Universidade de Aveiro's Voice Evaluation Protocol

Luis M. T. Jesus[1], *Anna Barney*[2], *Ricardo Santos*[3], *Janine Caetano*[4], *Juliana Jorge*[5], *Pedro Sá Couto*[6]

[1] IEETA and ESSUA, Universidade de Aveiro, Portugal; [2] ISVR, University of Southampton, UK; [3]Hospital Privado da Trofa, Portugal; [4]Agrupamento de Escolas Serra da Gardunha, Fundão, Portugal; [5]RAIZ, Esmoriz, Portugal; [6]Departamento de Matemática da Universidade de Aveiro, Portugal

lmtj@ua.pt, ab3@soton.ac.uk, ricardosantos_tf@hotmail.com,janine_caetano@hotmail.com, julianarjorge@hotmail.com, p.sa.couto@ua.pt

Abstract

This paper presents Universidade de Aveiro's Voice Evaluation Protocol for European Portuguese (EP), and a preliminary inter-rater reliability study. Ten patients with vocal pathology were assessed, by two Speech and Language Therapists (SLTs). Protocol parameters such as overall severity, roughness, breathiness, change of loudness (CAPE-V), grade, breathiness and strain (GRBAS), glottal attack, respiratory support, respiratory-phonotary-articulatory coordination, digital laryngeal manipulation, voice quality after manipulation, muscular tension and diagnosis, presented high reliability and were highly correlated (good inter-rater agreement and high value of correlation). Values for the overall severity and grade were similar to those reported in the literature.

Index Terms: Speech and Language Therapy, Voice, Protocol, Vocal Quality, Perception, Reliability.

1. Introduction

There has been a clear growth over the last twenty years of studies in the area of voice assessment, but there is still lack of objectivity in the description and evaluation of various aspects of vocal quality. Before establishing a diagnosis and planning an intervention, SLTs must perform an assessment that covers various aspects of voice quality.

Protocols used by most SLTs, include an evaluation of roughness, breathiness, aesthenia, strain, instability and nasality [1-4]. The Buffalo II Voice Screening Profile [5] and the Vocal Profile Analysis Protocol (VPA) [6, 7] have perceptual parameters that are also often used by SLTs. To assess more functional aspects, SLTs also use protocols such as The Boone Voice Program for Adults [8] or the Voice Assessment Protocol [9].

In practice in Portugal, the assessment of vocal pathologies is not uniform, because each Speech and Language Therapist (SLT) uses individual and diverse forms to evaluate. The patient is assessed with scales that try to include the most important and appropriate parameters for each patient, based in existing protocols which have been developed for languages other than EP [8-11].

Universidade de Aveiro's Voice Evaluation Protocol seeks to standardise subjective evaluation of voice quality for EP speakers, and to provide a working tool for SLTs, which brings together a range of essential information, thus preparing patients for a therapeutic intervention. It is intended that SLTs involved in future studies use the same evaluation instrument to acquire data that is comparable, and thereby normalize the nomenclature for this area of intervention, allowing also better inter-professional communication.

The aim of this project was to develop the first standardised and evaluated protocol for voice assessment in EP. A pilot study was conducted to test the reliability of the protocol, including the analysis of inter-rater correlation and using a group of patients with various vocal disorders.

1.1. Evaluation Protocol

The evaluation protocol includes the assessment of voice quality, glottal attack, respiratory support, respiratory-phonatary-articulatory coordination (see Figure 1), digital laryngeal manipulation (laryngeal crepitation) and laryngeal tension. It also allows the self-assessment of voice quality and instrumental evaluation. All the parameters that could be registered in a visual analogue scale were thus recorded.

Perceptual analysis of voice quality includes a EP version of the CAPE-V developed as part of this project [4, 10] and GRBAS [11] scales. Other areas of evaluation include verbal articulation and related disturbances. Voice aerodynamics, respiratory endurance, maximum phonation time (see Figure 1), s/z coefficient, and orofacial motricity, are all assessed and provide a wide range of information to plan an intervention.

