Doi: 10.1590/2317-1782/20162015163

Artigo Original

Original Article

Influência do treinamento dos avaliadores no julgamento perceptivo da hipernasalidade

Influence of listener training on the perceptual assessment of hypernasality

Adriana Cristina de Almeida Santos Furlan de Oliveira[1]
Rafaeli Higa Scarmagnani[1]
Ana Paula Fukushiro[1,2]
Renata Paciello Yamashita[1]

Descritores

Fissura Palatina
Insuficiência Velofaríngea
Fala
Distúrbios da Fala
Percepção da Fala

RESUMO

Introdução: Alto índice de concordância no julgamento perceptivo da hipernasalidade entre diferentes avaliadores é difícil de ser alcançado. O treinamento prévio dos avaliadores e a padronização dos critérios de análise podem ser estratégias eficazes para minimizar o efeito da subjetividade do julgamento perceptivo e aumentar a concordância entre os avaliadores. **Objetivo:** Investigar a influência do treinamento prévio sobre a concordância entre diferentes avaliadores no julgamento perceptivo da hipernasalidade. **Métodos:** Três fonoaudiólogas experientes analisaram 77 amostras de fala, de indivíduos com fissura de palato reparada. Na primeira etapa, as avaliadoras classificaram a hipernasalidade utilizando seus próprios critérios, em uma escala de quatro pontos. Setenta dias depois, foram submetidas a um treinamento para a definição das amostras utilizadas como referências para o julgamento na etapa seguinte. Na segunda etapa as avaliadoras julgaram as mesmas amostras e classificaram a hipernasalidade com a mesma escala, utilizando como critério as referências definidas no treinamento. Índices de concordância intra e interavaliadores foram estabelecidos nas duas etapas utilizando-se o coeficiente Kappa e foram comparados por meio do teste Z. **Resultados:** Os índices de concordância intra-avaliadores obtidos entre as duas etapas variou de 0,38 para 0,92, com diferença estatisticamente significativa para uma das avaliadoras (p=0,004). O índice de concordância quanto ao grau de hipernasalidade obtido entre as três avaliadoras após o treinamento (0,54) foi significativamente maior do que o obtido antes do treinamento (0,37; p=0,044). **Conclusão:** O treinamento das avaliadoras e a definição de critérios para a classificação da hipernasalidade levam ao aumento do índice de concordância intra e interavaliadores.

Keywords

Cleft Palate
Velopharyngeal Insufficiency
Speech
Speech Disorders
Speech Perception

ABSTRACT

Introduction: A high agreement in the perceptual assessment of hypernasality among different listeners is difficult to achieve. Prior listener training and the standardization of analysis criteria may be effective strategies to decrease the effect of perceptual assessment subjectivity and increase the agreement among listeners. **Objective:** To investigate the influence of prior training on agreement among different listeners in the perceptual assessment of hypernasality. **Methods:** Three experienced speech–language pathologists analyzed 77 audio-recorded speech samples of individuals with repaired cleft palate. During the first phase, the listeners classified hypernasality according to their own criteria, using a 4-point scale. Seventy days later, they were required to complete the training to define the *stimuli* to be used as anchors for the assessment in the following phase. During the second phase, the listeners analyzed the same samples and rated hypernasality in a 4-point scale, using the anchors defined during training as the criteria. Intra- and interrater agreement in both the phases were calculated by the kappa coefficient. These values were statistically compared using the *Z*-test. **Results:** The intrarater agreement obtained between the two phases of the study ranged from 0.38 to 0.92, with a statistically significant difference for one of the listeners (p=0.004). The agreement for the hypernasality degree obtained among the three listeners after training (0.54) was significantly higher than that obtained before training (0.37; p=0.044). **Conclusion:** Listener training and the definition of criteria to rate hypernasality lead to the increase of intra- and interrater agreement.

Endereço para correspondência:
Renata Paciello Yamashita
Laboratório de Fisiologia, Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo – USP
Rua Silvio Marchione 3-20, Vila Universitária, Bauru (SP), Brasil, CEP: 17012-900.
E-mail: rezeyama@usp.br

Recebido em: Junho 13, 2015

Aceito em: Julho 11, 2015

Trabalho realizado no Laboratório de Fisiologia do Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo – USP - Bauru (SP), Brasil.

[1] Laboratório de Fisiologia, Hospital de Reabilitação de Anomalias Craniofaciais, Universidade de São Paulo – USP - Bauru (SP), Brasil.

[2] Departamento de Fonoaudiologia, Faculdade de Odontologia de Bauru, Universidade de São Paulo – USP - Bauru (SP), Brasil.

Fonte de financiamento: bolsa concedida pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Conflito de interesses: nada a declarar.

INTRODUÇÃO

A fissura palatina é a causa mais frequente da disfunção velofaríngea (DVF) e a correção cirúrgica primária do palato deve priorizar o estabelecimento de condições anatômicas e funcionais para o fechamento velofaríngeo adequado. Apesar disso, sintomas de fala decorrentes da DVF podem permanecer mesmo após a palatoplastia primária e, na dependência de sua gravidade, tais sintomas podem prejudicar, e muito, a inteligibilidade da fala. A hipernasalidade é considerada o sintoma mais representativo da DVF e é definida como o excesso de ressonância nasal observada durante a produção de sons orais, ou seja, a perda de energia acústica para a cavidade nasal[1,2].

