CoDAS
ISSN 2317-1782 (Online version)

Original Article
Artigo Original

Effect of synthesized voice anchors on auditory-perceptual voice evaluation

Efeito de emissões âncoras de vozes sintetizadas na avaliação perceptivo-auditiva da voz

Priscila Campos Martins dos Santos[1]
Maurílio Nunes Vieira[2]
João Pedro Hallack Sansão[3]
Ana Cristina Côrtes Gama[1]

Keywords

Voice
Voice Disorders
Voice Quality
Dysphonia
Auditory Perception
Voice Training

ABSTRACT

Purpose: To analyze if the use of synthesized voice anchor emissions in auditory-perceptual evaluation improves intra- and inter-rater agreement. **Methods:** This is a quantitative study. Thirty-two inexperienced evaluators were selected and performed two activities on a Programming Interface created by the authors: Active Calibrator Activity — auditory-perceptual evaluation of the roughness and breathiness parameters as 0–no deviation, 1–slight deviation, 2–moderate deviation, or 3–intense deviation of 25 voices with the support of anchored emissions of synthesized voices; and Inactive Calibrator Activity — auditory-perceptual evaluation of these same voices without the support of anchored vocal emissions. The voices were randomized for each activity, and the order of the activities was drawn randomly for each evaluator. The second activity was performed 15 days after the first. The Kappa coefficient was used to analyze intra- and inter-rater agreement, and the confidence interval (CI) was employed to compare concordances. **Results:** Inter-rater agreement was higher for the intense degree of the breathiness parameter in the Active Calibrator Activity when compared to the Inactive Calibrator Activity, as well as the intra-rater agreement of the roughness parameter.333 **Conclusion:** Use of anchor emissions of synthesized voices directly in the evaluation improves intra- and inter-rater agreement in auditory-perceptual voice analysis.

Descritores

Voz
Distúrbios da Voz
Qualidade da Voz
Disfonia
Percepção Auditiva
Treinamento da Voz

RESUMO

Objetivo: Analisar se a utilização de emissões âncoras de vozes sintetizadas na avaliação perceptivo-auditiva melhora a concordância intra e interavaliador. **Método:** Trata-se de um estudo de natureza quantitativa. Foram selecionados 32 avaliadores inexperientes que realizaram, em um aplicativo criado pelos autores, duas atividades: Atividade Calibrador Ativo – avaliação perceptivo-auditiva dos parâmetros rugosidade e soprosidade como 0-ausência de desvio, 1-desvio leve, 2-desvio moderado ou 3-desvio intenso de 25 vozes com o apoio de emissões âncoras de vozes sintetizadas; e Atividade Calibrador Inativo – avaliação perceptivo-auditiva dessas mesmas vozes sem o apoio de emissões vocais âncoras. As vozes foram aleatorizadas em cada atividade, e a ordem de realização das atividades foi sorteada para cada avaliador, sendo que a segunda atividade foi realizada 15 dias após a primeira. Para análise da concordância intra e interavaliadores foi utilizado o coeficiente Kappa, e para comparação entre as concordâncias foi utilizado o intervalo de confiança (IC). **Resultados:** A concordância interavaliadores foi maior para o grau intenso do parâmetro soprosidade na Atividade Calibrador Ativo quando comparada à Atividade Calibrador Inativo, assim como a concordância intra-avaliadores do parâmetro rugosidade. **Conclusão:** O uso de emissões âncoras de vozes sintetizadas diretamente na avaliação melhora a concordância intra e interavaliadores na análise perceptivo-auditiva da voz.

Correspondence address:
Priscila Campos Martins dos Santos
Departamento de Fonoaudiologia da Faculdade de Medicina da Universidade Federal de Minas Gerais – UFMG
Av. Professor Alfredo Balena, 190, sala 249, Santa Efigênia, Belo Horizonte (MG), Brasil, CEP: 30130-100.
E-mail: priscila.fonoaudiologia@gmail.com

Received: August 13, 2019

Accepted: March 25, 2020

Study conducted at Faculdade de Medicina, Universidade Federal de Minas Gerais – UFMG - Belo Horizonte (MG), Brasil

[1] Departamento de Fonoaudiologia, Faculdade de Medicina, Universidade Federal de Minas Gerais – UFMG - Belo Horizonte (MG), Brasil.

[2] Departamento de Engenharia Eletrônica, Escola de Engenharia, Universidade Federal de Minas Gerais – UFMG - Belo Horizonte (MG), Brasil.

[3] Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades, Universidade Federal de São João Del Rei – UFSJ - Ouro Branco (MG), Brasil.

Financial support: Fundação de Amparo à Pesquisa do Estado de Minas Gerais – Fapemig (APQ-02594-15) and Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brasil – CNPq (nº309108/2019-5).

Conflict of interests: nothing to declare.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Perceptual-auditory analysis has been the main tool for assessing voice quality in Speech-Language Pathologists clinics and research due to its advantages: it allows perceptual descriptions that cover various vocal parameters; it is a quick, painless and comfortable method for the patient; it does not depend on equipment, and so is low cost[1]. However, the vocal quality characterized by more than one concomitant parameter is a frequent factor and makes this assessment complex. The evaluator needs to distinguish aurally the parameters in the same voice and isolate them so that they can make their analyses, which can be influenced by their internal standards, built from previous experiences and training[2-5]. This subjectivity, which is a disadvantage of auditory-perceptual analysis, generates high variability in intra- and inter-rater agreement, impairing the reliability of this assessment[6-8].

