Perturbation Measurements on the Degree of Naturalness of Synthesized Vowels

***Rosiane Yamasaki, †Arlindo Montagnoli, *Emi Z. Murano, ‡Eloisa Gebrim, *Adriana Hachiya, §Jorge Vicente Lopes da Silva, ¶Mara Behlau, and *Domingos Tsuji,** *†‡¶*São Paulo, and* §*Campinas, Brazil*

Summary: Objective. To determine the impact of jitter and shimmer on the degree of naturalness perception of synthesized vowels produced by acoustical simulation with glottal pulses (GP) and with solid model of the vocal tract (SMVT).
Study Design. Prospective study.
Methods. Synthesized vowels were produced in three steps: 1. Eighty GP were developed (20 with jitter, 20 with shimmer, 20 with jitter+shimmer, 20 without perturbation); 2. A SMVT was produced based on magnetic resonance imaging (MRI) from a woman during phonation-/ɛ/ and using rapid prototyping technology; 3. Acoustic simulations were performed to obtain eighty synthesized vowels-/ɛ /. Two experiments were performed. First Experiment: three judges rated 120 vowels (20 humans+80 synthesized+20% repetition) as "human" or "synthesized". Second Experiment: twenty PowerPoint slide sequences were created. Each slide had 4 synthesized vowels produced with the four perturbation condition. Evaluators were asked to rate the vowels from the most natural to the most artificial.
Results. First Experiment: all the human vowels were classified as human; 27 out of eighty synthesized vowels were rated as human, 15 of those were produced with jitter+shimmer, 10 with jitter, 2 without perturbation and none with shimmer. Second Experiment: Vowels produced with jitter+shimmer were considered as the most natural. Vowels with shimmer and without perturbation were considered as the most artificial.
Conclusions. The association of jitter and shimmer increased the degree of naturalness of synthesized vowels. Acoustic simulations performed with GP and using SMVT demonstrated a possible method to test the effect of the perturbation measurements on synthesized voices.
Key Words: Synthesized voices–Acoustical measurements–Auditory-perceptual evaluation–Naturalness perception–Vocal tract model.

INTRODUCTION

Naturalness and intelligibility are two important factors that affect the quality of synthesized speech. Naturalness refers to how closely the output sounds like human speech. Intelligibility refers to the ease with which the production is understood.[1] The perception of naturalness of synthesized voices has a direct impact in its acceptability and use in a wide range of technological applications.[2] To improve the quality of synthesized voices, it is important to know which acoustic characteristics of the human voice may contribute to increase the naturalness of the stimulus.

The human vocal signal has small perturbations in its acoustic wave that are considered normal phonation variations.[3–5] These perturbations occur because the successive glottal cycles produced by the vibration of the vocal folds have variations in both pitch and amplitude. Jitter and shimmer are traditional short-term perturbation measurements of the cycle-to-cycle variations of the fundamental frequency (f_0) and amplitude, respectively.[3,4,6] These measurements are widely used in voice assessment and in scientific research. Acoustic analysis software provides measurements that produce objective information about the laryngeal function, including irregular vocal fold vibration.[6] Generally, individuals with dysphonia have higher levels of jitter and shimmer when compared with individuals without dysphonia. This is due to difficulties in maintaining the regularity of vibratory cycles because of damage to the tissues or muscles or to neurologic disorders. These perturbation measures have been used to differentiate normal voices from dysphonic voices,[7,8] to compare the vocal evaluation before and after vocal therapy,[9,10] and also to evaluate phonosurgery outcomes.[11–13]

Synthesized voices can be produced using mathematical models of the voice source and of the vocal tract. Different mathematical models of the voice source can be chosen to develop glottal pulses. In a study by Kreiman et al,[14] the fit of five models of the voice source were examined—the Rosenberg model, the Fujisaki-Ljungqvist model, the Liljencrants-Fant model, and two models proposed by Alwan and colleagues—to provide information about which fits are perceptually important to listeners. In addition to mathematical models of the voice source, the inclusion of some perturbation measurements on glottal pulses may contribute to the production of synthesized voices with vocal quality closely aligned to human voice. Jitter and shimmer measurements have been used in the development of synthesized voice stimuli.[15–18] Rozsypal and Millar[19] investigated synthesized vowels with different amounts of jitter and shimmer and observed that some degree of roughness is necessary for vowels to be perceived as natural. Hillenbrand[20] studied the perception of aperiodicity in synthesized vowels. He found that the stimuli generated from sequences with a certain degree of correlation between adjacent pitch and amplitude values sounded rougher than the ones generated from uncorrelated sequences, especially for stimuli with variation in amplitude perturbation. The results also suggested that it is jitter that more

