A Perceptual Method to Rate Dysphonic Voices

*Jorge A. Gurlekian, *Humberto M. Torres, and †Melissa Rincón Cediel, **Buenos Aires, Argentina, and* †*Bucaramanga, Colombia*

Summary: Objective: To present and test a production-matching method with external references, looking at the improvement of inter-rater variability of expert evaluations.

Method: It consists of adjusting quality attribute levels of a synthetic vowel for a simultaneous matching with the natural patient vowel (NPV) attributes. In an initial experiment, seven speech-language pathology (SLP) experts performed this task with the new method and evaluated the same NPV with the standard method. Targets were twelve NPVs with a variety of quality attribute combinations. In a second experiment, we employed the proposed method to assess the evaluation performance of 65 SLP students.

Results: Expert evaluations show less dispersion for the proposed method than those obtained using the standard rating method. Student individual responses were compared with overall responses from their own group and were cross referenced with expert responses. A Kappa index is proposed as a measure of SLP students' performance.

Conclusions: The proposed method was readily accepted by both SLP experts and students. Experts' consensus was improved. SLP students could benefit by quickly learning to discriminate complex attributes, which usually demands years of experience.

Key Words: Dysphonic voice−Evaluation−Production−Matching−Inter-rater variability.

INTRODUCTION

There are several reasons why speech-language pathology (SLP) experts need to reach a consensus when they evaluate dysphonic voices: 1) to produce more consistent professional evaluations; 2) to have more stable references to evaluate SLP students; and lastly, 3) to satisfy the need for similar criteria from human expert evaluations to validate automatic acoustic evaluations and serve as reliable gold standards of new techniques.

While looking for agreement and reliability in judgments of dysphonic voices between SLP experts,[1] several experiments for natural patient vowel (NPV) evaluation based on unidimensional artificial external references and matching have been presented. Gerrat and Kreiman[2] presented results pertaining to listeners who judged the noisiness in a separated task using a traditional visual-analog rating scale of natural stimuli in comparison with the synthesis judgments. They concluded that “listeners can, in fact, agree with their perceptual assessments of voice quality, and that analysis-synthesis can measure perception reliably.” Furthermore, it was found[3] that anchors made up of synthesized signals combined with training were more effective than natural voice anchors in improving reliability to judge perceptual roughness and breathiness.

Standard psychophysical methods were compared to evaluate breathiness[4]; they showed less dispersion for a matching task. In the same direction, instead of using direct magnitude estimation, Patel et al[5] demonstrated that a matching task produces reliable estimates of roughness. The psychoacoustic relevance of matching has also been demonstrated by an improved magnitude estimation method.[6] Synthetic external references with varying levels of jitter and an intramodal matching procedure were used to evaluate vowel roughness. Results presented high inter-rater agreement among groups of SLP experts and inexperienced raters alike when compared with a numerical rating scale. Like in the case of roughness and breathiness, raters could focus on one dimension and produce better matches than when they used their own internal references. One question arises: Can SLP experts easily prepare synthetic external references at the clinic?

Speech synthesis by control of acoustic features requires a deep understanding of related acoustic parameters and requires enough time to adjust them. The original Klatt's synthesizer[7] took at least 100 times longer than real time to produce natural-like voices. Thirty years later the integrated software for the analysis and synthesis of voice quality[8] has improved the interface between analysis and synthesis and made it an invaluable tool for research and teaching.

Acoustic correlates for roughness and breathiness were examined by various scholars. Both were found to be dependent on fundamental frequency variations. Roughness, in particular, depended on jitter, shimmer, or a combination of the two. But roughness can also be perceived when high noise levels are present, due to random amplitude variations on periodic peak amplitudes and/or shifts on fundamental periods as shown for complex sounds.[9] Currently, it is considered that breathiness is produced by aspirated noise and open quotient variations.[10] Breathiness loudness is a result of an interaction of noise level and harmonic level. Shrivastav[11] established a model to predict breathiness loudness, where the relation of noise level to partial loudness—which is related to pure harmonic loudness—is a power function.

We propose that the interface's input parameters should be the quality attributes themselves, which are best known by

Accepted for publication January 8, 2018.

This work was funded by MINCYT-CONICET and COLCIENCIAS Argentine-Colombian Bilateral Project CO1402.

From the *Laboratorio de Investigaciones Sensoriales (LIS), INIGEM-CONICET-UBA, Buenos Aires, Argentina; and the †Universidad Manuela Beltrán (UMB), Bucaramanga, Colombia.

Address correspondence and reprint requests to Jorge Gurlekian, Hospital de Clínicas, Laboratorio de Investigaciones Sensoriales (LIS), INIGEM CONICET, Av. Córdoba 2351. Piso 9, Sala 2., 1120, Buenos Aires, Argentina. E-mail: anagraf99@hotmail.com

Journal of Voice, Vol. 33, No. 4, pp. 453−464
0892-1997

https://doi.org/10.1016/j.jvoice.2018.01.007

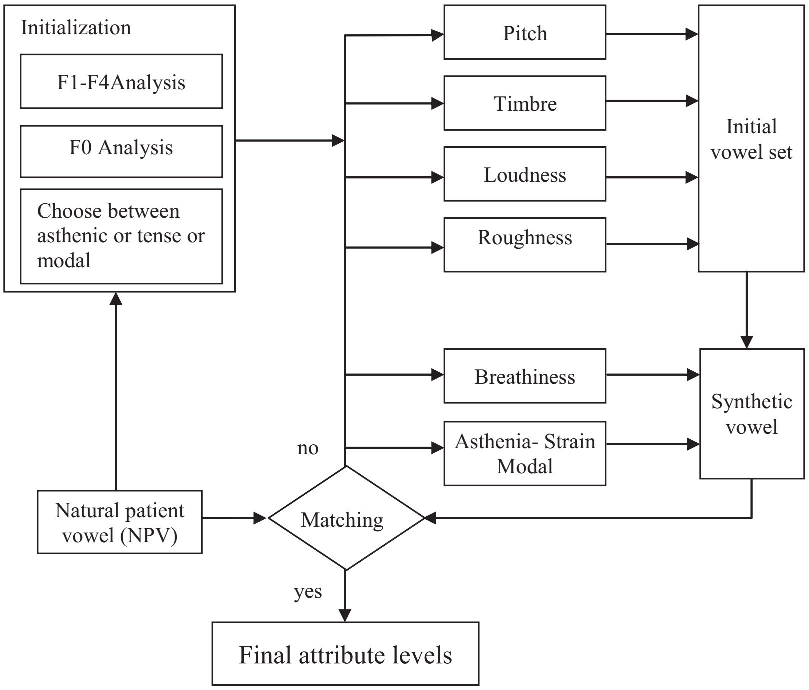

FIGURE 1. Operation flow scheme.

most SLP experts, making it possible to perform a total perceptual match between the artificial vowel produced and the NPV. Taking into account those experimental backgrounds, we have developed a clinical tool that helps evaluate NPVs in a simplified manner. This approach will control each quality attribute, mixing it with other attributes in real time, and considering the resulting percept as a whole.

