Doi: 10.1590/2317-1782/20162015163

Original Article

Artigo Original

The influence of listener training on the perceptual assessment of hypernasality

Influência do treinamento dos avaliadores no julgamento perceptivo da hipernasalidade

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

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.

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.

Correspondence address:
Renata Paciello Yamashita
Laboratory of Physiology, Hospital for Rehabilitation of Craniofacial Anomalies, Universidade de São Paulo – USP
Rua Silvio Marchione 3-20, Vila Universitária, Bauru (SP), Brazil, CEP: 17012-900.
E-mail: rezeyama@usp.br

Received: June 13, 2015

Accepted: July 11, 2015

Study carried out at Laboratory of Physiology, Hospital for Rehabilitation of Craniofacial Anomalies, Universidade de São Paulo – USP - Bauru (SP), Brazil.

[1]Laboratory of Physiology, Hospital for Rehabilitation of Craniofacial Anomalies, Universidade de São Paulo – USP - Bauru (SP), Brazil.

[2]Department of Speech-Language Pathology and Audiology, Bauru School of Dentistry, Universidade de São Paulo – USP - Bauru (SP), Brazil.

Financial support: grant awarded by the São Paulo Research Foundation (FAPESP).

Conflict of interests: nothing to declare.

INTRODUCTION

Cleft palate is the most common cause of velopharyngeal dysfunction (VPD), and the primary surgical correction of the palate should prioritize the establishment of anatomical and functional conditions for an adequate velopharyngeal closure. Nevertheless, the symptoms of speech impairment resulting from VPD may remain even after the primary palatoplasty; depending on their severity, these symptoms can critically harm the intelligibility of speech. Hypernasality is considered the most representative symptom of VPD and defined as the excessive nasal resonance observed during the production of oral sounds, i.e. the loss of acoustic energy into the nasal cavity[1,2].

The diagnosis of the symptoms of speech impairment is often conducted through the auditory perceptual assessment of speech, considered the "gold standard" for the clinical assessment of individuals with cleft palate and/or VPD. This is the method that enables the detection of disorders, the determination of their severity, and the assessment of the effectiveness of performed treatments, even if subjectively; therefore, it should be carried out by an experienced speech–language pathologist[3-6]. Over the years, literature has been concerned with improving the auditory perceptual assessment to make it less susceptible to errors resulting from its subjectivity. The introduction of technological resources that enable the capture and storage of speech samples on electronic media has been one of the strategies used for many years to decrease the perceptual assessment subjectivity for speech characteristics. Recordings, in both audio and video formats, facilitated the access to the data for future reference and the orientation of listener assessment to the desired aspects of speech[7]. The benefit of these procedures is the possibility of reassessing the same speech sample, multiple listeners assessing the same sample, and minimizing factors that could distract the listener during live assessment, which significantly improves the reliability of the subjective assessment[7,8]. Another strategy adopted for this purpose was the use of scores to represent the judgment of the evaluator. From there, different scales, such as equal-appearing interval scaling, direct magnitude estimation, the visual analog scaling, and ordinal scale, were introduced to rate and classify speech characteristics, thereby reducing the possibility of variation in judgments. The most popular one among them is the equal-appearing interval scaling, where the listener assigns a score to the assessed aspect, indicating their severity level. In this scale, the endpoints are fixed, and from a finite set of numbers or categories assigned by the listener, integers are adopted. The lowest value refers to the absence of the disorder and the highest to the maximum degree of the disorder[5,6,8]. Historically, equal-appearing interval scaling are commonly used, because they are more appropriate to the clinical setting and are, to this day, preferred by clinicians and researchers, as they are intuitive, and the rates obtained are relatively easy to compare among different scales and listeners[2,9,10].

Although the advantages regarding the advances in minimizing the perceptual judgment subjectivity are recognized, there is a consensus that such assessment is subject to variations and errors, even among the experienced listeners. This is mainly owing to the influence of the internal standard, i.e. the personal criteria each listener has and uses in their judgments, which differ from listener to listener[3,7,11]. In addition, factors such as previous experience, listener expectations, the patient's articulation, the severity of the speech symptoms, the type of *stimulus* presentation, the type of speech sample, speech intelligibility, prior listener training, vocal pitch and loudness, the phonetic context, and compensatory articulations and dysphonia may also affect listener judgments[3,4]. As for the latter variables, some authors believe that it is difficult for listeners to isolate hypernasality from other coexisting aspects of speech during the perceptual assessment. Speech samples that have hypernasality associated with compensatory articulations, for example, may be perceptually assessed as more nasalized[12]. Therefore, to reach a high rate of agreement in the perceptual assessment of hypernasality can be an arduous task.