O diagnóstico dos sintomas de fala é feito, frequentemente, por meio da avaliação perceptivo-auditiva da fala, considerada "padrão ouro" na avaliação clínica da fala de indivíduos com fissura palatina e/ou DVF. É o método que permite identificar as alterações presentes, aferir a sua gravidade e avaliar a efetividade dos tratamentos realizados, ainda que subjetivamente e, portanto, deve ser realizada por um fonoaudiólogo experiente[3-6]. Ao longo dos anos, a literatura preocupou-se em aperfeiçoar a avaliação perceptivo-auditiva a fim de torná-la menos suscetível a erros decorrentes de sua subjetividade. A introdução de recursos tecnológicos para a captura e armazenamento de amostras de fala em mídias eletrônicas, há muitos anos, é uma das estratégias utilizadas para minimizar a subjetividade do julgamento perceptivo das características de fala. As gravações, tanto em áudio quanto em vídeo, tornaram possível o acesso aos dados para consultas posteriores, bem como o direcionamento da análise dos avaliadores para os aspectos de fala desejados[7]. A vantagem desses procedimentos é a possibilidade da reavaliação da mesma amostra de fala, o julgamento da mesma amostra por múltiplos avaliadores, além da redução de fatores que poderiam distrair o ouvinte durante uma avaliação presencial, o que melhora significativamente a confiabilidade da avaliação subjetiva[7,8]. Outra estratégia adotada para esse fim foi o uso do critério de escores para representar o julgamento do fonoaudiólogo. A partir de então, diferentes escalas, tais como, escalas numéricas com intervalos iguais, estimativa de magnitude direta, escala visual analógica e comparações pareadas foram introduzidas para graduar e classificar as características de fala reduzindo, assim, a possibilidade de variação nos julgamentos. A mais popular dentre elas é a escala numérica com intervalos iguais, na qual o avaliador atribui uma nota ao aspecto avaliado, indicando o seu nível de gravidade. Nesse tipo de escala, os pontos das extremidades são fixos, adotando-se números inteiros de um conjunto finito de números ou categorias atribuídas pelo avaliador. O menor valor refere-se à ausência da alteração e o maior, ao grau máximo de alteração[5,6,8]. Historicamente, as escalas com intervalos iguais são comumente utilizadas por serem mais adequadas ao contexto clínico e são, até os dias de hoje, preferidas por clínicos e pesquisadores, uma vez que são intuitivas e as classificações obtidas são relativamente fáceis de comparar entre diferentes escalas e avaliadores[2,9,10].

Ainda que se reconheçam os avanços no sentido de minimizar a subjetividade do julgamento perceptivo, é consenso que ele está sujeito a variações e erros, mesmo entre ouvintes experientes. Isso se deve, principalmente, à influência do padrão interno, ou seja, dos critérios próprios que cada avaliador possui e utiliza no seu julgamento, os quais diferem de um para outro[3,7,11]. Além desse, fatores tais como a experiência prévia de cada um, as expectativas do julgador, o padrão articulatório do paciente, a gravidade do sintoma de fala, o modo de apresentação do estímulo, o tipo de amostra de fala que está sendo julgada, a inteligibilidade da fala, o treinamento prévio dos avaliadores, a tonalidade e a intensidade vocal, o contexto fonético e a presença de articulações compensatórias e disfonias também podem influenciar o julgamento dos ouvintes[3,4]. Quanto a essas últimas variáveis, há autores que acreditam que é difícil para o ouvinte isolar a hipernasalidade de outros aspectos de fala coexistentes durante o julgamento perceptivo. Amostras de fala contendo hipernasalidade associada a articulações compensatórias, por exemplo, podem ser julgadas perceptivamente como mais nasalizadas[12]. Por isso, alcançar um alto índice de concordância no julgamento perceptivo da nasalidade pode ser uma tarefa árdua.

Há muitos anos estudiosos da área vêm discutindo os diferentes aspectos que influenciam o julgamento perceptivo dos sintomas de fala decorrentes da fissura labiopalatina. Em um recente levantamento sobre os estudos publicados há 50 anos abordando o tema[10], os autores mostraram que, ainda em 1964, pesquisadores discutiam a necessidade de melhorar a confiabilidade do julgamento perceptivo da hipernasalidade, e já propunham estratégias como o treinamento dos ouvintes e o uso de modelos e referências para esse fim. Há autores que acreditam que os ouvintes deveriam ser treinados previamente à avaliação da hipernasalidade a fim de ajustar a sua escala interna, ou seja, deveriam classificar amostras de fala com vários graus de nasalidade até alcançarem a concordância entre seus próprios julgamentos[13]. Estudos envolvendo a avaliação perceptivo-auditiva da qualidade vocal já mostraram que o uso de referências, o treinamento prévio dos avaliadores e sua experiência favorecem a confiabilidade dos resultados[7,14-19].

No que se referem aos aspectos vocais, vários autores já demonstraram a eficácia tanto do treinamento do avaliador quanto do uso de referências no julgamento perceptivo dos sintomas de disfonia. Pesquisadores mostraram[14] que essas duas estratégias ajudaram a melhorar a confiabilidade da avaliação perceptiva da voz realizada por sujeitos inexperientes submetidos a um treinamento. Os autores concluíram, ainda, que os padrões internos referentes às qualidades de vozes patológicas não são estáveis e que tanto o treinamento quanto o uso de referências são estratégias necessárias para se estabelecer esses padrões internos. Mais tarde, esses mesmos autores[20] compararam dois programas de treinamento auditivo, um método de referência de correspondência, no qual os ouvintes deveriam relacionar o estímulo apresentado com uma das referências fornecidas, e um método de comparação pareada, no qual os ouvintes deveriam comparar se o nível de gravidade de soprosidade era idêntico para cada par de estímulos apresentado. Constataram que ambos os tipos de treinamento foram eficazes, pois os ouvintes inexperientes melhoraram significativamente sua capacidade de identificar a soprosidade vocal. Outros autores[15,17] também demonstraram melhora na confiabilidade intra e interavaliadores