This paper is organized as follows: After this introduction, the second section describes in detail the new method and attributes definitions that can be varied. NPVs to be evaluated are presented at the end of this section. In the third section two experiments are analyzed. In the first one, SLP experts employ the new method and their results are compared with the standard method when they evaluate the same stimuli. In the second experiment, inexperienced SLP students employ the new method to evaluate voice attributes for the first time. Results and statistical analyses are presented in the fourth section. In the fifth section, the design strategy of the method and results are discussed. A way to evaluate expert and student performances using the Kappa index is proposed. In the sixth section, we conclude that the new method could be employed at the clinic because it is capable of improving SLP expert agreements and the acceptability shown by SLP experts and students.

SYSTEM DESCRIPTION

Raters will be instructed to perform perceptual matches between the NPV and a synthetic vowel created by setting levels in a continuum of breathiness, roughness, and asthenia-strain. Once all attributes are set, an overall quality will emerge as a consequence, and raters could still decide to correct any individual match or finish the evaluation.

The strategy consists of an easy production of artificial stimuli by setting perceptual attributes of vowel quality and iteratively performing a match between the resulting auditory percept and the NPV percept until the best match is obtained for each attribute (see Figure 1). To do this, raters can choose a value from the continuum for each attribute and perform the perceptual match without time limits. These selections in turn control the acoustic parameters listed in Table 1. The sequence of settings is simple and does not demand a high cognitive load. The following sequence will be part of users' instructions. After loading and listening to the NPV, choose a vowel among the initial vowel set with a pitch (1) and timbre (2) that best approximate the NPV by listening to it and comparing it. Adjust the (3) loudness and modify your choices if necessary. Then, select the roughness (4) level and listen to how your actual vowel sounds, compare it with the NPV, and modify it if it is necessary to

TABLE 1.
Acoustic Correlates of Perceptual Attributes, Steps, and Ranges

Attribute	Acoustic Parameter
Pitch	*F*0 (10 Hz steps, from 80 to 300 Hz)
Timbre	Formant frequency structures (four types, see text)
Loudness	Harmonic relative intensity levels (0.01 steps, from 0 to 3)
Roughness	*F*0 irregularity (0.5% jitter steps, from 0% to 3%)
Breathiness	Noise bands + six noise intensities + harmonic relative intensities
Asthenia-strain	Filtered harmonic structure + harmonic/noise intensities
Tremor	Amplitude modulation (frequency and modulation index)
Breaks	Pauses (number and duration)

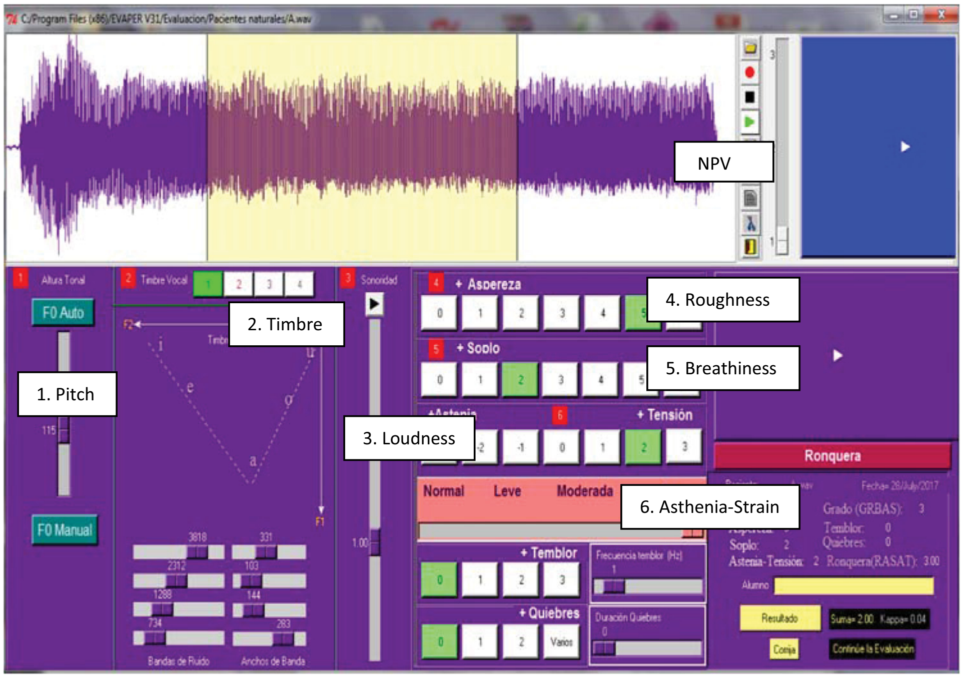

FIGURE 2. Graphical user interface of the Evaper method.

repeat the comparison. Do the same for breathiness (5), and finally for the asthenia-strain (6) dimension. Numbers in parentheses are also visible in the graphic user interface (GUI) shown later in Figure 2.

The initial vowel set and attributes definitions

Vowels were created by a Linear Prediction Coding (LPC) formant synthesizer at a sampling rate of 50 KHz and 16 bits. Two formant frequency structures were defined as representatives of female and male vowels /a/,[12,13] as indicated in Table 1. Twenty-three $F0$ values were selected from the range of 80 to 300 Hz at 10 Hz steps, which remained constant throughout the 3-second vowel duration.

Roughness is associated with cycle-to-cycle variations of fundamental frequency or cycle-to-cycle variations of amplitude or the two together. Different alternatives have been proposed to produce roughness. One possible way to do this is to introduce random noise in the glottal source to create controlled variations,[14] or by modulation functions.[5,15] In our implementation, we used a statistical model of jitter[16,17] which was presented and tested in a previous work.[6] For each vowel, we produced a new vowel set with jitter values ranging from 0% to 3% according to Equation 1. Fundamental frequency variation over time was created to produce stimuli with these intended jitter values. We chose the definition of percent jitter (Equation 1) as the average of the difference between two $F0$ values, normalized to the average $F0$ and multiplied by 100.

$$J\% = 100\frac{\frac{1}{N-1}\sum_{i=1}^{N}|F0_i - F0_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}F0_i} \tag{1}$$

Where $F0_i$ is the ith fundamental frequency cycle to cycle and N is the number of cycles. By using this method for the creation of vowel stimuli, $F0$ variations are made independent from amplitude variation, which remained constant. According to Titze[18] and Torres et al,[19] $F0_i$ values have a Gaussian behavior with normal density probability functions ($F0$, s$F0$). As a result, it is possible to synthesize vowels from reference average $F0$ values and each intended jitter value. The $F0_i$ values set have a Gaussian noise distribution with an average $F0$ and standard deviation (SD) given by Equation 2.

$$\sigma_{F0} = \frac{\sqrt{\pi}}{200}F0J\% \tag{2}$$

Once the $F0$ set that represents the glottal source was defined, 644 stimuli were synthesized using the LPC method[20] implemented in *MATLAB* 7.4 (Math Works, Inc., Natick, MA).

