os dez falantes, foram segmentados com o auxílio do programa de análise acústica Praat (BOERSMA, 2001) em unidades VV. Seguimos os critérios recomendados pela literatura para a identificação das fronteiras fonéticas relevantes, examinando simultaneamente o oscilograma e o espectrograma de banda larga do sinal acústico para guiar a segmentação (MACHAČ; SKARNITZL, 2009; TURK; NAKAI; SUGAHARA, 2006).

Como nos restringimos à análise da taxa de articulação, a duração correspondente a pausas silenciosas não foi incluída no cálculo cumulativo da taxa, mas elas foram marcadas em uma camada separada na segmentação para permitir o cálculo do número de pausas presentes no intervalo entre o início de cada amostra e o ponto de estabilização. Seguindo Künzel (1997), adotamos 100 ms como limiar para a identificação de uma pausa silenciosa. Foram ignorados no cômputo da taxa de articulação ocorrências de hesitações e alongamentos disfluentes (“é” e “aí” alongados, por exemplo). Casos desse tipo ocorreram apenas no estilo entrevista semiespontânea.

As marcações temporais das unidades VV e das pausas foram armazenadas em arquivos de metadados (objetos *TextGrid* do programa Praat) em camadas separadas para serem processadas posteriormente por meio de *scripts* do Praat escritos para essa finalidade.

5. Resultados e discussão

5.1. Análise considerando a duração total de cada amostra

Nesta seção apresentamos os resultados da análise dos tempos de estabilização das amostras de fala do corpus, tomadas em sua duração completa. No caso das amostras de entrevista semiespontânea, essa duração corresponde aos seus 300 segundos iniciais, e no caso das amostras de fala lida, à sua duração total, com valor máximo de 300 segundos. Começamos pela apresentação dos resultados gerados pela aplicação da técnica que usa o critério estatístico aos dados.

A tabela 1 apresenta o efeito do sexo feminino e masculino dos falantes sobre as variáveis dependentes, agrupando os dois estilos de elocução. Não há efeito significativo causado pelo sexo no tempo de estabilização conforme resultado de um teste *t* de amostras independentes [$t(18) = -0{,}8$ *ns*].

Sexo do falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Feminino	80	269	-0.5	25
Masculino	103	321	0.4	34

TABELA 1 – Dados obtidos pela aplicação do critério estatístico para determinação do ponto de estabilização: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro por sexo.
Fonte: elaborada pelo autor

A tabela 2 apresenta o efeito do estilo de fala sobre as variáveis dependentes, com os dados de todos os falantes agrupados. Há efeito significativo causado pelo estilo de elocução sobre o tempo médio de estabilização conforme resultado de um teste *t* de amostras independentes [$t(18) = 1{,}4$; $p < 0{,}001$]. O estilo de elocução entrevista apresenta tempo de estabilização médio mais alto do que o estilo leitura.

Estilo	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Entrevista	111	329	0.6	37
Leitura	72	261	-0.7	21

TABELA 2 – Dados obtidos pela aplicação do critério estatístico para determinação do ponto de estabilização: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro por estilo de fala.
Fonte: elaborada pelo autor

A tabela 3 apresenta a interação entre o efeito dos sexos masculino e feminino e os estilos de fala entrevista e leitura. O teste ANOVA de dois fatores não indica presença de efeito significativo dos fatores individualmente, nem sua interação [Sexo: $F(1, 16) = 0{,}75$ *ns*; Estilo: $F(1, 16) = 2{,}1$ *ns*; Interação: $F(1, 16) = 2{,}2$ *ns*].

Estilo	Sexo do falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Entrevista	Feminino	120	371	-0.8	40
	Masculino	103	288	2.0	34
Leitura	Feminino	40	167	-0.2	9
	Masculino	104	356	-1.1	34

TABELA 3 – Dados obtidos pela aplicação do critério estatístico para determinação do ponto de estabilização: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro entre sexos e estilo de fala.
Fonte: elaborada pelo autor

A tabela 4 apresenta as médias das variáveis por falante, agregando os estilos de elocução. O teste de ANOVA de um fator não aponta evidência de efeito significativo dos falantes sobre o tempo médio de estabilização [$F(9, 10) = 0{,}87$ *ns*].

Falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
F1	64	293	-1,2	27
F2	86	257	0,1	33
F3	28	85	0,9	5
F4	134	432	-1,7	31
F5	87	276	-1,0	25
M1	147	479	-1,5	42
M2	115	322	-0,8	36
M3	138	457	1,9	58
M4	89	247	0	21
M5	27	102	2,6	12

TABELA 4 – Dados obtidos pela aplicação do critério estatístico para determinação do ponto de estabilização: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro por cada falante. Na coluna "Falante", os códigos iniciados por "F" indicam falantes do sexo feminino e os iniciados por "M", falantes masculinos.
Fonte: elaborada pelo autor

Passamos agora a apresentar os resultados da análise gerados pela aplicação do critério baseado no limiar perceptual para definir o ponto de estabilização. A tabela 5 apresenta o efeito entre os sexos feminino e masculino dos falantes sobre as variáveis, no qual não há efeito significativo do sexo sobre o tempo de estabilização determinado pelo método baseado no limiar de percepção proposto por Kendall (2013) conforme resultado de um teste *t* de amostras independentes [$t(18) = -0{,}8$ *ns*].

Sexo do falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Feminino	68	208	-3.1	19

Masculino	76	233	-0.3	24

TABELA 5 – Dados obtidos através do enunciado fonético como escopo e do critério perceptual: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro por sexo.
Fonte: elaborada pelo autor

A Tabela 6 apresenta o efeito do estilo de elocução sobre os tempos de estabilização. Não há efeito significativo da variável estilo conforme resultado de um teste *t* de amostras independentes [*t*(18) = 1,4 *ns*].

Estilo	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Entrevista	86	238	-1.2	27
Leitura	58	204	-2.2	16

TABELA 6 – Dados obtidos através do enunciado fonético como escopo e do critério perceptual: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro por estilo de fala.
Fonte: elaborada pelo autor

A Tabela 7 apresenta a interação entre o efeito do sexo dos falantes e o efeito dos estilos de fala sobre os tempos de estabilização. A ANOVA de dois fatores não indica presença de efeito significativo dos fatores individualmente, nem de sua interação [Sexo: *F*(1, 16) = 0,76 *ns*; Estilo: *F*(1, 16) = 2,1 *ns*; Interação: *F*(1, 16) = 2,2 *ns*].

Estilo	Sexo do falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
Entrevista	Feminino	104	295	-4.8	31
	Masculino	69	182	2.2	23
Leitura	Feminino	32	122	-1.4	7
	Masculino	84	285	-3.0	25

TABELA 7 – Dados obtidos através do enunciado fonético como escopo e do critério perceptual: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro entre sexos e estilo de fala.
Fonte: elaborada pelo autor

A Tabela 8 apresenta as médias por falante, agregando os estilos de fala. A ANOVA de um fator não aponta evidência de efeito significativo dos falantes sobre o tempo médio de estabilização [*F*(9, 10) = 0,86 *ns*].

