Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Curso de Especialização em Estatística

# TREINAMENTO AUDITIVO COM ESTÍMULOS VOCAIS ÂNCORAS SINTETIZADOS: EFEITO NA CONCORDÂNCIA DOS AVALIADORES

Ana Cristina Côrtes Gama

Belo Horizonte
2020

Ana Cristina Côrtes Gama

# TREINAMENTO AUDITIVO COM ESTÍMULOS VOCAIS ÂNCORAS SINTETIZADOS: EFEITO NA CONCORDÂNCIA DOS AVALIADORES

Versão final da monografia apresentada ao Programa de Pós-Graduação em Estatística, da Universidade Federal de Minas Gerais, como requisito à obtenção do título de Especialista em Estatística.

Orientador: Prof. Dr. Roberto da Costa Quinino

Belo Horizonte
2020

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

Gama, Ana Cristina Côrtes

G184t Treinamento auditivo com estímulos vocais âncoras sintetizados: efeito na concordância dos avaliadores / Ana Cristina Côrtes Gama— Belo Horizonte, 2020.
64.f.. il.; 29 cm.

Monografia (especialização) - Universidade Federal de Minas Gerais – Departamento de Estatística.
Orientador: Roberto da Costa Quinino.

1. Estatística. 2. Qualidade da voz. .3. Distúrbios da voz.4. Disfonia.5. Treinamento da voz.I. Orientador. II. Título.

CDU 519.2 (043)

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

# ATA DO 208º. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE ANA CRISTINA CÔRTES GAMA.

Aos quinze dias do mês de junho de 2020, às 10:00 horas, com utilização de recursos de videoconferência à distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Ana Cristina Côrtes Gama**, intitulado: "Treinamento Auditivo com Estímulos Vocais Âncoras Sintetizados: Efeitos na Concordância dos Avaliadores", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Roberto da Costa Quinino – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 15 de junho de 2020.


Prof. Roberto da Costa Quinino (Orientador)
**Departamento de Estatística / UFMG**

Prof. Frederico Rodrigues Borges da Cruz
**Departamento de Estatística / UFMG**

Profa. Adriane Mesquita de Medeiros
**MED/UFMG**

## RESUMO

**Objetivo:** Analisar o resultado do treinamento auditivo, com vozes sintetizadas, na concordância intra-avaliadores da análise perceptivo-auditiva de rugosidade e soprosidade. **Método:** A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa (37872314.2.0000.5149). Trata-se de um estudo experimental composto por quatro sessões de treinamento auditivo com vozes humanas e estímulos âncoras sintetizados. A amostra foi composta por vinte alunos do curso de Fonoaudiologia que possuíam contato prévio com a avaliação perceptivo-auditiva, com idade de 21 a 37 anos, com média de 24 anos, sendo três homens e dezessete mulheres. Os avaliadores participaram de quatro sessões de treinamento com o intervalo de sete dias entre eles. Cada treinamento consistiu em três tarefas: 1) Atividade pré-treinamento: julgamento de 20 vozes naturais neutras e disfônicas, onde os participantes avaliaram os parâmetros de rugosidade e soprosidade e o grau de desvio vocal (0 – neutro, 1 – leve, 2 – moderado, 3 – intenso); 2) Atividade de treinamento: foram apresentados quatro estímulos âncoras sintetizados de rugosidade (R) e quatro de soprosidade (B) com o grau geral de desvio vocal variando de zero a três. Os participantes ouviram quatro estímulos de vozes naturais e um estímulo âncora, e foram orientados a parear a voz natural que mais se assemelhava ao estímulo âncora sintetizado. Os participantes realizaram o treinamento com 20 vozes naturais; e 3) Atividade pós-treinamento: as 20 vozes da atividade pré treinamento foram randomizadas e os indivíduos julgaram as mesmas vozes sem conhecimento prévio de que as vozes foram repetidas. A análise estatística dos dados foi realizada pelo teste $AC_2$ para avaliação da concordância intra-avaliadores e o Teste não-paramétrico de Friedman para comparação entre as sessões de treinamento. O software utilizado foi o R (versão 3.5.1) para realização do teste $AC_2$ e o Minitab® 19 para o teste de Friedman. Foi considerado um nível de significância de 5%. **Resultados:** Na análise da concordância intra-avaliador para o parâmetro perceptivo- auditivo de rugosidade, os resultados variaram de 79%, 85%, 85% e 86% entre a primeira e a quarta sessão de treinamento auditivo, respectivamente, com melhora da concordância intra-avaliador a partir da quarta sessão (p=0,005). Para a análise da concordância intra-avaliador no parâmetro auditivo de soprosidade, os resultados foram 88%, 90%, 90% e 92% da primeira à quarta sessão de treinamento auditivo, respectivamente. Na quarta

sessão de treinamento os juízes apresentaram uma maior concordância (p=0,036). **Conclusão:** O parâmetro perceptivo-auditivo de soprosidade apresentou indicador $AC_2$ maior do que para rugosidade indicando parecer mais concordante. Como no caso da rugosidade, a concordância intra-avaliador apresentou melhora a partir da quarta sessão de treinamento auditivo para avaliação da soprosidade.

**Descritores:** Voz; Qualidade da Voz; Distúrbios da Voz; Disfonia, Percepção Auditiva; Treinamento da voz.

## ABSTRACT

**Objective:** to analyze the results from the perceptual training, with synthesized auditory anchors, in the intrarater agreement of the perceptual rating of roughness and breathiness. **Method:** The research was approved by the Research Ethics Committee (37872314.2.0000.5149). This is an experimental study consisting on four perceptual training sessions with human voices and synthesized anchors. The sample consisted on twenty Speech Language Pathologist students who had previous contact with auditory-perceptual assessment, aged 21 to 37 years old, with an average of 24 years old, three men and seventeen women. The evaluators participated on four training sessions with an interval of seven days between them. Each training consisted of three tasks: 1) Pre-training activity: judgment of 20 normal and dysphonic voices, where the participants evaluated the parameters of roughness and breathiness and the degree of vocal deviation (0 - neutral, 1 - mild, 2 - moderate, 3 - intense); 2) Training activity: four synthesized auditory anchors for roughness (R) and four synthesized auditory anchors for breathiness (B) were presented with the general degree of vocal deviation ranging from zero to three. Participants heard four natural voice stimuli and an anchor stimulus, and were instructed to pair the natural voice that most closely resembled the synthesized anchor stimulus. Participants performed the training with 20 natural voices; and 3) Post-training activity: the 20 voices from the pre-training activity were randomized and the individuals judged the same voices without prior knowledge that the voices were repeated. Statistical analysis of the data was performed using the $AC_2$ test to assess intrarater agreement and the Friedman non-parametric test for comparison between training sessions. The software used was R (version 3.5.1) to perform the $AC_2$ test and Minitab® 19 for the Friedman test. A significance level of 5% was considered. **Results:** In the analysis of the intrarater agreement for the perceptual-auditory parameter of roughness, the results varied between 79%, 85%, 85% and 86% between the first and the fourth auditory training session, respectively, with improved intrarater agreement from the fourth session ($p = 0.005$). For the analysis of the intrarater agreement on the breathiness, the results were 88%, 90%, 90% and 92% from the first to the fourth auditory training session, respectively. In the fourth training session, the judges showed a greater

agreement (p = 0.036). **Conclusion:** The perceptual-auditory parameter of breathiness showed a higher indicator $AC_2$ than for roughness, suggesting that it was more consistent. As in the case of roughness, breathiness also showed improvement in intrarater agreement from the fourth session of auditory training.

**Key Words:** Voice; Voice Quality; Voice Disorders; Dysphonia; Auditory Perception; Voice Training.

## SUMÁRIO

## AGRADECIMENTOS

*À Maria Vanda Ferreira por ter sempre cuidado com carinho de mãe dos amores da minha vida: Lara e Taís. Vanda, com sua ajuda e apoio eu consegui trabalhar com tranquilidade, tendo a certeza que você estaria ao lado das meninas, dando a mesma atenção e carinho que eu daria. Muito obrigada por estar conosco todos estes anos.*

*Ao meu orientador, Prof. Dr. Roberto da Costa Quinino, pela disponibilidade, atenção e orientações que tornaram possíveis este trabalho. Foi um privilégio tê-lo como orientador.*

*Aos professores do curso de Especialização em Estatística pelos ensinamentos e disponibilidade.*

*A todos os colegas do curso de Especialização, pelos ótimos momentos de convívio e de trocas, em especial às colegas sempre presentes, Carla e Sara.*

*Aos colegas da linha de pesquisa de Análise perceptivo-auditiva da voz, Maurílio, João Pedro, Priscila e Sabrina. Obrigada pela parceria e aprendizados.*

*Ao Marco Aurélio, Lara e Taís, meu agradecimento pelo amor, compreensão, e incentivo durante o período do curso de Especialização. O apoio e carinho de vocês foi imprescindível para eu conseguir vencer este grande desafio. Para vocês todo o meu amor!!*

*Aos meus pais, irmãos, cunhada e sobrinhos, por darem um sentido tão importante e especial à palavra "família". Muito obrigada pelo amor, carinho e, por estarem sempre ao meu lado.*

*Aos meus alunos do curso de graduação e pós-graduação em Fonoaudiologia da UFMG. É o desafio da docência que me incentiva a procurar novos conhecimentos.*

*À Fundação de Desenvolvimento da Pesquisa (Fundep) pela bolsa de estudos.*

## LISTA DE ILUSTRAÇÕES

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| UFMG | Universidade Federal de Minas Gerais |
| COEP | Comitê de Ética em Pesquisa |
| PPVV | Pregas vocais |
| TV | Trato Vocal |
| TA | Músculo tireoaritenoideo |
| PV | Prega vocal |
| $f_0$ | Frequência Fundamental |
| G | Grau geral de desvio vocal |
| R | Rugosidade |
| B | Soprosidade |

# CONSIDERAÇÕES INICIAIS

Trata-se de um projeto de pesquisa de monografia de Especialização em Estatística que tem como objetivo analisar o resultado do treinamento auditivo com estímulos âncoras sintetizados, na concordância intra-avaliadores da análise perceptivo-auditiva da rugosidade e da soprosidade, utilizando os métodos de estatística de análise de concordância $AC_2$ de Gwet, e o teste não-paramétrico de Friedman. Este projeto de pesquisa é um estudo experimental, comparativo intra-sujeito, com amostra de conveniência, realizado na Faculdade de Medicina e no Instituto de Ciências Exatas da Universidade Federal de Minas Gerais (UFMG).

Neste primeiro capítulo serão abordadas as principais temáticas que envolvem este projeto de pesquisa, ancoradas pela literatura das áreas de conhecimento.

## 1. Produção Vocal

A voz permite a comunicação do ser humano, transmite informações e revela características emocionais do discurso (Behlau et al., 2001a). Nos países desenvolvidos, cerca de 60% da população economicamente ativa depende de sua voz e habilidades comunicativas para desempenhar seu trabalho (Roy et al, 2004).

A produção vocal é resultado da vibração das pregas vocais (PPVV), e amplificação do som laríngeo pelo trato vocal (TV). A prega vocal é composta por corpo e cobertura. O corpo é formado pela parte profunda da lâmina própria e pelo músculo tireoaritenoideo (TA). Atuando em harmonia com a cobertura, composta por epitélio e camada superficial da lâmina própria, a vibração das PPVV permite a propagação da onda mucosa verticalmente, responsável pela produção do som da voz humana (Alipour, Titze, 1991) (Figura 1).

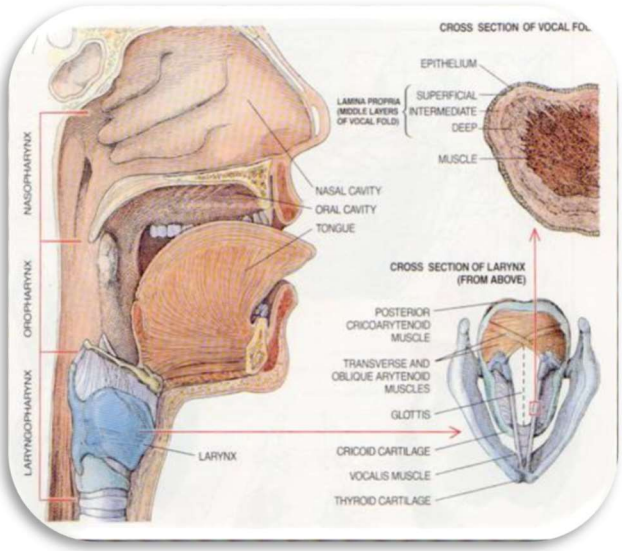



Figura 1: Imagem das estruturas laríngeas e cortes axial e sagital da prega vocal.

Fonte: Professional Voice – Robert Sataloff

A vibração das PPVV produz sucessivos ciclos glóticos (Figura 2), e estes se iniciam quando a pressão subglótica do ar expirado supera a resistência do músculo TA das PPVV. O ciclo glótico é composto pelas fases aberta e fechada. Na fase de abertura a pressão subglotal faz com que ocorra o movimento lateral da onda mucosa, caracterizada pela diferença de fase da prega vocal (PV), que pode ser avaliada por valores obtidos da onda mucosa (Lohscheller, 2008). Na fase fechada as bordas livres das PPVV entram em contato, interrompendo o fluxo aéreo.

Figura 2: Ciclo glótico composto pelas fases fechada e aberta.

Fonte: Professional Voice – Robert Sataloff

O som produzido pelas PPVV será amplificado e modificado pelo TV, que é limitado anteriormente pelos lábios e posteriormente pelas PPVV. O TV é, portanto, responsável pela ressonância da voz, definida como a amplificação dos sons produzidos ao nível das PPVV (Behlau et al.,2001a) (Figura 3).

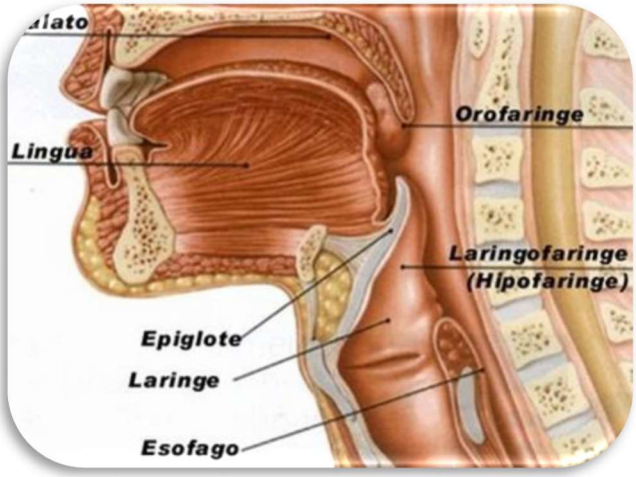



Figura 3: Representação do trato vocal e dos articuladores.

Fonte: https://acordesprajesus.files.wordpress.com/2017/07/voz-humana-aparelho-fonador.jpg

São as características anatômicas e fatores individuais que determinam os diferentes tipos de voz. As vozes masculinas, femininas e infantis apresentam características específicas em decorrência das diferenças anatômicas da laringe e das PPVV entre os sexos nas diferentes faixas etárias (Behlau et al., 2001b). A frequência fundamental ($f_0$) é a velocidade na qual as PPVV vibram por unidade de tempo, o que é indicado por Hz (Hertz). A $f_0$ da voz feminina varia de 150 a 250 Hz, caracterizando a formação de 150 a 250 ciclos glóticos em um segundo de vibração das PPVV. Nos homens a $f_0$ varia de 80 a 150 Hz, e nas crianças a $f_0$ é acima de 250 Hz. (Alipour, Titze, 1991).

A voz do indivíduo se modifica ao longo da vida, e as mudanças são determinadas por fatores relacionados ao desenvolvimento físico; mudanças emocionais e sociais; e aos diferentes usos da voz (Behlau et al., 2001a).

O choro é a primeira forma de comunicação da criança, e sua qualidade pode demonstrar sinais de dor, fome ou prazer (Vargas et al., 2015). A qualidade vocal das crianças é caracterizada por valores de $f_0$ que decrescem à medida que a idade aumenta, e com diferenças em relação à voz do adulto,

caracterizadas por uma leve soprosidade e instabilidade à emissão (Ramos et al., 2017).