Jitter continuum of seven steps (levels 0 to 6) is shown in Table 2. Level 0 presents no roughness. Level 6, corresponding to a jitter value of 3%, produced a roughness judged as high. The remaining roughness steps were distinguished clearly as was demonstrated in Gurlekian et al.[6]

TABLE 2.
Roughness Levels and Associated Jitter Values; Breathiness Levels and Associated Decibel Values of Noise

	Roughness Levels						
Jitter (%)	0	1	2	3	4	5	6
	0	0.5	1	1.5	2	2.5	3
	Breathiness Levels						
Noise intensities	0	1	2	3	4	5	6
Decibel SPL	0	8.8	13.7	17.3	20.2	23.1	27

TABLE 3.
Correspondences Between Asthenia-Modal-Strain Levels and Type, Cutoff Frequency (CF), and Gain of the Corresponding Filters

	Asthenia			Modal	Strain		
Levels	−3	−2	−1	0	1	2	3
Filter		Low Pass		—		High Pass	
CF (hertz)	500	1,000	2,000	—	1,000	2,000	3,000
Gain	0.1	0.5	0.8	1	1.2	1.5	2

When using the GUI, an automatic *F*0 measurement of the NPV is available within the interface to help raters to choose a vowel from the set with the closest pitch. Nevertheless, *F*0 can be adjusted manually according to raters' perceptual impression. Then a vowel with the nearest *F*0 is chosen from the vowel set. Also, NPV formant values are automatically estimated and used to shape noise bands. Vowel intensity is controlled by the GUI independently of further added noise to adjust harmonic loudness continuously from 0 to 80 dB.

At this step, raters can choose the vowel with the combination of pitch, timbre, loudness, and roughness from the vowel set that best approximates the NPV, then they could add levels of breathiness and asthenia-strain.

Acoustic correlates of breathiness were also selected from a variety of alternatives. A complete model has been created in which noise loudness relative to harmonic loudness has been shown to be a good predictor of perceived breathiness.[4,21] We chose to create breathiness levels from white noise that was band-pass filtered at central frequencies corresponding to the NPV formant frequencies. Seven fixed noise levels were created and measured at the maximum peak of the spectral representation. Noise loudness steps were between 3 and 5 dB, as shown in Table 2. In line with Patel et al's model,[21] we allowed users to freely combine noise loudness and harmonic loudness by separated controls, mixing harmonic samples with noise samples using the Mix command of the Snack Sound Toolkit (Swedish Royal Technical University, KTH, Stockholm, Sweden). Breathiness at level 0 introduces no noise; in this case, synthetic vowels will sound with the previously set roughness level. Alternatively, one of the other six breathiness levels will be added to the roughness level initially chosen. As breathiness is perceived relative to noise-to-harmonic loudness, if more breathiness is needed, harmonic loudness could be set at minimum or even zero. In the presence of high breathiness, an increase in roughness can occur. This side effect occurs because noise can produce a perturbation of both vowel periods and amplitudes. In these cases, it is possible to modify previous settings at any time.

Asthenia-strain is presented as one dimension of seven steps, from −3 to +3. They were combined in the same continuum because SLP experts found it difficult to evaluate them as two separate dimensions. They cannot coexist simultaneously, both physiologically and perceptually. Nevertheless, patient voice examples could have alternations, ie, starting as asthenic and finishing with strain. We recommended that the dominant percept be evaluated and the secondary percept be annotated as an observation.

By selecting main control buttons, users can modify the vowel harmonic structure with its corresponding roughness level and mix it with the noise already selected in the breathiness step. Level 0 corresponds to modal phonation. For asthenia, level settings are −1, level −2 until the maximum level of −3. Positive values correspond to strain levels, with +3 as the maximum.

Acoustic correlates of asthenia and strain are not univocally described in the literature. We took into account harmonic structure, intensity, and noise as acoustic correlates of these sensations. To control harmonic structure, we followed descriptions of several authors who identified differences in the spectral levels at regions of low and high frequencies.[22−24] Following these descriptions, asthenia levels were obtained thanks to low-pass filtering. Because fold contact is not reached, maximum pressure is not obtained and the glottal waveform looks more like a pure tone with few low-frequency harmonics. Strain levels were obtained by high-pass filtering to produce more flat spectra. Discrete-time IIR filters were created offline with 5-second order sections with *MATLAB* 7.4 and the Signal Processing Toolbox 6.7 (Math Works, Inc., Natick, MA). Cutoff frequencies are indicated in Table 3. Each level is amplified or attenuated as indicated in the same table. Table 4 shows an example for *F*0 equal to 160 Hz and the resulting harmonic-to-noise relations corresponding to asthenia-strain levels.

As a rule of thumb, asthenia is associated with breathiness and low loudness. Strain is associated with roughness and high loudness. But this general rule doesn't always apply because loudness also depends on what happens behind the glottis—subglottal pressure value—and above the glottis due to energy amplification by harmonic-formant frequency coincidences.

NPVs

As a component of the new system, we include the natural vowel set to be evaluated, NPVs, because expert performances will be used as references for further evaluations. The 12 NPVs were provided by voice therapists who recorded patient voices. Those sounds represent a variety of roughness, breathiness, and asthenia/strain combinations. Acoustic analysis of NPVs is presented in Table 5 using a speech analysis system.[25] Several measurements are shown: cepstrum maximum peak amplitude (normal from 1 to 0.3); jitter (normal from 0 to1); HNR (normal > 5), and Lyapunov maximum exponent (LME) (normal < 0) for an approximate ordering of dysphonic vowels.

TABLE 4.
Examples of HNR Measurements for Vowel /a/ and *F*0 Equal to 160 Hz for Breathiness and Asthenia-Strain Levels; Harmonic Loudness Was Fixed and Set at the Middle of the Scale

		Breathiness Levels						
		0	1	2	3	4	5	6
Asthenia	−3	>17	0.99	0.92	−1.18	−2.02	—	—
	−2	>15	6.67	4.13	2.58	0.97	−0.35	−1.07
	−1	>14	5.48	3.27	2.72	1.38	0.39	−0.24
Modal	0	>11	9.53	7.80	4.93	3.01	1.37	0.60
Strain	1	>10	10.66	10.50	10.29	9.81	7.37	6.99
	2	>10	10.61	9.64	9.04	7.68	3.25	2.78
	3	>10	9.02	8.08	7.28	5.71	2.08	1.79

EXPERIMENTAL PROCEDURE

First-time training

A set of 12 synthetic vowels with different attribute combinations were created using the proposed system to emulate NPVs for training. They were chosen as examples of 1) high roughness and strain; 2) high breathiness and asthenia; 3) high roughness and breathiness; 4) medium roughness and medium breathiness; 5) high roughness and asthenia; and 6) modal voices. These combinations were duplicated for male and female fundamental frequencies for a total of 12 examples.

Each attribute level used to create the examples defined the file name; ie, F0120T2L0.10R4B1S2, which represents the settings $F0 = 120$, Timbre = 2, Loudness = 0.10, Roughness = 4, Breathiness = 1, and Strain = 2. Trainees could load a file name like this as a synthetic “NPV” and recreate it following the level settings. By completing this training, they learn to a) fix attributes; b) associate attributes with physiological events; c) learn how attributes interact; and d) sense their discrimination and make subtle changes to the original settings.