Falante	Ponto de estabilização (s)	Nº de unidades VV	Erro (%)	Nº de pausas
F1	33	140	-4.8	16

F2	55	169	-4.8	16
F3	115	316	-4.9	28
F4	90	271	-0.7	22
F5	46	145	-0.2	12
M1	95	293	-0.1	27
M2	69	194	-0.0	20
M3	99	310	-0.6	41
M4	41	109	-0.4	7
M5	78	260	-0.4	23

TABELA 8 Dados obtidos através do enunciado fonético como escopo e do critério perceptual: ponto no tempo da amostra em que a taxa de articulação se estabiliza, a quantidade de segmentos e pausas presentes até esse ponto e a taxa percentual de erro para cada falante. Na coluna "Falante", os códigos iniciados por "F" indicam falantes do sexo feminino e os iniciados por "M", falantes masculinos.
Fonte: elaborada pelo autor

Para testar a presença de um possível efeito do método de determinação do ponto de estabilização sobre o tempo médio de estabilização, rodamos um teste-t de amostras pareadas, comparando os tempos médios de estabilização por falante. O resultado do teste mostra que não há diferença significativa nos tempos de estabilização causada pelos dois métodos conforme resultado de um teste t de amostras independentes [$t(19) = 1{,}3$ *ns*]. Esse resultado pode parecer surpreendente, dada a diferença nos tempos médios de estabilização associados aos dois métodos: 111 s (critério estatístico) e 86 s (critério perceptual) para o estilo de elocução entrevista e 72 s (critério estatístico) e 58 s (critério perceptual) para o estilo leitura de frases. A comparação das médias gerais parece sugerir que o critério estatístico toma mais tempo para estabilizar. No entanto, quando se observa os tempos de estabilização separadamente por falante e por método (tabela 4 para o critério estatístico e tabela 8 para o critério perceptual), observa-se que, embora para a maioria dos falantes o método perceptual tenha gerado tempo de estabilização menor, para dois falantes ocorreu o contrário (falantes F3 e M5). Essa variabilidade individual pode ser um dos fatores que colabora para o não aparecimento de uma diferença significativa no teste estatístico.

5.2. Variabilidade das estimativas do tempo de estabilização em função da duração da amostra de fala

Apresentamos agora uma análise da variabilidade das estimativas de tempo de estabilização em função da duração da amostra de fala, em especial comparando as possíveis diferenças causadas pelo uso dos dois critérios para determinação do ponto de estabilização (técnica *change point analysis* e o critério perceptual adotado por Kendall), estilos de elocução e sexo do falante. Esta análise consistiu em avaliar um possível efeito da duração total da amostra submetida à análise na determinação do tempo de estabilização da taxa de articulação.

Cada amostra do corpus analisado aqui tem duração total de aproximadamente 300 segundos, incluída aí a duração das pausas silenciosas. Para a análise reportada aqui, cada amostra de áudio original foi editada de forma a gerar uma sequência de amostras, na qual a primeira continha os 30 s do áudio original e as seguintes continham incrementos sucessivos de 30 s até alcançar a duração total da amostra original. Cada amostra da sequência foi em seguida submetida separadamente aos dois procedimentos de identificação do ponto de estabilização, aquele baseado no critério estatístico e o baseado no critério perceptual. O mesmo procedimento foi aplicado para as amostras do estilo de entrevista semiespontânea e leitura de frases. As figuras 4 e 5 mostram os resultados da aplicação do procedimento descrito. Nas figuras, o eixo horizontal indica a duração da amostra submetida à detecção do ponto de estabilização e o eixo vertical mostra o tempo de estabilização encontrado em cada caso. A figura 4 mostra o resultado para a técnica estatística e a figura 5 o resultado para o método perceptual. Nas duas figuras, as cores das linhas indicam os resultados para os dois estilos de elocução.

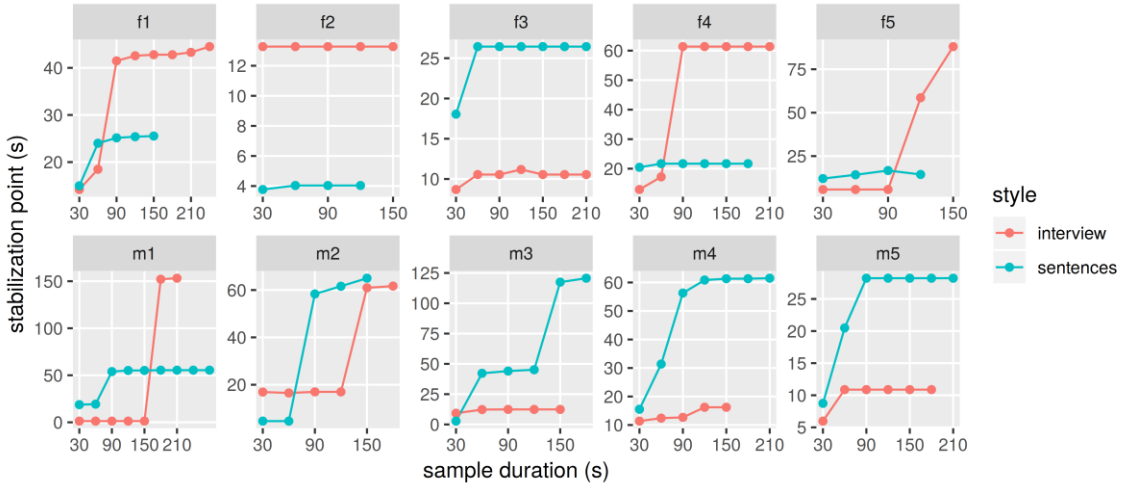

FIGURA 4 — Tempos de estabilização determinados pela técnica estatística change point analysis em função do aumento incremental da duração da amostra de fala submetida à análise. Os resultados são apresentados para cada falante e a cor das linhas representa os dois estilos de elocução.
Fonte: elaborada pelo autor

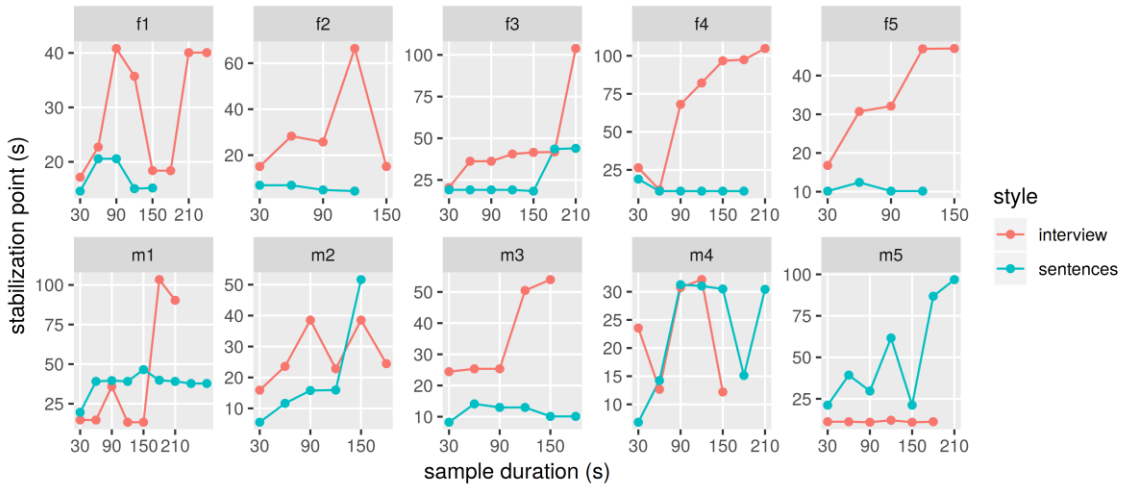

FIGURA 5 – Tempos de estabilização determinados pelo critério do limiar perceptual em função do aumento incremental da duração da amostra de fala submetida à análise. Os resultados são apresentados para cada falante e a cor das linhas representa os dois estilos de elocução.
Fonte: elaborada pelo autor

O exame visual das figuras 4 (técnica *change point analysis*) e 5 (critério do limiar perceptual) indica que a duração da amostra exerce efeito sobre a localização do ponto de estabilização, isto é, o ponto de estabilização em uma amostra pode mudar conforme a sua duração aumente. Esse efeito pode ser observado independentemente do critério usado para a determinação do ponto de estabilização. Podemos identificar dois comportamentos diferentes que são recorrentes nos resultados: (1) alguns falantes apresentam pouca ou nenhuma variabilidade nos tempos de estabilização em função da duração da amostra – por exemplo, o falante *f2* na figura 4 nos dois estilos e os falantes *f2*, *f4* e *f5* na figura 5 no estilo leitura de frases; (2) para um grupo de falantes, a série de dados começa num patamar mais baixo, depois há um ponto em que há uma subida e a série se estabiliza nesse patamar mais alto; exemplos desse padrão são os falantes *f1*, *m2*, *m4* e *m5* na figura 4 em ambos os estilos; na figura 5, os exemplos são o falante *m1* no estilo leitura de frases e, em certa medida, o falante *f3* no mesmo estilo.