For many years, researchers have been discussing the different aspects that influence the perceptual judgment of speech symptoms resulting from the cleft lip and palate. In a recent analysis of studies published 50 years ago addressing the issue[10], the authors indicated that, in 1964, researchers already discussed the need to improve the reliability of perceptual assessment of hypernasality and proposed strategies such as the training of listeners and the use of anchors for that purpose. Some authors believe that listeners should be trained prior to the assessment of hypernasality to adjust their internal scales, i.e., they should classify speech samples with different degrees of nasality until they reach an agreement in their own judgments[13]. Studies contemplating the auditory perceptual assessment of voice quality have shown that the use of anchors, prior listener training, and experience of the listeners favor the reliability of results[7,14-19].

Regarding the vocal aspects, multiple authors have demonstrated the efficacy in both the listener training and the use of anchors in the perceptual assessment of dysphonia symptoms. Researchers showed[14] that these two strategies have helped to improve the reliability of the perceptual assessment of voice carried out by inexperienced subjects who underwent training. Furthermore, the authors reached the conclusion that the internal standards related to the quality of pathological voices are not stable and that both training and the use of anchors are necessary strategies to establish these internal standards. Later, these same authors[20] compared two auditory training programs. The first was reference-matching test, in which listeners had to match the *stimulus* presented to one of the anchors, and the second was a paired-comparison test, in which the listeners had to compare the levels of the severity of breathiness to determine if they were identical for each pair of the supplied *stimuli*. They established that both types of training were effective, because inexperienced listeners significantly improved their ability to detect breathiness of voice. Other authors[15,17] also indicated the improvement in intra- and interrater reliability with training.

They used four different training programs (without anchor, with textual anchor, with auditory anchor, and the combined textual-auditory anchors) and found that the use of anchors, especially auditory anchors, in addition to training, led to a significant improvement in interrater reliability for perceptual assessments. Subsequently, the effect of the use of anchors in perceptual assessments was compared between the experienced and inexperienced listeners with respect to voice quality[18]. They found that the three groups of listeners judged the speech samples with significantly minor severity under the conditions in which the anchors were presented and that the listeners with and without experience showed improvements in interrater agreement for tasks that used anchors. That led them to the conclusion that listeners change their voice quality judgments systematically in response to auditory anchors and that the use of anchors reduces interrater variability; thus, it can improve the agreement among listeners.

In contrast, with the studies comprising voice quality, literature specifically concerning hypernasality is scarce. Lee et al.[7] present one of the best-known studies, in which the authors investigated the effect of training and feedback on intra- and interrater reliability in the assessment of hypernasality carried out by Speech–Language Pathology students. These students were divided into three training groups; the first group underwent simple exposure to speech samples presenting hypernasality; another underwent the practice of assessment of hypernasality but without any feedback, and the last group underwent the practice of assessment with feedback. All listeners attended a session in which they were presented examples of resonance, articulation, and voice disorders. The same set of speech samples was used to train the last two groups in the assessment of hypernasality. Only the group that underwent practice and feedback received comments after each assessment. After training, the listeners rated hypernasality using the direct magnitude estimation. The authors found that there was a significant difference regarding interrater agreement between the two groups that underwent practice and the group that only underwent exposure to samples, which led them to the conclusion that a planned training for perceptual assessments is useful to improve reliability when classifying hypernasality. Shortly before, researchers[16] had drawn attention to the need for listener training, even for experienced ones, using anchor *stimuli* to increase the reliability of results. The authors investigated the reliability of the perceptual speech assessment of children with cleft palate, considering several characteristics features, such as hypernasality, hyponasality, nasal air emission, weak intraoral air-pressure and articulation, and verified reliability rates ranging from moderate to good, with lower values for hypernasality.