Evaluation

Before using the tool, raters must listen to NPVs using binaural headphones (HD 407; Sennheiser electronic GmbH & Co. KG, Wedemark, Lower Saxony, Germany) and decide if the NPV is asthenic, tense, or modal, considering that in general, one condition excludes the others. That initial decision helps raters choose levels more quickly. They are guided by their perception and not just by trial and error. A fixed sequence of operations must be followed, as indicated in Figure 1. Starting with a generic /a/ vowel, the rater must fix all attributes available. This fixed sequence could be reinitiated at any step until the best match is found. Final levels used for each attribute will define their evaluation. Figure 2 shows the actual GUI written in TCL/TK, which guides the sequence of quality attribute settings and listening both the result and the reference NPV.

TABLE 5.
NPVs Showing Main Acoustic Parameters, Ordered Approximately from High to Low Levels of Perturbation

Patient	F0	LME	CPA	HNR	Jitter
J	136	0.46	0.07	−1.7	—
L	120	0.30	0.12	0.71	10
G	175	0.80	0.16	1.96	10
A	117	0.81	0.02	2.40	3.23
C	105	0.78	0.02	2.04	—
E	272	0.48	0.03	1.20	3.5
F	231	0.42	0.14	0.49	5
B	170	0.03	0.02	1.42	—
I	122	0.41	0.29	5.93	0.64
K	113	0.39	0.00	7.39	1.04
D	129	−0.33	0.45	6.27	1.36
H	258	−0.25	0.57	7.55	0.77

Abbreviation: CPA, cepstral peak amplitude.

Experiment 1

The goal was to quantify expert performance during an evaluation.

The subjects were seven SLP experts without any hearing impediments with more than 5 years of work at the voice clinic of Hospital de Clínicas—University of Buenos Aires (four experts) and at the voice clinic at the University of Bucaramanga (three experts). They were 35 years old on average.

The stimuli were 12 NPV with different grades of dysphonia presented binaurally through HD 407 headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Lower Saxony, Germany) in a silent room not acoustically treated.

The methods were as follows: 1) Evaper and 2) Standard (GRBAS scale).

We used GRBAS as the standard scale because it is the rating scale most widely used by SLP experts in Argentina and Colombia.

As explained in the section “Firsrt-time training,” SLP experts were provided with training in Evaper.

Instructions

Evaper method

Using headphones you will hear a patient vowel /a/. Your task is to produce the best match in roughness, breathiness, and asthenia-strain between the NPV and a synthetic vowel to be created by you using a GUI. First, you should adjust the pitch, loudness, and timbre of the synthetic vowel like you did in the training session. Each attribute will be adjusted selecting a value from a continuum of seven values

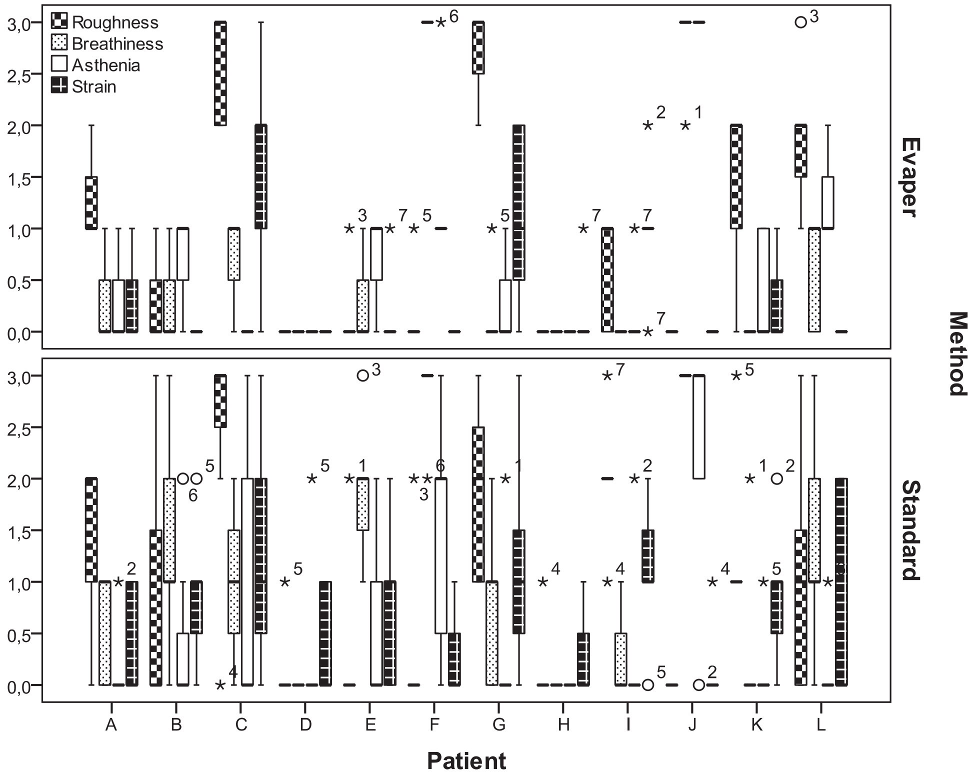

FIGURE 3. Box plots for four attributes (R, B, A, S) evaluated in 12 patients (A-L) by seven (1−7) experts; atypical values are indicated with its expert rater ID number.

and perceptually matched with the correspondent attribute of the NPV in that moment. Please modify the attributes in the following order: roughness, breathiness, and asthenia-strain. You can modify your settings any time. There is no time limit for this task.

Standard method

Using headphones you will hear a patient vowel /a/. Your task is to use a standard scale (0, 1, 2, 3) and assign the vowel a number that reflects the level of roughness, breathiness, asthenia, and strain. There is no time limit for this task.

Experiment 2

The goal was to quantify students' perceptual evaluations.

The subjects were 65 inexperienced listeners: SLP students from the University of Buenos Aires (21 students), Universidad del Salvador in Buenos Aires (20 students), and Universidad Manuela Beltran (24 students). They were 24 years old on average.

The stimuli were the same as above for the Evaper method.

The method and instructions were the same as above for the Evaper method.

RESULTS

Evaper method responses to roughness and breathiness (0, 6), asthenia (0, −3), and strain (0, +3) and responses using the standard rating scale (0, 3) were all mapped to a range from 0 to 3 so they could be compared. To perform the normalization, a plot of roughness and breathiness responses using both scales made it possible to define the following mapping: {0, 1} –> 0; {2, 3} –> 1; 4 → 2; {5, 6} –>3.

As shown in Figure 3, deviations of most quality attributes were lower for the Evaper method than for the standard method; ie, less dispersion is observed for qualities of breathiness and strain. A comparison of a number of outliers for both methods is shown in Figure 4. When using the standard method, there were one or two more outliers per rater. As seen in Figure 4, we can compare the total number of outliers: 11 for Evaper and 23 for the standard method.