Na adolescência as PPVV dos meninos dobram de tamanho, em torno dos 13 aos 15 anos de idade. Como resultado, a voz fica mais grave e se caracteriza como uma voz de adulto. Nas meninas o aumento das PPVV é menos significativo, a voz fica levemente mais grave, ao redor dos 12 aos 14 anos de idade (Gama et al., 2012). Essas mudanças ocorrem no período do desenvolvimento denominado muda vocal e, juntamente com as demais características sexuais secundárias, permitem a diferenciação de sexo por meio da voz, algo que não ocorria na infância (Behlau et al., 2001a).

Por volta dos 18 anos a voz atinge seu estágio maduro, caracterizada pela presença de um maior controle, e possibilidade de identificação sexual, emocional e social do indivíduo (Behlau et al., 2001b).

No idoso, após os 65 anos, ocorrem modificações funcionais e anatômicas na laringe, que modificam a qualidade vocal e caracterizam a presbifonia. A voz presbifônica (*prebys*, do grego = homem velho; *phoneo*, do grego = vocalizar ou emitir sons) caracteriza-se por qualidade vocal rouca/soprosa e instável, redução do tempo máximo de fonação (TMF), extensão vocal reduzida, alterações na $f_0$ e, ressonância nasal (Cerceau et al.,2009).

As mudanças vocais acompanham o indivíduo ao longo de sua vida, resultado das modificações anatomofuncionais e, das demandas comunicativas que ocorrem da infância à senescência.

## 2. Disfonia

A disfonia é caracterizada por toda e qualquer dificuldade ou alteração na emissão vocal que impeça a produção natural da voz (Behlau et al., 2001a), elas podem surgir quando ocorre uma alteração no padrão de vibração das PPVV, seja por ajustes musculares inadequados; por presença de lesões nas PPVV; ou pela associação destes fatores (Alipour, Titze, 1991).

A prevalência de disfonia na população brasileira é de 7,5% (Behlau et al., 2012), sendo mais frequentes nas mulheres do que nos homens (63,8% e 36,2% respectivamente) (Van Houtte et al., 2010).

Nos indivíduos que utilizam a voz profissionalmente a prevalência da disfonia é de 41%, e os docentes são os profissionais mais acometidos pelos

quadros de disfonia (Van Houtte et al., 2010). Nos professores, o tempo de uso da voz dobra, quando comparados aos não profissionais da voz (Hunter et al., 2010), e a prevalência de disfonia é de 52% (Medeiros et al., 2008).

Na população infantil a prevalência de disfonia varia de 6% a 38%, e o comportamento vocal inadequado é o principal fator etiológico, em 92% das crianças disfônicas, sendo mais prevalente nos meninos (Ramos et al., 2017).

Na população idosa a disfonia possui uma prevalência de 29%, com a presença de impacto negativo na qualidade de vida (Nichols et al., 2015).

As disfonia são mais comuns nas mulheres, justificado pelo fato destas possuírem em relação aos homens: 1) pregas vocais menores e com menor área de contato, sujeitas a maior força de atrito entre as mesmas (Miranda et al., 2011), 2) grande variabilidade da quantidade de ácido hialurônico em suas PPVV, relacionada à fase do ciclo menstrual, podendo ocasionar menor proteção de fonotraumas durante a vocalização, e menor viscoelasticidade da túnica mucosa (Korn et al, 2011) e 3) maior $f_0$ provocando a produção de um maior número de ciclos glóticos no tempo (Bridger, Epstein,1983).

A literatura classifica as disfonias pelo aspecto etiológico, sendo definidos três tipos: funcionais, organofuncionais e orgânicas (Behlau et al., 2001a).

As disfonias funcionais são decorrentes do uso da voz, onde não se observa lesões nas PPVV, também conhecidas como disfonias comportamentais.

As disfonias organofuncionais são definidas pela presença de lesões nas PPVV decorrentes do uso da voz, e os exemplos mais prevalentes são os nódulos (Figura 4) e os pólipos (Figura 5) (Behlau et al.,2001a).


Figura 4: Nódulos de pregas vocais em laringe feminina.

Fonte: arquivo pessoal da autora.

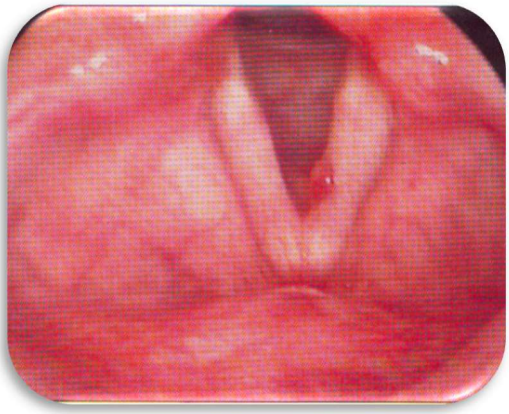


Figura 5: Pólipo de prega vocal esquerda em laringe feminina.
Fonte: Voz: o Livro do Especialista – Mara Behlau

Disfonias orgânicas são os quadros de disfonia que independem do uso da voz, e são caracterizadas principalmente pelos quadros neurológicos, como as paralisias de PPVV (Behlau et al., 2001a).

As disfonias são distúrbios decorrentes de alterações funcionais, aerodinâmicas e/ou acústicas sendo, portanto, de natureza multidimensional e de origem multifatorial, presente em todos os ciclos de vida (Roy et al, 2004).

Disfonias de base comportamental são as mais prevalentes, correspondendo 30% da clínica vocal (Van Houtte et al., 2010), e são mais frequentes nas mulheres e nos profissionais da voz. Os docentes são, entre os profissionais da voz, os mais vulneráveis para a ocorrência de disfonia, devido principalmente ao uso continuado da voz em condições desfavoráveis de ambiente de trabalho (Medeiros et al., 2008).

## 3. Avaliação da Voz

Devido ao aspecto multidimensional da disfonia, o Comitê de Foniatria da Sociedade Europeia de Laringologia sugere um protocolo para avaliação da disfonia que inclui a avaliação perceptivo-auditiva e acústica da voz; videoestroboscópica da laringe; aerodinâmica da fonação; e avaliação da autopercepção da disfonia por meio de protocolos de qualidade de vida (Dejonckere et al., 2002).

A análise perceptivo-auditiva é uma avaliação não invasiva, de baixo custo e de rápida realização, considerada exame padrão na clínica vocal por ser

capaz de perceber nuances da qualidade da voz que análises instrumentais não são capazes de avaliar (Yamasaki, Gama, 2019).

A avaliação acústica da voz é um exame instrumental, objetivo, que fornece informações qualitativas sobre o formato da onda sonora, e medidas quantitativas sobre o grau de periodicidade das ondas acústicas (Figura 6) (Felippe, et al., 2006).

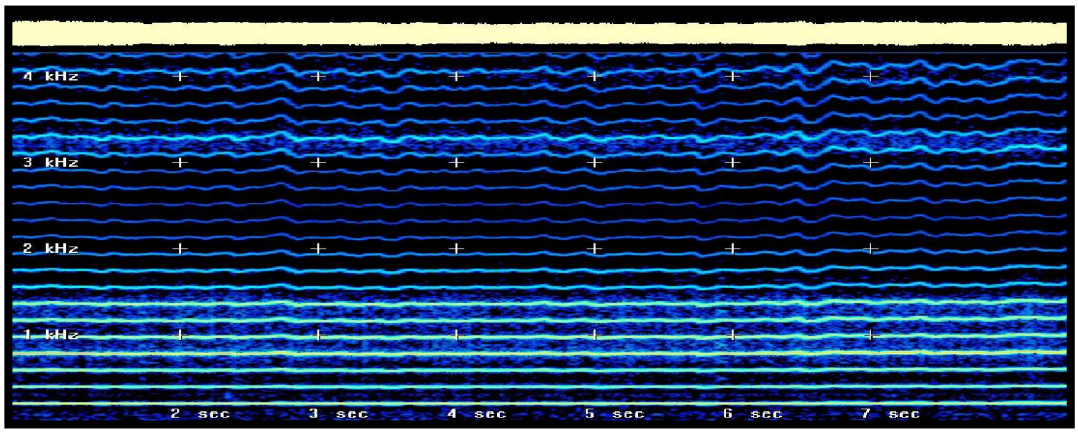



Figura 6: Espectrograma tempo X frequência da vogal /a/ sustentada.
Fonte: arquivo pessoal da autora.

O exame da laringe, denominado videolaringoscopia, é usado para se avaliar as estruturas anatômicas da laringe e, a função das PPVV durante a fonação. A análise visual destas estruturas envolve a avaliação das bordas livres das PPVV; da mobilidade da prega vocal (abdução / adução); da atividade supraglótica durante a fonação; e de manobras laríngeas em diferentes comportamentos vocais, como emissões cantadas e reflexas (Patel et al., 2018). A videolaringoestroboscopia utiliza uma fonte de luz estroboscópica, sendo um método convencional para a análise das propriedades vibratórias das PPVV, que permitem caracterizar os aspectos mucoondulatórios da túnica mucosa (Figura 7) (Wittenberg et al., 2000).

Figura 7: Exame laríngeo por videolaringoscopia de alta velocidade.
Fonte: arquivo de Elisa Meiti Ribeiro Lin Plec.

As medidas aerodinâmicas da voz, como a pressão subglótica e o fluxo oral, fornecem informações sobre as propriedades respiratórias e musculares da produção da voz. A laringe converte a energia aerodinâmica do ar expirado dos pulmões em energia acústica produzida pela vibração das PPVV (Genilhú, Gama, 2018). A avaliação aerodinâmica é objetiva e não invasiva, e auxilia na análise das propriedades aerodinâmicas e mioelásticas da fonação (Figura 8) (Patel et al., 2018).

Figura 8: Avaliação aerodinâmica da emissão da fala.
Fonte: arquivo de Patricia de Freitas Lopes Genilhú.

A qualidade de vida é definida na literatura como “a percepção do indivíduo de sua posição na vida, no contexto da cultura e do sistema de valores

nos quais ele vive e em relação aos seus objetivos, expectativas, padrões e preocupações" (Gill, Feinstein, 1994). Os protocolos de qualidade de vida em voz são questionários de autoavaliação que analisam o impacto da disfonia na qualidade de vida dos indivíduos, e envolvem dimensões emocionais, sociais, físicas e profissionais (Behlau et al., 2009).

Devido ao aspecto multidimensional do disfonia, a avaliação da voz deve conter aspectos relacionados ao clínico, com respaldo na análise da função muscular, respiratória, articulatória e ressonantal, além de uma avaliação centrada no paciente, com os protocolos de autoavaliação vocal.

## 4. Avaliação Perceptivo-Auditiva

A voz pode ser avaliada de várias formas, mas o que motiva o paciente a procurar por tratamento fonoaudiológico é a alteração auditiva que ele ou os outros percebem na sua voz, ou seja, as modificações de natureza perceptiva da qualidade vocal (Oates, 2009).

A avaliação perceptivo-auditiva da qualidade da voz é de natureza holística e integrativa, já que sua análise depende da impressão que um ouvinte tem da voz de um falante como um todo (Behlau et al., 2001a).

É a avaliação padrão na clínica vocal, porém há muitas críticas à sua subjetividade e limitações (Oates, 2009). Embora as medidas da avaliação perceptivo-auditiva sejam fáceis de serem obtidas e não necessitem de instrumentalização, possuem baixa sensibilidade e concordância entre avaliadores (Yamasaki, Gama, 2019).

Um dos grandes desafios do fonoaudiólogo na análise perceptivo-auditiva da qualidade vocal é melhorar os resultados da confiabilidade desta avaliação. As baixas concordâncias dos avaliadores no julgamento perceptivo-auditivo da voz humana podem ser explicadas por três fatores principais: 1) aspectos ligados ao avaliador (tempo de experiência e realização de treinamento auditivo); 2) o tipo de estímulo vocal a ser avaliado (vogal sustentada ou fala encadeada); e 3) o tipo de escala perceptivo-auditiva utilizada (Freitas et al., 2014).

A análise perceptivo-auditiva se baseia no modelo de protótipo da Ciência Cognitiva, ou seja, na utilização de padrões internos (Ghio et al., 2015). Este modelo pode ser explicado pelo fato de o avaliador classificar uma voz a partir de seu padrão interno, ao considerar a similaridade de um parâmetro vocal com

a referência que possui desta categoria de voz (Ghio et al., 2015). Neste sentido, ao analisar uma voz, o avaliador compara o novo estímulo com o padrão interno que possui, e classifica o novo estímulo vocal a partir desta comparação.

O desenvolvimento do padrão interno do avaliador pode ocorrer por múltiplas vias. O tempo de experiência do avaliador é um dos aspectos que podem afetar o grau de representatividade do padrão interno. Pesquisas evidenciam uma correlação positiva entre o tempo de experiência do avaliador e a confiabilidade da análise perceptivo-auditiva (Sofranko, Prosek, 2012; Oliveira et al., 2016) sugerindo que avaliadores experientes possuem maior concordância na análise perceptivo-auditiva que avaliadores inexperientes, que por terem mais vivência, possuem mais tempo para aperfeiçoar a construção de seus padrões internos para os diferentes tipos de qualidade vocal.

Outro aspecto que afeta a representatividade do padrão interno é o treinamento auditivo. Vários são os tipos de estratégias utilizadas nos programas de treinamento auditivo para o desenvolvimento dos padrões internos dos avaliadores, podendo destacar: 1) apresentação dos conceitos dos parâmetros perceptivo-auditivos treinados; 2) uso de referências externas com estímulos âncoras de vozes naturais ou sintetizadas; e 3) descrição dos dados laríngeos relacionados à voz avaliada (Chan, Yiu, 2002).

O uso de referências externas é sugerido na literatura (Brinca et al., 2015; Santos et al., 2018) para melhorar a confiabilidade da análise perceptivo-auditiva, já que este substitui o padrão interno do avaliador, e pode conduzir a uma avaliação mais confiável. Pesquisas evidenciam que a utilização de âncoras sintetizadas ou naturais, como referências externas, aumentam a concordância intra e interavaliador (Chan, Yiu, 2006; Gurlekian et al., 2016; Santos et al., 2018), e que as âncoras de vozes sintetizadas são mais eficientes, por serem capazes de manipular um único parâmetro perceptivo-auditivo, simplificando o julgamento auditivo da qualidade da voz humana (Chan, Yiu, 2006; Gurlekian et al., 2016).

Programas de treinamento auditivo com estímulos âncoras sintetizados, baseados no Método psicofísico de estimação de categorias (*Intramodal matching procedure*) (Gurlekian et al., 2016), onde o avaliador deve parear uma voz a um estímulo âncora que mais se assemelhe a ela, mostram impacto

positivo na concordância intra (Gurlekian et al., 2016) e interavaliador (Gurlekian et al., 2016; Santos et al., 2018) da análise perceptivo-auditiva.

Apesar de a literatura apresentar resultados promissores sobre os efeitos positivos do treinamento auditivo na confiabilidade da análise perceptivo-auditiva, pesquisas ainda são necessárias para se compreender qual o melhor formato de Programas de Treinamento auditivo para a formação de avaliadores pouco experientes, como alunos de graduação ou fonoaudiólogos recém-formados, considerando-se a grande importância da análise perceptivo-auditiva na clínica vocal, relacionada ao processo diagnóstico e ao tratamento das disfonias.

## 5. Teste Estatístico de Concordância $AC_1$ e $AC_2$ de Gwet

Análise de concordância é medir a capacidade de se obter resultados idênticos, aplicados ao mesmo sujeito/fenômeno, em situações de medições realizadas por instrumentos diferentes; pelo mesmo instrumento em tempos distintos; por avaliadores diferentes; ou pela combinação de alguma dessas situações (Miot, 2016).

Diversos exemplos podem ser utilizados para enumerar estas situações como a calibração de instrumentos; avaliação de equivalência entre distintas ferramentas de medidas; julgamento em provas de aptidão; e análise diagnóstica (concordância intra e interavaliador) e psicométrica (reprodutibilidade do teste diagnóstico) (Kottner et al. 2011).