The EP version of the CAPE-V (Request to Translate & Distribute granted by ASHA on 28/1/2008), includes six new EP sentences designed to provide production of every oral vowel in EP (/6 "ma4t6 i u 6"vo "viv6~i~ n6"kel1 k6z6"46~u~ "4Oz6 "vELu/), easy onset with /s/, only voiced phonemes, hard glottal attack, nasal phonemes and voiceless stops, as described in [10]. The proposed new sentences (thoroughly reviewed by a Speech and Hearing Scientist, a Linguist and three experienced SLTs) were:

a. "A Marta e o avô vivem naquele casarão rosa velho."
b. "Sofia saiu cedo da sala."
c. "A asa do avião andava avariada."
d. "Agora é hora de acabar."
e. "A minha mãe mandou-me embora."
f. "O Tiago comeu quatro peras."

Results from various instrumental evaluation techniques (videostroboscopy, aerodynamics, EGG and EMG) can be registered by the protocol, including an extensive acoustic analysis based on sustained productions of /a, i, u, O/, CAPE-V sentences and reading a passage based on the EP version of the "The North Wind and the Sun", recently used as a standard text for "Advanced Voice Assessment" [12].

1.2. Reliability

In the present study, the reliability of the protocol was tested, using only two independent raters, who evaluated at the same time, in a single moment, a group of patients who exhibited some change in voice quality.

Previous studies [13] claimed that the ordinal GRBAS scale had higher reliability amongst observers when compared to visual analogue scales. However, other studies [1] showed that the reliability of visual analogue scales was higher than ordinal and range ratio scales. It is possible that these

contradictory results were due to the difference in the size of the sample between both studies. Karnell et al. [4] studied the reliability of two pairs of raters, who used the CAPE-V and GRBAS scales. They obtained high values of reliability, but only compared values for the parameter overall severity (CAPE-V) and for the parameter grade (GRBAS).

2. Method

2.1. Participants and Recording

In this study, the voice quality of 10 patients was assessed, with the Universidade de Aveiro's Voice Evaluation Protocol. These patients had been admitted to the Department of Otolaryngology of the Hospital de São João, Porto, Portugal (with full ethics committee approval). The sample included 10 patients with vocal pathology (9 females and 1 male), aged between 39-57 years old, with several clinical diagnoses: nodules, polyps, hypotonia of the vocal folds, swelling of Reinke, musculo-skeletal syndrome and dysfunctional dysphonia. The sample was evaluated (once) simultaneous by two SLTs.

CAPE-V tasks (sustaining vowels /a, i/, reading sentences and spontaneously speaking), endurance counting (respiratory support), sustaining vowels /a, u, i/ to determine maximum phonation time (see Figure 1), sustaining fricatives /s, z/ to determine the s/z coefficient, were recorded directly onto a PC, using Praat 5.0.22 [14] and in a quiet environment.

Recording equipment included an external sound card Edirol UA-25, set to 16 bits and 44.1 kHz sampling frequency, and a Sennheiser e815S microphone. During the recordings, the microphone was held on a tripod placed 25-30 degrees to the left of the patient's mouth, at a distance of 30-40 cm.

Both SLTs were instructed on how to use the protocol prior to applying to patients. This was followed by a testing phase, during which the raters were thoroughly trained on patients that did constitute the sample.

2.2. Statistical Analysis

A database was developed using SPSS 11, where all the variables used for the reliability study were set. This consisted of 366 variables, among which 10 were ordinal, 136 quantitative and 220 nominal.

Variables overall severity, roughness, breathiness, strain, change of pitch, loudness and resonance (CAPE-V), and the variables grade, breathiness and strain (GRBAS), glottal attack, respiratory-phonotary-articulatory coordenation and respiratory support of the pneumo-phono-articulatory parameter were analysed. Posture related variables position at rest, digital laryngeal manipulation (rigidity, laryngeal crepitation and focal pain) and vocal quality after manipulation of the laryngeal region, and the variables standing and seated position (speech and rest), were included in the database. Muscle tension at cervical level, the scapula, the larynx and the face (during speech production and at rest) were statistically analysed. Finally, the variable diagnosis was examined. The other parameters evaluated by the protocol have not been analysed statistically due to a high number of non-responses, and the fact that SPSS 11 does not calculate the correlation between two raters with non square tables (for ordinal and nominal variables). The variables that were analysed have a number of non-responses less than or equal to 30%. The reliability study of the protocol used Spearman's correlation coefficient for ordinal and quantitative variables and Cohen's Kappa test for the nominal and ordinal ones. Spearman's correlation coefficient was used for both quantitative and ordinal variables so that we could compare the results of this study with those of Karnell et al. [4]. However, it was deemed relevant, to also apply Cohen's Kappa test to ordinal variables. Spearman's correlation coefficient measures the intensity of the correlation of answers between two observers, while Cohen's Kappa test measures its agreement, and the two tests are related. For the analysis of nominal variables, only the Cohen's Kappa test was used.