Recent studies have pointed out the use of anchor voice emissions in perceptual-auditory training of voice assessment as a useful tool to increase the reliability of this assessment[8,9]. Anchor vocal emissions are voice stimuli selected in agreement between at least two evaluators to be used as references for a given parameter and degree of vocal deviation[10-12]. The voices used as anchors can be natural, that is, human voices; or synthesized, which are created from mathematical calculations. The main advantage of using human voices as anchor emissions is their naturalness. However, this naturalness is associated with the fact that voices are generally characterized by more than one parameter concomitantly, which can be pointed out as the main disadvantage of using this type of emission, as it makes it difficult to classify the voices. In contrast, despite presenting the artificiality of the voices as a disadvantage, sometimes with robotic and unnatural features, synthesized vocal emissions have as their main advantage the possibility of manipulating acoustic parameters as desired or needed, allowing analysis of each vocal parameter separately. Therefore, it is believed that synthesized vocal emissions are the ideal type to be used as anchors in perceptual-auditory voice training[7].

Several studies have used synthesized voice anchor emissions in auditory-perceptual training and analyzed their effect on intra- and inter-rater reliability in the assessment of vocal quality[6,8,13]. A survey conducted with inexperienced evaluators[13] has shown that the use of anchor vocal emissions in training improved intra- and inter-rater reliability in post-training evaluation.

When comparing use of anchors to the pairing method in the training of experienced assessors, researchers observed that both methods facilitated auditory-perceptual assessment, with a significant improvement in the accuracy of assessment after training[8]. However, they realized that use of anchor vocal emissions in training allows this reference to be memorized and retrieved during auditory-perceptual assessment tasks, as it is a method more similar to the assessment of vocal quality than the pairing method.

These same authors analyzed, in another study[6], the effect of anchor emissions of both natural and synthesized voices on perceptual-auditory training, and pointed out that, when anchors are associated with training, they stabilize the internal standards of the evaluators, improving evaluation reliability. They also concluded that anchor emissions from synthesized voices proved to be more reliable than natural voice anchors.

Inexperienced raters showed the same degree of intra- and inter-rater reliability as experienced raters in a study that used synthesized anchor stimuli in two different types of training: one grading vocal stimuli according to the magnitude of the deviation, from the most to the least altered; and another organizing vocal stimuli into categories according to degree of deviation[14].

Given the abovementioned observations, anchor vocal emissions have often been associated with perceptual-auditory training for further analysis of their effect on voice assessment[9,10]. However, few studies have analyzed the use of anchor vocal emissions directly in voice assessment[11,15]. It is reasonable to assume that use of these anchor emissions during auditory-perceptual voice assessment would eliminate the need for prior memorization of reference voices through previous or periodic training, as well as reduce the influence of evaluators' internal standards on the vocal classification, as raters would have reference emissions at their disposal[15], just as an instrumentalist uses the stimuli offered by a tuner as a reference when tuning their instrument. Synthesized anchor voice emissions would facilitate differentiation of the evaluated parameters and their respective degrees of deviation, as they allow analysis of an isolated parameter, which is generally not possible with human voice anchors[8,16]. Therefore, the present study aimed to analyze whether the use of synthesized anchor voice emissions improves intra- and inter-rater reliability in auditory-perceptual assessment.

METHODS

This research was approved by the Research Ethics Committee (COEP) under number 920866. This is a quantitative study.

Before starting, the evaluators read the Free and Informed Consent Form (ICF) and selected the option "I Accept" to participate in the form. Then they answered a brief questionnaire providing data on their experience in auditory training and age, and received an initial presentation of the research. Finally, the 32 evaluators performed the auditory-perceptual evaluation of 30 vocal emissions.

Two activities were created by the researchers for auditory-perceptual assessment and provided in an application, designed by the researchers for this study and provided only to participants at the time of collection. In the so-called Active Calibrator Activity, evaluators assessed the voices with support from anchor emissions from synthesized voices; and in the Inactive Calibrator Activity evaluators assessed the voices without this support. A four-point scale was used in both activities to gauge roughness (R) and breathiness (B): (0–absence of deviation, 1–slight degree of deviation, 2–moderate degree of deviation and 3–intense degree of deviation). Vocal quality was considered as roughness when there was any noticeable irregularity during vocal production, and as breathiness when there was an audible air leak during voice production[17].

The activities were named as Auditory Calibrator, as the synthesized voice anchor emission available during the perceptual-

auditory evaluation is similar to the stimuli offered by a tuner as a reference for a musician when tuning their instrument. Therefore, in an activity in which synthesized voice anchor emissions are present, the Calibrator is Active — and it was named Active Calibrator Activity, while in an activity in which synthesized voice anchor emissions are absent the Calibrator is Inactive — and it was named Inactive Calibrator Activity.

The order in which activities were carried out was drawn randomly for each participant, and the second activity was performed precisely 15 days after the first activity (Figure 1). The literature records the use of an interval of at least one week between assessment activities in order to avoid any memorization[18-20].

The activities will be described below.

Active Calibrator Activity

The activity that used synthesized voice anchor emissions for the auditory-perceptual assessment was named Active Calibrator Activity.

Process

During this activity, each voice was evaluated first according to the R parameter and then according to the B parameter. For this, evaluators were instructed to perform the following procedures: 1. Listen to the natural voice to be evaluated; 2. Listen to the anchor emissions of synthesized voices for each degree of parameter R; 3. Listen again to the voice to be evaluated; 4. Indicate in the field in front of the "degree of roughness" icon the number corresponding to the degree of voice classification for parameter R, where 0–no deviation, 1–slight deviation, 2–moderate deviation or 3–intense deviation (Figure 2). Repeat the same procedures to classify the same voice for parameter B.

The written definition of the parameters was available at all times during the Active Calibrator Activity.