Pesquisadores de vários campos do conhecimento precisam avaliar a qualidade de uma coleta de dados de um projeto de pesquisa. Em muitos estudos, uma ferramenta de coleta de dados, como um questionário; um procedimento laboratorial; ou um sistema de classificação de uma determinada variável, é usada por diferentes avaliadores, observadores ou juízes (Gwet, 2008). Em um esforço para minimizar o efeito do fator avaliador ou do teste diagnóstico na qualidade dos dados coletados, os pesquisadores necessitam comprovar se todos os juízes aplicam o método de coleta de dados de maneira consistente (concordância intra-avaliador) e/ou concordante entre seus pares (concordância interavaliador), ou se o teste diagnóstico apresenta repetibilidade das respostas (Miot, 2016).

Os testes estatísticos de concordância quantificam a proximidade das pontuações atribuídas pelos avaliadores a um conjunto de dados de pesquisa, e quanto mais próximas estas pontuações, mais concordantes os avaliadores, e consequentemente, maior pode ser a confiabilidade do método de coleta de dados (Gwet, 2008).

Existem várias abordagens estatísticas usadas na medição da concordância, e a escolha do teste estatístico se baseia em suposições relativas ao tipo de variável (nominal, ordinal, intervalar ou contínuas), tipo de amostragem (aleatória, consecutiva, conveniência) e no tratamento de erros aleatórios e sistemáticos (Kottner et al., 2011).

Um fluxograma proposto por Gwet (2014) e apresentado na Figura 9, leva em consideração os tipos de variáveis analisadas, na escolha do teste estatístico de concordância.

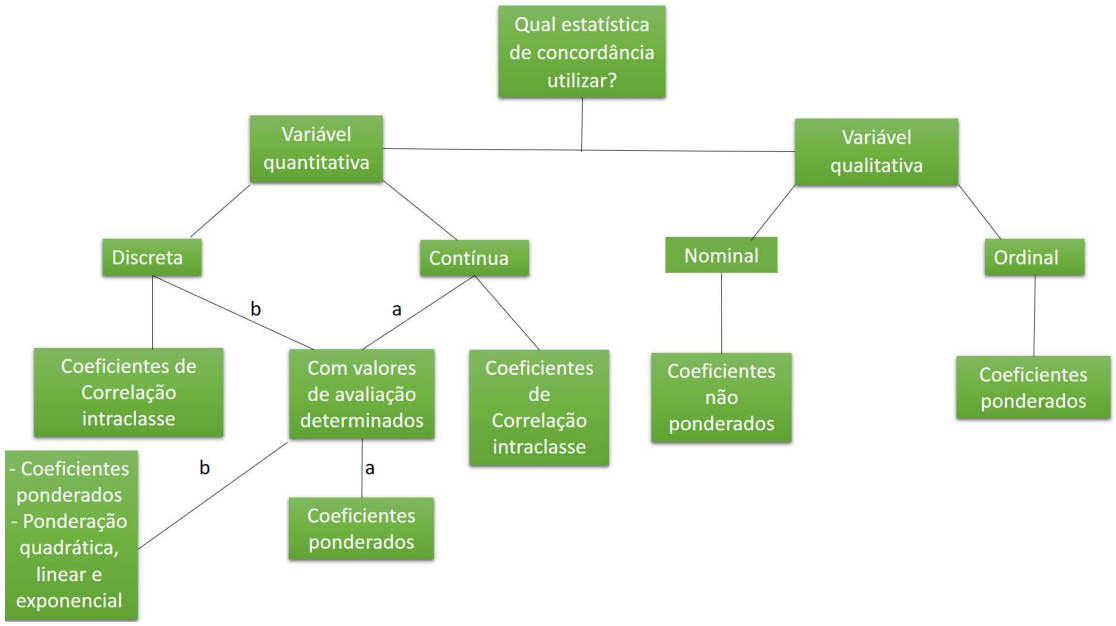



Figura 9: Fluxograma para escolha do teste estatístico de concordância proposto por Gwet (2014).

Considerando a escolha do teste estatístico de concordância sugerido por Gwet (2014), pode-se analisar os seguintes grupos:

1) Coeficientes não ponderados: englobam os testes estatísticos de concordância para variáveis nominais. Podem-se citar os testes Kappa Cohen e $AC_1$ de Gwet para experimento com dois avaliadores, e Kappa Fleiss e $AC_1$ de Gwet para experimentos com três ou mais avaliadores (Wongpakaran et al, 2013);

2) Coeficientes ponderados: indicados para situações de pesquisa onde as variáveis são ordinais, ou quantitativas contínuas e discretas quando os valores das medidas são pré-determinados no experimento. Neste grupo a literatura descreve os testes Kappa Cohen e $AC_2$ de Gwet para experimentos com dois avaliadores, e $AC_2$ de Gwet para experimentos com três ou mais avaliadores (Gwet, 2014);

3) Coeficientes de correlação intraclasse: indicados para pesquisas que utilizam variáveis quantitativas quando os valores das respostas da avaliação não são pré-determinados. Os Coeficientes de Correlação Intraclasse (ICC do inglês *Intraclass correlation coeficient*) ou coeficiente de reprodutibilidade (R) são baseados nos modelos de análise de variância (ANOVA) e estatística Kappa utilizando ponderação exponencial. O ICC é uma estimativa da variabilidade total de medidas devido a variações entre os indivíduos (Wongpakaran et al, 2013).

A literatura aponta limitações da estatística Kappa, apesar da sua popularidade como medida de concordância (Feinstein, Cicchetti, 1990). Sua limitação é a dependência na prevalência real da condição que está sendo avaliada (Santos, 2015). Situações em que existem totais marginais desbalanceados, a proporção bruta de concordância entre os avaliadores (concordância global) é alta, e a estatística Kappa pode apresentar valores baixos.

Os coeficientes de concordância $AC_1$ (Gwet, 2008) e $AC_2$ (Gwet, 2014) foram então propostos, na tentativa de se estimar valores de concordância mais robustos, com cálculos de concordância ao acaso mais adequados.

O primeiro coeficiente é para experimentos com qualquer número de avaliadores que utilizam um sistema de classificação categórica, e foi desenvolvido para superar as limitações associadas com o teste Kappa Cohen em experimentos com dois avaliadores, ou o Kappa Fleiss, nas situações de três ou mais avaliadores.

O primeiro coeficiente de concordância é chamado de coeficiente de concordância de primeira ordem ou estatística $AC_1$. Este indicador estatístico de concordância ajusta a probabilidade geral de concordância com a probabilidade de concordância ao acaso (Gwet, 2014). A concordância devido ao acaso ocorre quando os avaliadores concordam com uma classificação devido a um ou ambos os avaliadores atribuírem uma classificação de forma aleatória. Uma

classificação aleatória é observada quando um avaliador não tem certeza de como classificar uma variável, porque as características da variável não correspondem às instruções de classificação, ou porque a variável em análise é um caso raro ou atípico, portanto desconhecida do avaliador (Gwet, 2014).

A concordância devido ao acaso pode aumentar a probabilidade geral do teste, mas não deve contribuir para o valor de concordância real entre os avaliadores. Gwet (2008) afirma que "*a estatística $AC_1$ foi desenvolvida de tal forma que a concordância ao acaso é proporcional à ocorrência das classificações aleatórias, portanto, pontua o acaso na magnitude correta*".

A estatística $AC_1$ para experimentos com dois avaliadores e medida por escala nominal de q-níveis, é apresentada e denotada por $\widehat{\gamma_1}$ na equação 5.1.

$$\widehat{\gamma_1} = \frac{p_a - p_e}{1 - p_e} \tag{5.1}$$

Onde:

$$p_a = \sum_{k=1}^{q} p_{kk}$$

$$p_e = \frac{1}{1-q} \sum_{k=1}^{q} \pi_k (1 - \pi_k)$$

Na equação 5.1 temos:

- O símbolo $p_{kk}$ é o número relativo de classificações na categoria $k$ por ambos avaliadores;

- $\pi_k$ representa a probabilidade de um avaliador aleatoriamente classificar uma variável na categoria $k$, denotado por $\pi_k = \frac{(p_{k+} + p_{+k})}{2}$.

- Observe que $p_{k+}$ e $p_{+k}$ representam o número relativo das classificações na categoria $k$ pelos avaliadores A e B respectivamente.

- A probabilidade de concordância $p_e$ é o produto de duas quantidades:

1) A probabilidade de dois avaliadores concordarem, dado que a variável que está sendo avaliada é um caso raro e, portanto, recebeu uma classificação ao

acaso, é a probabilidade condicional $1/q$ desde que as avaliações ao acaso são consideradas aleatórias, com igual chance em todas as $q$ categorias.

2) A propensão de um avaliador apresentar uma análise ao acaso, é estimada por: $\sum_{k=1}^{q} \pi_k \, (1 - \, \pi_k)/(1 - 1/q)$. Nesta equação, Gwet (2014) enfatiza que uma distribuição de variáveis avaliadas de forma preponderante para algumas categorias, diminuirá a propensão aleatória da classificação.

Na Tabela 1 é apresentado o exemplo de Gwet (2014), ilustrando o indicador da estatística $AC_1$.

Tabela 1: Classificações de dor espinhal segundo os avaliadores.

| **Avaliador 1** | **Avaliador 2** | | |
|---|---|---|---|
| | **Síndrome Dessarranjo** | **Síndrome de Disfunção** | **Síndrome Postural** |
| **Síndrome de Desarranjo** | 55 | 10 | 2 |
| **Síndrome de Disfunção** | 6 | 4 | 10 |
| **Síndrome Postural** | 2 | 5 | 6 |

O teste $AC_1$ (equação 5.1) é calculado como:

$$\widehat{\gamma_1} = \frac{0{,}65 - 0{,}257725}{1 - 0{,}257725} = 0{,}5285$$

Observe que $p_{kk} = (55 + 4 + 6)/100$ ; $q$=3; $\pi_1 = \frac{0{,}67+0{,}63}{2}$; $\pi_2 = \frac{0{,}2+0{,}19}{2}$; $\pi_3 = \frac{0{,}13+0{,}18}{2}$.

Segundo Gwet (2002) identificar todas as respostas decorrentes da concordância ao acaso é impossível, portanto, é necessário calcular o valor da concordância ao acaso da forma mais adequada possível. Por meio de simulações e cálculos matemáticos, o autor concluiu que o valor razoável para a

probabilidade de concordância ao acaso não deve exceder a 0,5, e que a propensão a classificações ao acaso é definida em termos da proporção da máxima variância observada no experimento, o que torna a estatística $AC_1$ uma abordagem mais coerente e robusta em comparação à estatística Kappa (Gwet, 2014).

A estatística $AC_1$ para experimentos com três ou mais avaliadores é apresentada na equação 5.2 e denotada por $\widehat{\gamma_1}$.

$$\widehat{\gamma}_1 = \frac{p_a - p_e}{1 - p_e} \quad \text{onde,} \tag{5.2}$$

$$p_a = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \frac{r_{ik}\,(r_{ik} - 1)}{r_i (r_i - 1)}$$

$$p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k \,(1 - \pi_k)$$

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \frac{r_{ik}}{r_i}$$

Na equação 5.4 temos que:

- $q$ é o número de categorias;

- $r_{ik}$ é o número de avaliadores que classificaram a variável $i$ na categoria $k$;

- $r_i$ é o número de avaliadores que avaliaram a variável $i$;

- $n$ é o total de variáveis, e;

- $n'$ o número de sujeitos que foram avaliados por dois ou mais avaliadores de forma concordante.

O coeficiente de concordância de segunda ordem, ou estatística $AC_2$, foi desenvolvido como uma alternativa ao teste $AC_1$ para variáveis ordinais ou

quantitativas. Na análise de variáveis ordinais e quantitativas algumas respostas discordantes são mais sérias que outras, e apenas testes estatísticos de concordância ponderados, que definem pesos distintos para diferentes tipos de respostas, conseguem definir adequadamente os diferentes tipos de discordância/concordância (Gwet, 2014). Portanto, o teste $AC_2$ é uma versão ponderada do teste $AC_1$. Na estatística $AC_2$ discordâncias "menores" são tratadas como acertos parciais, e "grandes discordâncias" como erros (Gwet, 2014).

O teste $AC_2$ também se ajusta à concordância ao acaso e considera classificações faltantes (Gwet, 2014).

A estatística $AC_2$ para experimentos com dois avaliadores é apresentada na equação 5.3 e denotada por $\widehat{\gamma_2}$.

$$\widehat{\gamma_2} = \frac{p_a - p_e}{1 - p_e} \quad \text{onde,} \qquad (5.3)$$

$$p_a = \sum_{k,l}^{q} w_{kl}\ p'_{kl}$$

$$p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k\,(1 - \pi_k)$$

Sendo:

1) $q$ o número de pontuações usadas no experimento;

2) $w_{kl}$ é o peso associado com as duas categorias $k$ e $l$;

3) $p'_{kl} = p_{kl/\theta}$ , onde $p_{kl}$ é o número relativo de variáveis que os avaliadores 1 e 2 pontuaram como $x_k$ e $x_l$ respectivamente, e $\theta$ o número relativo de variáveis pontuadas pelos avaliadores, isto é, sem classificação ausente;

4) $T_w$ é a soma de todos os pesos $w_{kl}$ associados a todas as categorias;

5) $\pi_k$ é a probabilidade de um avaliador atribuir pontuação $x_k$ para uma variável, e é calculado como $\pi_k = (p'_{k+} + p'_{+k})/2$. Note que $p'_{k+} = \frac{p_{k+}}{\theta_1}$ com $\theta_1$ sendo o número relativo de variáveis que o avaliador 1 pontuou, e $p_{k+}$ o número relativo de variáveis que o avaliador 1 pontuou como $x_k$. Da mesma forma, $p'_{+k} = \frac{p_{+k}}{\theta_2}$ com $\theta_2$ sendo o número relativo de variáveis que o avaliador 2 pontuou, e $p_{+k}$ o número relativo de variáveis que o avaliador 2 pontuou como $x_k$.

6) Neste trabalho utilizamos $w_{kl} = 1 - (k-l)^2/(q-1)^2$ conhecido como peso quadrático.

A estatística $AC_2$ para experimentos com três ou mais avaliadores é o coeficiente $AC_1$ apresentado na equação 5.2, porém de forma ponderada, ou seja, considerando a ordem e a distância entre as respostas dos avaliadores. A estatística $AC_2$ é apresentada na equação 5.4 e também denotada por $\widehat{\gamma_2}$.

$$\widehat{\gamma_2} = \frac{p_a - p_e}{1 - p_e} \quad \text{onde,} \tag{5.4}$$

$$p_a = \frac{1}{n'} \sum_{i=1}^{n'} \sum_{k=1}^{q} \frac{r_{ik}\,(r_{ik}^{*} - 1)}{r_i(r_i - 1)}$$

$$p_e = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k\,(1 - \pi_k)$$

Sendo:

- $n'$ o número de variáveis que foram avaliadas por dois ou mais avaliadores, e $q$ o número de categorias diferentes que os juízes usaram no experimento para avaliar as variáveis;

- $r_i$ é o número de avaliadores que classificaram a variável $i$, e $r_{ik}$ é o número de avaliadores que avaliaram o $k$–th valor $x_k$ para a variável $i$.

- O termo $r_{ik}^{*}$ representa o valor ponderado dos avaliadores que analisaram a variável $i$, o $k$–th valor $x_k$ ou outra concordância parcial do experimento com $k$–th valores. $r_{ik}^{*}$ é definido como: $r_{ik}^{*} = \sum_{l=1}^{q} w_{kl}\ r_{il}$.

- $T_W$ é o somatório de todos os pesos $w_{kl}$ associados com todas as $q$ categorias, e $\pi_k$ é a propensão para atribuir pontuação $x_k$ para a variável, e é calculado como mostrado na equação 5.2.