Accepted for publication September 19, 2016.
From the *Department of Otorhinolaryngology, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil; †Department of Electrical Engineering, Universidade de São Paulo—São Carlos, Brazil; ‡Department of Radiology, InRad, Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Brazil; §Division of Tridimensional Technologies, Centro de Tecnologia da Informação Renato Archer, Campinas, Brazil; and the ¶Department of Speech-Language Pathology and Audiology—Universidade Federal de São Paulo—UNIFESP; Centro de Estudos da Voz—CEV, São Paulo, Brazil.
Address correspondence and reprint requests to Rosiane Yamasaki, Rua Oscar Freire, 2250 5o andar/502, CEP 05409-011, SP, Brazil. E-mail: r.yamasaki@uol.com.br
Journal of Voice, Vol. 31, No. 3, pp. 389.e1–389.e8
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.09.020

closely approximates the roughness that is present in human dysphonic voices. Shimmer, when added alone, did not produce a natural vocal quality. Thus, synthesized voice signals that are absolutely stable are not perceived as human.

Mathematical models of concatenated tubes representing the trachea and the vocal tract have been used to produce synthesized vowels.[3,21,22] However, the mathematical models do not represent the actual three-dimensional (3D) geometric configuration of the vocal tract, which is quite complex. In the literature, there are few reports that used solid models to explore the acoustic characteristics of the vocal tract. Some studies have been addressed to examine the acoustic characteristics of the piriform fossa using solid models.[23,24] In 2005, Fujita and Honda[25] produced synthesized vowels by acoustic simulations with solid models of the vocal tract and the Rosenberg model of the voice source. These procedures allowed the authors to study the acoustic characteristics of the hypopharyngeal cavities with real 3D configuration of the region. Clearly, there remains a need to better understand the output of the vocal signal under conditions similar to human voice production.

The development of synthesized vowels using glottal pulses generated by mathematical models and a solid model of the vocal tract can be a specific resource to test the auditory-perception impact of short-term perturbation measurements related to the degree of naturalness of synthesized vowels. The aim of this research was to determine the impact of short-term perturbation measures, jitter and shimmer, on the degree of naturalness perception of the synthesized vowels, produced by acoustic simulations, using auditory-perceptual evaluation.

METHODS

This study was approved by the Ethics Committee from Comissão de Ética para Análise de Projetos de Pesquisa (CAPPesq)—number 188.183.

Production of synthesized vowels

The three steps we used to develop the synthesized vowels were the following: (1) development of glottal pulses with and without jitter and shimmer; (2) production of a 3D solid model of the vocal tract; and (3) performance of acoustic simulations to obtain synthesized phonations of vowel /ɛ/.

Glottal Pulses

The synthesized glottal pulses were generated from 20 human productions of vowel /a/: 9 from vocally healthy women, and 11 from women with dysphonia, aged from 20 to 40 years. All human voices were recorded in similar conditions: quiet environment, voices recorded directly on a notebook computer (Dell Inspiron 15R– 5537; Dell Inc, USA); microphone headset (Karsect HT-9; Karsect, Brazil) connected to an external sound card (Andrea PureAudio USB; Andrea Electronics, USA). The microphone was located at a 45-degree and 5-cm distance from the speaker's mouth. The voice recorder was performed using *FonoView* software (CTS Informática, Brazil). The value of jitter (local), obtained using *Praat* software, ranged from 0.10% to 0.39% in vocally healthy women, and from 0.22% to 0.62% in women with dysphonia. The value of shimmer (local) ranged from 0.65% to 1.31% in vocally healthy women and from 1.07% to 2.80% in women with dysphonia. Traditional digital signal processing algorithms[26] were developed using C# language in Visual Studio platform, generating a graphical program that extracted values of the f_0, jitter, and shimmer of each human vowel. From those extracted values, glottal pulses based on the Rosenberg Mathematical Model were created.[27] The human vowel /a/ was chosen to avoid any interference on the production of synthesized vowel /ɛ/ during acoustical simulations.

With the knowledge that the periodicity of the correlation function is equal to the periodicity of the sequences involved, the partial autocorrelation uses a simple cross-correlation function of the voice signal with a single sample containing a fundamental period:

$$r_{xy[l]} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\,y[n-l] \tag{1}$$

Considering that:

$$\begin{cases} y[n] = x[n] & para \quad n < n_o \\ y[n] = 0 & para \quad n \geq n_o \end{cases} \tag{2}$$

where:

$n_o = F_s/F_0$, number of samples related to the f_0
F_s = sound card sampling frequency
F_0 = f_0 of the vocal signal

The fact that peaks can be seen on the correlated signal indicates clearly that each period can easily be identified. This period contains the information of the respective perturbations, as shown in the example of Figure 1.