Podemos identificar casos isolados, como o falante *m3* na leitura de frases, na figura 4, em que parece haver dois patamares, um mais baixo entre as durações de 60, 90 e 120 segundos e outro mais alto nas durações 150 e 180 segundos. Outro exemplo desse padrão é o falante *f5* no estilo entrevista, que permanece num patamar baixo nas durações de 30 a 90 segundos e depois entra em uma trajetória ascendente nas duas durações seguintes.

Um padrão que chama a atenção na observação da figura 4 é a diferença entre os tempos de estabilização dos dois estilos de elocução, principalmente quando são comparados os valores nos patamares estáveis. São exemplos disso os falantes *f1*, *f2*, *f3*, *f4*, *m4* e *m5*, embora os estilos se alternem, a depender do falante, em apresentar tempos de estabilização mais longos ou curtos.

Podemos identificar algumas diferenças relevantes entre os comportamentos dos tempos de estabilização mostrados nas figuras 4 e 5, indicando um potencial efeito do critério para determinação do ponto de estabilização. De modo geral, o uso do critério estatístico (figura 4) parece gerar um

número maior de casos de tempos de estabilização que mudam menos conforme a duração da amostra aumenta. Em contraste, a aplicação do critério perceptual (figura 5) gera padrões que simplesmente não ocorrem na figura 4, como os casos em que os tempos de estabilização oscilam para cima e para baixo mais de uma vez à medida que a duração da amostra de fala aumenta. Os dados dos falantes *f1*, *f2*, *m1*, *m4* e *m5* são exemplos desse comportamento oscilante. Esse padrão é inesperado, uma vez que a observação visual de muitas séries temporais da taxa de elocução cumulativa mostra que o mais comum é que o aumento na duração da amostra analisada esteja associado a uma diminuição marcante na variabilidade dos valores da taxa cumulativa, padrão que pode ser observado na figura 2.

Para verificar se o critério estatístico gera mesmo tempos de estabilização menos variáveis do que o critério perceptual, usamos o desvio mediano absoluto (MAD, *median absolute deviation*) como estimador estatístico da variabilidade das séries dos tempos de estabilização em função da duração das amostras. O MAD foi calculado em função dos dois critérios para determinação do ponto de estabilização (técnica *change point analysis* e o critério perceptual), dos estilos de elocução, do sexo dos falantes e pela combinação dessas variáveis. Apresentaremos os resultados a seguir.

Os valores de MAD são 13,2 s para a técnica *change point analysis* e 15 s para o critério perceptual. Não é uma diferença expressiva, mas vai na direção da observação visual, sugerindo que os resultados produzidos pelo critério perceptual são ligeiramente mais variáveis.

Quando acrescentamos a variável estilo de elocução, os resultados mostram, na tabela 9, uma interação complexa. A variabilidade nos tempos de estabilização no estilo entrevista é 3 vezes menor do que a do estilo leitura quando se usa o o critério estatístico. Quando se usa o critério perceptual, o estilo menos variável é a leitura de frases, mas a diferença entre os estilos é muito menos expressiva.

Estilo	**Critério estatístico**	**Critério perceptual**
Entrevista	5,5 s	18 s
Leitura	17,9 s	12,5 s

TABELA 9 – Variabilidade (desvio mediano absoluto) das estimativas de tempo de estabilização em amostras com duração progressivamente maior, apresentadas em função das técnicas de detecção da estabilização e dos estilos de elocução.
Fonte: elaborada pelo autor

Os resultados do efeito do sexo dos falantes sobre a variabilidade das estimativas do tempo de estabilização, apresentados na Tabela 10, indicam menor variabilidade para falantes do sexo feminino, independentemente do critério para determinação do ponto de estabilização. O uso do critério estatístico produz resultados menos variáveis para as falantes do sexo feminino; para o sexo masculino, o uso do critério perceptual gera a menor variabilidade.

Sexo	Critério estatístico	Critério perceptual
Feminino	10,3 s	13,5 s
Masculino	21,8 s	17,1 s

TABELA 10 – Variabilidade (desvio mediano absoluto) das estimativas de tempo de estabilização em amostras com duração progressivamente maior, apresentadas em função das técnicas de detecção da estabilização e do sexo dos falantes.
Fonte: elaborada pelo autor

Cruzando as três variáveis, os resultados, mostrados na Tabela 11, indicam uma interação complexa entre as variáveis. O estilo entrevista é aquele no qual se observa a maior diferença na variabilidade das estimativas produzidas pelas duas técnicas de detecção da estabilidade, independentemente do sexo dos falantes. A técnica estatística é a que produz menor variabilidade. No caso do estilo leitura de frases, as duas técnicas produzem resultados comparáveis em termos de variabilidade, mas há uma diferença relativamente grande causada pelo sexo do falante: as estimativas extraídas das falantes do sexo feminino são menos variáveis, com a exceção do estilo entrevista para os falantes masculinos.

Sexo	Estilo	Critério estatístico	Critério perceptual
Feminino	Entrevista	9,63 s	18 s
	Leitura	7,03 s	7,55 s
Masculino	Entrevista	5,65 s	16,1 s
	Leitura	20,4 s	21,1 s

TABELA 11 – Variabilidade (desvio mediano absoluto) das estimativas de tempo de estabilização em amostras com duração progressivamente maior, apresentadas em função das técnicas de detecção da estabilização, do sexo dos falantes e do estilo de elocução.
Fonte: elaborada pelo autor

5.3. Variabilidade do erro da estimativa da taxa de articulação no ponto de estabilização em função da duração da amostra de fala

Analisamos aqui o comportamento do erro de estimativa, isto é, a diferença entre a taxa de articulação medida no ponto de estabilização e a taxa de articulação estimada usando todos os dados da amostra de fala (taxa global) à medida que a duração da amostra aumenta e avaliar se as variáveis controladas no estudo exercem algum efeito sobre o comportamento do erro. A figura 6 mostra a série dos erros de estimativa de erro gerados pela aplicação do critério estatístico e a figura 7, a série gerada pela aplicação do critério do limiar perceptual.