**5. 1 Terminologia**

Quando se fala em testes estatísticos de concordância, é necessário que alguns termos, muitas vezes utilizados de forma intercambiável, ou até mesmo incorreta, sejam bem especificados. Neste sentido, esta seção objetiva conceituar termos comuns da área da Estatística e da pesquisa clínica que se ocupam da análise de concordâncias e validade de testes e avaliadores:

1) Concordância interavaliadores (*inter-rater reliability/agreement*): definem o quão dois ou mais avaliadores concordam no julgamento de uma mesma variável, utilizando uma mesma escala de medida. Em uma pesquisa, se dois avaliadores têm uma alta concordância, isto significa que os resultados de ambos avaliadores podem ser utilizados de forma intercambiável, que o julgamento do avaliador não irá interferir no resultado da pesquisa (Gwet, 2014). Na Figura 9 são apresentados os possíveis métodos estatísticos na análise desta medida.

2) Concordância intra-avaliadores (*intra-rater reliability/agreement)*: também conhecida como concordância teste-reteste, indica o quanto uma medida apresenta repetibilidade (variabilidade no sistema de medição causada pelo instrumento de medida) ou reprodutibilidade (variabilidade no sistema de medição causada pelas diferenças de julgamento dos juízes). Relaciona-se também a analisar o quanto o avaliador é capaz de avaliar mais de uma vez a mesma variável da mesma forma. Em outras palavras, a concordância intra-avaliador pode ser entendida como o quanto o juiz é capaz de apresentar uma consistência nos seus julgamentos. Em termos estatísticos, Gwet (2014) aponta que os cálculos destas medidas não requerem nenhum desafio especial, apenas

uma adaptação nos cálculos de medidas de concordância interavaliadores, que podem ser consultados na Figura 9.

3) Validade (*validity*): refere-se ao grau em que um instrumento realmente mensura a variável que pretende mensurar, também descrito na literatura como acurácia. Enquanto a concordância é um item necessário para garantir uma adequada validade, um sistema pode ser concordante, mas não apresentar acurácia. Resultados válidos envolvem resultados que são ao mesmo tempo concordantes e coincidentes com as referências "padrão-ouro" (Gwet, 2014). Testes que avaliam validade devem levar em consideração resultados considerados "corretos" ou referências padrão-ouro.

4) Consistência interna (*internal consistency*): nas áreas da saúde e humanas, existem questionários que apresentam duas ou mais questões que coletam a mesma informação, mas de formas distintas, em diferentes perspectivas. Se um grupo de questões que se propõe a medir o mesmo construto geral produzem resultados semelhantes, conclui-se que a consistência interna é alta. O teste coeficiente alfa, que foi descrito em 1951 por Lee J. Cronbach, é indicado nestes casos (Gwet, 2014).

## 6. Teste Estatístico Paramétrico ANOVA com medidas repetidas

### 6.1 Método Paramétrico

A análise ANOVA com medidas repetidas pode ser representada por meio de um modelo de regressão linear múltipla com (n-1) variáveis *Dummys* para diferenciar os $n$ participantes e outras ($k$-1) variáveis *Dummys* para diferenciar as $k$ medidas repetidas (Montgomery, 2001).

Neste modelo os indivíduos serão tratados como blocos e objetivamos verificar se as $k$ medidas repetidas são longitudinalmente diferentes. Para um melhor entendimento, considere a representação na Tabela 2 com n=4 participantes e k=3 medidas repetidas. A generalização para o caso geral com $n$ participantes e $k$ repetições entendemos que será similar. Observe que se k=2 então teremos o tradicional teste T pareado.

Tabela 2: Exemplo de experimentos com medidas repetidas

| | Avaliações Repetidas | | |
|---|---|---|---|
| **Participantes** | **1** | **2** | **3** |
| 1 | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ |
| 2 | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ |
| 3 | $Y_{31}$ | $Y_{32}$ | $Y_{33}$ |
| 4 | $Y_{41}$ | $Y_{42}$ | $Y_{43}$ |

Neste cenário podemos rearranjar os dados para uso do modelo de regressão linear múltipla sendo expresso pela Tabela 3:

Tabela 3: dados rearranjados para Regressão

| | | | Participante | | | Repetição | |
|---|---|---|---|---|---|---|---|
| **Avaliação** | **Participante** | **Repetição** | **$D_1$** | **$D_2$** | **$D_3$** | **$D_4$** | **$D_5$** |
| $Y_{11}$ | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| $Y_{12}$ | 1 | 2 | 1 | 0 | 0 | 0 | 1 |
| $Y_{13}$ | 1 | 3 | 1 | 0 | 1 | -1 | -1 |
| $Y_{21}$ | 2 | 1 | 0 | 1 | 0 | 1 | 0 |
| $Y_{22}$ | 2 | 2 | 0 | 1 | 0 | 0 | 1 |
| $Y_{23}$ | 2 | 3 | 0 | 1 | 0 | -1 | -1 |
| $Y_{31}$ | 3 | 1 | 0 | 0 | 1 | 1 | 0 |
| $Y_{32}$ | 3 | 2 | 0 | 0 | 1 | 0 | 1 |
| $Y_{33}$ | 3 | 3 | 0 | 0 | 1 | -1 | -1 |
| $Y_{41}$ | 4 | 1 | -1 | -1 | -1 | 1 | 0 |
| $Y_{42}$ | 4 | 2 | -1 | -1 | -1 | 0 | 1 |
| $Y_{43}$ | 4 | 3 | -1 | -1 | -1 | -1 | -1 |

Assim, o modelo de Regressão linear múltipla pode ser representado por:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 + \varepsilon \qquad (1)$$

Onde $\varepsilon$ é componente aleatória do erro e supostamente possui distribuição normal, e variância constante (homocedasticidade) (Rocha, Júnior, 2018).

Estas hipóteses precisam ser verificadas para o uso da metodologia paramétrica clássica. Estamos supondo também que a variância das diferenças entre todos os pares possíveis da repetição é igual.

Nosso objetivo é verificar se os valores médios da repetição são iguais ou não. Assim, podemos representar o objetivo por meio do teste de hipóteses:

$$\begin{cases} H_0: \beta_4 = \beta_5 = 0 \\ H_1: pelo\ menos\ uma\ diferença \end{cases}$$

$$\begin{cases} H_0 : \beta_4 = \beta_5 = 0 \\ H_1 : \text{Pelo menos uma diferença} \end{cases}$$

Sob $H_0$ a estatística:

$F = \frac{(R^2 - R_*^2)/(k-1)}{(1-R^2)/(n+k-1)} \sim F_{k-1;n+k-1}$. Assim, se o valor de F calculado com os dados do experimento for maior que o valor tabelado $F_{1-\alpha;k-1;n+k-1}$ então rejeitamos $H_0$ e concluímos com nível de significância 5% que as repetições não são todas iguais. A variável $R^2$ é o coeficiente de determinação do modelo (1) e a variável $R_*^2$ é o coeficiente de determinação do modelo incompleto

$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon^*$, ou seja assumindo $H_0$ como verdadeira.

Se rejeitarmos H0 então as comparações entre as repetições podem ser feitas com o uso do Teste T pareado com as probabilidades de significância ajustadas por meio da correção para testes múltiplos de Bonferroni. No nosso caso basta multiplicar pelo número de repetições. Se o valor for maior do que 1 então adotamos que a probabilidade de significância é 1.

### 6.2 Método Não Paramétrico: Teste de Friedman

O teste de Friedman é uma alternativa não paramétrica para a Anova com medidas repetidas. O objetivo é verificar se as repetições possuem resultados

iguais (hipótese H0) ou não. O teste deve ser usado quando a abordagem paramétrica não é recomendável por violar as hipóteses do modelo. Como o teste de Friedman não faz suposições sobre a distribuição, ele não é menos poderoso do que a versão paramétrica. Este teste utiliza os *ranks* dos dados ao invés de seus valores brutos para o cálculo da estatística de teste.

Para calcular a estatística de teste de Friedman, ordenamos as $k$ observações da menor para a maior de forma separada em cada um dos $n$ blocos e atribuímos os *ranks* {1, 2, ..., k} para cada bloco da tabela de observações. Assim, a posição esperada de qualquer observação sob H0 é $(k + 1) /2$. Sendo $r(Yij)$ o *rank* da observação $Yij$, $i$=1,2,…,n e $j$=1,…,$k$, definimos a soma de todos os *ranks* da coluna $j$ (ou seja, de cada repetição) por:

$$R_j = \sum_{i=1}^{n} Y_{ij}$$

Se $H_0$ é verdadeira, a estatística S possui distribuição aproximadamente Qui-quadrado com $k$-$1$ graus de liberdades. Assim, H0 deve ser rejeitada se o valor observado de S for superior ao valor $\chi^2_{k-1;\,1-\alpha}$ com nível de significância $\alpha$.

Quando rejeitamos $H_0$ podemos considerar significativas os pares de repetições cujas diferenças absolutas de $R_j$ forem maior ou igual a $Z_{1-\frac{\alpha}{k(k-1)}}\sqrt{\frac{nk(k-1)}{6}}$, em que Z indica a distribuição normal padrão o valor de Z. Maiores detalhes podem ser obtidos em Siegel, Castellan (2006).

## ESTUDO DE CASO

Trata-se de um estudo experimental, de natureza quantitativa e de comparação intra-sujeitos. A pesquisa foi aprovada pelo Comitê de Ética em Pesquisa (COEP) da Universidade Federal de Minas Gerais (UFMG) sob o parecer CAAE – 37872314.2.0000.5149 (Anexo 1).

Os participantes leram o Termo de Consentimento Livre e Esclarecido (TCLE) e, concordaram em participar da pesquisa. Posteriormente responderam um breve questionário informando idade, sexo e se tinham experiência no julgamento perceptivo-auditivo de vozes.

Para a realização da pesquisa foi desenvolvido e disponibilizado pelos pesquisadores o Programa de treinamento auditivo, que consistiu de quatro sessões de treinamento em avaliação perceptivo-auditiva com vozes naturais e estímulos âncoras sintetizados.

### 1. Amostra

A amostra foi composta por 20 alunos do curso de Fonoaudiologia da UFMG, com idade de 21 a 37 anos, e média de 24 anos (DP=3,87), sendo três homens e 17 mulheres. Foi considerado critério de inclusão o participante ter experiência em análise perceptivo-auditiva e não apresentar queixa autorreferida de perda de audição. Considerou-se a presença de experiência em avaliação perceptivo-auditiva os juízes já terem formação acadêmica e treino na aplicação de protocolos de análise auditiva da voz. Foram excluídos os participantes que não realizaram todas as etapas das quatro sessões do programa de treinamento auditivo. Participaram da pesquisa 24 alunos, e quatro foram excluídos.

### 2. Programa de treinamento auditivo

O Programa de treinamento auditivo foi composto por quatro sessões, com tempo médio de duração de cada sessão de 45 minutos. Todos os avaliadores participaram de todas as sessões, com o intervalo de sete dias entre cada uma, durante o período de um mês.

Todas as sessões do Programa foram compostas por três atividades, assim organizadas:

1) Atividade pré-treinamento: constou na avaliação de 20 vozes naturais neutras e disfônicas, onde os participantes avaliaram os parâmetros de rugosidade (R- percepção de irregularidade durante a produção vocal) e soprosidade (B-percepção de escape de ar audível durante a produção vocal), e marcaram o grau de desvio vocal em uma escala de *Likert* de quatro pontos (0 – neutro, 1 – leve, 2 – moderado, 3 – intenso) para cada um dos dois parâmetros perceptivo-auditivos avaliados;

2) Atividade de treinamento: utilizou o Método psicofísico de estimação de categorias (Gurlekian et al., 2016). Foram apresentados quatro estímulos âncoras sintetizados de R e quatro de B para cada grau de desvio vocal, variando de zero a três. Os participantes ouviram os estímulos âncoras e um estímulo de voz natural, e foram orientados a parear a voz natural com o estímulo âncora sintetizado que mais se assemelhava à voz apresentada. Durante toda a atividade, os participantes tiveram acesso à definição escrita dos parâmetros perceptivo-auditivos que estavam avaliando. Esta etapa constou de treinamento com 20 vozes naturais neutras e disfônicas de graus de desvio leve a intenso.

3) Atividade pós-treinamento: as 20 vozes utilizadas na atividade pré-treinamento foram aleatorizadas e, os indivíduos julgaram novamente as mesmas vozes, sem conhecimento prévio de que estas eram repetidas, mantendo o mesmo protocolo de avaliação da Atividade pré-treinamento.

Os participantes podiam escutar as vozes naturais ou os estímulos âncoras sintetizados quantas vezes desejassem, em todas as atividades do treinamento. As três atividades foram organizadas com o mesmo grupo de vozes, em todas as quatro sessões de treinamento, se diferenciando apenas na ordem em que as vozes eram apresentadas. Os participantes não tiveram acesso a esta informação durante o desenvolvimento da pesquisa.

Para a realização das atividades do Programa se utilizou fone de ouvido supra-auricular modelo *Multilaser Vibe Headphone* estéreo e, as sessões foram realizadas individualmente em ambiente silencioso, com nível de ruído inferior a 50 dBNPS, aferido por um medidor de nível de pressão sonora da marca RadioShack®.

**3. Seleção dos estímulos vocais das atividades pré e pós-treinamento.**

Para constituir a amostra de vozes naturais da avaliação pré e pós-treinamento, utilizou-se o banco de vozes do Ambulatório de Voz do Hospital das Clínicas da UFMG (AV/HC-UFMG), formado por 381 vozes da emissão da vogal /a/ sustentada de forma habitual, de indivíduos de ambos os gêneros com idade superior a 18 anos.

Quatro fonoaudiólogas, especialistas em voz, com cinco a 10 anos de experiência em avaliação perceptivo-auditiva, analisaram individualmente as vozes, utilizando o fone de ouvido supra-auricular modelo *Multilaser Vibe Headphone* estéreo. Todas as vozes foram classificas conforme o parâmetro predominante (R ou B) e o grau geral de desvio vocal (0 – neutro, 1 – leve, 2 – moderado, 3 – intenso).

Foram selecionadas 20 vozes naturais, sendo 13 de sujeitos do sexo feminino e sete do sexo masculino, quatro neutras e 16 disfônicas que apresentaram concordância na análise das quatro avaliadoras.

**4. Seleção dos estímulos vocais da atividade de treinamento.**

Para a construção das vozes sintetizadas neutras, R, ou B, com diferentes graus de desvio vocal, utilizou-se como fonte (fluxo glótico) um modelo paramétrico que permite o controle da frequência fundamental, do *jitter*, do *shimmer* e, da relação sinal ruído. Como filtro, utilizou-se um trato vocal que modela a vogal /a/, extraído de voz natural por técnica de predição linear. Os estímulos foram construídos por um engenheiro, totalizando 300 vozes sintetizadas (Vieira et al., 2014).

Para a análise do grau de naturalidade das vozes sintetizadas, foram selecionados três fonoaudiólogos com mais de cinco anos de experiência em avaliação perceptivo-auditiva, que realizaram individualmente a análise da naturalidade das vozes (relacionado ao quanto o ouvinte percebe a voz como humana) por meio de uma escala visual analógica (EVA). Posteriormente se realizou a classificação das vozes em neutras (ausência de desvio vocal), rugosas ou soprosas, e da mensuração, também por meio de uma EVA, do grau de desvio vocal de cada uma das emissões. Os valores encontrados para o

desvio vocal das vozes classificadas por meio da EVA, foram convertidos segundo valores sugeridos pela literatura (Baravieira et al., 2016): 1) para a rugosidade: grau 0 até 8,5 mm; grau 1 de 8,5 a 28,5 mm; grau 2 de 28,5 a 59,5 mm; e grau 3 a partir de 59,5 mm; e 2) para a soprosidade: grau 0 até 8,5 mm; grau 1 de 8,5 a 33,5 mm; grau 2 de 33,5 a 52,5 mm; e grau 3 a partir de 52,5 mm. Foram selecionadas como âncoras as vozes sintetizadas de diferentes graus de desvio, para cada parâmetro, classificadas com o maior grau de naturalidade por pelo menos dois avaliadores (Santos et al., 2018), totalizando oito vozes, sendo quatro vozes sintetizadas para o parâmetro R, e quatro para o parâmetro B.