Assuming a significance level (α) of 0.05 and an agreement level between the two raters of 0.80, the sample size necessary to achieve a power of 80% for the correlation between evaluators was estimated. Thus, for a small difference ($|\rho| = 0.1$), the sample size should be 783, for an intermediate difference ($|\rho| = 0.3$), the sample size should be 82, and for a large difference ($|\rho| = 0.5$), the size should include 27 elements. For this study, an intermediate difference would be ideal. In a similar project Karnell et al. [4] used a sample size of 103, which is close to the value, presented above. We are currently expanding our database to reach 100 patients, so the method presented in this section can be easily applied.

3. Results

3.1. CAPE-V

For the variables of the CAPE-V scale, results showed high and statistically significant correlation values for the overall severity (ρ=0.964, p=0.000), roughness (ρ=0.834, p=0.010), breathiness (ρ=0.991, p=0.000) and loudness change (k=1.000, p=0.000), which showed a good agreement between the two evaluators. The variables resonance (k=0.608, p=0.047) and strain (ρ=0.659, p=0.076), had an intermediate correlation value, the first one being statistically significant but not the second one. For the variable change of pitch (k=0.500, p=0.032), a low correlation value was found, but statistically significant, which indicates low inter-rater correlation. It is pertinent to note that no data was analysed for the quantitative variables pitch and loudness of CAPE-V, due to a large amount of non-responses.

There was a high correlation value for overall severity, roughness, breathiness and loudness change, probably because these parameters are assessed regularly in the practice of a SLT. These concepts are part of clinical practice of the two raters, so these parameters were more easily evaluated. The maximum correlation for the variable loudness change could be explained by the fact that it is often reported as one of the easiest parameters to evaluate.

It should also be noted that the value of correlation for the parameter overall severity obtained in this study (ρ=0.964) is similar to that found by Karnell et al. [4] (ρ=0.897).

The correlation value for the parameter pitch change was relatively low, possibly due to the evaluation of this parameter not being a simple perception of the fundamental frequency, but also the perception of other very complex voice properties, that need to be considered as a whole.

3.2. GRBAS

For the variables of the GRBAS scale, results of an analysis based on the Spearman correlation coefficient showed a high and statistically significant value for grade (ρ=0.815, p=0.004) and breathiness (ρ=0.846, p=0.002), which reveals a good agreement between raters, and an intermediate and statistically significant value for strain (ρ=0.662, p=0.037), which shows a reasonable agreement between the two evaluators.

Results of the Cohen's Kappa test, showed a high and statistically significant values for strain (k=0.848, p=0.000), which means there's a good agreement between the raters, and an intermediate and statistically significant value for grade (k=0.667, p=0.000) and breathiness (k=0.571, p=0.003), i.e., a reasonable agreement. It was not possible to run the tests for variables grade and asthenia due to lack of representativeness of classes (100% agreement between raters).

Analysing results for Spearman's correlation coefficient, it appears that the intensity of relation between grade and breathiness is high. However, the agreement between evaluators is only reasonable, when analysed by a Cohen's Kappa test. Results for strain were precisely the opposite, namely a median intensity relation and a high level of agreement between evaluators. This high agreement shows a great familiarity of both evaluators with this term (strain). It should be noted that the results of this study, for the parameter grade using Spearman's correlation coefficient (ρ=0.815), were similar to those (ρ=0.854) reported by Karnell et al. [4].

3.3. Comparing the CAPE-V and GRBAS Scales

The variables that are common to CAPE-V and GRBAS scales were compared using Spearman's correlation coefficient and the Cohen's Kappa test. Results showed, for the variables grade and strain, very similar values. For the variable breathiness it should also be noted that there was some relationship between the values of the Spearman's correlation coefficient and the Cohen's Kappa test.

Studies such as Bele's [13], that used a sample of 71 patients, indicate that the ordinal GRBAS scale has a higher reliability amongst observers when compared with visual analogue scales, which is not confirmed by this study, and probably results from the difference in sample size. However, studies such as those of Gerratt et al. [1], which used a sample of 22 patients, show that the inter-rater's reliability of the visual analogue scales is higher than the reliability of ordinal and range ratio scales. It should be noted that the sample size of Gerratt et al.'s [1] study is closer to the present study.