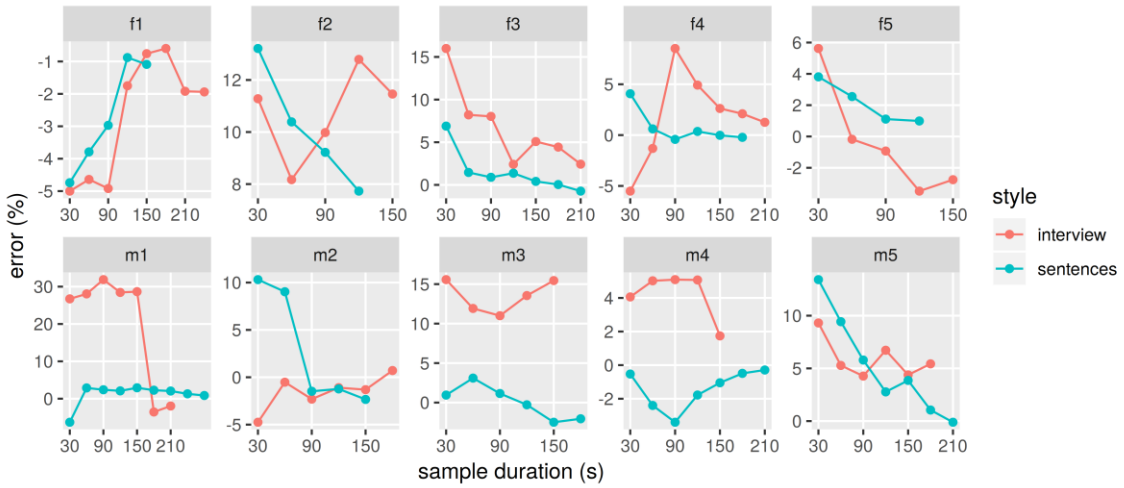

FIGURA 6 – Erro de estimativa gerado pela aplicação do critério estatístico em função do aumento da duração da amostra de fala e do estilo de elocução.
Fonte: elaborada pelo autor

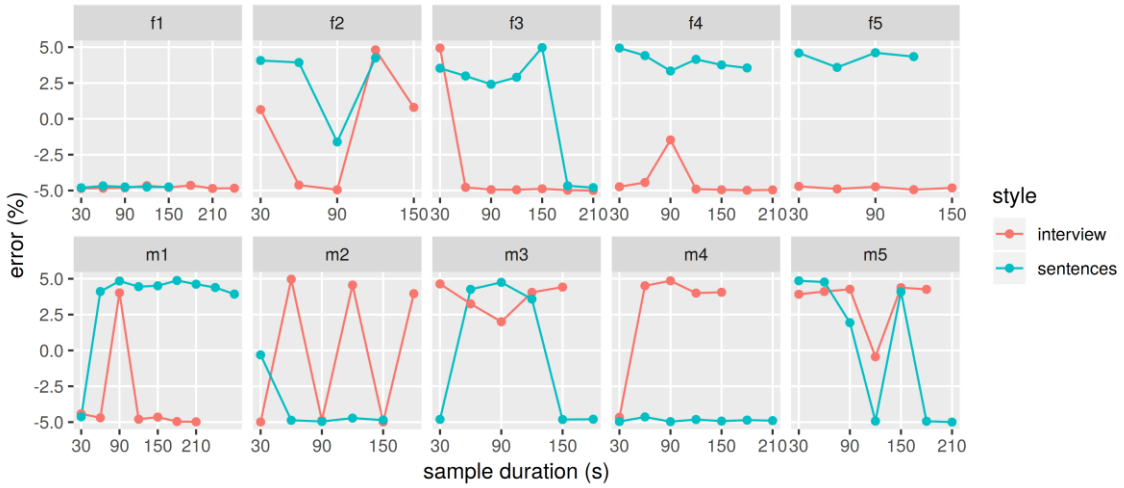

FIGURA 7 – Erro de estimativa gerado pela aplicação do critério do limiar perceptual em função do aumento da duração da amostra de fala e do estilo de elocução.
Fonte: elaborada pelo autor

As séries geradas pela aplicação do critério do limiar perceptual (figura 7) apresentam um nível de variabilidade menor (estimado por meio do desvio absoluto mediano) do que as geradas pelo critério estatístico (figura 6): 0,34 contra 3,02. A mediana do valor absoluto do erro, no entanto, indica que a magnitude do erro é menor no caso da técnica estatística em comparação com a técnica do limiar perceptual: 2,88% contra 4,71%. A observação visual comparativa das duas figuras corrobora os números – as séries mostradas na Figura 7 são realmente mais estáveis, mas concentram-se nos valores entre -5 e 5%, que

são os limites definidos pelo critério perceptual; em contraste, os dados da figura 6 são realmente mais variáveis, mas os movimentos nas séries temporais correspondem, em muitos casos, a aproximações a patamares de erro próximos a 0% (os falantes *f*1, *f*3 e *f*4 exemplificam esse padrão).

A conclusão sugerida pelos dados é que a técnica estatística para detecção do ponto de estabilização tende a gerar erros menores conforme aumenta a duração da amostra. No caso da técnica baseada no critério perceptual, o aumento na duração da amostra tende a influenciar menos a magnitude do erro, que tende a ser mais estável, mas em nível absoluto maior do que os gerados pela técnica estatística.

5.4. Efeito do falante sobre a variação nos tempos de estabilização

Analisamos nesta seção o efeito do falante sobre os tempos de estabilização. A questão é saber se essas estimativas são muito dependentes do falante que produziu a amostra ou os valores podem ser considerados relativamente independentes do falante. Para fazer essa análise, as amostras dos dois estilos de elocução de cada falante foram subdivididas em cinco partes com durações aproximadamente iguais, como ilustra a figura 8. Cada subamostra foi analisada pelas duas técnicas de detecção de estabilização, de modo que para cada falante foi possível obter 20 estimativas de ponto de estabilização, cinco para cada estilo e uma para cada técnica de detecção.

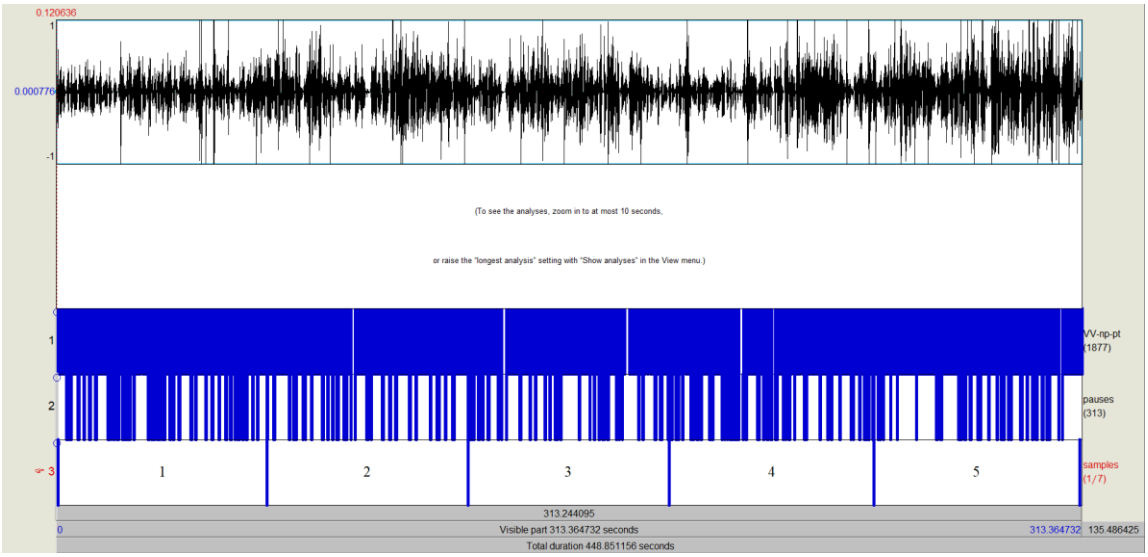

FIGURA 8 – Oscilograma acompanhado do TextGrid da amostra de um falante do corpus, no estilo entrevista semiespontânea. A camada 3 mostra a divisão da amostra em cinco subamostras com durações semelhantes, identificadas pelos números de 1 a 5.
Fonte: elaborada pelo autor