Para a elaboração do grupo de vozes naturais desta atividade, foram selecionadas outras 20 vozes do banco do AV/HC-UFMG, utilizando o mesmo protocolo de avalição das Atividades pré e pós-treinamento. Este grupo de vozes foi composto por três vozes neutras e 17 disfônicas, com diferentes graus de desvio, sendo 14 vozes femininas e seis masculinas.

## 5. Análise Estatística

Para análise estatística primeiramente foi realizada uma análise descritiva dos dados, e depois foi realizado o teste $AC_2$ de Gwet para avaliação da concordância intra-avaliadores. O teste $AC_2$ foi realizado no software R (versão 3.5.1). Para comparação entre as respostas das concordâncias intra-avaliadores, nas quatro sessões de treinamento, inicialmente foi utilizado o Teste paramétrico ANOVA para medidas repetidas no programa Minitab® 19. Como as hipóteses do modelo não foram satisfeitas então usamos o teste de Friedman. Consideramos o nível de significância de 5%.

### 5.1. Análise descritiva

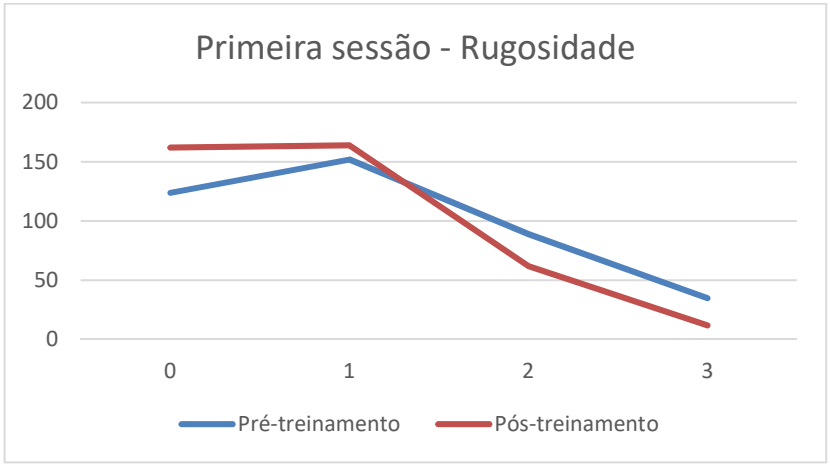
As respostas dos 20 avaliadores na análise das 20 vozes humanas no momento pré-treinamento totalizaram 400 respostas. Esta mesma análise e número de respostas foram obtidos no momento pós-treinamento, para cada sessão do Programa de Treinamento Auditivo.

Os gráficos 1 a 4 representam as respostas das avaliações vocais dos dois momentos do treinamento, para cada uma das quatro sessões do Programa, na avaliação do parâmetro perceptivo-auditivo de R.

Observa-se uma predominância de respostas para os graus neutro/normal (zero) e com desvio vocal leve (um). Tais resultados eram esperados porque representam o banco de vozes avaliadas, que foi criado a partir da caracterização da clínica vocal, que possui uma maior predominância de desvios leves, e com menor frequência de desvios vocais intensos (três).

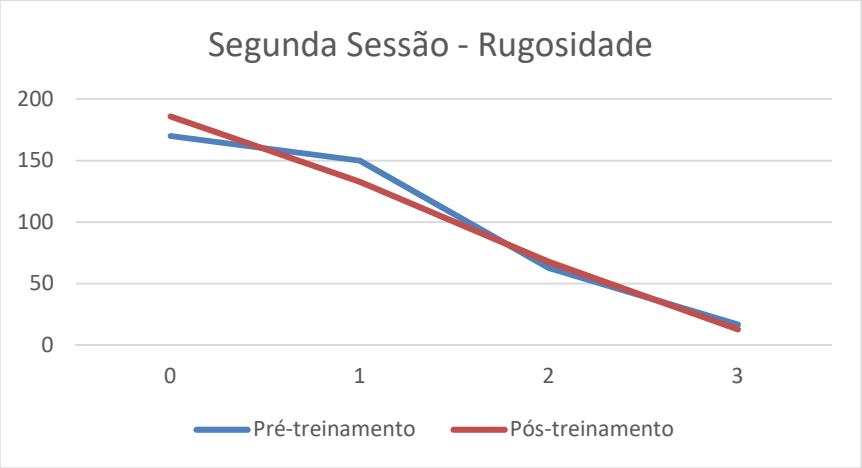
Após o treinamento auditivo da primeira sessão (Gráfico 1), os avaliadores passaram a avaliar as vozes com um desvio menor. Observa-se um aumento de respostas para os graus zero (normal) e um (leve), e diminuição de respostas para os graus dois (moderado) e três (intenso).

Gráfico 1: Frequência de respostas na análise da rugosidade na primeira sessão.



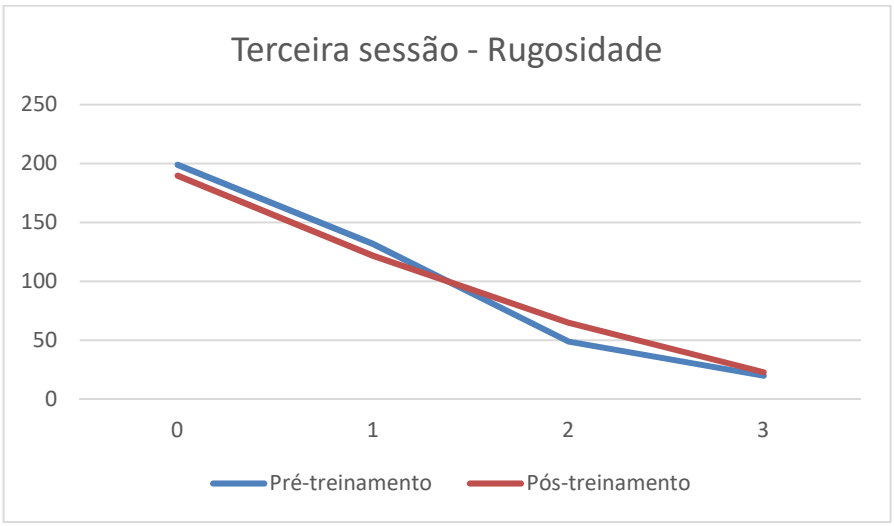

Na segunda sessão, os juízes modificaram os julgamentos das vozes entre os graus neutro e leve, após o treinamento auditivo (Gráfico 2).

Gráfico 2: Frequência de respostas na análise da rugosidade na segunda sessão.



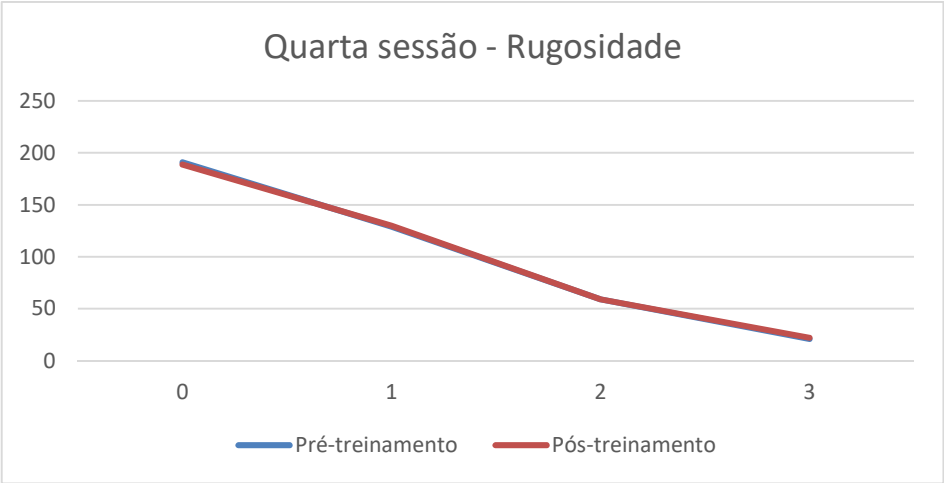

No julgamento do parâmetro R, os juízes mantiveram uma leve tendência de similaridade de respostas, antes e após o treinamento auditivo, na terceira sessão do Programa de Treinamento auditivo (Gráfico 3).

Gráfico 3: Frequência de respostas na análise da rugosidade na terceira sessão.



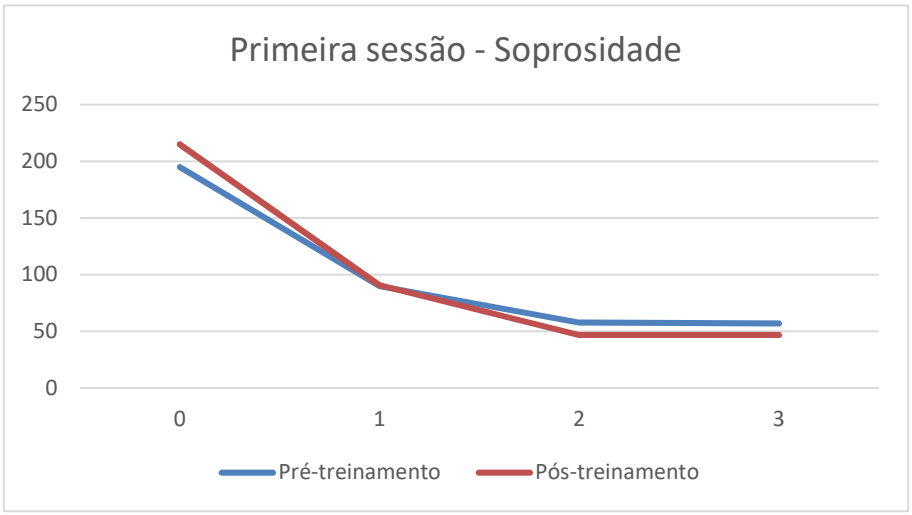

A quarta sessão do treinamento auditivo (Gráfico 4) evidencia que os juízes apresentaram respostas coincidentes antes e após o treinamento, sugerindo que os avaliadores passaram a ser menos influenciados pelo treinamento.

Gráfico 4: Frequência de respostas na análise da rugosidade na quarta sessão.



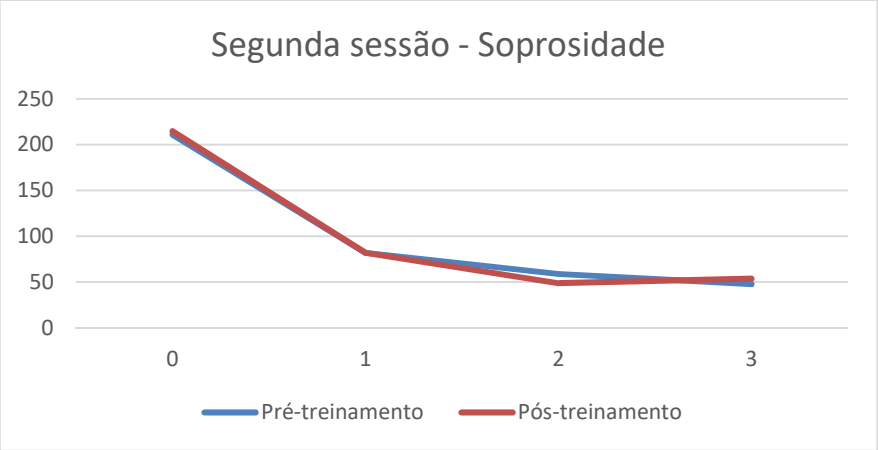

As respostas dos juízes na avaliação do parâmetro B são apresentadas nos gráficos 5 a 8. No gráfico 5 se observa uma leve tendência de os avaliadores analisarem o parâmetro B das vozes com menos desvio após o treinamento auditivo, de forma similar ao que ocorreu na análise do parâmetro R, mas com menor variabilidade nas respostas.

Gráfico 5: Frequência de respostas na análise da soprosidade da primeira sessão.




Na segunda sessão, a análise do parâmetro B já evidencia uma certa tendência de respostas similares, antes e após o treinamento, principalmente para os graus de desvio zero e um (Gráfico 6).

Gráfico 6: Frequência de respostas na análise da soprosidade da segunda sessão.




A análise do parâmetro B evidencia respostas coincidentes dos avaliadores, antes e após o treinamento, na terceira (Gráfico 7) e quarta (Gráfico 8) sessões.

Gráfico 7: Frequência de respostas na análise da soprosidade da terceira sessão.

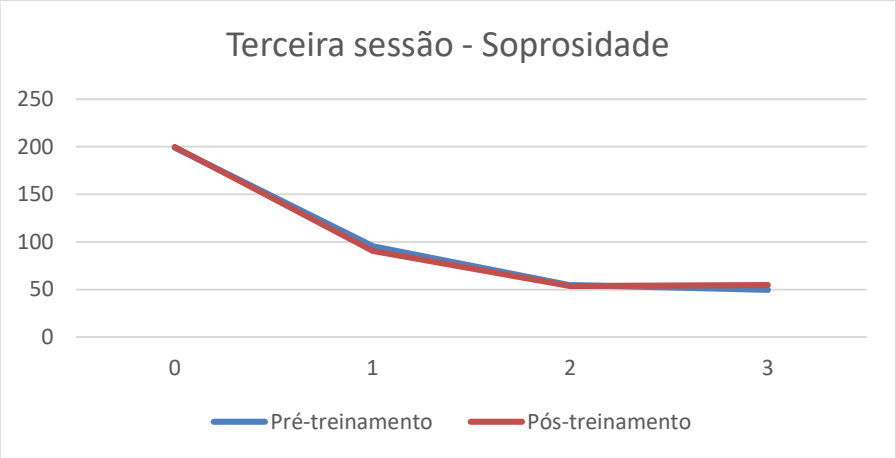
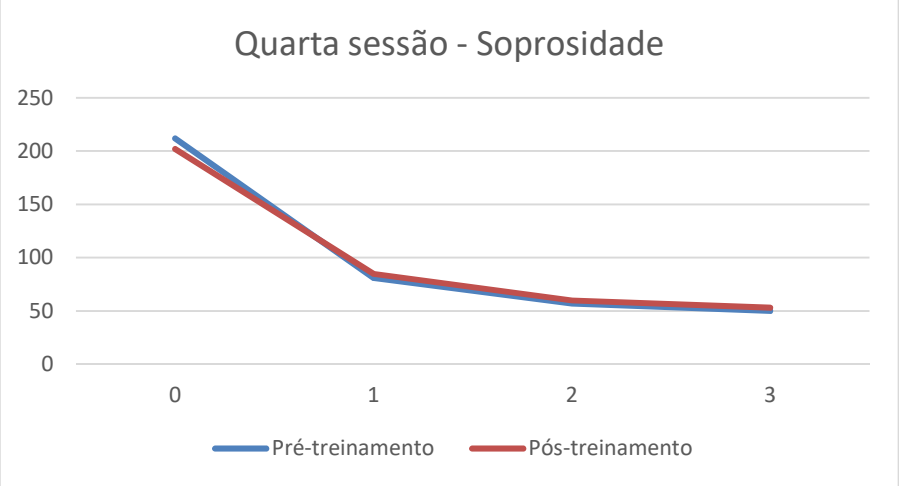



Gráfico 8: Frequência de respostas na análise da soprosidade da quarta sessão.



A análise dos Gráficos 1 e 5 sugere que a primeira sessão do treinamento auditivo torna a avaliação dos juízes menos “patologizante”, ou seja, eles tendem a valorizar menos a presença de desvios na qualidade vocal, ou por considerarem estes desvios inerentes às características normais da voz, ou por considerarem os desvios com menor grau de alteração. Este aspecto esteve presente na avaliação de ambos os parâmetros perceptivo-auditivos, de forma mais evidente na análise do parâmetro R.

Na segunda sessão (Gráficos 2 e 6) os juízes passam a apresentar, após o treinamento auditivo, trocas de julgamentos entre pares próximos de desvio vocal. No caso do parâmetro R, entre os graus zero e um, e para o parâmetro B, entre os graus 2 e 3. Tais resultados poderiam ser considerados como “concordâncias parciais”, e demostram que os avaliadores estão refinando suas categorizações nas análises das vozes.