com o treinamento. Utilizaram quatro diferentes tipos de treinamento (sem modelos, com modelos definidos por escrito, com modelos auditivos e a combinação de modelos auditivos e definidos por escrito) e constataram que o uso de modelos, em particular os auditivos, em conjunto com treinamento levou à melhora significativa da confiabilidade intervaliadores dos julgamentos perceptivos. Mais tarde, o efeito do uso de referências em julgamentos perceptivos foi comparado entre ouvintes experientes e inexperientes no que se refere à qualidade vocal[18]. Verificaram que os três grupos de ouvintes julgaram as amostras de fala com significativa menor gravidade nas condições nas quais foram apresentados os modelos e, ainda, que os ouvintes com e sem experiência mostraram melhora na concordância interavaliadores nas tarefas com os modelos. Isso os levou a concluir que ouvintes modificam seus julgamentos de qualidade vocal sistematicamente em resposta a modelos auditivos e que o uso de modelos reduz a variabilidade interavaliadores podendo, assim, melhorar a concordância entre os ouvintes.

Diferentemente dos estudos que envolvem a qualidade vocal, a literatura referente à hipernasalidade, especificamente, é mais escassa. Um dos trabalhos mais conhecidos é o de Lee et al.[7], no qual os autores investigaram o efeito do treinamento e do *feedback* sobre a confiabilidade intra e interavaliadores do julgamento da hipernasalidade feito por estudantes de fonoaudiologia. Esses foram distribuídos em três grupos de treinamento, um grupo submetido à simples exposição de amostras de fala com hipernasalidade; outro submetido à prática com julgamentos da hipernasalidade, porém, sem *feedback* e, um último grupo, submetido à prática com julgamentos de hipernasalidade e com *feedback*. Todos os avaliadores participaram de uma sessão na qual foram apresentados exemplos de distúrbios de ressonância, articulação e voz. O mesmo conjunto de exemplos de fala foi utilizado para treinar os dois últimos grupos no julgamento da hipernasalidade. Somente o grupo submetido à prática e *feedback* recebeu uma devolutiva após cada julgamento. Após o treinamento, os avaliadores classificaram a hipernasalidade utilizando a escala de estimativa de magnitude direta. Os autores verificaram que houve diferença significativa quanto à concordância interavaliadores, entre os dois grupos submetidos à prática e o grupo submetido somente à exposição das amostras, o que os levou a concluir que um treinamento programado para o julgamento perceptivo é útil para melhorar a confiabilidade da classificação da hipernasalidade. Um pouco antes disso, pesquisadores[16] já haviam chamado a atenção para a necessidade de treinamento dos avaliadores, até mesmo os experientes, utilizando amostras de referência a fim de aumentar a confiabilidade do resultado. As autoras investigaram a confiabilidade da avaliação perceptiva da fala de crianças com fissura palatina considerando várias características, tais como, hipernasalidade, hiponasalidade, emissão de ar nasal, fraca pressão aérea intraoral e articulação, e verificaram índices de confiabilidade que variaram de moderado a bom, com valores mais reduzidos para a hipernasalidade.

Em um trabalho recente[19], estudiosos defenderam essas mesmas estratégias afirmando que, embora o julgamento perceptivo dos sintomas de fala decorrentes da fissura palatina apresente limitações e muita variabilidade, o treinamento prévio dos ouvintes pode ser utilizado para se alcançar níveis aceitáveis de confiabilidade, independentemente do grau de experiência do avaliador. Esses autores descreveram os índices de confiabilidade da avaliação de fala de dois protocolos de avaliação perceptiva (*Cleft Audit Protocol for Speech-Augmented* – CAPS-A e *Cleft Audit Protocol for Speech-Augmented-Americleft Modification* – CAPS-A-AM) desenvolvidos para avaliar resultados de fala em estudos de colaboração intercentros e, também, investigaram o efeito do treinamento sobre a concordância entre diferentes ouvintes. Mostraram que as avaliações de fala de crianças com fissura palatina podem ser realizadas de forma confiável para a maioria dos parâmetros de fala analisados nesses protocolos. Segundo os autores, à semelhança de outros estudos da literatura, os índices de concordância interavaliadores podem aumentar após um sistemático programa de treinamento do avaliador. Concluíram que o treinamento pode e deve ser utilizado para melhorar a concordância entre diferentes avaliadores e, assim, melhorar a confiabilidade do julgamento perceptivo dos sintomas da fala tanto para pesquisadores quanto para clínicos envolvidos no tratamento de indivíduos com fissura palatina.

Ainda que o caráter subjetivo do julgamento de um ouvinte quanto à presença ou, ainda, quanto à gravidade de um sintoma de fala jamais seja anulado, cada vez mais, altos índices de confiabilidade dos métodos de avaliação perceptiva de fala são exigidos, tanto para fins de pesquisa quanto na prática clínica. Ao que tudo indica, uma das maneiras mais eficazes de se alcançar esse objetivo é a definição e o estabelecimento dos parâmetros que devem ser considerados durante a análise do ouvinte para a classificação do sintoma, com a finalidade de padronizar os critérios internos de diferentes avaliadores e, assim, minimizar a subjetividade dessa tarefa. Acredita-se, e esta foi a hipótese que motivou a realização deste estudo, que o índice de concordância entre diferentes avaliadores pode aumentar quando os mesmos são submetidos a um treinamento prévio e à utilização de modelos de referência para a classificação da hipernasalidade. Assim, o presente estudo teve por objetivo investigar a influência do treinamento dos avaliadores sobre a concordância no julgamento perceptivo da hipernasalidade comparando-se os índices de concordância intra e interavaliadores obtidos antes e após o treinamento prévio dos avaliadores.