A figura 9 mostra a variabilidade individual das estimativas dos pontos de estabilização por técnica de detecção de estabilidade, agregando os dois estilos de elocução. O exame dos gráficos de caixa sugere que as distribuições dos pontos de estabilização dos diferentes falantes são parecidas. Um teste

de homogeneidade de Fligner-Killeen para comparação de variâncias indica que as amostras são heterocedásticas tanto no caso do critério estatístico [$\chi^2(16) = 1105$; $p < 0{,}001$] quanto do critério do limiar perceptual [$\chi^2(16) = 1105$; $p < 0{,}001$]. Dada a heterocedasticidade, o teste não paramétrico de Kruskal-Wallis foi usado para testar a hipótese nula de ausência de diferença entre as médias do tempo de estabilização dos 10 falantes. Não se verificou efeito significativo quer se considere o critério *change point* [$\chi^2(9) = 11{,}8$; ns], quer se considere o critério perceptual [$\chi^2(9) = 12{,}9$; ns]. Esse resultado sugere que os tempos de estabilização não são uma propriedade idiossincrática de cada falante e provavelmente refletem propriedades mais gerais da estrutura de organização temporal da língua.

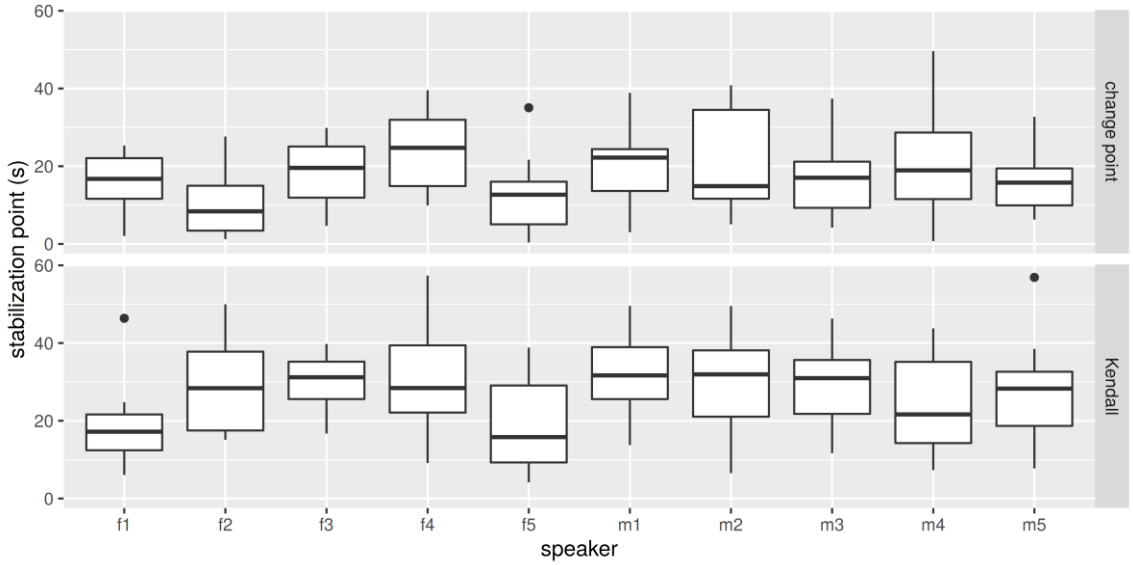

FIGURA 9 – Gráficos de caixa mostrando a distribuição dos valores dos pontos de estabilização dos falantes da amostra em função técnica de detecção. Os dois estilos de fala foram agregados.
Fonte: elaborada pelo autor

Conclusões

A literatura prévia a respeito da duração mínima que uma amostra de fala deve ter para a estimação estável da taxa de produção da fala é restrita, dado que há poucos trabalhos e eles usam metodologias ligeiramente diferentes, que chegam a resultados divergentes. Uma das referências no tema é Kendall (2013), que sugere durações mínimas entre 3 e 9 minutos para a obtenção de estimativas estáveis ou, expressando em termos de quantidade de material linguístico, entre 80 e 200 enunciados fonéticos. Outro ramo nessa literatura são os trabalhos de Arantes e colaboradores (ARANTES, 2015; ARANTES; LIMA, 2017), que de forma independente em relação a Kendall (2013), conduziram pesquisas iniciais apontando tempos de estabilização muito mais curtos, em torno de 30 segundos. Em função de diferenças relevantes em termos de metodologia em relação a Kendall (2013), trabalhos posteriores do grupo liderado por Arantes incorporou aspectos da metodologia usada por Kendall (2013) em sua pesquisa (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) para investigar a influência dessas escolhas metodológicas nos

resultados. As principais diferenças entre as linhas de pesquisa são os critérios diferentes adotados pelos autores para a identificação da estabilidade na série temporal formada pelos valores da taxa de produção da fala calculada de forma cumulativa e também a granularidade do cálculo da taxa cumulativa. Kendall (2013) adota um critério baseado em um limiar de percepção de mudança na taxa de produção derivado empiricamente e Arantes e colegas usam uma técnica estatística que identifica mudança na variabilidade subjacente à série temporal. Em termos da granularidade de medida da taxa, Kendall (2013) calcula a taxa em unidades de escopo mais largo, os enunciados fonéticos, enquanto Arantes e colegas inicialmente adotaram a unidade VV, uma unidade de escopo mais restrito (ARANTES, 2015; ARANTES; LIMA, 2017). As duas linhas de pesquisa têm em comum o fato de investigarem a fala lida. Conforme dissemos na seção 2, os trabalhos posteriores de Arantes e colegas (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) sugerem que a adoção do critério baseado no limiar perceptual para a identificação dos pontos de estabilização e do uso do enunciado fonético como escopo para o cálculo da taxa de produção não resultam em mudanças drásticas nos resultados obtidos principalmente em Arantes (2015) e Arantes e Lima (2017), que ainda assim giram em torno de 30 segundos.

A falta de efeito significativo observada nos tempos de estabilização pelo grupo de Arantes e colegas mesmo com a adoção das escolhas metodológicas de Kendall (2013) sugere que talvez a diferença entre os resultados se deva ao fato de que, em Kendall (2013), as séries temporais de taxa de produção cumulativa tinham uma extensão mínima de 20 enunciados fonéticos, com taxa de incremento de 10 enunciados. Segundo se pode inferir a partir dos dados relatados em seu trabalho, isso significa que a duração mínima de suas séries temporais era de aproximadamente um minuto. A assunção de que a taxa de produção não pode se estabilizar em um intervalo mais curto do que um minuto não parece sustentável tendo em vista os resultados do grupo de Arantes. Isso se justifica porque Arantes e colegas (ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) chegam a tempos de estabilização, em média, próximos a 30 segundos usando os mesmos procedimentos de Kendall (2013), isto é, o critério baseado no limiar de percepção e cálculo da taxa por enunciado fonético, apenas relaxando a exigência mínima de 20 enunciados fonéticos para a amostra inicial, usando séries que começavam com um enunciado e adotando incrementos também de um enunciado. Seus resultados mostram que é possível obter taxas de articulação e elocução com um número muito menor de enunciados fonéticos, entre 3 e 4. Os trabalhos de Arantes e colegas (ARANTES, 2015; ARANTES; LIMA, 2017; ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018) também mostraram que a adoção de diferentes unidades linguísticas para o cálculo da taxa de produção não gera tempos de estabilização sensivelmente diferentes. O mesmo vale para o tipo de taxa investigado, de forma que tanto a taxa de articulação quanto de elocução estabilizam em tempos bastante similares.

