Comparison of Two Perceptual Methods for the Evaluation of Vowel Perturbation Produced by Jitter

*Jorge A. Gurlekian, *Humberto M. Torres, and †María E. Vaccari, *†*Buenos Aires, Argentina*

Summary: Objectives. To explore perceptual evaluation of jitter produced by fundamental frequency ($F0$) variation in a sustained vowel /a/, using two different methods. One is based on listener's internal references and the other is based on external references provided by the experimenter.

Methods. We used two methods: one is magnitude estimation-converging limits (ME-CL), which is close to the standard approach used by speech therapists when they use numerical estimations and their own standards, and other is intramodal matching procedure (IMP), where each matched stimulus is to be compared with a fixed-set matching stimuli. Systematic variations were introduced in vowel /a/ by Linear Prediction Coding synthesis using an $F0$ contour function obtained from a statistical jitter model. Six jitter values were used for each of two reference $F0$ values. Three groups of listeners were tested: expert speech therapists, speech therapy students, and naïve listeners.

Results. Perceptual functions appear to be similar and linear for both methods as the theory predicts. The answers of all groups of listeners tested with ME-CL present higher standard deviations than for IMP. When subjects were tested with IMP, intrareliability and interreliability measurements show a significant improvement for both expert and naïve listeners.

Conclusions. Both intraindividual and interindividual differences for expert speech therapists could be better managed when tested with an IMP than when they use numerical estimations and internal standards to evaluate vowel perturbation produced by jitter. This procedure could be the basis for the development of a clinical evaluation tool.

Key Words: Perception–Voice evaluation–Magnitude estimation–Matching.

INTRODUCTION

The lack of agreement and reliability regarding the perceptual evaluation of dysphonic voices is well known among speech therapists, in particular when variations of fundamental frequency (jitter) and amplitude (shimmer) occur.[1] Experts at the clinic make use of their internal references to judge perturbation. One approach for training young speech pathologists is that they develop their own internal standard earned with the experience of years of listening. Experts noted that when this training is in progress, internal references may also change over time influenced by the level of severity of their patients. To perform this task in a reliable manner, the professional must develop a kind of absolute ear for frequency perturbation over time. But, even if they accomplish the task with repeatability, reaching interrater consensus is much less probable because each professional develops his/her own reference. Perceptual judgments at the clinic, using a variety of experimental tasks such as visual analog[2] and magnitude estimation such as RBH (Rauheit/Roughness, Behauchtheit/Breathiness, and Heisekeitsome/Hoarseness)[3] or GRBAS (Grade, Roughness, Breathiness, Asthenia, and Tension),[4] always require that listeners rely on an internal reference or scale. Regarding these methods, some doubts arise about the scale that different clinicians are using to judge perturbation: is it a category scale or a proportion scale? Guirao[5] introduced the method of Magnitude Estimation-Converging Limits (ME-CL) indicating that this method is a compromise between both scales. Eisler and Guirao[6] claim that this procedure allows more consistent judgments and produces less individual scatter in the data than the conventional magnitude estimation method. In this article, ME-CL method was used as one of the methods to represent the use of internal references.

Nevertheless, sensation thresholds to $F0$ variations could change according to listener's levels and types of experience. In a recent work,[7] the effect of types of experience was shown to be more important than levels on the judgment of voice quality. The use of numerical scales and a variable remembering of the standards are also some of the conditions that contribute to the variability at the clinic, that is, new professionals with less exposure to pathologic voices are likely to have different standards. Even at the laboratory, some experienced listeners using a four- or seven-point categorical scale show that the internal standard is inherently unstable and judgments shift relative to the set of stimuli used in the experiment. The use of a reference standard could help but is not a guarantee to reduce variability. Paired comparisons did not increase reliability between listeners either.[8] In a tutorial review article,[9] an external reference was recommended as a possible solution to cope with the sources of variability. Following this advice, Gerratt and Kreiman[10] proposed a multiparameter matching task to quantify voice quality.

For clinical practice, sorting stimuli has been proposed to avoid the use of internal references.[11] An interesting scheme for the purpose of clinical practice was investigated and compared with the GRBAS scale.[12] In this approach, raters

Accepted for publication May 15, 2015.

From the *Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Buenos Aires, Argentina; and the †Facultad de Medicina, University of Buenos Aires, Fonoaudiología, Buenos Aires, Argentina.