MÉTODOS

Amostras de fala

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa em Seres Humanos do Hospital de Reabilitação de Anomalias Craniofaciais da Universidade de São Paulo (HRAC-USP) conforme parecer nº 941.709. Foram incluídas no estudo 77 amostras de fala gravadas em áudio pertencentes a indivíduos com fissura de palato reparada, com ou sem DVF, selecionadas dentre as gravações de ótima qualidade armazenadas na base de dados da instituição. As gravações são realizadas de rotina, em sala acusticamente tratada, utilizando-se o programa WaveStudio (Sound Blaster, Creative) com placa de som modelo Audigy 2 (Sound Blaster, Creative). Os pacientes permanecem sentados, com um microfone de cabeça (*headset*) modelo PRA-30

XLR (Superlux) posicionado lateralmente a uma distância de 5 centímetros da boca e acoplado a um microcomputador.

As gravações utilizadas no presente estudo foram recuperadas da base de dados, salvas no formato MP3, com 44100 Hz e 16 Bits estéreos e editadas utilizando-se as ferramentas do programa WaveStudio (CreativeLabs), excluindo-se a participação do profissional interlocutor do registro da fala e padronizando o tempo de gravação em aproximadamente 1 minuto. Neste estudo foram incluídos trechos contendo contagem de 1 a 10 e repetição de frases com fones plosivos e fricativos. Após a edição, as amostras foram numeradas e copiadas em *compact discs* (CDs).

Avaliadores

Três fonoaudiólogas com experiência na avaliação perceptiva de sujeitos com fissura palatina participaram deste estudo como ouvintes e classificaram o grau de hipernasalidade das amostras de fala.

Julgamento perceptivo da hipernasalidade

Os CDs com as amostras de fala foram entregues às avaliadoras juntamente com uma carta explicativa e uma ficha para preenchimento dos resultados. As avaliadoras foram instruídas para que as análises fossem feitas individualmente, de preferência em sala acusticamente tratada ou utilizando fones de ouvido estéreo SHP1900 (Philips) disponibilizados para este estudo. Foi permitido que as avaliadoras escutassem as amostras de fala quantas vezes julgassem necessário antes de emitir o seu julgamento. A tarefa de julgamento perceptivo da hipernasalidade foi realizada em duas etapas: antes do treinamento (pré-treinamento) e após o treinamento (pós-treinamento).

Etapa pré-treinamento

As avaliadoras classificaram a hipernasalidade de acordo com seus critérios próprios (referências internas), utilizando uma escala ordinal de 4 pontos, sendo: 1 = ausência de hipernasalidade (ressonância oronasal equilibrada); 2 = hipernasalidade leve; 3 = hipernasalidade moderada; e 4 = hipernasalidade grave. As análises foram finalizadas num intervalo de 20 dias a partir da entrega dos CDs.

Treinamento das avaliadoras

Setenta dias depois de finalizada a primeira etapa do estudo, as avaliadoras foram submetidas ao treinamento. Esse treinamento consistiu na definição dos critérios a serem utilizados para a classificação da hipernasalidade na próxima etapa do estudo a fim de uniformizar a tarefa do julgamento perceptivo da hipernasalidade. Para tanto, as três avaliadoras foram reunidas e, em conjunto, analisaram uma série de amostras de fala previamente selecionadas de estudo anterior realizado no Laboratório de Fisiologia[21], cuja classificação da hipernasalidade resultou da concordância total (100%) entre diferentes avaliadores experientes. Nenhuma das amostras utilizadas no treinamento foi incluída nas análises do estudo, em nenhuma das duas etapas. Tomou-se o cuidado de selecionar, no mínimo, duas amostras correspondentes a cada uma das quatro categorias da escala de classificação: hipernasalidade ausente, hipernasalidade leve, hipernasalidade moderada e hipernasalidade grave. As amostras foram apresentadas às três avaliadoras simultaneamente, utilizando-se um duplicador de som (*plug* adaptador duplo estéreo p2) conectado à saída de som do computador, contendo três entradas para fones de ouvido, o que possibilitou às três avaliadoras analisarem a mesma gravação de fala ao mesmo tempo. Cada avaliadora, utilizando um fone de ouvido estéreo (SHP1900-Philips) conectado ao duplicador de som, classificou a hipernasalidade da fala manifestando oralmente o seu julgamento. Os casos em que não houve concordância entre as três avaliadoras foram discutidos até que se obtivesse um resultado final quanto ao grau de hipernasalidade que representou, dessa forma, o consenso entre as três avaliadoras. Finalizado o treinamento, essas amostras de fala foram definidas como modelos de referência a serem utilizados durante o julgamento da hipernasalidade na segunda etapa do estudo.

Etapa pós-treinamento

As três avaliadoras analisaram as mesmas amostras de fala da primeira etapa, utilizando a escala de quatro pontos para classificação da hipernasalidade. Nessa etapa, contudo, o julgamento perceptivo foi feito com base nas referências (modelos) dos quatro graus de hipernasalidade definidos no treinamento. Os CDs com as amostras gravadas continham, também, as amostras de referências definidas no treinamento. As avaliadoras foram instruídas a consultar as referências a cada amostra de fala analisada, antes de emitir o seu julgamento final.

Análise dos dados

A hipernasalidade foi expressa em escores. Do total de 77 registros de fala, 20% (17 amostras) foram duplicadas aleatoriamente para a análise de concordância intra-avaliadores. Os índices de concordância intra-avaliadores e interavaliadores foram estabelecidos para as duas etapas do estudo. A concordância intra e interavaliadores foi analisada por meio do teste Kappa ponderado e a interpretação do índice de concordância foi baseada em Altman[22]. A comparação entre os índices de concordância pré e pós-treinamento foi analisada por meio do teste Z considerando o nível de significância de 5% ($p<0,05$).