Dados os resultados conseguidos nas pesquisas anteriores, o presente trabalho teve como objetivo principal investigar se o estilo de elocução tem um papel importante na determinação dos tempos de estabilização da taxa de produção. Nos propusemos, ainda, a analisar o possível efeito da duração total da amostra de fala sobre a estimativa dos tempos de estabilização e a variação individual dessas estimativas. Os resultados trazidos pelo presente trabalho mostram, em primeiro lugar, que os tempos médios de estabilização derivados dos dados do corpus analisado aqui são mais longos do que os

obtidos no corpus de leitura analisado pelos trabalhos de Arantes e colegas. A média geral para o estilo leitura de frases no corpus do presente trabalho é de 72 s quando se usa o critério estatístico para a determinação do ponto de estabilização e 58 s quando se usa o critério do limiar de percepção. Nos trabalhos anteriores de Arantes e colegas, a média geral para a taxa de articulação é 9,4 s e 10,6 s respectivamente para os dois critérios. Dois fatores podem ajudar a explicar essa diferença entre os resultados. O primeiro é a composição dos corpora. Nos trabalhos anteriores de Arantes e colegas, todos os participantes (oito no total) leram o mesmo texto. No corpus analisado no presente trabalho (dez no total), cada falante leu um material com composição segmental diferente. A invariância na composição segmental do material pode ter colaborado para a obtenção de tempos de estabilização mais curtos e menos variáveis nos trabalhos de Arantes e colegas em comparação com os obtidos agora. O segundo fator é a duração total da amostra. No caso do corpus do presente trabalho, as amostras todas têm uma duração próxima a 300 s, enquanto no corpus analisado nos trabalhos de Arantes e colegas as amostras têm duração média de 30 s. A seção 5.2 mostrou que o tempo de estabilização pode mudar conforme a amostra analisada tem sua duração total aumentada. É possível que se tivéssemos limitado a duração das amostras do corpus analisado aqui a pouco mais de 30 s, os tempos de estabilização obtidos fossem mais curtos e mais próximos aos dos trabalhos anteriores de Arantes e colegas. Esse resultado mostrou a importância de investigar amostras com duração mais elevada para observar o comportamento da taxa de produção cumulativa em trechos de fala mais longos e o desempenho dos dois critérios de detecção de estabilidade nesse cenário.

Um segundo resultado relevante trazido pelo presente trabalho é a observação de um efeito do estilo de elocução sobre o tempo de estabilização. Observamos uma diferença significativa no tempo de estabilização entre o estilo fala semiespontânea (111 s) e a leitura de frases (72 s) quando o critério estatístico para a detecção da estabilidade foi usado. A diferença observada entre os dois estilos – 86 s para a fala semiespontânea e 58 s para a leitura de frases – não se mostrou estatisticamente significativa quando o critério baseado no limiar perceptual foi usado. De todo modo, o estilo entrevista semiespontânea gerou médias mais altas, resultado que confirma achados prévios na literatura que sugerem a existência de diferenças na organização temporal de estilos mais espontâneos em comparação à leitura. Não se observou diferenças importantes devidas ao sexo dos falantes, quer os estilos fossem agrupados, quer considerados separadamente. Esse resultado é útil porque dispensa o uso de limiares de duração diferentes para falantes dos dois sexos.

Um terceiro conjunto de resultados que cabe destacar é a evidência segundo a qual a variabilidade intrafalante nos tempos de estabilização não é marcadamente diferente da variação apurada entre as estimativas do conjunto dos falantes do corpus. Esse resultado é interessante porque indica que os tempos de estabilização refletem menos propriedades individuais dos falantes e mais, possivelmente, aspectos das características duracionais dos fones da língua e de suas propriedades distribucionais.

Os resultados apresentados aqui complementam de maneira relevante aqueles já reportados na literatura anterior. Os tempos de estabilização obtidos no presente experimento ocupam uma região intermediária entre a gama de valores bastante elevados (entre 3 e 9 minutos) sugeridos por Kendall (2013) e os valores consideravelmente mais curtos (até 30 segundos) reportados por Arantes e colegas

(ARANTES, 2015; ARANTES; LIMA, 2017; ARANTES; ERIKSSON; LIMA, 2018; LIMA; ARANTES, 2018). A comparação de metodologias diferentes, a inclusão do estilo de elocução e da variação da duração total da amostra como variáveis, bem como a comparação da variabilidade intra e interfalante nas estimativas, permite dizer que os tempos de estabilização intermediários obtidos aqui (médias entre 30 e 130 segundos) são estimativas seguras e robustas. Pensando em sua aplicabilidade, os resultados relatados têm consequência direta para prática da fonética forense. Em cenários como o exame de comparação de locutores, no qual as amostras de fala a serem comparadas tendem a ter duração reduzida, é preciso dispor de um limiar de duração para a amostra que assegure que a estimativa da taxa de produção da fala derivada dessa amostra seja estável e representativa. Nesse sentido, as condições testadas no presente estudo refletem de forma mais realista as condições encontradas no cenário relevante para a fonética forense.

Informações complementares

Avaliação e resposta do autor

Avaliação: https://doi.org/10.25189/rabralin.v23i2.2234.R
Resposta do autor: https://doi.org/10.25189/rabralin.v23i2.2234.A

Editoras

Luma Miranda
Afiliação: Universidade Eötvös Loránd
ORCID: https://orcid.org/0000-0002-5529-0338

Manuella Carnaval
Afiliação: Universidade Federal do Rio de Janeiro
ORCID: https://orcid.org/0000-0002-4321-5859

Carolina Gomes da Silva
Afiliação: Universidade Federal da Paraíba
ORCID: https://orcid.org/0000-0002-1490-0814

RODADAS DE AVALIAÇÃO

Avaliador 1: Saulo Mendes Santos
Afiliação: Universidade Federal de Minas Gerais

Avaliador 2: Ubiratã Kickhofel Alves
Afiliação: Universidade Federal do Rio Grande do Sul
ORCID: https://orcid.org/0000-0001-6694-8476

AVALIADOR 1

O trabalho expande o conhecimento sobre o assunto, já estudado em trabalhos anteriores do mesmo autor, considerando os efeitos de novas variáveis sobre a estimação do tempo de estabilização da taxa de produção da fala (TPF). Além disso, o assunto explorado tem implicações bastante práticas para análises na área de linguística forense e merecem ser publicados.

De modo geral, o texto está bem escrito, é bem estruturado e é suficientemente claro em termos de metodologia, dados, análises e conclusões. Existe, no entanto, uma aparente contradição entre a existência de efeitos do tipo de elocução sobre os tempos de estabilização da TPF - não significativa na seção de análise e significativa nas conclusões. Recomenda-se, portanto, ao autor verificar a possível divergência ou esclarecer os diferentes escopos das conclusões de forma a deixar o texto ainda mais claro.

No arquivo em anexo, o autor poderá encontrar pequenas correções textuais (em controle), além de alguns apontamentos em comentários marginais.

AVALIADOR 2

O artigo reporta os resultados de um estudo que visou a investigar se o estilo de elocução e o sexo dos locutores exerceram efeitos sobre o tempo de estabilização da "taxa de produção" oral (termo esse cunhado e devidamente caracterizado no artigo).