Address correspondence and reprint requests to Jorge A. Gurlekian, Laboratorio de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Av Cordoba 2351, 9 piso Sala 2, 1120 C.A.B.A., Argentina. E-mail: jag@fmed.uba.ar

Journal of Voice, Vol. 30, No. 4, pp. 506.e1-506.e8
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2015.05.009

must reorder a sequence of voices with different degrees of perturbation as a way of training. Despite encouraging results as evaluators ordered pathological voices, they complained about the difficulty of order in one dimension, stimuli with complex variations of roughness and breathiness.

Psychophysical methods were compared to evaluate breathiness[13,14] showing less dispersion for a matching-production task.

The present work explores an experimental procedure following the idea of using external standards. Our main interest is to focus in a method to improve the interrater agreement and reliability. Experiments were conducted with several external references on the basis of the matching task[15] to evaluate the capacity to identify percentages of jitter using this method. Agreement between expert listener responses for matching will be compared with results obtained from the ME-CL method.

We chose jitter as one of the acoustic correlates of roughness, considering that psychoacoustic functions of roughness use frequency modulation to obtain the physical continuum.[16,17] Another reason for this selection is that abnormal jitter values are more frequently found in normal voices when compared with shimmer[18] although the presence of shimmer produces a similar effect than jitter to produce roughness sensation,[7] and finally because only few works produced systematic jitter variations.

Experimental procedure

Stimuli. On the basis of the acoustic parameters of both female and male[19,20] vowels /a/ indicated on Table 1, two synthesized vowel versions were created by an Linear Prediction Coding (LPC) formant synthesizer at a sampling rate of 50 KHz and 16 bits. Once the average $F0$ was selected, either 120 Hz or 240 Hz, it remained constant for all the vowel duration of 3 seconds for a jitter value of 0%. This supranormal stimulus was located at one extreme of the continuum. At the other extreme, we produced a stimulus with a jitter value of 3% according to Equation 1. Five intermediate jitter values completed the linear sequence of values with variation percentages indicated on Table 2.

Fundamental frequency variation over time was created to produce stimuli with these intended jitter values. One possible way to do this is to introduce random noise in the glottal source to create controlled variations.[21,22] The method used in this article uses a statistical model of jitter.[23] 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[24] and Torres et al,[25] $F0$i values have a Gaussian behavior with normal density probability functions ($F0,\sigma_{F0}$). 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 $F0$, either 120 Hz or 240 Hz, and standard deviation given by the Equation (2).

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

Once the $F0$ set which represents the glottal source is defined, stimuli are synthesized using the LPC method.[26] Stimuli are further verified to insure that they effectively have the programmed perturbation. Differences between programmed and real values do not exceed a maximum value of 4% as summarized in Table 2. Figure 1 presents an example of original and modified stimulus signal, with a jitter of 0 and a jitter of 3, respectively.

TABLE 1.
***F0* and Formant Values, in Hz**

Parameter	Male	Female
F0	120	240
F1	700	900
F2	1250	1450
F3	2500	2500
F4	3500	3500
F5	4000	4000

TABLE 2.
Jitter in % for Intended Values and Their Real Production and the Difference Between Them

	120 Hz			240 Hz		
No.	Planned	Real	Dif	Planned	Real	Dif
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.50	0.52	0.02	0.50	0.52	0.02
3	1.00	1.02	0.02	1.00	1.06	0.06
4	1.50	1.56	0.06	1.50	1.51	0.01
5	2.00	2.08	0.08	2.00	1.99	0.01
6	2.50	2.55	0.05	2.50	2.53	0.03
7	3.00	3.06	0.06	3.00	2.97	0.03

Test protocols

Magnitude estimation-converging limits method. For test 1, we used the ME-CL method.[6] Six stimuli indicated in Table 2 (from 2 to 7) were presented in a random order for each listener, with four repetitions each and 4 seconds of interstimulus interval for a total of 24 trials. The test was implemented in a graphical software interface for orientation, testing, and response collection. Instruction to subjects was "you will hear the vowel /a/, your task is to evaluate the degree of perturbation by giving a number to each stimuli according to a numerical scale of your preference. Always assign high numbers to stimuli with high perturbation. Assign low numbers