RESULTADOS

Pré-treinamento – concordância intra e interavaliadores

Antes do treinamento, os índices de concordância intra-avaliadores quanto ao grau de hipernasalidade obtidos foram de 0,38 para a avaliadora 1 e de 0,39 para a avaliadora 2, ambos indicando concordância regular; e de 0,76 para a avaliadora 3, indicando concordância boa.

No que se refere à concordância interavaliadores, os índices obtidos entre as avaliadoras 1 e 2 e entre as avaliadoras 2 e 3 foram de 0,35 e de 0,26, respectivamente, indicando concordância regular nos dois casos. Entre as avaliadoras 1 e 3 o índice foi 0,52,

indicando concordância moderada. O índice de concordância Kappa entre as três avaliadoras juntas foi de 0,37, indicativo de concordância regular.

Pós-treinamento – concordância intra e interavaliadores

Após o treinamento, os índices de concordância intra-avaliadores variaram de moderado a muito bom. Para a avaliadora 1, o índice Kappa foi de 0,61 e para a avaliadora 2, o índice Kappa foi de 0,92, indicando concordância boa no primeiro caso e muito boa no segundo. Para a avaliadora 3, o índice Kappa foi de 0,50, indicando concordância moderada.

Quanto à concordância interavaliadores, os índices obtidos entre as avaliadoras 1 e 2 e entre as avaliadoras 1 e 3 foram de 0,57 e de 0,44, respectivamente, indicando concordância moderada em ambos os casos. Já para as avaliadoras 2 e 3, foi de 0,63, indicando concordância boa. A análise das três avaliadoras mostrou que o índice de concordância foi de 0,54, indicativo de concordância moderada.

Pré versus pós-treinamento – comparação entre os índices de concordância

A comparação estatística entre os coeficientes de concordância intra-avaliadores obtidos nas duas etapas do estudo mostrou que, para a avaliadora 1, o índice Kappa aumentou de 0,38 (regular) para 0,61 (bom), porém sem significância estatística. Para a avaliadora 2, houve aumento significante do índice Kappa de 0,39 (regular) para 0,92 (muito bom). Para a avaliadora 3, verificou-se ligeira redução do índice Kappa de 0,76 (bom) para 0,50 (moderado), porém, essa diferença não foi significante. Estes resultados estão demonstrados na Tabela 1.

No que se refere aos índices de concordância interavaliadores, houve aumento significativo após o treinamento entre as avaliadoras 1 e 2, que passou de 0,35 (regular) para 0,57 (moderado) e entre as avaliadoras 2 e 3, que passou de 0,26 (regular) para 0,63 (boa). Não houve diferença significante entre os índices de concordância obtidos entre as avaliadoras 1 e 3, que passou de 0,52 para 0,44, permanecendo moderado. O índice de concordância entre as três avaliadoras aumentou significativamente de 0,37 (regular) antes do treinamento para 0,54 (moderado) após o treinamento, conforme mostra a Tabela 2.

DISCUSSÃO

A hipernasalidade é um sintoma comum em indivíduos com fissura palatina, sendo considerado o mais representativo da disfunção velofaríngea. A identificação e, mais ainda, a classificação da gravidade desse sintoma é feita de maneira subjetiva, utilizando como instrumento o ouvido humano e a percepção do ouvinte. Quando ouvintes classificam um determinado aspecto de voz, eles comparam o estímulo apresentado a um padrão interno. Esses padrões internos desenvolvem-se ao longo do tempo, são mantidos na memória do indivíduo e acabam por diferir de ouvinte para ouvinte. Além disso, são inerentemente instáveis e podem ser influenciados por fatores como lapsos de memória e atenção, e por variáveis externas, como o padrão articulatório, a gravidade do sintoma de fala, o modo de apresentação do estímulo, a intensidade vocal, o contexto fonético e outras já descritas anteriormente[23]. A fim de minimizar o efeito de padrões internos sobre o julgamento da fala, o presente estudo utilizou como estratégia o treinamento dos ouvintes, conforme recomenda a literatura, e comparou os resultados obtidos antes e após esse treinamento. Ressalta-se que o tipo de análise realizada no presente estudo, a de classificar um único sintoma de fala (hipernasalidade) utilizando uma escala de quatro categorias, é uma das mais complexas de se

Tabela 1. Porcentagem de concordância intra-avaliadores, coeficiente Kappa e sua interpretação, obtidos na análise perceptiva da hipernasalidade: comparação estatística dos coeficientes Kappa pré e pós-treinamento

Concordância Intra-avaliadores							
Avaliadoras	Pré-treinamento			Pós-treinamento			Valor de p
	% de concordância	Coeficiente Kappa	Interpretação	% de concordância	Coeficiente Kappa	Interpretação	
1	53	0,38	Regular	71	0,61	Boa	0,330
2	59	0,39	Regular	94	0,92*	Muito boa	0,004
3	82	0,76	Boa	65	0,50	Moderada	0,234

*Pré versus pós-treinamento: teste Z

Tabela 2. Porcentagem de concordância, coeficiente Kappa e sua interpretação, obtidos entre as avaliadoras na análise perceptiva da hipernasalidade: comparação estatística dos coeficientes Kappa pré e pós-treinamento

Concordância interavaliadores							
Avaliadoras	Pré-treinamento			Pós-treinamento			Valor de p
	% de concordância	Coeficiente Kappa	Interpretação	% de concordância	Coeficiente Kappa	Interpretação	
1 e 2	52	0,35	Regular	69	0,57*	Moderada	0,009
1 e 3	65	0,52	Moderada	58	0,44	Moderada	0,398
2 e 3	44	0,26	Regular	73	0,63*	Boa	0,001
1, 2 e 3	31	0,37	Regular	51	0,54*	Moderada	0,044

*Pré versus pós-treinamento: teste Z

executar, ou seja, a avaliadora precisou não somente identificar o sintoma, mas também graduá-lo e, em alguns casos, fazê-lo na presença de outros sintomas concorrentes.