A partir da terceira sessão, se observa uma menor interferência do treinamento auditivo na análise das vozes, de forma mais pronunciada para o parâmetro B. A quarta sessão já evidencia, para a avaliação dos parâmetros R e B, similaridades dos resultados de análises antes e após o treinamento auditivo.

### 5.2. Teste $AC_2$ de Gwet para avaliação da concordância dos avaliadores

A concordância intra-avaliador foi medida comparando-se as respostas das análises das 20 vozes antes e após o treinamento auditivo de cada um dos 20 juízes da pesquisa. Para esta análise se utilizou o teste $AC_2$ proposto por Gwet e seu *script* (Anexo 2) foi rodado no programa R.

Na Tabela 4 são apresentados os resultados das concordâncias intra-avaliadores, nas quatro sessões do Programa de Treinamento Auditivo, para o parâmetro perceptivo-auditivo de R. Observa-se um aumento dos valores das médias da concordância entre as sessões do Programa.

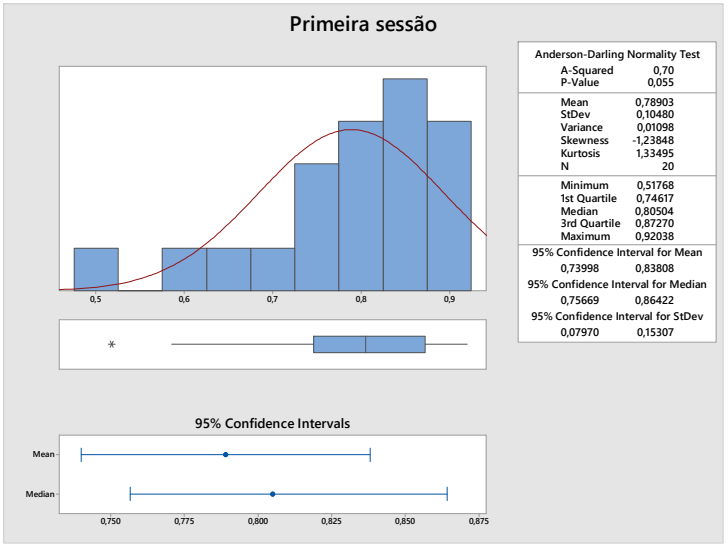
Tabela 4 - Concordância intra-avaliadores nas quatro sessões de Treinamento auditivo, para o parâmetro rugosidade, por meio do coeficiente $AC_2$.

| Participante | Sessão do Treinamento | | | |
|---|---|---|---|---|
| | 1ª | 2ª | 3ª | 4ª |
| **1** | 0,79 | 0,86 | 0,86 | 0,87 |
| **2** | 0,81 | 0,85 | 0,91 | 0,92 |
| **3** | 0,83 | 0,91 | 0,71 | 0,73 |
| **4** | 0,87 | 0,81 | 0,78 | 0,55 |
| **5** | 0,52 | 0,76 | 0,72 | 0,63 |
| **6** | 0,84 | 0,97 | 0,99 | 0,98 |
| **7** | 0,89 | 0,91 | 0,96 | 0,92 |
| **8** | 0,92 | 0,92 | 0,83 | 0,92 |
| **9** | 0,88 | 0,89 | 0,96 | 0,92 |
| **10** | 0,80 | 0,92 | 0,96 | 0,92 |
| **11** | 0,84 | 0,76 | 0,86 | 0,90 |
| **12** | 0,72 | 0,92 | 0,67 | 0,92 |
| **13** | 0,75 | 0,84 | 0,83 | 0,69 |
| **14** | 0,66 | 0,65 | 0,98 | 0,92 |
| **15** | 0,74 | 0,85 | 0,85 | 0,93 |
| **16** | 0,80 | 0,77 | 0,84 | 0,84 |
| **17** | 0,87 | 0,78 | 0,77 | 0,91 |
| **18** | 0,89 | 0,94 | 0,94 | 0,97 |
| **19** | 0,59 | 0,90 | 0,84 | 0,92 |
| **20** | 0,76 | 0,90 | 0,86 | 0,91 |
| **Média** | 0,79 | 0,85 | 0,85 | 0,86 |

Os Gráficos 9 a 12 apresentam as análises descritivas dos valores de concordância intra-avaliadores nas quatro sessões de treinamento, para o parâmetro R. Na primeira sessão (Gráfico 9), os valores de concordância variaram de 0,5176 a 0,9203, com desvio-padrão de 0,1048 e média de 0,7890.

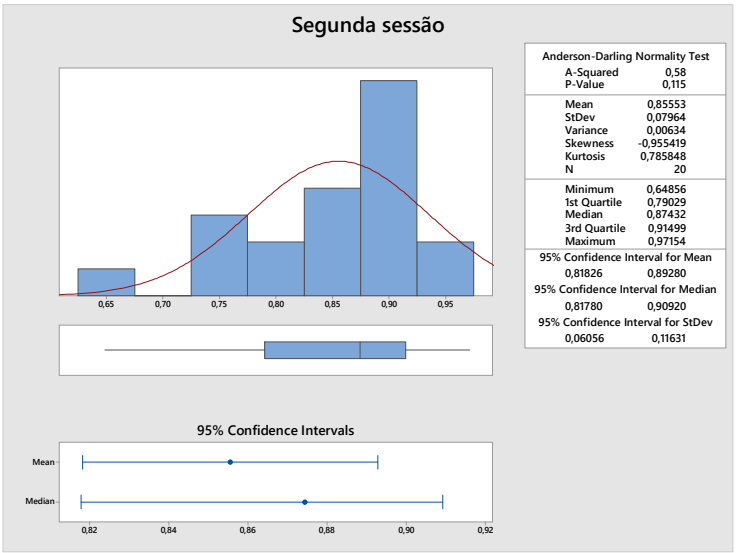
Observa-se também a presença de *outlier* (0,5176), e distribuição normal dos dados calculado pelo teste de Anderson-Darling (p=0,055).

Gráfico 9: Análise descritiva da concordância intra-avaliador para o parâmetro rugosidade.




No Gráfico 10 é apresentada a análise descritiva dos valores da segunda sessão do Programa. Os resultados variaram de 0,6485 a 0,9715, com desvio-padrão de 0,0796 e média de 0,8555. Observa-se distribuição normal dos dados calculado pelo teste de Anderson-Darling (p=0,115).

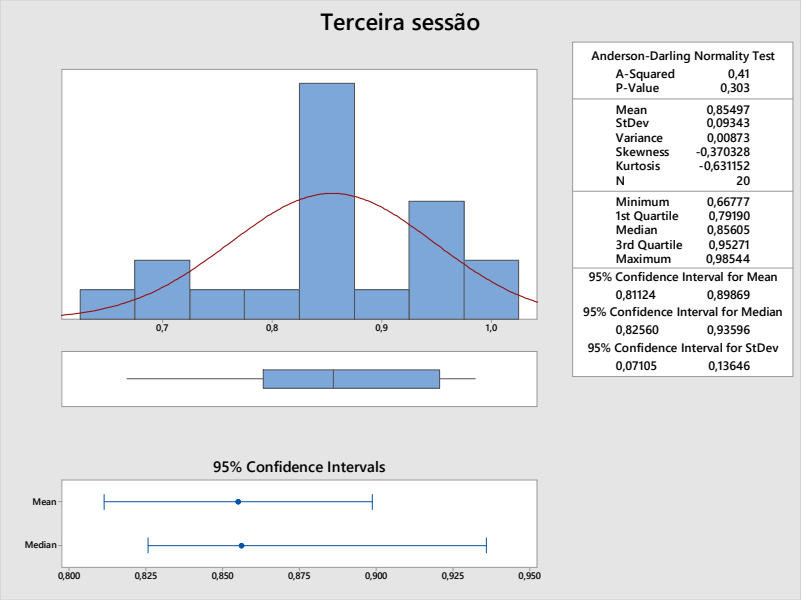
Gráfico 10: Análise descritiva da concordância intra-avaliador para o parâmetro rugosidade.




Na terceira sessão do Programa (Gráfico 11), os valores variaram de 0,6677 a 0,9854, com desvio-padrão de 0,0934 e média de 0,8550. Os

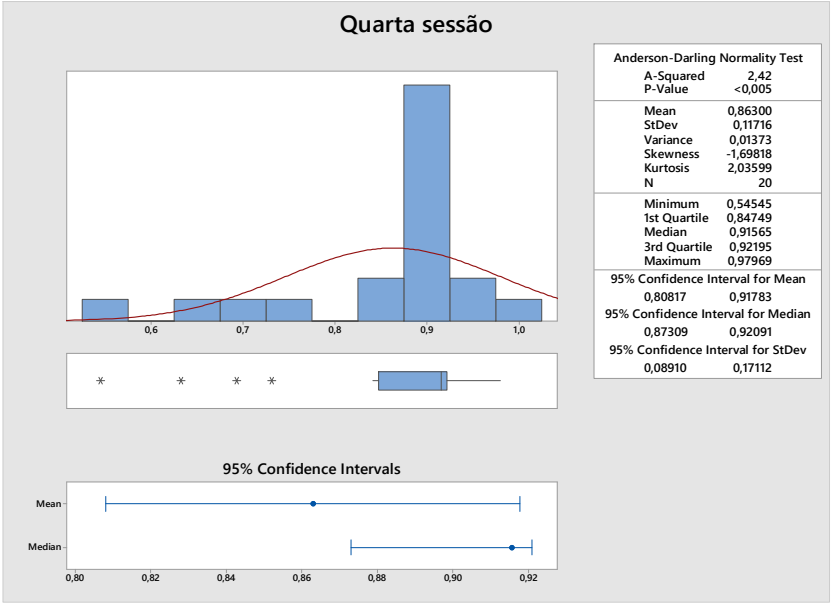
resultados apresentam distribuição normal calculado pelo teste de Anderson-Darling (p=0,303).

Gráfico 11: Análise descritiva da concordância intra-avaliador para o parâmetro rugosidade.




No Gráfico 12 é apresentada a análise descritiva dos valores da quarta sessão. Os resultados variaram de 0,5454 a 0,9796, com desvio-padrão de 0,1172 e média de 0,863. Observa-se distribuição assimétrica à esquerda calculada pelo teste de Anderson-Darling (p=0,005), e presença de *outliers*.

Gráfico 12: Análise descritiva da concordância intra-avaliador para o parâmetro rugosidade.



Na Tabela 5 são apresentados os resultados das concordâncias intra-avaliadores, nas quatro sessões do Programa de Treinamento Auditivo, para o parâmetro perceptivo-auditivo de B. Observa-se uma certa constância dos valores médios de concordância entre as sessões do Programa.

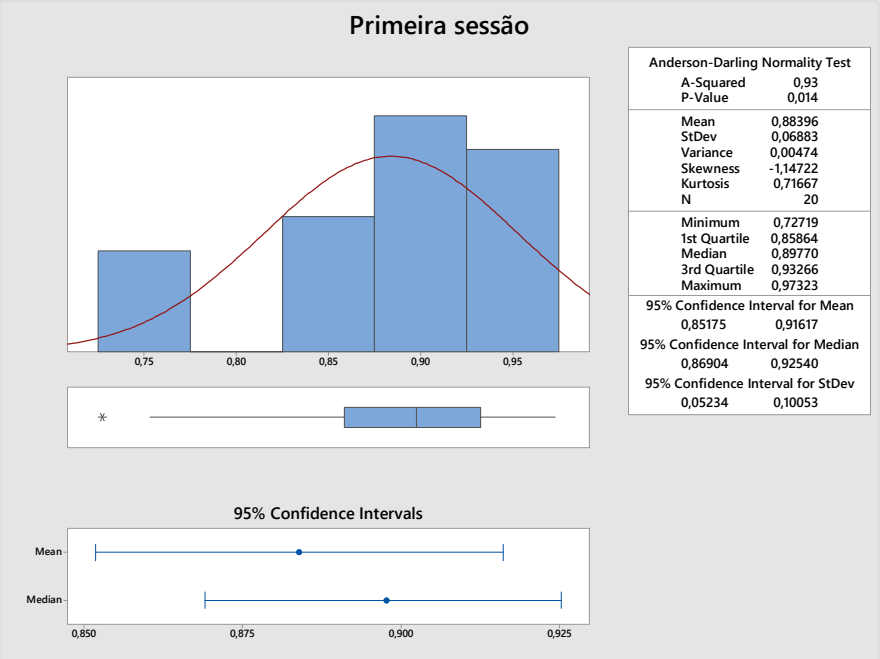
Tabela 5 - Concordância intra-avaliadores nas quatro sessões de Treinamento auditivo, para o parâmetro soprosidade, por meio do coeficiente $AC_2$.

| Participante | Sessão do Treinamento | | | |
|---|---|---|---|---|
| | 1ª | 2ª | 3ª | 4ª |
| **1** | 0,86 | 0,85 | 0,76 | 0,94 |
| **2** | 0,90 | 0,88 | 0,88 | 0,92 |
| **3** | 0,93 | 0,94 | 0,95 | 0,98 |
| **4** | 0,97 | 0,96 | 0,87 | 0,95 |
| **5** | 0,94 | 0,98 | 0,95 | 0,94 |
| **6** | 0,91 | 0,95 | 0,90 | 0,96 |
| **7** | 0,97 | 0,92 | 0,94 | 0,95 |
| **8** | 0,93 | 0,94 | 0,83 | 0,64 |
| **9** | 0,90 | 0,88 | 0,95 | 0,92 |
| **10** | 0,87 | 0,85 | 0,91 | 0,98 |
| **11** | 0,91 | 0,88 | 0,79 | 0,93 |
| **12** | 0,87 | 0,89 | 0,96 | 0,89 |
| **13** | 0,85 | 0,79 | 0,93 | 0,88 |
| **14** | 0,94 | 0,95 | 0,97 | 0,99 |
| **15** | 0,75 | 0,97 | 0,96 | 1,00 |
| **16** | 0,73 | 0,75 | 0,90 | 0,82 |
| **17** | 0,89 | 0,90 | 0,89 | 0,89 |
| **18** | 0,92 | 0,89 | 0,93 | 0,99 |
| **19** | 0,75 | 0,97 | 0,96 | 0,96 |
| **20** | 0,89 | 0,94 | 0,77 | 0,91 |
| **Média** | 0,88 | 0,90 | 0,90 | 0,92 |

Os Gráficos 13 a 16 apresentam as análises descritivas dos valores de concordância intra-avaliadores nas quatro sessões de treinamento, para o parâmetro B.

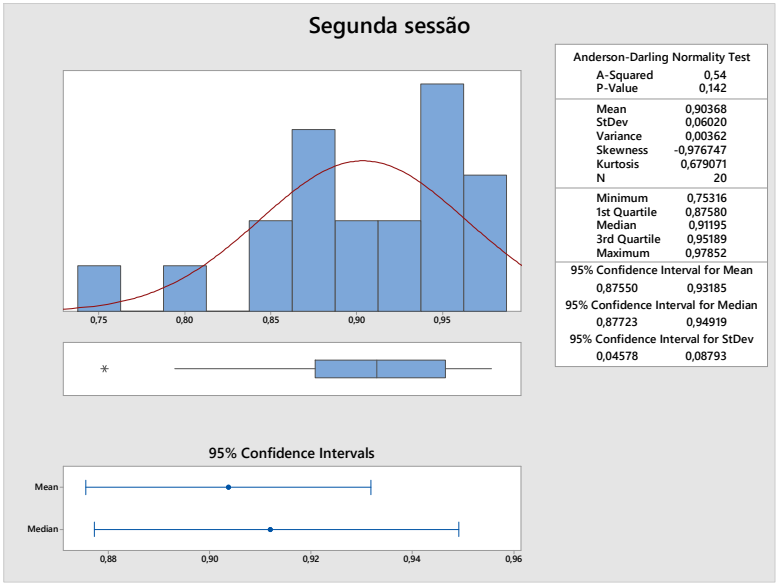
Na avaliação do parâmetro perceptivo-auditivo de B, na primeira sessão (Gráfico 13), o valor mínimo foi de 0,7272, máximo de 0,9732, desvio-padrão de 0,0688 e média de 0,8840. Os dados possuem distribuição assimétrica à esquerda calculada pelo teste de Anderson-Darling (p=0,014), e presença de *outlier* (0,7272).