We performed a single-sample chi-square test for each of the 12 NPVs to retain or refuse the null hypothesis of linearity for raters' responses to each of the four attributes. The results were all nonsignificant ($P > 0.05$), supporting the use of linear statistics. SDs were calculated and compared for Evaper and the standard method for each attribute in

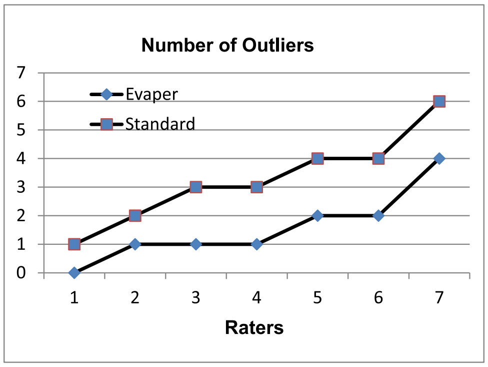

FIGURE 4. Number of outliers for both methods.

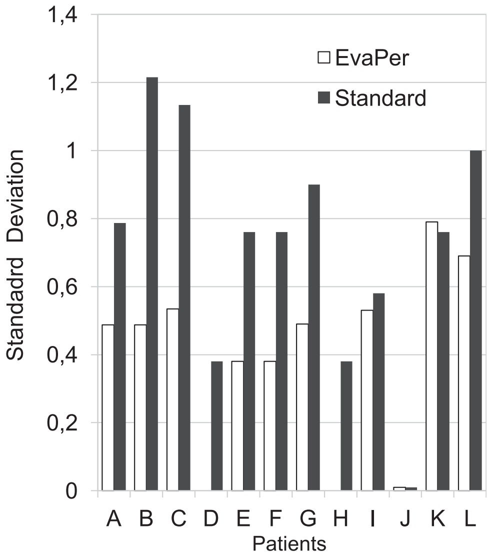

FIGURE 5. Average SD of roughness for seven experts using both methods.

Figures 5 to 8. Figures show higher deviations for the standard method when compared with Evaper. An analysis of variance verified these differences when raters evaluated each attribute using both methods. Considering a normal distribution for responses to each NPV, values *F* and *P* for a general linear model are shown in Table 6. The nonparametric Kruskal-Wallis test was also performed, and the results are presented in the same table.

As a measure of judgment variability between experts, exact judgment coincidences for all experts were calculated using the Kappa index.[26–28] Table 7 show results for each patient and attribute when using both methods. Kappa indexes vary between 0 and 1, where 1 corresponds to the exact match. Values close to 0.7 are considered an acceptable agreement. All Kappa indexes were higher when experts used the Evaper method than when they used the standard method; breathiness 0.73/0.46 and strain 0.68/0.3 were the attributes with the highest differences. When using the Evaper method, the lowest Kappa index was for roughness evaluation (0.58) and the highest was for breathiness (0.73). When the seven experts used the new method, an overall Kappa of 0.67 was obtained, compared with 0.44 obtained with the standard method. The closest agreements were in breathiness, strain, asthenia, and roughness, in that order.

In order to evaluate experts' performance when using both methods, the exact agreement percentage for each expert was calculated. This is shown in Table 8. This measure could be used to create a ranking of experts.

Figure 9 contains the results for student evaluations including median values of roughness, breathiness, and asthenia-strain for the Evaper method.

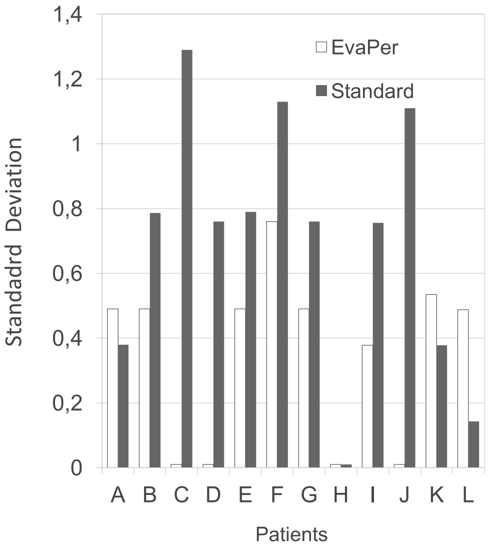

FIGURE 7. Average SD of asthenia for seven experts using both methods.

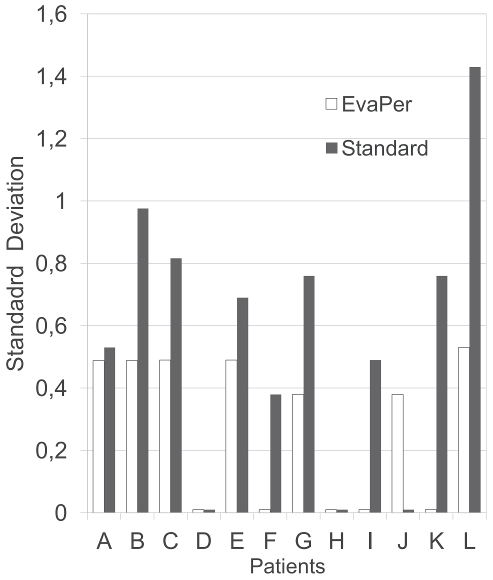

FIGURE 6. Average SD of breathiness for seven experts using both methods.

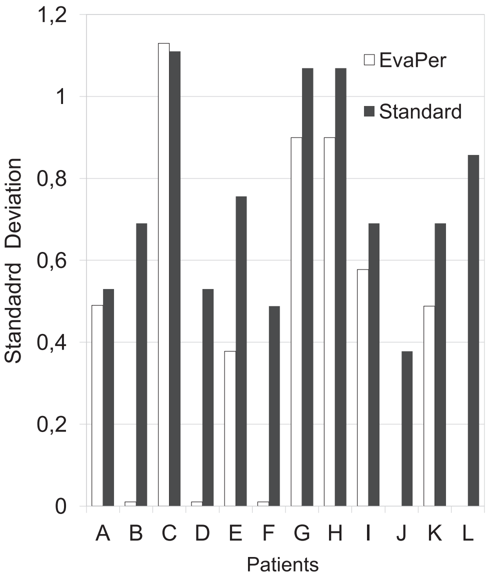

FIGURE 8. Average SD of strain for seven experts using both methods.

TABLE 6.
A Parametric and Nonparametric Analysis of Variance; *F* Values and Kruskall-Wallis Test Are Presented for Both Methods

	R		B		A		S	
	F	*P*<	*F*	*P*<	*F*	*P*<	*F*	*P*<
Evaper	32.89	0.000	63.84	0.000	28.47	0.000	7.75	0000
Standard	7.33	0.000	18.34	0.000	5.04	0.000	1.83	0.064

	R		B		A		S	
	$\chi 2$	*P*<	$\chi 2$	*P*<	$\chi 2$	*P*<	$\chi 2$	*P*<
Evaper	68.48	0.000	59.33	0.000	57.83	0.000	42.93	0.000
Standard	46.63	0.000	58.87	0.000	30.50	0.001	17.34	0.098

TABLE 7.
Kappa Index for Each Attribute and Each NPV Evaluated by Seven Experts Using Evaper and the Standard Method