O artigo se destaca pela sua clareza didática, ao reportar um estudo bastante complexo de forma compreensível e bem elaborada do ponto de vista organizacional. A fundamentação teórica que embasa as escolhas metodológicas é bastante clara. Os procedimentos analíticos são apresentados com grande rigor. Considero que o artigo se mostra de grande importância não somente para a área de Fonética Forense (como ressaltado ao longo do próprio texto), mas também para as áreas de Fonética Experimental e Fonologia de Laboratório como um todo, uma vez que apresenta caráter inovador para tratar de uma questão que implica decisões metodológicas importantes para os estudiosos dessas áreas. Destaco, também, o caráter de inovação do artigo, sobretudo no que diz respeito à análise de Change Points.

Visando a contribuir com a versão final do artigo, trago, neste parecer, discussões referentes a dois aspectos estruturais do texto: (i) os objetivos e (ii) o mote atribuído à Fonética Forense.

No que diz respeito aos objetivos do artigo, esses são expressos na seção 3. Na referida seção, lê-se que "o objetivo do trabalho reportado aqui é investigar o efeito da variação no estilo de elocução sobre o tempo mínimo necessário para a estimativa de taxa de produção da fala". Entretanto, tal objetivo contempla apenas uma das duas variáveis independentes cujos efeitos são testados no estudo.

Além disso, a redação do objetivo não condiz com o próprio título do trabalho. Uma vez que tal título menciona o "estilo de elocução" e o papel do "falante", imaginaria que a discussão dos efeitos referentes à variável "sexo" também fosse incluída no objetivo geral.

Ainda no que diz respeito à seção dos objetivos, é preciso considerar que o estudo também visou a analisar o possível efeito da duração total da amostra de fala sobre a estimativa dos tempos de estabilização e a variação individual dessas estimativas. Tais análises complementares, ainda que descritas no resumo e na conclusão, precisam ser apresentadas de forma explícita na seção dos objetivos.

No que diz respeito ao segundo aspecto, referente à pertinência do estudo para a área de Fonética Forense, considero que há uma certa “oscilação” no teor dado a tal questão ao longo do artigo. Por um lado, a seção 1.1 apresenta toda uma contextualização da área de Fonética Forense, de modo a fundamentar (de forma bastante adequada, a meu ver) a pertinência do estudo para a referida área. Por sua vez, as seções de análise e (principalmente) de conclusão carecem de uma retomada mais explícita dessa contextualização realizada na seção 1.1. É preciso que a seção de conclusão faça menção explícita às implicações dos resultados para a área. Além disso, verifiquei que, apesar do forte mote referente à Fonética Forense instaurado na seção 1.1, tal campo não é contemplado no resumo/abstract. Penso que, havendo espaço suficiente, o resumo e o abstract podem investir, também, no referido mote, uma vez que explicita a pertinência da realização do estudo. Nesse sentido, penso que sobretudo o resumo para o leitor leigo deve deixar clara a pertinência da realização do presente estudo para a referida área do conhecimento.

No arquivo Word em anexo, apresento uma série de comentários pontuais, sobretudo de caráter formal. Dada a pertinência do estudo e o rigor científico com que esse foi conduzido, sou plenamente favorável à sua aprovação. Parabenizo o(s) autor(es) pela contribuição prestada através do artigo.

RESPOSTA DO AUTOR

São Carlos, 29 de janeiro de 2024

Conforme instruções recebidas da equipe editorial em comunicação no dia 9 de janeiro, segue a presente carta para apresentar a versão revisada do manuscrito “Efeito do estilo de elocução e do falante sobre o tamanho mínimo de amostra para estimativa da taxa de produção da fala”, submetido por mim a esta revista.

Inicio agradecendo os dois pareceristas pela leitura cuidadosa e rigorosa da versão do texto submetida inicialmente. As considerações dos dois certamente ajudaram a melhorar a qualidade do texto e melhorar sua coesão.

Descrevo a seguir as modificações feitas ao texto inicial para atender as sugestões e considerações dos dois pareceristas. Mudanças mais importantes introduzidas no texto em função dos comentários dos pareceristas estão destacadas no manuscrito revisado em fundo amarelo para identificação rápida por parte dos pareceristas e editores.

Foram introduzidas mudanças no resumo em português, inglês e no resumo para não especialistas, conforme sugestões, de modo a deixar mais clara a implicação dos resultados para aplicações em fonética forense. Destaco que essas mudanças precisaram ser relativamente sutis para não aumentar demasiadamente o número de palavras nessas seções.

Em relação às considerações do parecerista Ubiratã Kickhofel Alves, destaco as seguintes mudanças:

- Revisei a seção 3 para incluir de forma explícita como objetivos do trabalho a análise do efeito do falante (sexo e variabilidade intrafalante) sobre os tempos de estabilização e efeito da duração total da amostra de fala sobre a estimativa dos tempos de estabilização e a variação individual dessas estimativas.
- Acrescentei à seção "Introdução" um exemplo de taxa de produção local, cobrado por ele.
- Acrescentei um parágrafo final de fechamento na seção 4.1 para destacar as diferentes implicações de se adotar o critério de limiar perceptual ou a análise de change point.
- Acrescentei um parágrafo final na seção 6 (Conclusões) discutindo de forma mais explícita o que considero serem as implicações dos resultados do trabalho especificamente para o campo da fonética forense. Em relação às considerações do parecerista Saulo Mendes Santos, destaco as seguintes mudanças:
- Acrescentei a informação em diferentes pontos ao longo da seção 5 (Resultados e discussão) a respeito dos testes estatísticos usados.
- Fiz modificações no parágrafo final da seção 5.1 para comentar a respeito de possíveis explicações para a falta de efeito estatístico significativo na comparação entre os tempos de estabilização dos dois métodos para identificação do ponto de estabilização.
- Em relação ao comentário feito pelo parecerista no início quarto parágrafo da seção 6 (Conclusões), a descrição da diferença estatisticamente significativa entre os tempos de estabilização em função do estilo de elocução quando o critério estatístico para a detecção da estabilidade é usado aparece na seção 5.1, no parágrafo que antecede a tabela 2 ([$t(18) = 1,4$; $p < 0,001$]). Esse é o resultado relatado no início do quarto parágrafo das conclusões. Portanto, não há divergência entre o que aparece nas duas seções. As incorreções gramaticais e as sugestões pontuais reformulação textuais introduzidas pelos pareceristas diretamente no corpo do texto foram todas aceitas. Todos os erros de digitação apontados foram corrigidos. Creio ter conseguido endereçar todas as considerações dos pareceristas na versão revisada do manuscrito que submeto agora. Coloco-me à disposição para atender a algum ponto que os colegas julguem não ter sido adequadamente tratado.

Cordialmente,
Pablo Arantes

Conflito de interesses

O autor declara não haver conflito de interesse na execução desta pesquisa.

Agradecimentos

O presente trabalho baseia-se parcialmente nos resultados do projeto de pesquisa "Efeito do estilo de elocução e do falante sobre o tamanho mínimo de amostra para estimativa da taxa de produção da fala" financiado pela FAPESP (processo 2019/01661-7) na modalidade Iniciação Científica, concedida a Verônica Gomes Lima e orientado pelo autor. O autor agradece a Anders Eriksson, da Universidade de Estocolmo, por ter cedido as gravações do português brasileiro analisadas no presente trabalho, que compõem o acervo do projeto "A typology for word stress and speech rhythm based on acoustic and perceptual considerations" coordenado por ele. Gostaria de expressar minha gratidão aos dois pareceristas e às editoras do dossiê pela leitura cuidadosa da versão inicial do manuscrito e pelas sugestões de mudanças que ajudaram a melhorar a legibilidade e qualidade da versão final do manuscrito.