In a recent study[19], experts defended these same strategies stating that, although the perceptual assessment of cleft speech symptoms presents limitations and great variability, prior listener training can be used to achieve acceptable levels of reliability, regardless of the level of listener experience. These authors described the reliability levels for the speech assessment of two perceptual assessment protocols (Cleft Audit Protocol for Speech-Augmented, CAPS-A, and Cleft Audit Protocol for Speech-Augmented-Americleft Modification, CAPS-A-AM) developed to assess speech results in intercenter collaborative studies and investigated the effect of training on the agreement among different listeners. They showed that the cleft speech assessment could be reliably performed for most of the speech parameters analyzed in these protocols. According to the authors, similar to other literature studies, the interrater agreement can increase after a systematic listener-training program. They reached the conclusion that training can and should be used to improve the agreement among different listeners and, thus, improve the reliability of the perceptual assessment of speech symptoms, for both the researchers and the clinicians treating individuals with cleft palate.

Although the subjective nature of a listener's judgment regarding the presence or, even more, the severity of a speech symptom is never canceled, increasingly high reliability in methods for the perceptual speech assessment are required, for research purposes and the clinical practice. Apparently, one of the most effective ways of achieving this goal is to determine and establish the parameters to be considered during listener analyses for classification of symptoms, in order to standardize different listeners' internal criteria and, thus, decrease the subjectivity of the task. It is believed, and this was the hypothesis that motivated this study the agreement among different listeners may increase when they undergo training and the use of anchor to classify hypernasality. Therefore, this study aimed at investigating the influence of listener training on agreement in the perceptual assessment of hypernasality and comparing the intra- and interrater agreement index obtained before and after listener training.

METHODS

Speech samples

The Human Research Ethics Committee of the Hospital for Rehabilitation of Craniofacial Anomalies of the University of Sao Paulo (HRAC-USP) approved this study (No. 941.709). The study included 77 audio-recorded speech samples of individuals with repaired cleft palate, with or without VPD, selected among high quality recordings stored in the database of the institution. The recordings are routinely performed in a soundproof room using the WaveStudio software (Sound Blaster, Creative) with the Audigy 2 sound card model (Sound Blaster, Creative). Patients remained seated using a PRA-30 XLR (Superlux) headset microphone laterally positioned at a five-centimeter distance from their mouths, connected to a microcomputer.

The recordings used in this study were retrieved from the database, saved in MP3 format, using 44,100 Hz and 16 bits, and edited using the tools of the WaveStudio software (Creative Labs), precluding the participation of the professional speaker from speech records and standardizing the recording time to approximately one minute. The speech samples included in this

study comprised counting from one to ten and the repetition of sentences with plosive and fricative phonemes. After editing, the samples were numbered and copied to compact discs (CDs).

Listeners

Three speech–language pathologists with experience in the perceptual assessment of individuals with cleft palate participated in this study as listeners and rated the degree of hypernasality in the speech samples.

Perceptual assessment of hypernasality

CDs with the speech samples were given to listeners along with a cover letter and an index card to be filled out with the results. Listeners were instructed so that the analyses were carried out individually, preferably, in a room with acoustic treatment or using SHP1900 stereo headphones (Philips), made available for this study. Listeners were allowed to listen to the speech samples as many times as necessary, before their judgment. The task of perceptual judgment of hypernasality was performed in two phases: before training (pre-training) and after training (post-training).

Pre-training phase

Listeners rated hypernasality according to their own criteria (internal standard), using an ordinal four-point scale, where 1 = absence of hypernasality (normal resonance); 2 = mild hypernasality; 3 = moderate hypernasality; and 4 = severe hypernasality. Analyses were completed within 20 days from the delivery of CDs.

Listener training

Seventy days after the end of the first phase of the study, the listeners underwent training. This training consisted of determining the criteria for the classification of hypernasality in the following phase of the study, to standardize the task of perceptual assessment of hypernasality. Therefore, the three listeners were gathered and, as a group, analyzed a series of speech samples previously selected from an earlier study conducted at the Laboratory of Physiology[21], in which the rating of hypernasality resulted from full agreement (100%) among different experienced listeners. None of the samples used in the training was included in the study analyses, in neither of the phases. For this study, special attention was given to the selection of at least two corresponding samples to each of the four categories in the rating scale: absence of hypernasality, mild hypernasality, moderate hypernasality, and severe hypernasality. The samples were simultaneously presented to the three listeners, using a sound replication device (stereo P2 plug) connected to the computer's sound output, with three inputs for headphones, which allowed the three listeners to analyze the same speech recording at the same time. Each listener, using a stereo headphone (Philips SHP1900) connected to the sound replicator, rated the hypernasality by orally expressing their judgment. For cases in which there was no agreement between the three listeners, a discussion was carried out until a result was reached regarding the degree of hypernasality that represented, therefore, a consensus among the three listeners. After the completion of the training, these speech samples were established as anchors to be used during the assessment of hypernasality during the second phase of the study.