Stimuli	Evaper/Standard Method				
	R	B	A	S	Ave
A	0.46/0.25	0.46/0.35	0.46/0.68	0.46/0.35	0.46/0.40
B	0.46/0.19	0.46/0.08	0.46/0.41	1/0.25	0.6/0.23
C	0.35/0.41	0.46/0.14	1/0.25	0.29/0.03	0.53/0.20
D	1/1	1/1	1/0.68	1/0.3	1/0.74
E	0.68/0.68	0.62/0.14	0.51/0.19	0.73/0.19	0.64/0.29
F	0.68/0.68	1/0.68	0.68/0.08	1/0.46	0.84/0.47
G	0.46/0.14	0.68/0.19	0.46/0.08	0.13/0.08	0.43/0.27
H	1/0.68	1/1	1/1	0.73/0.46	0.93/0.78
I	0.35/0.41	1/0.46	0.68/0.68	0.4/0.25	0.61/0.44
J	1/1	0.68/0.68	1/0.14	1/0.68	0.92/0.62
K	0.25/0.68	1/0.68	0.35/0.68	0.46/0.25	0.51/0.56
L	0.25/0.08	0.35/0.08	0.46/0.68	1/0.35	0.51/0.29
Ave	0.58/0.51	0.73/0.46	0.67/0.51	0.68/0.30	0.67/0.44

TABLE 8.
Percentage of Exact Agreement Between Evaluations in Both Methods for Each Expert and Each of the NPVs

Stimuli	Expert Number ID						
	1	2	3	4	5	6	7
A	41.7	54.17	54.2	54.17	45.83	41.67	66.7
B	62.5	54.17	75	75	50	50	50
C	54.2	62.5	58.3	66.67	66.67	62.5	62.5
D	62.5	95.83	75	54.17	66.67	54.17	54.2
E	75	70.83	54.2	54.17	79.17	70.83	54.2
F	91.7	91.67	91.7	91.67	70.83	70.83	91.7
G	62.5	33.33	45.8	62.5	41.67	45.83	58.3
H	95.8	95.83	95.8	95.83	95.83	100	79.2
I	70.8	54.17	75	75	75	70.83	37.5
J	75	95.83	95.8	95.83	95.83	95.83	95.8
K	54.2	41.67	54.2	54.17	66.67	66.67	62.5
L	62.5	50	54.2	66.67	62.5	58.33	45.8
Ave	67.4	66.67	69.1	70.49	68.06	65.63	63.2
SD	23.6	28.22	25.2	24.17	24.06	23.91	23.1

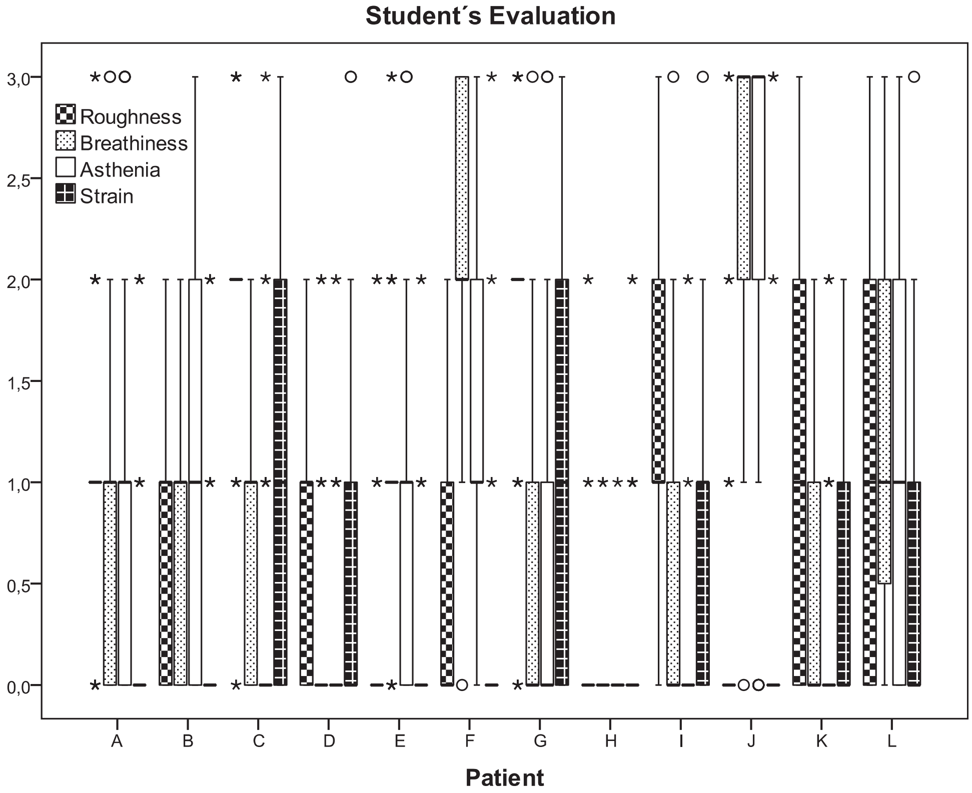

FIGURE 9. Box plots for four quality attributes, evaluated by 65 students in 12 patients (A-L) using Evaper. Circles represent atypical values and stars extreme values.

Proportions of correct student responses are calculated from answers that range from plus-minus 1 SD, 1.5 SD to 2 SD from expert raters' averages. Calculated individual Kappa indexes are presented in Figure 10, grouped as an accumulated percentage of students. The red line corresponds to an average expert Kappa of 0.67.

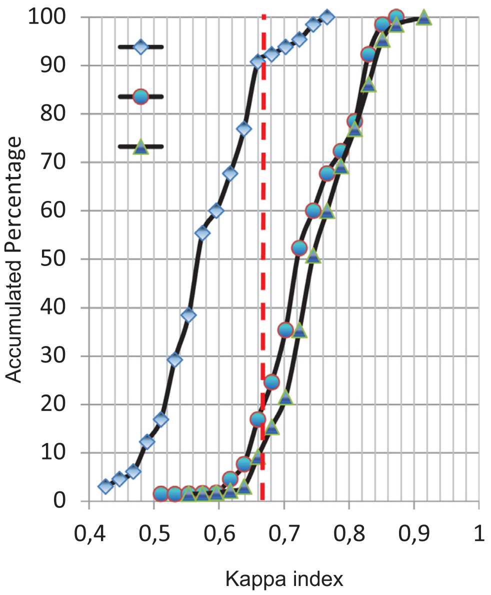

FIGURE 10. Accumulated percentage of students for kappa values calculated as plus minus 1 SD, 1.5 SD, and 2 SD of expert rater media values.

The grade was reported for the standard method. Because usually a degree of hoarseness or voice abnormality is represented, we explored grade correlations to pitch, roughness, breathiness, and strain when evaluated using both the Evaper and standard methods. Figure 11 and Table 9 show these correlations. For the standard method, roughness and breathiness appear to contribute to grade in the same high proportion, followed by asthenia and strain. In the case of the Evaper method, high positive correlations were found for roughness, followed by breathiness, asthenia, and strain. Figure 12 and Table 10 show grade correlations to acoustic measurements. Grade, which was negative with cepstrum peak amplitude and harmonic-to-noise relation, had a high correlation with almost all parameters. It was also positive

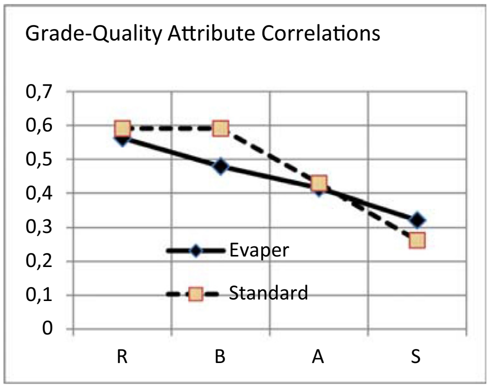

FIGURE 11. Correlations between grade and quality attributes estimated by both methods.