Selection of vocal emissions for evaluation

To compose the sample of natural voices to be assessed, the voice bank of a university outpatient clinic was used, consisting of 381 voices, samples of the emission of the vowel /a/ sustained habitually, from individuals of both genders aged over 18 years. Two evaluators, Speech-Language Pathologists and voice specialists, with over five years of experience in auditory-perceptual evaluation, individually analyzed the voices using the *Multilaser Vibe Headphone* model stereo supra-headset. They classified the voices according to the predominant parameter, R or B, and the general degree of vocal deviation (0–no deviation, 1–slight deviation, 2–moderate deviation, 3–intense deviation), using the GRBASI scale.

The following inclusion criteria were considered: natural voices from female and male subjects, aged 18 and over, with a predominant parameter of varying degrees of vocal deviation; voices that received the same classification from both evaluators.

Three vocal emissions were selected for each degree of the predominant parameters R and B, and a degree of one of the parameters was exemplified by four vocal emissions to reach the N previously found through sample calculation, with a total of 25 voices. In order to define the parameter and degree that would receive an additional sample, a draw was carried out, and the light degree of the breathiness parameter was selected. 20% of the voices were added in order to analyze intra-rater reliability, totaling 30 vocal emissions. The evaluators did not

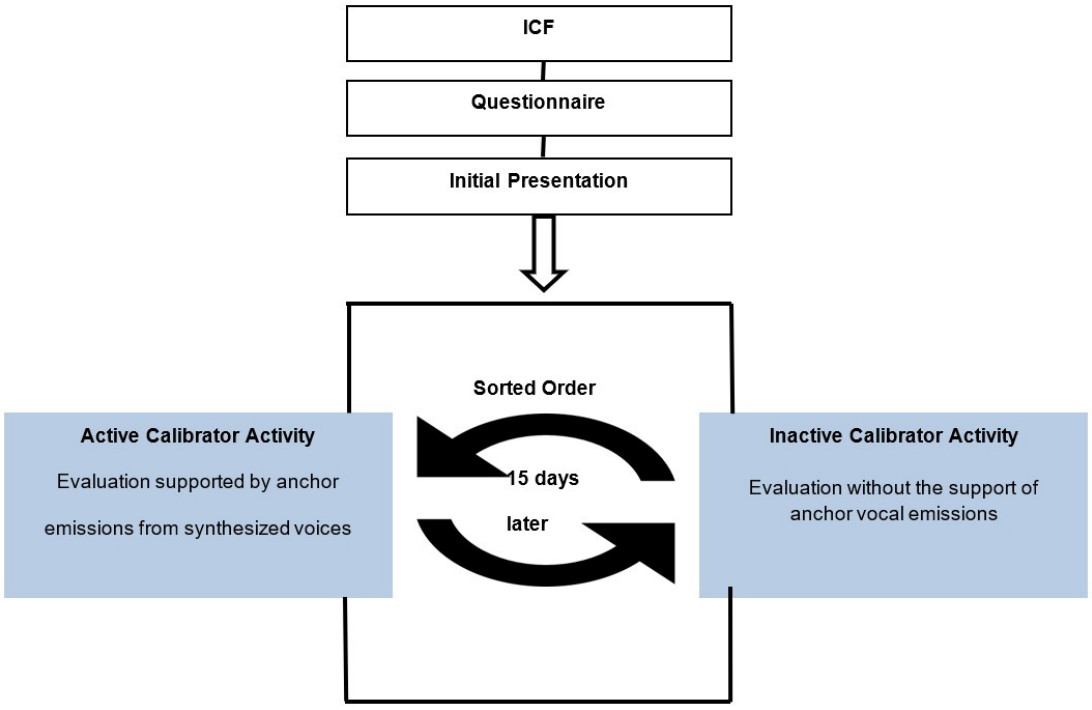

Figure 1. Auditory Calibrator Flowchart

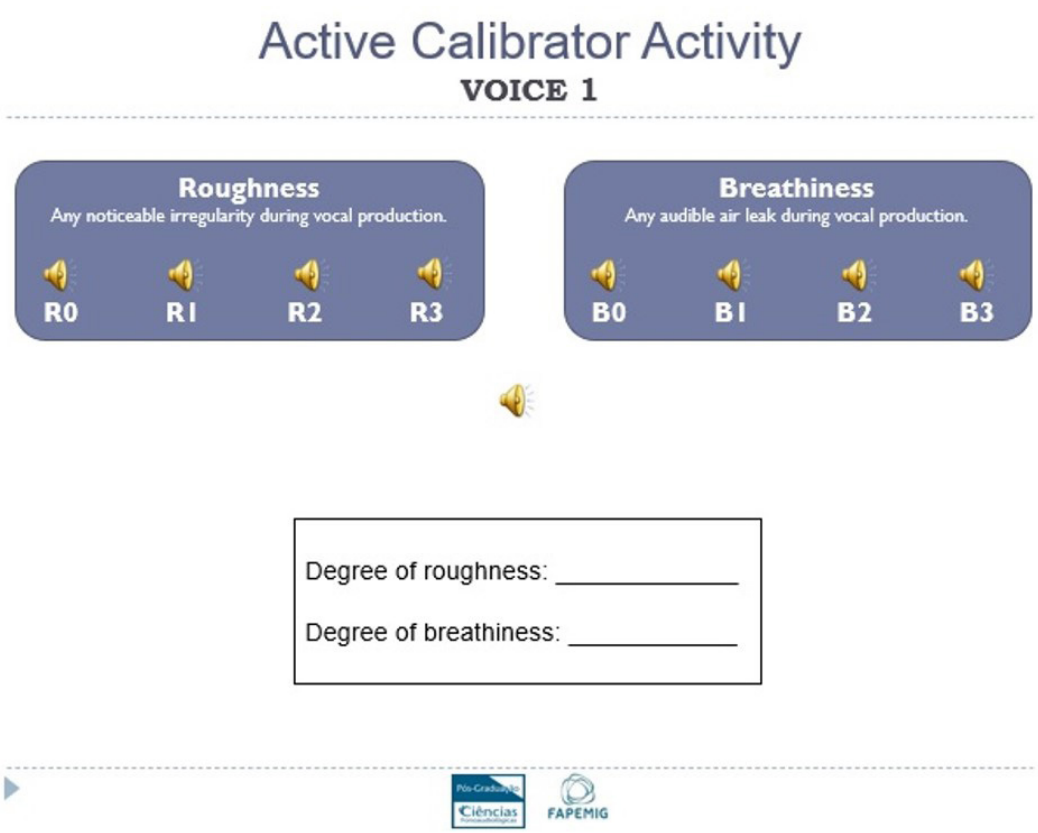

Figure 2. Application Active Calibrator Activity

know how many emissions there were in each grade, nor did they know that there were duplicate voices.

The voices were identified by numbers at all stages of the research.

Selection of anchor vocal emissions for training

The sample of anchor vocal emissions was composed of synthesized voices. A parametric model was used as the source (glottal flow) for creation of the synthesized neutral voices (N) or those containing the R or B parameter with different degrees of vocal deviation, allowing control of the fundamental frequency, jitter, shimmer and signal-to-noise-ratio. Manipulation of these measures gave the voices their characteristics of roughness or breathiness. A vocal tract model of the vowel /a/ was used as a filter, extracted from a natural voice using the linear prediction technique. The vocal emissions were created by an engineer, totaling 300 synthesized voices[21].

To analyze the degree of naturalness and the quality of synthesized voices, three evaluators were selected, Speech-Language Pathologists with over five years of experience in voice assessment, who individually performed the analysis of each voice according to three aspects. First, an auditory-perceptual analysis of the voice's naturalness (related to how much the listener perceives the voice as human) was done, indicating on a 100-mm visual analog scale (VAS) how much they considered that voice as natural, where zero was unnatural and 10, indicated maximum naturalness. The voice was then classified as neutral, rough or breathy. Finally, the degree of vocal deviation for the parameter in which it was previously classified (R or B) was also measured, using a 100 mm VAS. Values found for the vocal deviation of the voices classified as R or B using VAS were then converted as suggested by the literature[22], as shown in Table 1.