Post-training phase

The three listeners analyzed the same speech samples from the first phase, using an ordinal 4-point scale for the classification of hypernasality. At this stage, however, the perceptual assessment was carried out based on anchors (models) of the four degrees of hypernasality established in training. The CDs with the recorded samples also included the anchor *stimuli* established in training. Listeners were instructed to refer to the anchors after each speech sample analyzed, before emitting their final assessment.

Data analysis

The hypernasality was determined in scores. From 77 speech recordings, 20% (17) samples were randomly duplicated for the analysis of intrarater agreement. The intra- and interrater agreement coefficients were established for the two phases of the study. Intra- and interrater agreements were analyzed using the weighted kappa test, and the strength of the agreement was based on Altman[22]. The comparison between the pre-and post-training agreement was analyzed using the *Z*-test, considering a 5% significance level ($p<0.05$).

RESULTS

Pre-training: Intra- and interrater agreements

Before training, the intrarater agreement regarding the degree of hypernasality obtained were 0.38 for listener 1 and 0.39 for listener 2, both indicating fair agreement, and 0.76 for listener 3, indicating good agreement.

With respect to interrater agreement, the kappa values obtained between listeners 1 and 2 and between listeners 2 and 3 were 0.35 and 0.26, respectively, indicating a fair agreement in both the cases. Between the listeners 1 and 3, the index was 0.52, indicating moderate agreement. The kappa agreement among all the three listeners was 0.37, an indication of fair agreement.

Post-training: Intra- and interrater agreements

After training, the intrarater agreement ranged from moderate to very good. For listener 1, the kappa was 0.61, and for listener 2, the kappa was 0.92, indicating good agreement in the first case and very good in the second. For listener 3, the kappa was 0.50, indicating moderate agreement.

With respect to interrater agreement, the kappa values obtained between listeners 1 and 2 and between listeners 1 and 3 were 0.57 and 0.44, respectively, indicating moderate agreement in both the cases. As for listeners 2 and 3, it was 0.63, indicating

good agreement. The three listeners' analysis showed that the kappa value was 0.54, an indication of moderate agreement.

Pre-versus post-training: Comparison between the agreement index

The statistical comparison between the intrarater agreement coefficients obtained between the two phases of the study showed that, for listener 1, the kappa index increased from 0.38 (fair) to 0.61 (good) but without statistical significance. For listener 2, there was a significant increase in the kappa from 0.39 (fair) to 0.92 (very good). For listener 3, a slight reduction was noticed, from 0.76 (good) to 0.50 (moderate); however, this difference was not significant. These results are shown in Table 1.

With regard to interrater agreement, there was a significant increase after training between listeners 1 and 2, from 0.35 (fair) to 0.57 (moderate), and between listeners 2 and 3, from 0.26 (fair) to 0.63 (good). There was no significant difference between listeners 1 and 3, which agreement varied from 0.52 to 0.44, remaining moderate. The agreement among the three listeners increased significantly, from 0.37 (fair) before training to 0.54 (moderate) after training, as shown in Table 2.

DISCUSSION

Hypernasality is a common symptom in individuals with cleft palate, and it is considered the most representative symptom of velopharyngeal dysfunction. The detection and, more importantly, the classification of the severity of this symptom are made subjectively, using the human ear and the perception of the listener as tools. When listeners classify a certain aspect of voice, they compare the presented *stimulus* to an internal standard. Such internal standards are developed over time, are preserved in the memory of individuals and differ from one listener to another. Moreover, they are inherently unstable and may be influenced by factors, such as lapses in memory and attention, and by external variables such as articulation, severity of the speech symptoms, *stimulus* presentation, vocal intensity, and phonetic context, among others previously described[23]. To minimize the effect of internal standards on the speech assessment, this study used listener training as a strategy, as recommended in literature, and compared the results obtained before and after that training. It is noteworthy that the type of analysis performed in this study, which classifies a single speech symptom (hypernasality) by using a scale of four categories, is one of the most complex to implement; that is, the listener was required to not only identify the symptom but also to rate it and, in some cases, doing so in the presence of combined symptoms.