Uma vez que muitos autores afirmam que a experiência é um fator essencial para se obter um resultado confiável na avaliação perceptivo-auditiva[7,18], tomou-se o cuidado de convidar profissionais experientes na avaliação de fala de indivíduos com fissura labiopalatina para participar do estudo. As 3 fonoaudiólogas que julgaram as amostras de fala do presente estudo tinham, no mínimo, 12 anos de experiência na área e realizaram a tarefa de classificar a hipernasalidade, individualmente, nas 2 etapas do estudo. O índice de concordância encontrado entre as 3 avaliadoras na primeira etapa foi de 0,37, interpretado como concordância regular. Esse resultado confirma a dificuldade em se obter um alto índice de concordância no julgamento perceptivo da hipernasalidade entre diferentes ouvintes, como já demonstrado na literatura, muito provavelmente justificado pelo fato das avaliadoras terem utilizado seus próprios padrões para o julgamento do sintoma, os quais, sabe-se, podem diferir mesmo entre ouvintes experientes[7,11]. A utilização de critérios próprios pode explicar, também, os índices de concordância intra-avaliadores obtidos na primeira etapa do estudo. À exceção de uma avaliadora, que obteve um índice de concordância bom, para as outras duas esse índice foi regular. Alguns autores acreditam que os padrões internos dos avaliadores podem ser instáveis até mesmo para um mesmo avaliador, independente do nível de experiência[11,14]. Essa é uma das razões que levou alguns clínicos e pesquisadores a defenderem a utilização do treinamento prévio dos avaliadores como estratégia para aumentar a confiabilidade da avaliação perceptivo-auditiva[14,18,24].

Diversos tipos de treinamento foram documentados na literatura com resultados, na sua maioria, positivos. Grande parte deles foi utilizado na análise de aspectos vocais como rouquidão e soprosidade[14,15,17,18,20], e poucos na análise da nasalidade[6,7,16,19]. Uma das estratégias de treinamento mais eficientes apontadas na literatura é a utilização de referências de correspondência, como a que foi utilizada no presente estudo[20]. De acordo com esses autores, as referências são eficazes para se estabelecer padrões internos, pois os ouvintes se familiarizam com as referências utilizadas no treinamento e, eventualmente, armazenam esses modelos em sua memória como padrões internos. Em outras palavras, uma vez vivenciadas, essas representações são armazenadas na memória como exemplos.

Os resultados obtidos após o treinamento comprovaram essa teoria. Os achados mostram que o coeficiente de concordância intra-avaliadores aumentou de regular a muito bom. O mesmo ocorreu com a concordância interavaliadores, que aumentou significativamente de regular a moderada entre as três avaliadoras. Resultados semelhantes, tais como concordância moderada na classificação da hipernasalidade julgada por diferentes avaliadores[8] e índices de concordância de moderado a bom no julgamento da hipernasalidade e da hiponasalidade feito por avaliadores experientes[25] foram relatados por outros pesquisadores da área. Recente estudo realizado no HRAC-USP[26] também mostrou índices que variaram de moderado a bom. Contudo, nesse caso, a análise feita pelos avaliadores foi baseada na presença ou ausência da hipernasalidade e não na graduação do sintoma. Dos estudos que utilizaram a classificação da hipernasalidade em escala de quatro pontos, a maioria encontrou índices de concordância semelhante aos relatados no presente estudo, variando de regular a moderado[21], de moderado a bom[27] e moderado[28]. No que se refere ao uso de treinamento, assim como o presente estudo, outros pesquisadores também demonstraram melhora na confiabilidade intra e interavaliadores após o treinamento dos ouvintes[15]. À semelhança do que ocorreu no presente estudo, a combinação de treinamento e o uso de modelos melhoraram significativamente a confiabilidade interavaliadores das classificações perceptivas da voz[17,18]. Contudo, esses autores não verificaram efeitos significativos do uso de referências ou experiência do ouvinte sobre a concordância intra-avaliadores. No presente estudo, esse índice de concordância aumentou para duas das avaliadoras, sendo que, para uma delas, a diferença foi significante.

Como destacado anteriormente, a grande maioria dos trabalhos que estudaram modelos de referência e treinamento investigou o seu efeito sobre a análise de sintomas vocais, considerados por estudiosos estímulos que variam em termos de mudança na qualidade. O uso dessas estratégias sobre os sintomas de fala decorrentes da fissura labiopalatina, como a nasalidade, um estímulo que varia em termos de mudança na magnitude, foi pouco investigado, até hoje[6,7]. Ressalte-se, ainda, que nenhum outro estudo da literatura comparou o desempenho do mesmo grupo de ouvintes antes e após o treinamento na análise e classificação da nasalidade, como foi realizado no presente estudo. Verificou-se diferença significativa entre as etapas pré e pós-treinamento, com aumento dos índices de concordância após o treinamento, na maioria das comparações realizadas com o mesmo grupo de avaliadoras. A comparação dos índices intra-avaliadores revelou aumento para duas, das três avaliadoras, sendo que para uma delas esse aumento foi significante, mostrando que o treinamento realizado foi eficaz no estabelecimento do padrão interno do ouvinte. Entretanto, para uma das avaliadoras houve redução desse índice, embora diferença significante não tenha sido identificada nesse caso. Uma explicação para esse resultado pode ser a influência, comprovada na literatura, de fatores internos como atenção, memória e até mesmo, cansaço. No que se refere à comparação interavaliadores, os índices de concordância também aumentaram significantemente de regular a moderado e de regular a muito bom. Em uma das comparações houve uma ligeira redução após o treinamento, embora ainda interpretado como moderado, sendo esse o índice mais expressivo obtido nas comparações feitas entre essas mesmas avaliadoras na etapa pré-treinamento (0,52 – moderado).