Gráfico 13: Análise descritiva da concordância intra-avaliador para o parâmetro soprosidade.



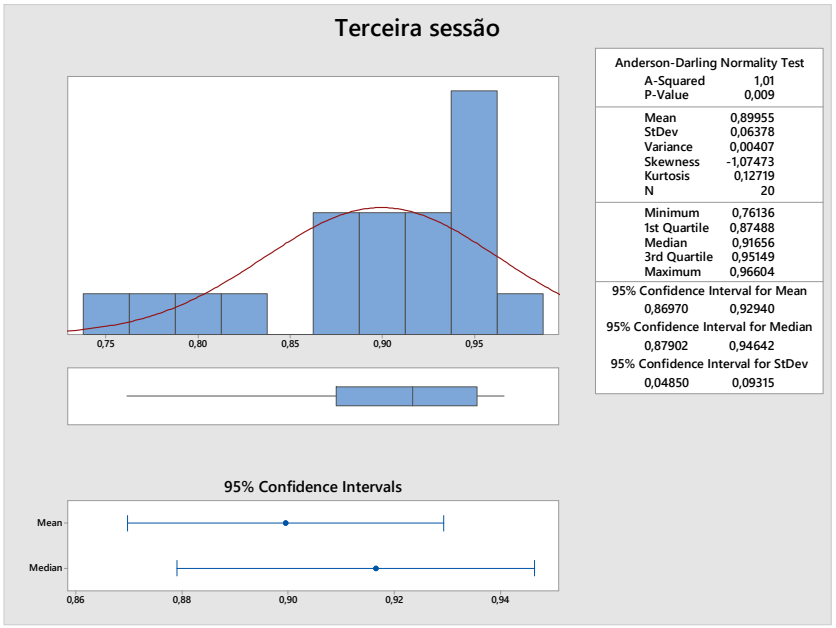

No Gráfico 14 é apresentada a análise descritiva dos valores da segunda sessão do Programa. Os resultados variaram de 0,7536 a 0,9785, com desvio-padrão de 0,0602 e média de 0,9037. O teste de Anderson-Darling evidencia que os dados apresentam distribuição normal (p=0,142).

Gráfico 14: Análise descritiva da concordância intra-avaliador para o parâmetro soprosidade.




Na terceira sessão do Programa (Gráfico 15), os valores variaram de 0,7614 a 0,9660, com desvio-padrão de 0,0638 e média de 0,8995. Os resultados apresentam distribuição assimétrica à esquerda calculada pelo teste de Anderson-Darling (p=0,009).

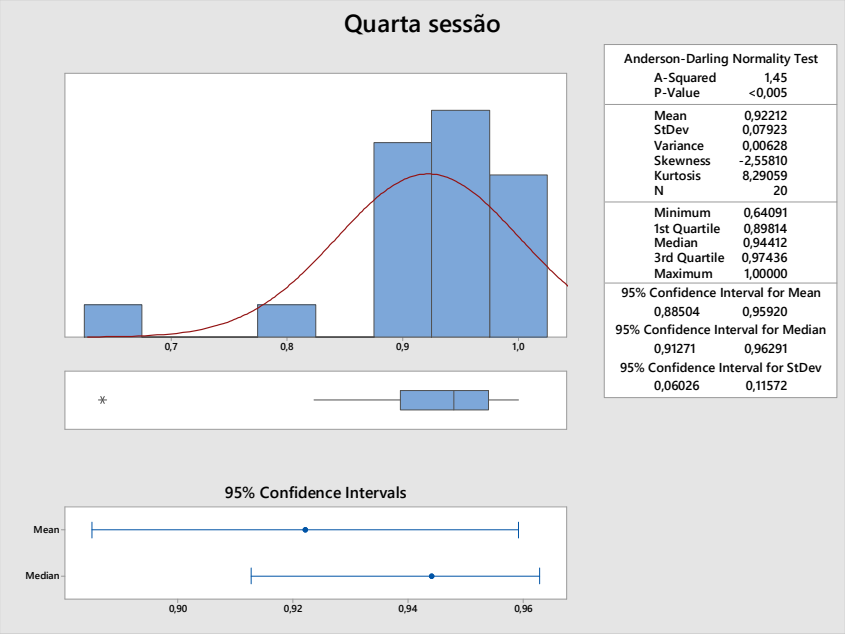
Gráfico 15: Análise descritiva da concordância intra-avaliador para o parâmetro soprosidade.




No Gráfico 16 é apresentada a análise descritiva dos valores da quarta sessão do Programa. Os resultados variaram de 0,6409 a 1,000, com desvio-padrão de 0,0792 e média de 0,9221. Observa-se distribuição assimétrica à

esquerda calculada pelo teste de Anderson-Darling (p=0,005), e presença de *outlier*.

Gráfico 16: Análise descritiva da concordância intra-avaliador para o parâmetro soprosidade.



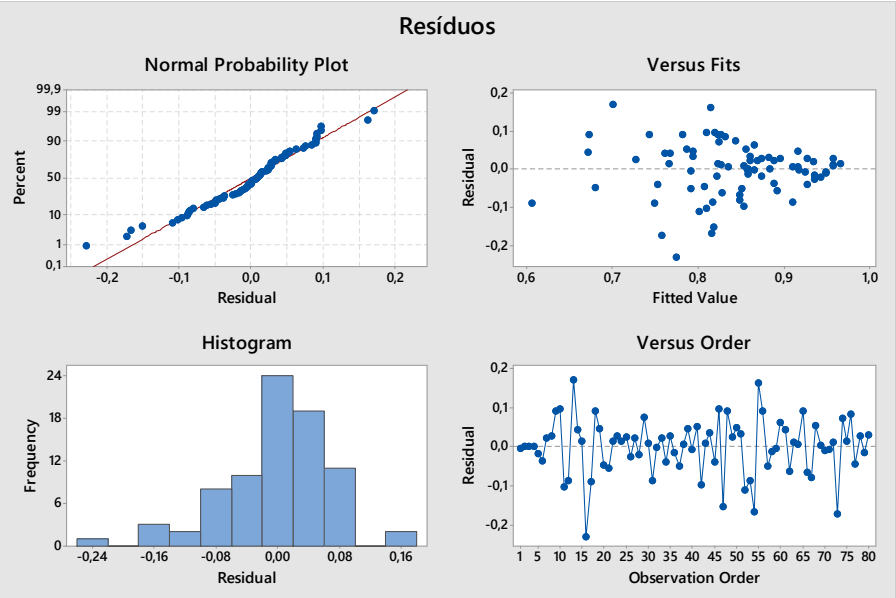

## 5.3. Teste para Verificação da Igualdade das sessões de avaliações

### 5.3.1. Anova com medidas repetidas

Para a comparação entre as quatro sessões do Programa de Treinamento auditivo, para os parâmetros R e B, foi realizado o teste ANOVA para medidas repetidas conforme explicado na seção Considerações Iniciais (item 6.1) desse trabalho.

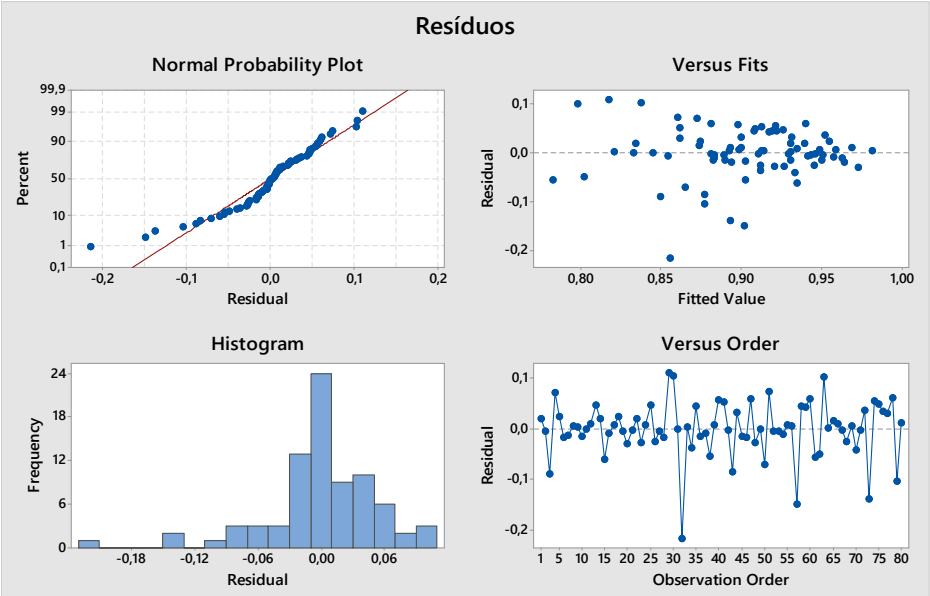
No Gráfico 17 observa-se as representações de distribuição assimétrica dos resíduos (Anderson-Darling, p=0,015) e heterocedasticidade uma vez que a distribuição dos resíduos *versus* os valores esperados não são aleatoriamente distribuídos, para a análise dos dados referentes à variável R.

Gráfico 17: Representações gráficas do resíduo e da variância.




A análise da ANOVA com medidas repetidas para os dados da variável B indicou uma distribuição assimétrica dos resíduos (Anderson-Darling, p=0,005) e heterocedasticidade uma vez que a distribuição dos resíduos *versus* os valores esperados não são aleatoriamente distribuídos (Gráfico 18).

Gráfico 18: Representações gráficas do resíduo e da variância.



Com as hipóteses do modelo de ANOVA com medidas repetidas violadas e não obtendo transformação da variável resposta adequada para viabilizar o uso do modelo paramétrico, optamos para realizar o teste não paramétrico de Friedman.

### 5.3.2. Teste Friedman

Nesta seção usamos as equações definidas na seção Considerações Iniciais (item 6.2) e o uso do *software* Minitab® 19.

Para a análise dos dados referentes à variável R, obtivemos $R_1 = 35$, $R_2 = 50$, $R_3 = 53$ e $R_4 = 62$ o que resultou no valor S=11,34 que é maior do que $\chi^2_{k-1;\,1-\alpha=}\ \ \chi^2_{4-1;\,1-0{,}05=7{,}81}$. Assim, rejeitamos ao nível de 5% de significância que as sessões possuem desempenhos iguais (p=0,01). Já nas comparações múltiplas observamos diferença significativa apenas entre a sessão 1 e 4 com probabilidade de significância 0,005. Em conjunto com o estudo descritivo podemos concluir que o treinamento melhorou o grau de concordância ao nível de significância 5%. As probabilidades relativas às comparações são descritas na Tabela 6.

Tabela 6: probabilidade de significância para comparações múltiplas para a rugosidade.

| Sessão | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1,0000 | 0,2559 | 0,1219 | 0,0052 |
| 2 | 0,2559 | 1,0000 | 0,9831 | 0,4559 |
| 3 | 0,1219 | 0,9831 | 1,0000 | 0,6881 |
| 4 | 0,0052 | 0,4559 | 0,6881 | 1,0000 |

Para a análise dos dados referentes à variável B, obtivemos $R_1 = 42$, $R_2 = 46{,}5$ , $R_3 = 47{,}5$ e $R_4 = 64$ o que resultou no valor S=8,40 que é maior do que $\chi^2_{k-1;\,1-\alpha=}\ \ \chi^2_{4-1;\,1-0{,}05=7{,}81}$. Assim, rejeitamos ao nível de 5% de significância que as sessões possuem desempenhos iguais (p=0,01). Já nas comparações múltiplas observamos diferença significativa apenas entre a sessão 1 e 4 com probabilidade de significância 0,036. Em conjunto com o estudo descritivo podemos concluir que o treinamento melhorou o grau de concordância ao nível de significância 5%. As probabilidades relativas às comparações são descritas na Tabela 7.

Tabela 7: probabilidade de significância para comparações múltiplas para a soprosidade.

| Sessão | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1,0000 | 0,9071 | 0,9463 | 0,0355 |
| 2 | 0,9071 | 1,0000 | 0,9993 | 0,1802 |
| 3 | 0,9463 | 0,9993 | 1,0000 | 0,1395 |
| 4 | 0,0355 | 0,1802 | 0,1395 | 1,0000 |

## CONCLUSÃO

O parâmetro perceptivo-auditivo de soprosidade apresentou indicador $AC_2$ maior do que para rugosidade, indicando ser mais concordante.

O treinamento auditivo com estímulos âncoras sintetizados melhora a concordância intra-avaliador no julgamento da rugosidade e da soprosidade a partir da quarta sessão.

## REFERÊNCIAS BIBLIOGRÁFICAS


1. Alipour F, Titze IR. Elastic models of vocal fold tissues J Acoust Soc Am. 1991; 90: 1326–1331.
2. Baravieira PB, Brasolotto AG, Montagnoli AN, Silvério KCA, Yamasaki R, Behlau M. Auditory-perceptual evaluation of rough and breathy voices: correspondence between analogical visual and numerical scale. CoDAS. 2016;28(2):163-167.
3. Behlau M, Azevedo R, Pontes P. Conceito de voz normal e classificação das disfonias. In: Behlau M. Voz: o livro do especialista. vol 1. Rio de Janeiro: Revinter; 2001a. p.53-79.
4. Behlau M, Madazio G, Feijo D. Avaliação da voz. Rio de Janeiro: Revinter; 2001b. Voz: o livro do especialista; p.84-246.
5. Behlau M, Oliveira G, dos Santos LMA, Ricarte A. Validação no Brasil de protocolos de auto-avaliação do impacto de uma disfonia. Pró-Fono R Atual Cient. 2009; 21(4): 326-332.
6. Behlau M, Zambon F, Guerrieri AC, Roy N. Epidemiology of voice disorders in teachers and nonteachers in Brazil: Prevalence and adverse effects. J Voice. 2012, 26(5) 665.e9-18.
7. Bridger MW, Epstein R. Functional voice disorders. A review of 109 patients. J Laryngol Otol. 1983; 97(12):1145-1148.
8. Brinca L, Batista AP, Tavares AI, Pinto PN, Araújo L. The Effect of Anchors and Training on the Reliability of Voice Quality Ratings for Different Types of Speech Stimuli. J Voice. 2015;29(6): 776.e7-14.
9. Cerceau JSB, Alves CFT, Gama ACC. Análise acústica da voz de mulheres idosas. Rev CEFAC. 2009; 11(1):142-149.
10. Chan KM, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. J Speech Lang Hear Res. 2002; 45(1):111-26.
11. Chan KMK, Yiu EML. A comparison of two perceptual voice evaluation training programs for naive listeners. J Voice. 2006; 20:229–241.
12. Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier--Buchman L, Friedrich G, Heyning VDEP, Remacle M, Woisard V. A Basic protocol for functional assessment of voice pathology, especially for investigating the

efficacy of (phonosurgical) treatments and evaluating new assessment techniques: guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol. 2001;258(2):77-82.

13. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990; 43(6):543-549.
14. Felippe ACN, Grillo MHMM, Grechi TH. Normatização de medidas acústicas para vozes normais. Rev Bras Otorrinolaringol. 2006;72(5):659-64.
15. Freitas SV, Pestana PM, Almeida V, Ferreira A. Audio-Perceptual Evaluation of Portuguese Voice Disorders—An Inter- and Intrajudge Reliability Study. J Voice. 2014; 28(2): 210-215.
16. Gama ACC, Mesquita GM, Reis C, Bassi IB. Análises perceptivo-auditiva e acústica da voz nos momentos pré e pós fonoterapia de pacientes com falsete mutacional. Rev Soc Bras Fonoaudiol. 2012;17(2):225-9.
17. Genilhú PFL, Gama ACC. Medidas acústicas e aerodinâmicas em cantores: comparação entre homens e mulheres. CoDAS. 2018;30(5): e20170240.
18. Ghio A, Dufour S, Wengler A, Pouchoulin G, Revis J, Giovanni A. Perceptual Evaluation of Dysphonic Voices: Can a Training Protocol Lead to the Development of Perceptual Categories? J Voice. 2015;29 (3): 304-311.
19. Gill TM, Feinstein AR. A critical appraisal of the quality of Quality-of life Measurements. JAMA. 1994;272(2):619-26.
20. Gurlekian JA, Torres HM, Vaccari ME. Comparison of Two Perceptual Methods for the Evaluation of Vowel Perturbation Produced by Jitter. J Voice. 2016;30(4): 506.e1-8.
21. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Brit J Math Stat Psy. 2008; 61: 29–48
22. Gwet KL. Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters. Statistical Methods for Inter-Rater Reliability Assessment. 2002; 1: 1-5.