As many authors assert that experience is a key factor to obtain a reliable result in the auditory perceptual assessment[7,18], in conducting this study, special attention was given to the invitation of professionals who are experienced in the assessment of speech of individuals with cleft lip and palate to participate. The three speech–language pathologists who assessed the speech samples in this study revealed at least 12 years of experience in the field and carried out the task of individually classifying hypernasality in the two phases of the study. The agreement found among the three listeners in the first phase was 0.37, which was interpreted as a fair agreement. This result confirms the difficulty in obtaining high agreement in the perceptual assessment of hypernasality among different listeners, as shown in the literature. That is most likely explained by the fact that listeners used their own standards to assess the symptom, which, as it is known, may differ even among experienced listeners[7,11]. The use of personal criteria may also explain the intrarater agreement obtained in the first

Table 1. Intrarater agreement percentage, kappa coefficient, and its interpretation obtained in the perceptual analysis of hypernasality: statistical comparison of pre- and post-training kappa coefficients

Listeners	Intrarater agreement						p-Value
	Pre-training			Post-training			
	Agreement %	Kappa coefficient	Interpretation	Agreement %	Kappa coefficient	Interpretation	
1	53	0.38	Fair	71	0.61	Good	0.330
2	59	0.39	Fair	94	0.92*	Very good	0.004
3	82	0.76	Good	65	0.50	Moderate	0.234

*Pre- versus post-training: Z-test

Table 2. Agreement percentage, kappa coefficient, and its interpretation obtained among listeners in the perceptual analysis of hypernasality: statistical comparison of pre- and post-training kappa coefficients

Listeners	Interrater agreement						p-Value
	Pretraining			Posttraining			
	Agreement %	Kappa coefficient	Interpretation	Agreement %	Kappa coefficient	Interpretation	
1 and 2	52	0.35	Fair	69	0.57*	Moderate	0.009
1 and 3	65	0.52	Moderate	58	0.44	Moderate	0.398
2 and 3	44	0.26	Fair	73	0.63*	Good	0.001
1, 2, and 3	31	0.37	Fair	51	0.54*	Moderate	0.044

*Pre- versus post-training: Z-test

phase of the study. Except for one listener, who obtained a good agreement, the other listeners achieved a fair agreement. Some authors believe that listeners' internal standards may be unstable even for single listener, regardless of the experience level[11,14]. That is one of the reasons that led some clinicians and researchers to support the use of prior listener training as a strategy to increase the reliability of auditory perceptual assessments[14,18,24].

Several programs of training have been documented in literature, mostly with positive results. A large number was used in the analysis of vocal aspects, such as roughness and breathiness[14,15,17,18,20], and only a few in the analysis of nasality[6,7,16,19]. One of the more effective training strategies mentioned in literature is the use of the reference-matching, as the one that used in this study[20]. According to the mentioned authors, the references are effective in the establishment of internal standards, as listeners become familiar with the references used in training and eventually store these references within their memories as internal standards. In other words, once experienced, these representations are stored within the memory as examples.

Results obtained after training proved this theory. Findings show that the intrarater agreement coefficient increased from fair to very good. The same happened with the interrater agreement, which significantly increased from fair to moderate among the three listeners. Similar results, such as the moderate agreement in the assessed classification of hypernasality by different listeners[8] and the moderate to good agreement in the assessment of hypernasality and hyponasality carried out by experienced listeners[25], were reported by other researchers in the field. A recent study conducted at the HRAC-USP[26] also showed agreement ranging from moderate to good. However, in this case, the analysis made by listeners was based on the presence or the absence of hypernasality and not on the grading of the symptom. From the studies that classified hypernasality using a four-point scale, most found agreement that are similar to those reported in this study, ranging from fair to moderate[21], moderate to good[27], and moderate[28]. With regard to the use of training and this study, other researchers also reported improvement in intra- and interrater reliability after listener training[15]. Similar to what occurred in this study, combining training and the use of models significantly improved interrater reliability in perceptual assessment of voice[17,18]. However, these authors did not observe any significant effects of the use of anchors or listener experience on intrarater agreement. In this study, this agreement increased for two of the listeners, and there was a significant difference for one of them.