Synthesized voices of different degrees of deviation, classified as most natural by at least two evaluators, were selected as anchors for each parameter. The sample of anchor vocal emissions was composed by an emission of each degree — absence of deviation, slight, moderate, and intense deviation of each parameter — R and B, totaling eight voices.

Neutral voices or those with less vocal deviation were classified as more natural for both parameters, their natural character decreasing as the degree of deviation increased (Table 2). For the R parameter, the voice classified as having no deviation was rated as more natural, followed by the voices classified with a slight, moderate, and intense degree of deviation. Regarding parameter B, the voice with a slight degree of deviation was classified as more natural, followed by the one with no deviation and, finally, by those with moderate and intense deviation. The voices selected for the light, moderate and intense degrees of parameter B were more natural than those selected for the same degrees of deviation of parameter R.

Inactive Calibrator Activity

The activity that did not use synthesized anchor voice emissions for the auditory-perceptual evaluation was named Inactive Calibrator Activity.

Table 1. Correlation of vocal deviation classification by visual analog scale and numerical scale

Degree of deviation	Correlation of vocal deviation classification by visual analog scale and numerical scale	
	Rough (mm)	Breathy (mm)
Neutral	0 – 8.5	0 – 8.5
Slight	8.5 – 28.5	8.5 – 33.5
Moderate	28.5 – 59.5	33.5 – 52.5
Intense	From 59.5	From 52.5

Statistical test: Roc curve

Table 2. Average degree of naturalness of the synthesized voices for each perceptual-auditory parameter selected for the sample

Degree of deviation	Classification of naturalness of voices (mm)		
	Neutral	Rough	Breathy
_____	97.3	_____	_____
Slight	_____	56	86
Moderate	_____	41	60
Intense	_____	37	40

Average of the markings made by evaluators in mm on the Visual Analogue Scale regarding the naturalness of the voices

Process

During this activity, each voice was also evaluated first according to the R parameter and then to B parameter. Once more, evaluators were instructed to perform the following procedures: 1. Listen to the natural voice to be evaluated; 2. Indicate in the field in front of the "degree of roughness" icon the number corresponding to the degree of voice classification for parameter R, where 0–no deviation, 1–slight deviation, 2–moderate deviation or 3–intense deviation. The same procedures were repeated to classify the same voice for parameter B.

Selection of vocal emissions for evaluation

The same vocal emissions used in the Active Calibrator Activity were used for the Inactive Calibrator Activity. The voices were randomized for each activity.

For the collection, schedules were arranged in computer labs in different buildings of the educational institution, to facilitate participation of students from the initial periods of the Speech-Language Pathologists course as evaluators, as they take classes in different buildings and full time. The evaluators performed the tasks outside of class hours, attending the laboratories exclusively to carry out the research activities. Prior scheduling was carried out with participants to ensure that each evaluator would have a computer at their disposal in which they would perform the activities individually by accessing the application using the Internet Explorer browser. One of the researchers accompanied the evaluators, providing guidance prior to performance of the activities but without intervening in the tasks themselves. *Stereo Multilaser Vibe Headphone* model headphones were used for all procedures. Evaluators could listen to the voices as many times as they deemed necessary, provided that they respected the order of procedures.