3.4. Other Parameters

Variables glottal attack (k=0.830, p=0.001), respiratory support (ρ=0.964, p=0.000), respiratory-phonotary-articulatory coordination (ρ=0.859, p=0.006), digital laryngeal manipulation– laryngeal crepitation (k=1.000, p=0.003), vocal quality after manipulation (k=1.000, p=0.003), scapula tension (ρ=0.910, p=0.002), laryngeal tension (ρ=0.891, p=0.003) and diagnosis (ρ=1.000, p=0.000) showed statistical significance and high correlation, i.e., there was a good agreement between evaluators. For the variables position at rest (k=0.608, p=0.047), digital laryngeal manipulation (k=0.608, p=0.047), values of intermediate and statistically significant correlation were found. However, the variables cervical tension (ρ=0.586, p=0.126) and facial tension (ρ=0.509, p=0.197), showed intermediate correlation values and no statistical significance. There were also values of low correlation and no statistical significance for the variables digital laryngeal manipulation (k=0.400, p=0.134), standing position (rest and speech) (ρ=0.266, p=0.564) and sitting position (ρ=0.216, p=0.641).

Some variables (e.g., glottal attack) had high correlation values and were statistically significant. This might be explained by the frequent assessment of these parameters by the majority of SLTs in cases of vocal pathology, which leads to greater familiarity with these concepts. High/maximum values of correlation were also found in the digital laryngeal manipulation (laryngeal crepitation), in the scapula tension and in the laryngeal tension. It can be argued that these parameters are evaluated more objectively since they are assessed not only by touch but also by listening.

In parameter vocal quality, the raters were able to perceive the vocal quality of the patient before and after manipulating the larynx, which facilitates the evaluation. This may justify the value of maximum correlation obtained.

The variable diagnosis also presents a maximum value of agreement between evaluators, which was consistent with the fact that the SLT diagnosis is aided by clinical diagnosis. The variables digital laryngeal manipulation (rigidity), standing and sitting position (rest and speech) and cervical and facial tension, were not scrutinised because the p-value was not statistically significant.

4. Conclusions

In the clinical practice of Portuguese SLTs, the large variety of subjective data and the lack of standardisation included in the vocal pathology protocols can make diagnosis insufficiently precise. The CAPE-V and GRBAS scales, and other parameters evaluated by the protocol have shown reliability in this study and in other studies [1, 4, 13], with high inter-rater correlation values. It seems therefore, reasonable to claim that the use of this instrument for evaluating vocal pathologies in the SLTs practice could facilitate the formulation of diagnosis, since there is a greater standardisation of parameters assessed by these professionals.

The similarity of Spearman's correlation values for the parameters overall severity (CAPE-V) and grade (GRBAS), with those of other studies [4] should also be noted. However, the results presented in this paper should be treated with some caution, since the sample size is relatively small and there is some frequency of non-responses, which may be reflected in the findings.

Summarising, we have developed an EP version of CAPE-V, the first standardised and evaluated protocol for voice assessment in EP, and shown that it produced similar results to assessments used in other languages.

Ongoing and future work will extend this study to a larger number of patients, so the protocol can be used with more confidence. Further validation of the EP version of CAPE-V will use the procedures presented in [15], including the production of a CD-ROM with the voice samples to be evaluated, voices used for training and samples of voices that represent specific grades of severity (MI – Mildly Deviant; MO – Moderately Deviant; SE – Severely Deviant).

5. Acknowledgements

This work was partially supported by Fundação para a Ciência e a Tecnologia, Portugal (PTDC/SAU-BEB/67384/2006).

6. References

[1] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. S. Berke, "Comparing Internal and External Standards in Voice Quality Judgments," Journal of Speech Hearing Research, vol. 36, pp. 14-20, 1993.

[2] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," Rev Laryngol Otol Rhinol, vol. 117, pp. 219-224, 1996.

[3] J. F. Piccirillo, C. Painter, D. Fuller, A. Haiduk, and J. M. Fredrickson, "Assessment of two objective voice function indices," Ann. Otol. Rhinol. Laryngol., vol. 107, pp. 396-400, 1998.