As previously noted, the majority of studies regarding anchor and training investigated their effects on vocal symptoms, considered by experts as *stimuli* that vary in terms of quality. The use of these strategies on speech symptoms resulting from cleft lip and palate, as nasality, a *stimulus* that varies in terms of magnitude, has been little investigated to date[6,7]. It should also be noted that no other study in literature compared the performance of the same group of listeners before and after training in the analysis and the rating of nasality, as it was carried out in this study. There was a significant difference between pre- and post-training phases, with increased agreement after training in most of the comparisons conducted with the same listener group. The comparison of intrarater rates showed an increase for two of the three listeners, revealing a significant increase for one of them, which indicates that the provided training was effective in establishing the listener's internal standard. However, one of the listeners showed a decrease of this index, although a significant difference was not determined in this case. An explanation for such result may be the influence, proven in literature, of internal factors such as attention, memory, and even fatigue. With regard to the comparison between interraters, the agreement also showed a significant increase from fair to moderate and from fair to very good. In one of the comparisons, there was a slight reduction after training, although still interpreted as moderate; that was the most expressive rate obtained in the comparisons made between these same listeners in the pre-training phase (0.52-moderate).

By considering the use of planned training to establish standards and anchors for the perceptual assessment of nasality, as conducted in this study, even better intra- and interrater agreement than those obtained after training could be expected. A explanation for such a result can be the use of the ordinal scale, which was applied in this study and used in clinical setting to rate nasality. Although this type of scale is the most widely used method, there are concerns as to their validity for the assessment of hypernasality both in research and in clinical practice. The reason for those is that the ordinal scale divides the categories without, however, to quantify the magnitude of the difference between each category and listeners tend to subdivide the lower end of the scale into smaller intervals. Several authors suggest that nasality is a sensation that is mentally processed as a prothetic dimension, i.e., it differs in terms of degree or quantity (magnitude). According to Stevens[29], when assessing prothetic *stimuli*, listeners do not perceive the intervals between categories as equal in different points of the scale. Therefore, "equal-appearing intervals" are not "necessarily equal" for the whole scale. Hence, the equal-appearing interval scaling may not be as effective to rate nasality, regardless of the use of prior training. Some authors argue that nasality would be better rated with ratio-based scales, as the direct magnitude estimation and the visual analog scaling, which enable more valid and reliable classifications for the perception of nasality[6,30]. However, there are authors who consider the direct magnitude estimation to be impractical for the clinical setting, because the speech sample to be classified should be compared with a standard sample[9]. Conversely, the visual analog scaling has been employed and supported by others. A recent study[2] showed that this type of scale offers a greater reliability than the equal-appearing interval scaling in the perceptual assessment of the cleft speech symptoms, as hypernasality and audible nasal air emission,

and suggests its use as an alternative method to assess these speech parameters.

In summary, despite the difficulty in reaching a consensus on the severity of hypernasality, this study has shown that prior training is an effective strategy to increase the agreement among different listeners and, thus, improve the reliability of the perceptual assessment of speech, a method that remains as the main indicator of the clinical significance of the speech symptoms. Additional studies are being done at the Laboratory of Physiology of HRAC-USP, using new methods of perceptual assessment, in the attempt to find those which enable more reliable and reproducible assessment of hypernasality.

CONCLUSION

Previous listener training leads to the increase of the intra-and interrater agreement and, consequently, the improvement of the reliability of the perceptual assessment of hypernasality in individuals with cleft palate. These results reinforce the importance of establishing standardized criteria to decrease the influence of individual internal standards in the perceptual assessment of the speech symptoms.

REFERENCES

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.
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.

Author contributions

ACASFO was in charge of the study, data analysis and collection, and article composition; RHS cooperated with the data collection and article composition; APF cooperated with the data analysis and article composition; RPY was in charge of the project, study design and overall orientation for the implementation phases, and manuscript preparation.