The researcher who accompanied the evaluators noted that the Inactive Calibrator Activity lasted approximately twenty minutes, although the session duration was not recorded. The Active Calibrator Activity had a slightly longer duration when compared to the Inactive Calibrator Activity.

Selection of evaluators

A sample calculation was performed to determine the number of 32 evaluators, considering 25 observations (voices to be evaluated) and eight variables (parameters R and B with no deviation, slight, moderate, and intense deviation), using the Kappa test proposed by Fleiss, with a statistical power of 80% and a significance level of 5%.

Thirty-two individuals were selected to evaluate the voices, 27 female and five male. All were students from the first to the third period of the undergraduate course in Speech-Language Pathologists, with no experience or previous training in perceptual auditory voice assessment, aged 17 to 24 years old (average = 19.66 years). The following inclusion criteria were considered: answering the initial questionnaire, participating in all activities, having no previous experience in perceptual auditory voice assessment, and absence of hearing complaints.

At no time were the evaluators identified.

The Kappa coefficient was used to analyze intra- and inter-rater agreement, and the confidence interval (CI) was used to compare reliability. The software Stata version 12 was used to perform the statistical analysis. A significance level of 5% was considered in all analyzes.

RESULTS

Although there is no difference, observing the CIs (Table 3) there was a tendency of increasing inter-rater reliability for grades 0, 1 and 2 of the R parameter as well as decreasing it for grade 3 of this same parameter in the Activity Active Calibrator — that performed with anchor emissions of synthesized voices — when compared to reliability in the Inactive Calibrator Activity, that done without voice anchor emissions, considering the same parameter and degrees of deviation (Table 3 and Figure 3).

As for breathiness, there was no difference when observing the CIs (Table 4) of grades 0, 1 and 2. However, it was also possible to see a tendency towards greater inter-rater reliability in the Active Calibrator Activity — that performed with anchored emissions of synthesized voices — than in the Inactive Calibrator Activity, done with no voice anchor emissions for these degrees. Inter-rater reliability for breathiness grade 3 was statistically higher in the Active Calibrator Activity when compared to the Inactive Calibrator Activity (Table 4 and Figure 4). It could be observed that inter-rater reliability was higher for grades 0 and 3 of the two parameters evaluated (Figures 3 and 4).

Intra-rater reliability was statistically higher for the roughness parameter in the Active Calibrator Activity when compared to the Inactive Calibrator Activity (Table 5). There was also greater

reliability in the Active Calibrator Activity for the breathiness parameter, although no difference was observed (Table 5 and Figure 5).

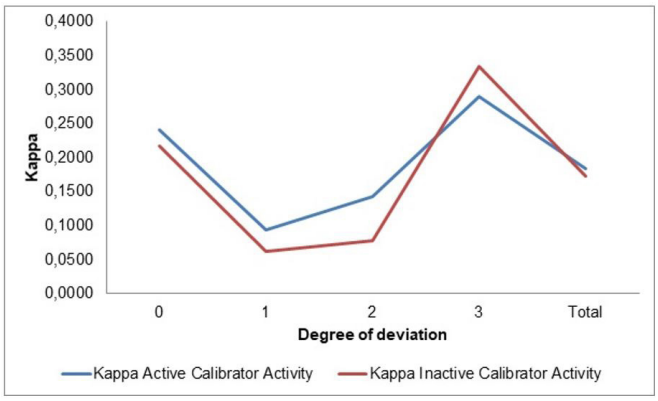

Figure 3. Comparison between inter-rater reliability in the Active Calibrator Activity — with anchor emissions of synthesized voices — and Inactive Calibrator Activity, without voice anchor emissions, for each degree of deviation regarding the Roughness parameter, using the weighted Kappa coefficient

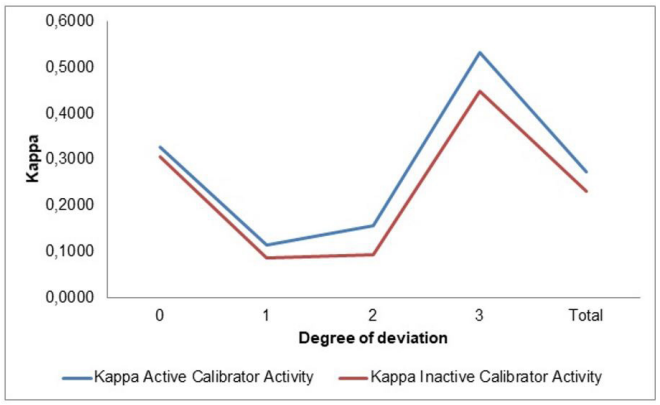

Figure 4. Comparison between inter-rater reliability in the Active Calibrator Activity — with synthesized voice anchor emissions — and Inactive Calibrator Activity, without anchor vocal emissions, for each degree of deviation regarding the Breathiness parameter, using the weighted Kappa coefficient

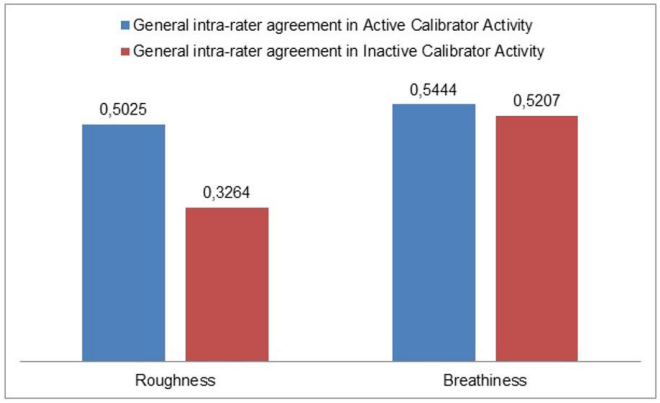

Figure 5. Comparison between intra-rater reliability in the Active Calibrator Activity — with anchor emissions of synthesized voices — and Inactive Calibrator Activity, without voice anchor emissions, for the parameters Roughness and Breathiness, using the weighted Kappa coefficient

DISCUSSION

In the present study we opted for using synthesized voices as anchors. Research suggests that it is possible to reduce variability in the classification of vocal quality by replacing the unstable internal patterns of the listeners with external patterns, such as anchor voices, or reference voices for different vocal qualities, which can be compared to the voice sample to be evaluated[4,7,9--12,23]. Use of synthesized voices allows listening to each vocal parameter in isolation during the assessment, facilitating their perception[7]. We also opted for inexperienced raters in order to eliminate the influence of any previous experience or training as well as internal standards, making it possible to analyze exclusively the effect of the anchor on the assessment.