[4] M. Karnell, S. Melton, J. Childes, T. Coleman, S. Dailey, and H. Hoffman, "Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient- Based (V-RQOL and IPVI) Documentation of Voice Disorders," Journal of Voice, vol. 21, pp. 576-590, 2007.
[5] J. Muñoz, E. Mendoza, M. Fresneda, G. Carballo, and I. Ramirez, "Perceptual analysis in different voice samples: agreement and reliability," Journal Percept Mot Skills, vol. 94, pp. 1187-9, 2002.
[6] P. Carding, E. Carlson, R. Epstein, L. Mathieson, and C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom," Logopedics Phoniatrics Vocology, vol. 25, pp. 133-138, 2000.
[7] S. Peppé, "Assessment of Prosodic Ability in Atypical Populations," Department of Clinical Language Sciences, Reading University. , 2007.
[8] D. R. Boone, The Boone Voice Program for Adults, 2nd Ed ed., 1982.
[9] W. Haynes, M. Moran, and R. Pindzola, "Voice Disorders. In: Haynes, W. Moran, M Pindzola, R. Communication Disorders in the Classroom: An Introduction for Professionals in Schools Settings," 2006, pp. 267-269.
[10] ASHA, CAPE-V Form and Procedures. ASHA Special Interest Division 3, Voice and Voice Disorders, 2006.
[11] M. Hirano, "Clinical Examination of Voice," Vienna: Springer-Verlag, 1982.
[12] M. Pedersen and K. Munck, "Advanced Voice Assessment," in MAVEBA, Florence, Italy, 2007, pp. 61-64.
[13] I. V. Bele, "Reliability in Perceptual Analysis of Voice Quality," Journal of Voice, vol. 19, pp. 555-573, 2004.
[14] P. Boersma, "Praat, a system for doing phonetics by computer," in Glot International, 2001, pp. 341-345.
[15] R. I. Zraick, "Results of the CAPE-V Validation Study," in Oral Presentation at ASHA Convention, Boston, 2007.

Coordenação Pneumo-Fono-Articulatória (PFA)

Endurance – Suporte Respiratório

Suporte Respiratório _____ C I ____/100
AL AM AS

Indicar tarefa(s) □ Contagem 1-50 □ Dias da semana □ Meses do Ano □ Alfabeto

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Notação: "PC": pausa curta; "PL": pausa longa; "FI": fonação inspiratória; "<": diminuição da intensidade.

Coordenação PFA _____ C I ____/100
AL AM AS

□ Adequada ao aumento da intensidade
□ Uso de ar residual
□ Pausas inspiratórias insuficientes
□ Fluxo de ar excessivo
□ Fluxo de ar insuficiente
□ Ciclos respiratórios curtos
□ Insp. insuficiente para completar as frases
□ Pausas para deglutição insuficientes
□ Fonação inspiratória
□ Outra:_____

Observações

Tempo Máximo de Fonação (TMF)

Vogais Sustentadas

/a/ _____ (s); _____ (s); _____ (s) média: /a/ _____ (s)
/u/ _____ (s); _____ (s); _____ (s) média: /u/ _____ (s)
/i/ _____ (s); _____ (s); _____ (s) média: /i/ _____ (s)

(Valores de referência: Homens: 23,7 s; Mulheres: 16,7 s)[*]

Coeficiente S/Z

/s/ _____ (s) ; _____ (s)
/z/ _____ (s); _____ (s)
(Valores de referência: aproximadamente 20-25 s)[†]

Coeficiente S/Z: _____
(/s/ mais longo//z/ mais longo)

[0.8 – 1.2]: fonação normal; <0.8: hipercinésia/hiperadução das pv; >1.2: hipocinésia, fenda glótica ou patologia vocal.

[*] Hirano, 1981.
[†] Pindzola, Rebekah et al. (1987) (Um coeficiente de 1.0, com duração normal da produção dos sons /s/ e /z/ - aproximadamente entre 20-25 s nos adultos – sugere uma capacidade respiratória normal e ausência de patologia nas pregas vocais; um coeficiente inferior a 1.0 indica uma possível ineficácia respiratória, podendo o paciente ter uma redução da capacidade vital ou fraco controlo na fase expiratória; um coeficiente de 1.2 ou superior, com duração superior da produção /s/ indica a existência de patologia nas pregas vocais.)

Figure 1: *A sample page of the protocol.*