TABLE 9.
Correlation Matrix for Grade, Estimated With the Standard Method, and Quality Attributes Estimated With Both Methods Evaper: R1, B1, A1, S1 and Standard: R2, B2, A2, S2

	R1	R2	B1	B2	A1	A2	S1	S2
Grade	0.563	0.591	0.479	0.591	0.4148	0.429	0.322	0.262
R1	—	0.835	−0.27	−0.28	−0.284	−0.23	0.708	0.533
R2	—	—	−0.26	−0.25	−0.371	−0.24	0.772	0.672
B1	—	—	—	0.877	0.8007	0.907	−0.26	−0.6
B2	—	—	—	—	0.8119	0.829	−0.23	−0.38
A1	—	—	—	—	—	0.829	−0.45	−0.48
A2	—	—	—	—	—	—	−0.06	−0.5
S1	—	—	—	—	—	—	—	0.532

with jitter and LME. Also, a mild negative correlation with fundamental frequency confirmed that voices with a low pitch sound hoarser.

DISCUSSION

This paper focused on a method to improve the evaluation of the NPV /a/. The use of NPVs resembles the real clinical situation where a patient's voice is evaluated.

Considering that external references proved to be more stable anchors (2, 3, 4, 5), we first concentrated on how to use those references in a matching task. A first question raised during the design of this method was the following: Can multidimensional matching of external references benefit from the results already obtained for the evaluation of breathiness and roughness separately? Should raters match her or his overall impression, or should they evaluate each particular attribute?

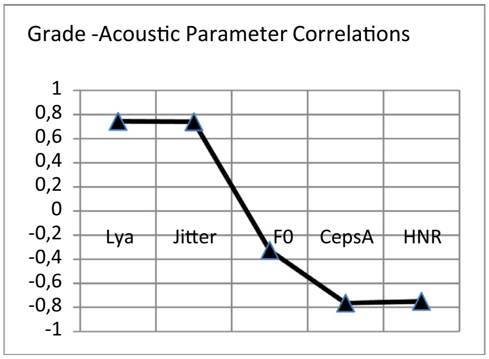

FIGURE 12. Correlation between grades estimated with the standard method and acoustic parameters.

It has been shown that multidimensional auditory matching is a more challenging task.[29,30] Looking at the problem of multidimensional parameters in complex signals, Bregman[31] argued that recognizing a timbre when it is mixed with other timbres depends on the perception of simultaneous temporal-spatial perceptual attributes that could be segregated and also integrated according to the scene analysis. The auditory system can group certain acoustic components into the same perceptual stream and discard others. Also, a number of situations have been described involving timbres of signals embedded in mixtures, which can be recognized even if there are no simple acoustic marks for segregating the partials from the mixture.[32] Attribute discriminations are indeed influenced by cognitive processes related to learning and training.[33,34] Then, based on the human ability to perform judgments in a single percept in stimuli characterized by multiple perceptual attributes, we adopted the matching of each attribute of the NPV with the same attribute in a synthetic vowel. As some acoustic interactions occur, raters could modify attribute levels any time to have the best overall coincidence.

A second important aspect was how to help raters to build up the matching vowel. Patel et al[5] warned that multiple acoustic parameter combinations that could reflect identical percepts should be avoided. To solve this inconvenience, raters were asked to a) manipulate perceptual attributes, not acoustic parameters; b) perceptually define in advance the dominant attribute between roughness and breathiness and to make a categorical decision between asthenic, strained, or modal attributes; and c) compare each attribute sequentially in line with previous settings so raters could concentrate on each attribute, one at a time.

TABLE 10.
Correlation Matrix for Grade, Estimated With the Standard Method, and Acoustic Attributes

	LME	Jitter	*F*0	CPA	HNR
Grade	0.742828	0.7389	−0.3268	−0.76395	−0.74935
LME	—	0.4601	−0.2551	−0.70973	−0.482351
Jitter	—	—	−0.0331	−0.35557	−0.757501
F0	—	—	—	0.278239	−0.050599
CPA	—	—	—	—	0.6017922

Abbreviation: CPA, cepstral peak amplitude.

In this work raters started setting levels of pitch, timbre, loudness, roughness, then breathiness and asthenia/strain. During training, expert raters agreed that it was difficult to discriminate attributes in some voices. One of the difficulties is that some dysphonic voices have mixed attributes that alternate during the utterance. Depending upon where a rater focuses his or her attention, a set of completely opposite observations could be obtained. Raters became aware of these ambiguities, and we chose perceptual dominance by majority votes as a way of attenuating this inconvenience. This positive training was reflected in the reduction of expert raters' dispersions in Figure 3 and Figures 5 to 8 when compared with the standard method.

Based on responses obtained from experts to NPVs, it was evident that the construction of a synthetic matching stimulus helps evaluate quality attributes more precisely, as indicated in previous works for individual attributes.[3–6]

The availability of the initial vowel set (see the section “The initial vowel set and attributes definitions”) helped to start quickly with the matching task. With the variety of pitch and roughness levels already made, raters should concentrate in a) breathiness levels and harmonic loudness and b) asthenia or strain levels.

Main differences with previous works are as follows: 1) raters had direct control over each attribute dimension—roughness, breathiness, and asthenia/strain—avoiding control of multiple acoustic parameters; 2) attributes are matched individually, and a final whole percept is also compared, not just a single dimension; 3) matching is performed between NPVs and an artificial vowel produced by the same raters; and 4) asthenia and strain dimensions were defined as one continuum to help rater decisions.

In order to quantify agreement between raters, we used the Kappa index. This average index was 0.67 for the matching method and 0.44 for the standard method. Mean interjudge correlation indexes presented by Patel et al[5] for matching experiments of roughness (0.67) and by Shrivastav et al[29] for breathiness (0.72) are similar to Kappa indexes obtained here. In the present evaluation subjects also found it more difficult to reach a consensus for roughness than for breathiness as reported by these last authors. A possible answer is the difficulty to agree in roughness when fundamental frequency is low. This happens for stimuli K, L, C, I, and A, where the lowest $F0$ of the whole set are present: 113, 120, 105, 122, and 117 Hz, respectively (see Tables 5 and 7). Best consensus was obtained for stimuli H and J, which are the extreme representatives of the set; that is, a normal vowel and one with high perturbation NPV. Attributes ordered from best to worst consensus were breathiness, strain, asthenia, and roughness, with Kappas of 0.58, 0.67, 0.68, and 0.73, respectively (see Figure 3 on top, and Figures 5–8). In all cases they showed better Kappas than those obtained with the standard method for this NPV set.

The proportion of equal responses for each expert when they employ both methods, new and standard, computed for each voice attribute was used to evaluate their performance; ie, experts 6 and 7 present the lowest percentages of concordance, and perhaps they need more training.