Despite the promising use of synthesized voices, this is still not a common practice due to the difficulty of producing voices that seem natural to the listener. Therefore, to select the synthesized voices, classification of the voices for naturalness was previously performed for each of the parameters, to ensure

Table 3. Inter-rater reliability of the Active Calibrator Activity — with anchor emissions of synthesized voices — and the Inactive Calibrator Activity, without anchor vocal emissions, for each degree of deviation regarding the Roughness parameter, using the Kappa coefficient

Degree	Active Calibrator Activity			Inactive Calibrator Activity		
	Kappa	CI		Kappa	CI	
0	0.2412	0.1947	0.2877	0.2177	0.1698	0.2656
1	0.0943	0.0388	0.1498	0.0619	0.0044	0.1194
2	0.1421	0.0895	0.1947	0.0778	0.0213	0.1343
3	0.2898	0.2463	0.3333	0.3346	0.2938	0.3754
Total	0.1846	0.1346	0.2346	0.1724	0.1216	0.2232

For statistical analysis, the weighted Kappa coefficient and the confidence interval (CI) were considered

Table 4. Inter-rater reliability of the Active Calibrator Activity — with anchor emissions of synthesized voices — and the Inactive Calibrator Activity, without anchor vocal emissions, for each degree of deviation regarding the Breathiness parameter, using the Kappa coefficient

Degree	Active Calibrator Activity			Inactive Calibrator Activity		
	Kappa	CI		Kappa	CI	
0	0.3279	0.2867	0.3691	0.3060	0.2635	0.3485
1	0.1147	0.0604	0.1690	0.0850	0.0289	0.1411
2	0.1572	0.1055	0.2089	0.0927	0.0371	0.1483
3	0.5321	0.5034	0.5608	0.4498	0.4161	0.4835
Total	0.2738	0.2293	0.3183	0.2313	0.1842	0.2784

For statistical analysis, the weighted Kappa coefficient and the confidence interval (CI) were considered.

Table 5. Intra-rater reliability of the Active Calibrator Activity — with anchor emissions from synthesized voices — and of the Inactive Calibrator Activity, without anchor vocal emissions, regarding the parameters Roughness and Breathiness, through the Kappa coefficient

	Active Calibrator Activity			Inactive Calibrator Activity		
	Kappa	CI		Kappa	CI	
Roughness	0.5025	0.4862	0.5188	0.3264	0.3105	0.3423
Breathiness	0.5444	0.5284	0.5604	0.5207	0.5047	0.5367

For statistical analysis, the weighted Kappa coefficient and the confidence interval (CI) were considered

that the most natural voices were selected for the present study. High-quality synthesized voice samples were achieved mainly in the degrees of absence of deviation and slight deviation of the roughness (R) and breathiness (B) parameters, but naturalness decreased as the degree of vocal deviation increased. Another study pointed out the high quality of the synthesized voices, showing greater accuracy in the classification of the voices as synthesized for more intense degrees of the same parameters[24]. Discrepancies between studies can be justified by methodological issues. These studies developed the synthesized voices using different mathematical methods; while the present research analyzed the degree of naturalness, the literature[24] reviewed evaluated which voices, taken from a bank of human and synthesized samples, were correctly identified. The different ways of assessing naturalness in the two investigations probably impacted the results. Future investigations are necessary for better understanding of the auditory perception of synthesized voices when compared to human vocal emissions.

A study in which anchor emissions were used directly in auditory perceptual assessment of voices[11] selected three groups of evaluators, both experienced and inexperienced. The parameters evaluated, general degree of vocal deviation and vocal effort, were classified as grades 1, 2 or 3. A 100 mm visual analog scale (VAS) was used to assess and anchor natural voice emissions. Two groups, composed of inexperienced and experienced evaluators, evaluated the voices along a VAS, first without the support of voice anchor emissions and later with the anchor; a third group, a control team of inexperienced evaluators, performed the evaluation only supported by anchors. Intra- and inter-rater reliability were significantly higher in the evaluation with anchor vocal emission support for the two parameters evaluated.

Another study[15], conducted with anchors in the evaluation, used synthesized voice emissions. Only the roughness parameter was analyzed by experienced evaluators in two evaluations. In the first assessment, the evaluators listened to the voices without support from synthesized voice anchor emissions and classified them on a five-point scale, in which one indicated a normal voice and five defined the intense degree of roughness. In the second assessment, each point on the five-point scale was represented by a synthesized voice, anchor emission. The participant would listen to the synthesized anchors twice and then to the voice to be evaluated. After that, they would select the synthesized voice anchor emission with the classification most similar to the voice under assessment. Evaluators could listen to the voices as many times as deemed necessary and were instructed to ignore other deviations present in the voice, focusing only on roughness. There was a high level of reliability for the two scales. However, intra- and inter-rater reliability were significantly higher in the assessment using the anchored scale. The study also showed that two evaluators will agree significantly more on the anchored scale than on the scale without anchors.