Considerando a utilização de um treinamento programado para estabelecer padrões e referências para o julgamento perceptivo da nasalidade, como o realizado no presente estudo, poderia se esperar índices de concordância intra e interavaliadores ainda melhores do que os obtidos após o treinamento. Uma justificativa para tal resultado pode ser a utilização da escala ordinal, empregada neste estudo e utilizada na rotina clínica para a classificação da nasalidade. Embora esse tipo de escala seja o método mais adotado, existem ressalvas quanto à sua validade para o julgamento da hipernasalidade tanto em pesquisas quanto na prática clínica. Isso porque a

escala ordinal divide as categorias do sintoma sem, contudo, quantificar a magnitude das diferenças entre cada categoria e os ouvintes tendem a subdividir o extremo inferior da escala em intervalos menores. Diversos autores sugerem que a nasalidade é uma sensação mentalmente processada como uma dimensão protética, ou seja, difere em termos de mudanças na quantidade ou magnitude. Segundo Stevens[29], ao julgar estímulos protéticos os ouvintes não percebem os intervalos entre as categorias como iguais em diferentes pontos da escala. Desse modo, intervalos "aparentemente iguais" não são "necessariamente iguais" para toda a escala. Por isso, a escala numérica de intervalos iguais pode não ser tão eficaz para a classificação da nasalidade, mesmo com o uso do treinamento prévio. Alguns autores defendem que a nasalidade seria melhor julgada utilizando-se escalas baseadas em proporção (relação), como a escala de magnitude direta e a escala visual analógica, as quais possibilitam classificações mais válidas e confiáveis para a percepção da nasalidade[6,30]. Entretanto, há autores que consideram a escala de magnitude direta impraticável na rotina clínica, visto que a amostra de fala a ser classificada deve ser comparada a uma amostra padrão[9]. Por outro lado, a escala visual analógica tem sido empregada e defendida por outros. Um estudo recente[2] demonstrou que esse tipo de escala oferece maior confiabilidade do que a escala numérica de intervalos iguais no julgamento perceptivo de sintomas de fala característicos da fissura palatina, como a hipernasalidade e a emissão de ar nasal audível, e sugere a sua utilização como um método alternativo à escala de intervalos iguais para a avaliação desses parâmetros de fala.

Em resumo, a despeito da dificuldade em se obter consenso quanto à gravidade da hipernasalidade, este estudo comprovou que o treinamento prévio é uma estratégia eficiente para aumentar a concordância entre diferentes ouvintes e, assim, melhorar a confiabilidade da avaliação perceptiva da fala, método esse que continua sendo o principal indicador da significância clínica dos sintomas de fala. Outros estudos vêm sendo conduzidos no Laboratório de Fisiologia do HRAC-USP, empregando novos métodos de avaliação perceptiva, na tentativa de encontrar aqueles que permitam julgamentos mais confiáveis e reprodutíveis da hipernasalidade.

CONCLUSÃO

O treinamento prévio das avaliadoras leva ao aumento dos índices de concordância intra e interavaliadores e, consequentemente, à melhora da confiabilidade do julgamento perceptivo da hipernasalidade de indivíduos com fissura palatina. Esses resultados reforçam a importância de se estabelecer critérios padronizados a fim de minimizar a influência de padrões internos individuais no julgamento perceptivo dos sintomas de fala.