REFERÊNCIAS

ARANTES, P. Estimativas de longo termo da frequência fundamental: implicações para a fonética forense. **Revista Virtual de Estudos da Linguagem – ReVEL**, v. 12, n. 23, p. 217–236, 2014.

ARANTES, P. Speech rate estimation: how long should the utterance be? **Anais do Colóquio Brasileiro de Prosódia da Fala**, v. 3, p. 1–4, 2015.

ARANTES, P.; ERIKSSON, A. **Temporal stability of long-term measures of fundamental frequency**. (N. Campbell, D. Gibbon, D. Hirst, Eds.) *In*: Proceedings of the 7th International Conference on Speech Prosody. **Anais** [...]. Dublin: ISCA, 2014.

ARANTES, P.; ERIKSSON, A.; GUTZEIT, S. **Effect of language, speaking style and speaker on long-term F0 estimation**. *In*: Interspeech 2017. **Anais** [...].Stockholm: ISCA, 2017. p. 3897–3901.
ARANTES, P.; LIMA, V. G. Towards a methodology to estimate minimum sample length for speaking rate. **Revista do GEL**, v. 14, n. 2, p. 183–197, 2017.

BARBOSA, P. A. **Incursões em torno do ritmo da fala**. Campinas: Pontes, 2006.

BARBOSA, P. A. **Prosódia**. 1. ed. São Paulo: Parábola, 2019.

BOERSMA, P. Praat, a system for doing phonetics by computer. **Glot International**, v. 5, n. 9/10, p. 341–345, 2001.

BÓNA, J. Temporal characteristics of speech: The effect of age and speech style. **The Journal of the Acoustical Society of America**, v. 136, n. 2, p. EL116–EL121, 1 ago. 2014.

CAO, H.; LEI, Y. Fundamental frequency statistics for young male speakers of Mandarin. **Journal of Forensic Science and Medicine**, v. 3, n. 4, p. 217–222, 2017.

CAO, H.; WANG, Y. **A forensic aspect of articulation rate variation in Chinese**. *In*: Proceedings of the XVIIth ICPhS. **Anais** [...]. Hong Kong, 2011. p. 396–399.

CRYSTAL, T. H.; HOUSE, A. S. Segmental durations in connected speech signals: Preliminary results. **The Journal of the Acoustical Society of America**, v. 72, n. 3, p. 705–716, 1 set. 1982.

EDWARDS, J.; BECKMAN, M. E.; FLETCHER, J. The articulatory kinematics of final lengthening. **The Journal of the Acoustical Society of America**, v. 89, n. 1, p. 369–382, 1 jan. 1991.

ERIKSSON, A. Aural/acoustic vs. automatic methods in forensic phonetic case work. *In*: NEUSTEIN, A.; PATIL, H. A. (Eds.). **Forensic Speaker Recognition: Law Enforcement and Counter-terrorism**. [s.l.] Springer, 2011. p. 41–70.

GFROERER, S. **Auditory-instrumental forensic speaker recognition**. *In:* Proceedings of Eurospeech 2003. **Anais** [...]. Geneva: ISCA, 2003, p. 705–708.

GOLD, E.; FRENCH, P. International practices in forensic speaker comparison. **The International Journal of Speech, Language and the Law**, v. 18, n. 2, p. 293–307, 2011.

GOLD, E.; FRENCH, P. International practices in forensic speaker comparisons: second survey. **International Journal of Speech Language and the Law**, v. 26, n. 1, p. 1–20, 2019.

HIROSE, K.; KAWANAMI, H. Temporal rate change of dialogue speech in prosodic units as compared to read speech. **Speech Communication**, v. 36, n. 1–2, p. 97–111, jan. 2002.

HOWELL, P.; KADI-HANIFI, K. Comparison of prosodic properties between read and spontaneous speech material. **Speech Communication**, v. 10, n. 2, p. 163–169, jun. 1991.

HUDSON, T. et al. **F0 statistics for 100 young male speakers of Standard Southern British English**. *In:* ICPhS XVI. **Anais** [...]. Saarbrücken: ISCA, 2007. p. 1809–1812.

JAFFE, J.; BRESKIN, S. Temporal Patterns of Speech and Sample Size. **Journal of Speech and Hearing Research**, v. 13, n. 3, p. 667–668, set. 1970.

JESSEN, M. Forensic reference data on articulation rate in German. **Science and Justice**, v. 47, p. 50–67, 2007.

JESSEN, M. Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. *In*: GREWENDORF, G.; RATHERT, M. (Eds.). **Formal linguistics and law**. Berlin: Mouton de Gruyter, 2009. p. 115–139.

KENDALL, T. **Speech Rate, Pause, and Sociolinguistic Variation: Studies in Corpus Sociophonetics**. London: Palgrave Macmillan, 2013.

KILLICK, R.; ECKLEY, I. A. changepoint: An R Package for Changepoint Analysis. **Journal of Statistical Software**, v. 58, n. 3, p. 1–19, 2014.

KÜNZEL, H. Some general phonetic and forensic aspects of speaking tempo. **Forensic Linguistics**, v. 4, n. 1, p. 48–83, 1997.

LEHISTE, I. **Suprasegmentals**. Cambridge, MA: MIT Press, 1970.

LINDH, J. Preliminary descriptive F0-statistics for young male speakers. **Lund Working Papers**, v. 52, p. 89–92, 2006.

MACHAČ, P.; SKARNITZL, R. **Principles of Phonetic Segmentation**. Prague: Epocha Publishing House, 2009.

MORRISON, G. Forensic voice comparison. *In*: I. FRECKELTON; H. SELBY (Eds.). **Expert Evidence**. Sydney, Australia: Thomson Reuters, 2010.

MORRISON, G. S. Forensic voice comparison and the paradigm shift. **Science and Justice**, v. 49, n. 4, p. 298–308, 2009.

OLIVEIRA, J. C. C. **Multiparametric analysis of phonetic-acoustic measures in genetically and non-genetically related speakers: implications for forensic speaker comparison**. Tese de doutorado–Campinas: Universidade Estadual de Campinas, 2021.

PETTORINO, M. et al. **VtoV: a perceptual cue for rhythm identification**. *In:* Prosody-Discourse Interface Conference 2013. **Anais** [...]. Leuven: 2013. p. 101-106.

PFITZINGER, H. R. **Two approaches to speech rate estimation**. *In:* Proceedings of the 6th Australian Internation Conference on Speech Science and Technology (SST '96). **Anais** [...]. 1996. Adelaide: Australasian Speech Science and Technology Association, 1996. p. 421–426.

PFITZINGER, H. R. **Local speech rate as a combination of syllable and phone rate**. *In:* Proceedings of the 5th ICSLP. **Anais** [...]. Sydney: ISCA, 1998. p. 1087–1090.

QUENÉ, H. On the just noticeable difference for tempo in speech. **Journal of Phonetics**, v. 35, p. 353–362, 2007.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2020.

SKARNITZL, R.; VAŇKOVÁ, J. Fundamental frequency statistics for male speakers of Common Czech. **AUC PHILOLOGICA**, v. 2017, n. 3, p. 7–17, set. 2017.

TURK, A.; NAKAI, S.; SUGAHARA, M. Acoustic segment durations in prosodic research: a practical guide. *In*: SUDHOFF, S. et al. (Eds.). **Methods in empirical prosody research**. Berlin: Walter de Gruyter, 2006. p. 2–27.

WIGHTMAN, C. W. et al. Segmental durations in the vicinity of prosodic phrase boundaries. **The Journal of the Acoustical Society of America**, v. 91, n. 3, p. 1707–1717, 1 mar. 1992.