In the present study, inter-rater reliability for the roughness parameter tended to increase in the Active Calibrator Activity — with anchored emissions of synthesized voices for grades 0, 1 and 2 of the R parameter — when compared to reliability in the Inactive Calibrator Activity — without voice anchor emissions for the same parameter and degrees, although there is no difference when observing the CIs. The result corroborates the literature[15] that points to a significantly higher inter-rater reliability for roughness in an analysis carried out by experienced evaluators with support from voice anchor emissions when compared to the evaluation without anchors, although the study did not quote the reliability by degree of vocal deviation for roughness. The literature[25] points out that the greater the degree of vocal deviation, the greater the reliability of the assessment. However, in the present study, grade 3 of the R parameter tended to be lower in the Active Calibrator Activity as compared to the Inactive Calibrator Activity. This finding may be related to the complexity of the R[19] parameter, which involves different vocal qualities, such as hoarseness, harshness, crackling and bitonality, which may have favored the different perception among evaluators regarding the parameter and contributed to reduce reliability between them.

As for breathiness, there was no difference in the present study when observing the CIs (Table 4) of grades 0, 1 and 2. However, there is also a tendency to increase inter-rater reliability in the Active Calibrator Activity when compared to the Inactive Calibrator Activity. Inter-rater reliability for breathiness grade 3 was statistically higher in the Active Calibrator Activity. No studies were found in the literature in which anchor emissions from synthesized voices were used directly for evaluation of the breathiness parameter. However, a study in which this same parameter was evaluated after training with anchor vocal emission found a significant increase in inter-rater reliability[13]. Moreover, according to the literature[25], intense vocal deviations favor greater inter-rater reliability, which corroborates this finding.

Intra-rater reliability was statistically higher in the Active Calibrator Activity when compared to that in the Inactive Calibrator Activity for the roughness parameter in the present study. This result corroborates the literature[15], which points out a significantly higher intra-rater reliability for roughness in evaluations carried out with the support of voice anchor emissions when compared to evaluation without anchors. This finding also shows that, despite the disagreement among evaluators in the perception of the R parameter, use of the anchor favors stabilization of internal standards, increasing intra-rater reliability.

In the present study there was also a tendency for increase in intra-rater reliability in the Active Calibrator Activity for the breathiness parameter, although no difference was observed. A study in which this same parameter was evaluated after training with anchor vocal emission found a tendency for increase of intra-rater reliability[13], although no difference was also observed. Use of chained speech tasks associated with the sustained vowel could improve perception of this parameter, helping to increase intra-rater reliability, since, according to the literature[26], breathiness is more easily identified in chained speech than in sustained vowels.

In the present study, the Kappa[27] coefficient classification showed a low inter-rater reliability for the R parameter and a regular one for the B parameter, with moderate intra-rater reliability for the two parameters. That is, intra-rater reliability

was greater than inter-rater reliability for the two parameters, a finding that corroborates the literature reviewed[26].

The professional experience of Speech-Language Pathologists impacts positively on inter-rater reliability, suggesting that being experienced in this analysis tends to standardize auditory judgment of dysphonic voices[28]. This relationship was verified in the present study by selecting inexperienced evaluators for the research and offering them the same voice references for evaluation; there was an improvement in inter-rater reliability in the analysis of breathy voices of intense degree and in intra-rater reliability in that of rough voices. However, other studies show that reliability on auditory-perceptual assessment is greater for experienced assessors, due to the previously developed internal standard. A previous study[11] pointed out that experienced evaluators showed less variance in reliability in the evaluation supported by anchor vocal emission. In a second study[29], experienced evaluators showed greater ability to classify human and synthesized voices. Another study[28] pointed out the positive impact of evaluators' experience on inter-rater reliability regarding perceptual-auditory analysis of voices. Still another study[30] showed that individuals experienced in perceptual-auditory analysis of voices seem to have increased capacity in using learning strategies to improve their performance in voice assessment, showing that professional experience positively influences this analysis. Therefore, the importance of carrying out further studies with synthesized voice anchor emissions in the perceptual-auditory assessment with experienced evaluators should be emphasized.

One study[22] points out that evaluators may be more critical in evaluating isolated parameters than in the assessment of the general degree of vocal quality. However, it is important to emphasize that, besides assessing the general degree of vocal quality, the majority of scales used in clinical practice and in Speech-Language Pathologists research, an assessment of the parameters is carried out in isolation. Thus, the use of instruments that improve the perception of isolated parameters through anchor emissions can facilitate the learning process during academic training in Speech-Language Pathologists, as well as help increase intra and inter-rater reliability, improving the reliability of this assessment.

We suggest improvement of the use of anchor emissions for auditory-perceptual evaluation of the voice based on adjustments in future studies, such as: use of connected speech in addition to sustained vowel tasks; definition of more complex parameters, such as roughness; as well as selection of experienced evaluators and application to a larger number of participants in order to obtain increased reliability for degrees and parameters not observed in the present study.

CONCLUSION

The use of synthesized voice anchor emissions in the auditory-perceptual evaluation of voices improved inter-rater reliability in the analysis of breathy voices of intense degree and intra-rater reliability of rough voices. However, we suggest adjustments in future studies to improve the use of anchor emissions and favor both teaching and the clinical practice of auditory-perceptual voice assessment.

ACKNOWLEDGEMENTS

To the support of the Fundação de Amparo à Pesquisa do Estado de Minas Gerais – Fapemig (APQ-02594-15) and of the Conselho Nacional de Desenvolvimento Científico e Tecnológico-Brasil – CNPq (nº309108/2019-5).