The evaluation of SLP students was calculated relative to average answers of expert raters. For 1 SD of expert responses, only 6.16% of students presented Kappa indexes greater than the average expert Kappa. For 2 SD, it increased to 84.62%. The best compromise was obtained for 1.5 SD, where 75.39% of the students were within the expert range. SLP students could auto-evaluate their progress, obtaining his or her individual Kappa calculated with reference to 1.5 SD of SLP expert averages, and compare it with the Kappa threshold of 0.67.

Grade was evaluated by the standard method. High correlations were obtained with roughness, breathiness, asthenia, and strain evaluated in the two tasks. The correlation of grade with jitter and LME (positive) and cepstral peak amplitude and HNR (negative) indicates that grade and hoarseness could be calculated with an equation.

CONCLUSIONS AND FUTURE WORK

A perceptual evaluation method was presented, which consists of the production of a synthetic vowel by direct adjustment of their quality attributes and its interactive matching against a patient's vowel. Results show that overall dispersion measured with the SD of mean judgments and a number of outliers for experts were reduced from previous works' results when isolated attributes were tested. A measure of agreement—the Kappa index—increased for all attributes relative to the standard method. Average expert Kappa could be used to evaluate SLP students and trainees, counting as correct responses those within a range of plus-minus 1.5 SD of an expert rater's responses. The method is also useful for teaching new students about the complex attribute combinations and evaluating their progress.

As future work, we propose a) an evaluation of patients with spasmodic dysphonia with different tremor levels; b) a more precise definition of the grade of dysphonia or hoarseness in order to create a model related to subject responses; and c) a study with a different set of SLP experts to assess for their intravariability.

Acknowledgments

The authors thank the group of experts and students at the University of Buenos Aires and Manuela Beltrán University for their participation in the evaluations. In particular, we would like to express our gratitude to speech therapists Clara Campos, Marcelo Ramos, and Gabriela Castelli. We are also grateful to the anonymous reviewers who helped improve this paper.

REFERENCES

1. Kreiman J, Gerrat BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial and frame work for future research. *J Speech Hear Res*. 1993;36:21–40.
2. Gerrat BR, Kreiman J. Measuring vocal quality with speech synthesis. *J Acoust Soc Am*. 2001;110(5 pt 1):2560–2566.
3. Chan MK, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126.

4. Patel SA, Shrivastav R, Eddins DA. Perceptual distances of breathy voice quality: a comparison of psychophysical methods. *J Voice*. 2011;24:168–177.
5. Patel SA, Shrivastav R, Eddins DA. Identifying a comparison for matching rough voice quality. *J Speech Lang Hear Res*. 2012;55:1407–1422.
6. Gurlekian JA, Torres HM, Vaccari ME. Comparison of two perceptual methods for the evaluation of vowel perturbation produced by Jitter. *J Voice*. 2016;30:506.e1–506.e8. https://doi.org/10.1016/j.jvoice.2015.05.009.
7. Klatt DH. Software for a cascade/parallel formant synthesizer. *J Acoust Soc Am*. 1980;67:971–995.
8. Kreiman J, Antoñanzas-Barroso N, Gerrat BR. Integrated software for analysis and synthesis of voice quality. *Behav Res Methods*. 2010;42:1030–1041. https://doi.org/10.3758/BRM.42.4.1030.
9. Therhard E. Noise induced shifts in the pitch of pure and complex tones. *J Acoust Soc Am*. 1981;7 0:1661–1668.
10. Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am*. 1990;87:820–857.
11. Shrivastav R. The use of an auditory model in predicting perceptual ratings of breathy voice quality. *J Voice*. 2003;17:502–512.
12. Borzone de Manrique AM. *Manual de Fonética Acústica*. Buenos Aires: Hachette; 1980.
13. Gurlekian JA, Elisei NG, Eleta M. Caracterización articulatoria de los sonidos vocálicos del español de Buenos Aires mediante técnicas de resonancia magnética. *Fonoaudiológica*. 2004;50:7–14 [in Spanish].
14. Ruinskiy D, Lavner Y. *Stochastic models of pitch jitter and amplitude shimmer for voice modification*. In: Proc. of 25th IEEE Convention of Electrical and Electronics Engineers. Dec. Israel. 2008.
15. Guirao M, Garavilla JM. Perceived roughness of amplitude-modulated tones and noise. *J Acoust Soc Am*. 1977;60:1335–1338.
16. Alzamendi GA, Schlotthauer G, Rufiner HR, et al. Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters. *Latin Am Appl Res*. 2013;43:225–230.
17. Rabiner LR. Digital formant synthesizer for speech synthesis studies. *J Acoust Soc Am*. 1968;43:822–828.
18. Titze IR. *Workshop on acoustic voice analysis: summary statement*. Technical report. Denver, USA: National Center for Voice and Speech. 1995.
19. Torres ME, Schlotthauer G, Rufiner HL, et al. Empirical mode decomposition. Spectral properties in normal and pathological voices. In: Vander Sloten J, Verdonck P, Nyssen M, et al., eds. *ECIFMBE 2008, IFMBE Proceedings*. Vol. 22, Berlin, Heidelberg: Springer-Verlag; 2009:252–255.
20. Rabiner LR, Shafer RW. *Digital Processing of Speech Signals*. Boston: Pearson Education; 2006.
21. Patel SA, Shrivastav R, Eddins DA. Developing a single reference signal for matching breathy voice quality. *J Speech Lang Hear Res*. 2012;55:639–647.
22. Gauffin J, Sundberg J. Spectral correlates of glottal voice source waveform characteristics. *J Speech Hear Res*. 1989;32:556–565.
23. Hammarberg B, Fritzell B, Gauffin J, et al. Acoustic and perceptual analysis of vocal dysfunction. *J Phonetics*. 1986;14:533–547.
24. Löqvist A, Shalén L. Perceptual and acoustic analysis of the voice in acute laryngitis. *Actes du XIIeme Congrès International dès Sciences Phonétiques*. 1991;4:342–345.
25. Gurlekian JA. El laboratorio de Audición y Habla del LIS. In: Guirao M, ed. *Procesos Sensoriales y Cognitivos*. Buenos Aires: Editorial Dunken; 1997:55–81.
26. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213–220.
27. Fleiss J. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378–382.
28. Carletta J. Assessing agreement on classification task: the kappa statistic. *Comput Linguist*. 1996;22:249–254.
29. Shrivastav R, Camacho A, Patel S, et al. A model for the prediction of breathiness in vowels. *J Acoust Soc Am*. 2011;129:1605–1615.
30. Kreiman J, Gerratt BR, Berke GS. The multidimensional nature of pathological voice quality. *J Acoust Soc Am*. 1994;96:1291–1302.
31. Bregman AS. *Auditory Scene Analysis, the Perceptual Organization of Sound*. Cambridge, MA: MIT Press; 1994.
32. Nordmark JO. Time and frequency analysis. Tobias JV, ed. *Foundations of Modern Auditory Theory*. Vol. 1, New York: Academic Press; 1970.
33. Chan MK, Yiu EM. A comparison of two perceptual voice evaluation training programs for naïve listeners. *J Voice*. 2006;20:229–241.
34. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544.