REFERÊNCIAS

1. Wermker K, Jung S, Joos U, Kleinheinz J. Objective assessment of hypernasality in patients with cleft lip and palate with the nasal view system: a clinical validation study. Int J Otolaryngol. 2012;2012:321319. http://dx.doi.org/10.1155/2012/321319. PMid:22518153.
2. Baylis A, Chapman K, Whitehill TL, Group TA. Validity and reliability of visual analog scaling for assessment of hipernasality and audible nasal emission in children with repaired cleft palate. Cleft Palate Craniofac J. 2015;52(6):660-70. http://dx.doi.org/10.1597/14-040. PMid:25322442.
3. Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. Am J Speech Lang Pathol. 1996;5(3):7-23. http://dx.doi.org/10.1044/1058-0360.0503.07.
4. Whitehill TL, Lee AS, Chun JC. Direct magnitude estimation and interval scaling of hypernasality. J Speech Lang Hear Res. 2002;45(1):80-8. http://dx.doi.org/10.1044/1092-4388(2002/006). PMid:14748640.
5. Lohmander A, Olsson M. Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature. Cleft Palate Craniofac J. 2004;41(1):64-70. http://dx.doi.org/10.1597/02-136. PMid:14697067.
6. Baylis AL, Munson B, Moller KT. Perceptions of audible nasal emission in speakers with cleft palate: a comparative study of listener judgments. Cleft Palate Craniofac J. 2011;48(4):399-411. http://dx.doi.org/10.1597/09-201. PMid:20572776.
7. Lee A, Whitehill TL, Ciocca V. Effect of listener training on perceptual judgement of hypernasality. Clin Linguist Phon. 2009;23(5):319-34. http://dx.doi.org/10.1080/02699200802688596. PMid:19399664.
8. John A, Sell D, Sweeney T, Harding-Bell A, Williams A. The cleft audit protocol for speech-augmented: a validated and reliable measure for auditing cleft speech. Cleft Palate Craniofac J. 2006;43(3):272-88. http://dx.doi.org/10.1597/04-141R.1. PMid:16681400.
9. Brancamp TU, Lewis KE, Watterson T. The relationship between nasalance scores and nasality ratings obtained with equal appearing interval and direct magnitude estimation scaling methods. Cleft Palate Craniofac J. 2010;47(6):631-7. http://dx.doi.org/10.1597/09-106. PMid:20500059.
10. Bressmann T, Sell D. Plus ça change: selected papers on speech research from the 1964 issue of the Cleft Palate Journal. Cleft Palate Craniofac J. 2014;51(2):124-8. http://dx.doi.org/10.1597/13-310. PMid:24446923.
11. Keuning KH, Wieneke GH, Dejonckere PH. The intrajudge reliability of the perceptual rating of cleft palate speech before and after pharyngeal flap surgery: the effect of judges and speech samples. Cleft Palate Craniofac J. 1999;36(4):328-33. http://dx.doi.org/10.1597/1545-1569(1999)036<0328:TIROTP>2.3.CO;2. PMid:10426599.
12. Starr CD, Moller KT, Dawson W, Graham J, Skaar S. Speech ratings by speech clinicians, parents and children. Cleft Palate J. 1984;21(4):286-92. PMid:6595084.
13. McWilliams BJ, Morris HL, Shelton RL. Cleft palate speech. 2nd ed. Philadelphia: BC Decker; 1990.
14. 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. http://dx.doi.org/10.1044/1092-4388(2002/009). PMid:14748643.
15. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. J Voice. 2006;20(4):527-44. http://dx.doi.org/10.1016/j.jvoice.2005.08.007. PMid:16324823.
16. Brunnegård K, Lohmander A. A cross-sectional study of speech in 10-year-old children with cleft palate: results and issues of rater reliability. Cleft Palate Craniofac J. 2007;44(1):33-44. http://dx.doi.org/10.1597/05-164. PMid:17214536.
17. Awan SN, Lawson LL. The effect of anchor modality on the reliability of vocal severity ratings. J Voice. 2009;23(3):341-52. http://dx.doi.org/10.1016/j.jvoice.2007.10.006. PMid:18346869.
18. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. J Speech Lang Hear Res. 2011;54(2):430-47. http://dx.doi.org/10.1044/1092-4388(2010/09-0205). PMid:20884782.
19. Chapman KL, Baylis A, Trost-Cardamone J, Cordero KN, Dixon A, Dobbelsteyn C, et al. The Americleft Speech Project: a training and reliability study. Cleft Palate Craniofac J. 2016;53(1):93-108. PMid:25531738.
20. Chan KM, Yiu EM. A comparison of two perceptual voice evaluation training programs for naive listeners. J Voice. 2006;20(2):229-41. http://dx.doi.org/10.1016/j.jvoice.2005.03.007. PMid:16139475.
21. Barbosa DA. Resultados de fala e de função velofaríngea do retalho faríngeo e da veloplastia intravelar na correção da insuficiência velofaríngea: estudo

comparativo [dissertação]. Bauru: Universidade de São Paulo, Hospital de Reabilitação de Anomalias Craniofaciais; 2011.

22. Altman DG. Practical statistcs for medical research. London: Chapman & Hall; 1991. 611 p.
23. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. J Speech Hear Res. 1993;36(1):21-40. http://dx.doi.org/10.1044/jshr.3601.21. PMid:8450660.
24. Watterson T, Mancini MC, Brancamp TU, Lewis KE. Relationship between the perception of hypernasality and social judgments in school-aged children. Cleft Palate Craniofac J. 2013;50(4):498-502. http://dx.doi.org/10.1597/11-126. PMid:22292671.
25. Brunnegård K, Lohmander A, van Doorn J. Untrained listeners' ratings of speech disorders in a group with cleft palate: a comparison with speech and language pathologists' ratings. Int J Lang Commun Disord. 2009;44(5):656-74. http://dx.doi.org/10.1080/13682820802295203. PMid:18821109.
26. Prado-Oliveira R, Marques IL, Souza L, Souza-Brosco TV, Dutka JC. Assessment of speech nasality in children with Robin Sequence. CoDAS. 2015;27(1):51-7. http://dx.doi.org/10.1590/2317-1782/20152014055. PMid:25885197.
27. Brandão GR, Souza Freitas JA, Genaro KF, Yamashita RP, Fukushiro AP, Lauris JR. Speech outcomes and velopharyngeal function after surgical treatment of velopharyngeal insufficiency in individuals with signs of velocardiofacial syndrome. J Craniofac Surg. 2011;22(5):1736-42. http://dx.doi.org/10.1097/SCS.0b013e31822e624f. PMid:21959422.
28. Scarmagnani RH. Correlação entre as dimensões do orifício velofaríngeo, hipernasalidade, emissão de ar nasal audível e ronco nasal em indivíduos com fissura de palato reparada [dissertação]. Bauru: Universidade de São Paulo, Hospital de Reabilitação de Anomalias Craniofaciais; 2013.
29. Stevens SS. Perceptual magnitude and its measurement. In: Carterette C, Friedman MP, editors. Handbook of perception: psychophysical judgment and measurement. New York: Academic Press; 1974. p. 22-40.
30. Zraick RI, Liss JM. A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. J Speech Lang Hear Res. 2000;43(4):979-88. http://dx.doi.org/10.1044/jslhr.4304.979. PMid:11386483.

Contribuição dos autores

ACASFO foi responsável pelo estudo, coleta e análise dos dados e redação do artigo; RHS colaborou na coleta de dados e redação do artigo; APF colaborou na análise dos dados e redação do artigo; RPY foi responsável pelo projeto, delineamento do estudo e orientação geral das etapas de execução e elaboração do manuscrito.