REFERENCES

1. Oates J. Auditory-perceptual evaluation of disordered vocal quality: pros, cons and future directions. Folia Phoniatr Logop. 2009;61(1):49-56. http://dx.doi.org/10.1159/000200768. PMid:19204393.
2. Behlau M. Voz: o livro do especialista. Vol. 1. Rio de Janeiro, RJ: Revinter; 2001.
3. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am. 2007;122(4):2354-64. http://dx.doi.org/10.1121/1.2770547. PMid:17902870.
4. Solomon NP, Helou LB, Stojadinovic A. Clinical versus laboratory ratings of voice using the CAPE-V. J Voice. 2011;25(1):e7-14. http://dx.doi.org/10.1016/j.jvoice.2009.10.007. PMid:20430573.
5. Chaves CR, Campbell M, Côrtes Gama AC. The influence of native language on auditory-perceptual evaluation of vocal samples completed by brazilian and Canadian SLPs. J Voice. 2017;31(2):258.e1-5. http://dx.doi.org/10.1016/j.jvoice.2016.05.021. PMid:27427162.
6. Chan KMK, Yiu EML. The effects 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.
7. Yiu EML, Murdoch B, Hird K, Lau P. Perception of synthesized voice quality in connected speech by Cantonese speakers. J Acoust Soc Am. 2002;112(3 Pt 1):1091-101. http://dx.doi.org/10.1121/1.1500753. PMid:12243157.
8. Chan KMK, Yiu EML. 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.
9. dos Santos PCM, Vieira MN, Sansão JPH, Gama ACC. Effect of auditory-perceptual training with natural voice anchors on vocal quality evaluation. J Voice. 2017;33(2):220-5. http://dx.doi.org/10.1016/j.jvoice.2017.10.020. PMid:29331406.
10. 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.
11. 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.
12. Sofranko JL, Prosek RA. The effect of the levels and types of experience on judgment of synthesized voice quality. J Voice. 2014;28(1):24-35. http://dx.doi.org/10.1016/j.jvoice.2013.06.001. PMid:24119637.
13. 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.
14. Gurlekian JA, Torre 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. http://dx.doi.org/10.1016/j.jvoice.2015.05.009. PMID: 26106070.
15. Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. J Speech Hear Res. 1993;36(1):14-20. http://dx.doi.org/10.1044/jshr.3601.14. PMid:8450655.
16. Goldstone RL. Perceptual learning. Annu Rev Psychol. 1998;49(1):585-612. http://dx.doi.org/10.1146/annurev.psych.49.1.585. PMid:9496632.
17. Hirano M. Clinical examination of voice. New York: Springer Verlag; 1981.

18. Helou LB, Solomon NP, Henry LR, Coppit GL, Howard RS, Stojadinovic A. The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. Am J Speech Lang Pathol. 2010;19(3):248-58. http://dx.doi.org/10.1044/1058-0360(2010/09-0012). PMid:20484704.

19. Silva RSA, Simões-Zenari M, Nemr NK. Impacto de treinamento auditivo na avaliação perceptivo-auditiva da voz realizada por estudantes de Fonoaudiologia. J Soc Bras Fonoaudiol. 2012;24(1):19-25. http://dx.doi.org/10.1590/S2179-64912012000100005. PMid:22460368.

20. 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. http://dx.doi.org/10.1016/j.jvoice.2015.01.007. PMid:25795348.

21. Vieira MN, Sansão JPH, Yehia HC. Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms. Speech Communication. 2014;61-62:17-32. http://dx.doi.org/10.1016/j.specom.2014.04.001.

22. Baravieira PB, Brasolotto AG, Montagnoli AN, Silvério KCA, Yamasaki R, Behlau M. Análise perceptivo-auditiva de vozes rugosas e soprosas: correspondência entre a escala visual analógica e a escala numérica. CoDAS. 2016;28(2):163-7. http://dx.doi.org/10.1590/2317-1782/20162015098. PMid:27191880.

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. Englert M, Madazio G, Gielow I, Lucero J, Behlau M. Perceptual error identification of human and synthesized voices. J Voice. 2016;30(5):639.e17-23. http://dx.doi.org/10.1016/j.jvoice.2015.07.017. PMid:26337775.

25. Eadie T, Sroka A, Wright DR, Merati A. Does knowledge of medical diagnosis bias auditory-perceptual judgments of dysphonia? J Voice. 2011;25(4):420-9. http://dx.doi.org/10.1016/j.jvoice.2009.12.009. PMid:20347262.

26. Law T, Kim JH, Lee KY, Tang EC, Lam JH, van Hasselt AC, et al. Comparison of Rater's reliability on perceptual evaluation of different types of voice sample. J Voice. 2012;26(5):666.e13-21. http://dx.doi.org/10.1016/j.jvoice.2011.08.003. PMid:22243971.

27. Altman DG. Some common problems in medical research. In: Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

28. Oliveira SB, Gama ACC, Chaves AR. Interferência do tempo de experiência na concordância da análise perceptivo-auditiva de vozes. Distúrb Comun. 2016;28(3):415-22.

29. Englert M, Madazio G, Gielow I, Lucero J, Behlau M. Perceptual error analysis of human an synthesized voices. J Voice. 2016;31(4): 516.E5-18. https://doi.org/10.1016/j.jvoice.2016.12.015.

30. Englert M, Madazio G, Gielow I, Lucero J, Behlau M. Influência do fator de aprendizagem na análise perceptivo-auditiva. CoDAS. 2018;30(3):e20170107. http://dx.doi.org/10.1590/2317-1782/20182017107. PMid:29898037.

Author contributions

The authors PCMS, ACCG, MNV and JPHS conceived and planned the project, as well as analyzed and interpreted the data and critically reviewed the content of the manuscript.