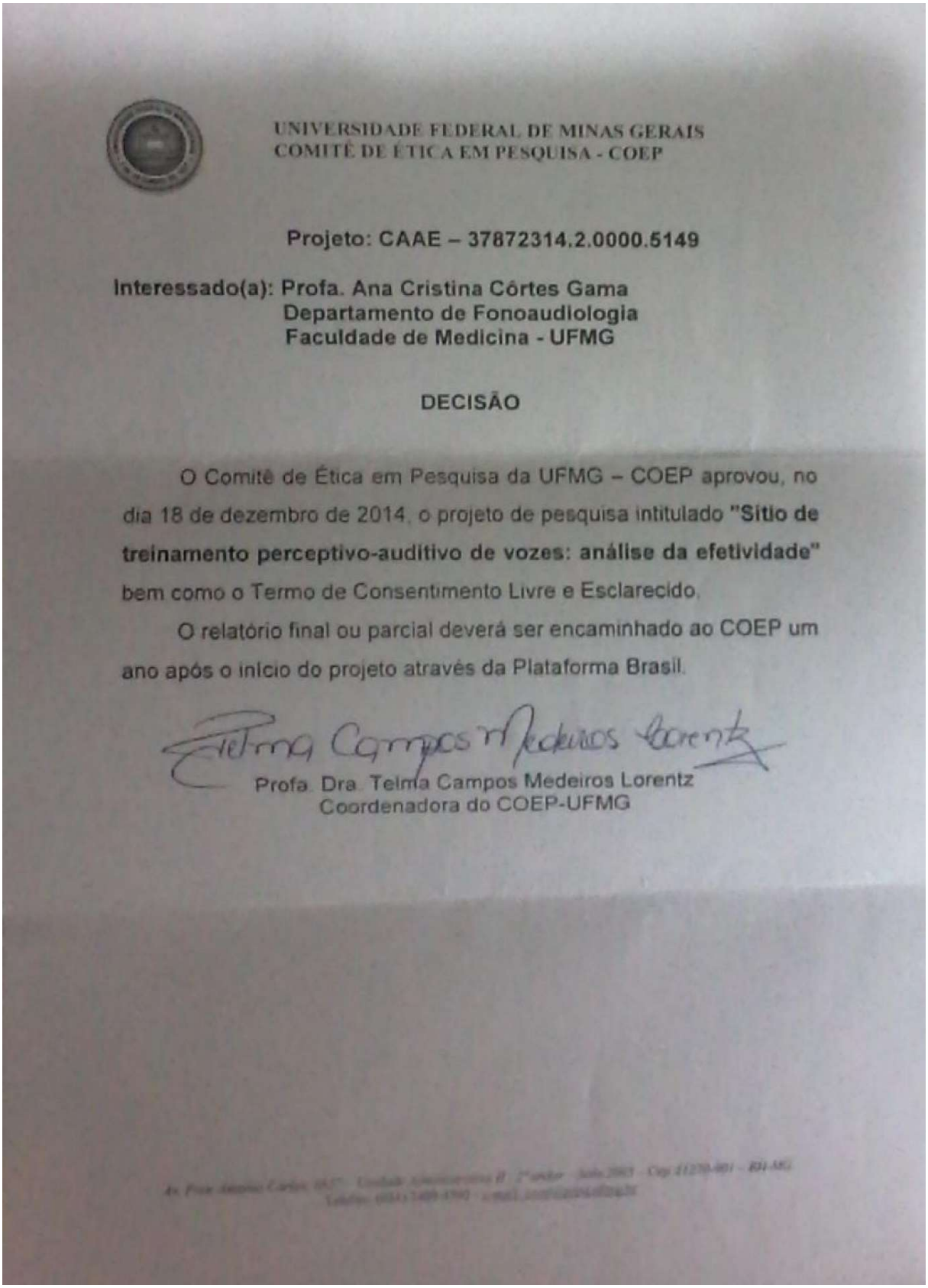
23. Gwet, KL. Handbook of Inter-Rater Reliability: The definitive guide to measuring the extent of agreement among raters. 2014: Advanced Analytics LLC, PO box 2696 Gaithersburg, MD. 20886-2696.
24. Hunter EJ, Titze IR. Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus non-occupational settings. J Speech Lang Hear Res. 2010; 53 (4):862-75.
25. Korn GP, Martins JR, Park SW, Mendes A, Kobayashi EY, Nader HB. Concentration of hyaluronic acid in human vocal folds in young and old subjects. Otolaryngol Head Neck Surg. 2011; 145:981-6.
26. Kottner J, Audige L, Brorson S, Donner A, Gajewski JB, Hro´bjartsson, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106.
27. Lohscheller J. Phonovibrography: Mapping High-Speed Movies of Vocal Fold Vibration into 2-D Diagrams for Visualizing and Analyzing the Underlying Laryngeal Dynamics. IEEE Transaction on Med. Imaging. 2008; 2(3):10-11.
28. Medeiros AM, Barreto SM, Assunção AA. Voice Disorders (Dysphonia) in public school female teachers working in Belo Horizonte: Prevalence and associated factors. J Voice. 2008; 22(6) 676-87.
29. Miot HA. Análise de concordância em estudos clínicos e experimentais. J Vasc Bras. 2016; 15(2):89-92.
30. Miranda SVV, Mello RJV, Silva HJ. Correlação entre o envelhecimento e as dimensões das pregas vocais. Rev. CEFAC. 2011; 13:444-51.
31. Montgomery, D. C. Design and Analysis of Experiments. Fifth Edition. John Wiley & Sons, INC. New York. 2001.
32. Nichols BG, Varadarajan V, Bock JM, Blumin JH. Dysphonia in Nursing Home and Assisted Living Residents: Prevalence and Association with Frailty. J Voice. 2015, 29 (1): 79-82.
33. Oates J. Auditory-perceptual evaluation of disordered vocal quality—pros, cons and future directions. Folia Phoniatr Logop. 2009; 61:49–56.
34. Oliveira SB, Gama ACC, Chaves AR. Interference of background experience on agreement of perceptive-auditory analysis of neutral and dysphonic voices. Distúrbios Com. 2016; 28(3): 415-422.

35. Patel RR, Awan SN, Barkmeier-Kraemer J, Courey M, Deliyski D, Eadie T, Paul D, Švec JG, Hillmani R. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. Am J Speech Lang Pathol. 2018; 27:887–905.
36. Ramos LA, Souza BO, Gama ACC. Análise vocal na infância: uma revisão integrativa. Distúrb Comun. 2017; 29(1): 20-32.
37. Rocha KR, Júnior AJB. ANOVA medidas repetidas e seus pressupostos: análise passo a passo de um experimento. Perspectivas da Ciência e Tecnologia. 2018; 10: 29:51.
38. Roy N, Merrill RM, Thibeault S, Parsa RA, Gray SD, Smith EM. Prevalence of voice disorders in teachers and the general population. J Speech Lang Hear Res. 2004; 47:281–293.
39. Santos HTA. Deficiências da Estatística Kappa na concordância entre avaliadores e medidas alternativas [trabalho de conclusão de curso]. Brasília (DF): Universidade de Brasília. Departamento de Estatística; 2015.
40. Santos PCM, Vieira MN, Sansão JPH, Gama ACC. Effect of Auditory-Perceptual Training With Natural Voice Anchors on Vocal Quality Evaluation. J Voice. 2018; 33: 220-225.
41. Siegel, S., Castellan, N. J. Estatística Não-Paramétrica para Ciências do Comportamento. Artmed-Bookman. 2006.
42. Sofranko JL, Prosek RA. The Effect of Experience on Classification of Voice Quality. J Voice. 2012; 26(3): 299-303.
43. Van Houtte E, Van Lierde K, D'Haeseleer E, Claeys S. The Prevalence of Laryngeal Pathology in a Treatment-Seeking Population with Dysphonia. Laryngoscope. 2010; 120: 306-312.
44. Vargas CL, Berwig LC, Steidl EMS, Prade LS, Bolzan G, Keske-Soares M, Weinmann ARM. Prematuros: crescimento e sua relação com as habilidades orais. CoDAS. 2015, 27(4), 378-383.
45. Vieira MN, Sansão JPH, Yehia HC. Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms. Speech Commun. 2014;61–62:17–32.

46. Wittenberg T, Tigges M, Mergell P, Eysholdt U. Functional imaging of vocal fold vibration: digital multislice high-speed kymography. J Voice. 2000;14(3):422-442.
47. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013: 13-61.
48. Yamasaki R, Gama ACC. Desafios e referências na avaliação perceptivo-auditiva da voz. In: Fundamentos e atualidades em voz Clínica. Rio de Janeiro: Thieme Revinter; 2019: 9-30.

**ANEXO 1**

## APROVAÇÃO DO COMITÊ DE ÉTICA EM PESQUISA (COEP) DA UFMG



UNIVERSIDADE FEDERAL DE MINAS GERAIS
COMITÊ DE ÉTICA EM PESQUISA - COEP

**Projeto: CAAE – 37872314.2.0000.5149**

**Interessado(a): Profa. Ana Cristina Côrtes Gama**
**Departamento de Fonoaudiologia**
**Faculdade de Medicina - UFMG**

**DECISÃO**

O Comitê de Ética em Pesquisa da UFMG – COEP aprovou, no dia 18 de dezembro de 2014, o projeto de pesquisa intitulado **"Sitio de treinamento perceptivo-auditivo de vozes: análise da efetividade"** bem como o Termo de Consentimento Livre e Esclarecido.

O relatório final ou parcial deverá ser encaminhado ao COEP um ano após o início do projeto através da Plataforma Brasil.

Profa. Dra. Telma Campos Medeiros Lorentz
Coordenadora do COEP-UFMG

## ANEXO 2 – *Script* do teste $AC_2$ de Gwet para concordância intra-avaliador.

```
gwet.ac1.raw <- function(ratings,weights=" quadratic ",conflev=0.95,N=Inf,print=TRUE){

ratings.mat <- as.matrix(ratings)

if (is.character(ratings.mat)){

ratings.mat <- trim(toupper(ratings.mat))

ratings.mat[ratings.mat==''] <- NA_character_

}

n <- nrow(ratings.mat) # number of subjects

r <- ncol(ratings.mat) # number of raters

f <- n/N # finite population correction

# creating a vector containing all categories used by the raters

categ.init <- unique(na.omit(as.vector(ratings.mat)))

categ <- sort(categ.init)

q <- length(categ)

# creating the weights matrix

if (is.character(weights)){

if (weights=="quadratic")

weights.mat<-quadratic.weights(categ)

else if (weights=="ordinal")

weights.mat<-ordinal.weights(categ)

else if (weights=="linear")

weights.mat<-linear.weights(categ)

else if (weights=="radical")

weights.mat<-radical.weights(categ)

else if (weights=="ratio")

weights.mat<-ratio.weights(categ)

else if (weights=="circular")
```

```
weights.mat<-circular.weights(categ)

else if (weights=="bipolar")

weights.mat<-bipolar.weights(categ)

else weights.mat<-identity.weights(categ)

}else weights.mat= as.matrix(weights)

# creating the nxq agreement matrix representing the distribution of raters by subjects and category

agree.mat <- matrix(0,nrow=n,ncol=q)

for(k in 1:q){

categ.is.k <- (ratings.mat==categ[k])

agree.mat[,k] <- (replace(categ.is.k,is.na(categ.is.k),FALSE)) %*% rep(1,r)

}

agree.mat.w <- t(weights.mat%*%t(agree.mat))

# calculating gwet's ac1 coefficient

ri.vec <- agree.mat%*%rep(1,q)

sum.q <- (agree.mat*(agree.mat.w-1))%*%rep(1,q)

n2more <- sum(ri.vec>=2)

pa <- sum(sum.q[ri.vec>=2]/((ri.vec*(ri.vec-1))[ri.vec>=2]))/n2more

pi.vec <- t(t(rep(1/n,n))%*%(agree.mat/(ri.vec%*%t(rep(1,q)))))

pe <- sum(weights.mat) * sum(pi.vec*(1-pi.vec)) / (q*(q-1))

gwet.ac1 <- (pa-pe)/(1-pe)

# calculating variance, stderr & p-value of gwet's ac1 coefficient

den.ivec <- ri.vec*(ri.vec-1)

den.ivec <- den.ivec - (den.ivec==0) # this operation replaces each 0 value with -1 to make the next ratio
calculation always possible.

pa.ivec <- sum.q/den.ivec

pe.r2 <- pe*(ri.vec>=2)

ac1.ivec <- (n/n2more)*(pa.ivec-pe.r2)/(1-pe)

pe.ivec <- (sum(weights.mat)/(q*(q-1))) * (agree.mat%*%(1-pi.vec))/ri.vec

ac1.ivec.x <- ac1.ivec - 2*(1-gwet.ac1) * (pe.ivec-pe)/(1-pe)
```

```
var.ac1 <- ((1-f)/(n*(n-1))) * sum((ac1.ivec.x - gwet.ac1)^2)

stderr <- sqrt(var.ac1)# ac1's standard error

p.value <- 2*(1-pt(abs(gwet.ac1/stderr),n-1))

lcb <- gwet.ac1 - stderr*qt(1-(1-conflev)/2,n-1) # lower confidence bound

ucb <- min(1,gwet.ac1 + stderr*qt(1-(1-conflev)/2,n-1)) # upper confidence bound

if(print==TRUE) {

if (weights=="unweighted") {

cat("Gwet's AC1 Coefficient\n")

cat('======================\n')

cat('Percent agreement:',pa,'Percent chance agreement:',pe,'\n')

cat('AC1 coefficient:',gwet.ac1,'Standard error:',stderr,'\n')

cat(conflev*100,'% Confidence Interval: (',lcb,',',ucb,')\n')

cat('P-value: ',p.value,'\n')

}

else {

cat("Gwet's AC2 Coefficient\n")

cat('==========================\n')

cat('Percent agreement:',pa,'Percent chance agreement:',pe,'\n')

cat('AC2 coefficient:',gwet.ac1,'Standard error:',stderr,'\n')

cat(conflev*100,'% Confidence Interval: (',lcb,',',ucb,')\n')

cat('P-value: ',p.value,'\n')

cat('\n')

if (!is.numeric(weights)) {

cat('Weights: ', weights,'\n')

cat('---------------------\n')

}else{

cat('Weights: Custom Weights\n')

cat('----------------------\n')

}
```

```
print(weights.mat)
}
}
invisible(c(pa,pe,gwet.ac1,stderr,p.value))
}
setwd(choose.dir()) # escolha a pasta onde está seu banco de dados AC1 INTRA
dados = read.table("conc_intra.txt", header=T)
str(dados)
#install.packages("rel") #caso não tenha o pacote instalado, instale aqui retirando o #
require(rel)
### intra juíz A
gac(dados[,1:2],weight=" quadratic ", conf.level = 0.95)
### intra juíz B
gac(dados[,3:4],weight=" quadratic ", conf.level = 0.95)
### intra juíz C
gac(dados[,5:6],weight=" quadratic ", conf.level = 0.95)
### intra juíz D
gac(dados[,7:8],weight=" quadratic ", conf.level = 0.95)
### intra juíz E
gac(dados[,9:10],weight=" quadratic ", conf.level = 0.95)
### intra juíz F
gac(dados[,11:12],weight=" quadratic ", conf.level = 0.95)
### intra juíz G
gac(dados[,13:14],weight=" quadratic ", conf.level = 0.95)
### intra juíz H
gac(dados[,15:16],weight=" quadratic ", conf.level = 0.95)
### intra juíz I
gac(dados[,17:18],weight=" quadratic ", conf.level = 0.95)
```