This information generates a so-called impulse train signal that has the same information regarding the amplitude or periods of the original signal of real voice. Subsequently, the impulse train is converted to a glottal pulse train by traditional convolution sum in Equation 3. The single glottal pulse *h[n]* is generated by the Rosenberg model in Equation 4. This function, denoted by f_D in Rosenberg's work, was our choice because it does not have slope discontinuities. These equations can be literally translated to a computational algorithm. Finally, the *y[n]* result is used to excite the 3D model that contains the same perturbation of the impulse.

$$y[n] = \sum_{k=0}^{N-1} x[k]\cdot h[n-k] \tag{3}$$

Considering that:

y[n]: glottal pulse train
x[n]: impulse train
h[n]: Rosenberg model of glottal pulse

$$\begin{cases} f_D = \dfrac{a}{2}\left[1-\cos\left(\dfrac{t}{t_P}\pi\right)\right]; & 0 \leq t \leq T_P \\ f_D = \dfrac{a}{2}\left[1+\cos\left(\dfrac{t-t_P}{t_N}\dfrac{\pi}{2}\right)\right]; & T_P \leq t \leq T_P + T_N \end{cases} \tag{4}$$

where:

T_P: portion of the pulse with positive slope
T_N: portion of the pulse with negative slope

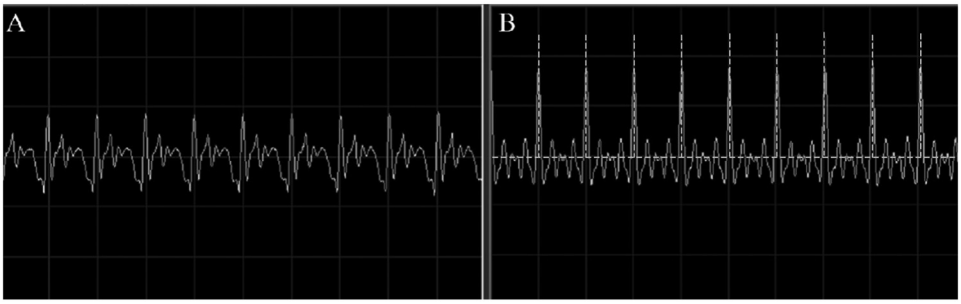

FIGURE 1. **(A)** Original vocal signal; **(B)** partial autocorrelation function (*gray*) and impulse train (*dashed line*).

In this research, the glottal pulses were generated using T_P/T of 0.45 and T_N/T of 0.09. The Voice Analysis Program, version 1.5,[28] was used to generate glottal pulses with four possibilities of manipulation: (1) without perturbation; (2) with the addition of jitter; (3) with the addition of shimmer, and (4) with the addition of jitter and shimmer. Without the addition of jitter, the period of impulse signal maintained a fundamental period equal to that of human voice, and without the addition of shimmer, the impulse signal amplitude was kept constant. The perturbation added was equal to the extracted variation of each period and its respective amplitude of human voice throughout the time of speech, as described above. Therefore, the jitter and shimmer added were the same as that found in each human voice. A total of 80 glottal pulses were generated.

The sound of the glottal pulses was similar to a "buzz" once there was no vocal tract involved. Figure 2 shows examples of glottal pulses and the spectral envelope of the four synthesis conditions. Supplementary audios of the four glottal pulses are presented in MP3 format.

Solid model of the vocal tract

Five solid models of the vocal tract were developed with different physical characteristics, such as material, thickness, and formfitting to the driver. The same glottal pulse train was used in the acoustical simulation. The selected solid model of the vocal tract was the one that generated the best auditory-perceptual analysis for the vowel /ɛ/; for this selection, a speech-language pathologist specializing in voice performed the auditory-perceptual evaluation of the five synthesized vowels, considering the intelligibility and the vocal quality. A solid model with 1 mm thickness generated vowels with low intensity. Thus, the simulations were performed with a solid model of 2 mm thickness (Figure 3). The solid model was developed from magnetic resonance images (MRI) of a 23-year-old woman without vocal complaints, with a normal laryngeal examination, a normal auditory-perceptual evaluation of voice, acoustic parameters within normal limits, and no disadvantages of self-assessment protocols for vocal disorder. The pharyngeal structures and the oral cavity structures such as tongue, hard and soft palate, and dental arches were normal. The magnetic resonance equipment was a GE 1.5-T scanner (General Electric Medical Systems, Waukesha, WI, USA). The MR protocol was to acquire a 3D image while the participant phonated the sustained vowel /ɛ/ for approximately 8 seconds. The vowel /ɛ/ is similar to /æ/ in American English. This vowel has less vocal tract modifications than the other vowels.[29]

MR images were segmented semiautomatically using the *InVesalius* program[30] and reconstructed three-dimensionally. The 3D reconstruction of the vocal tract in STL format allowed the solid model to be printed by rapid prototyping technology. The material used for printing was DuraForm PA (3D Systems, USA)

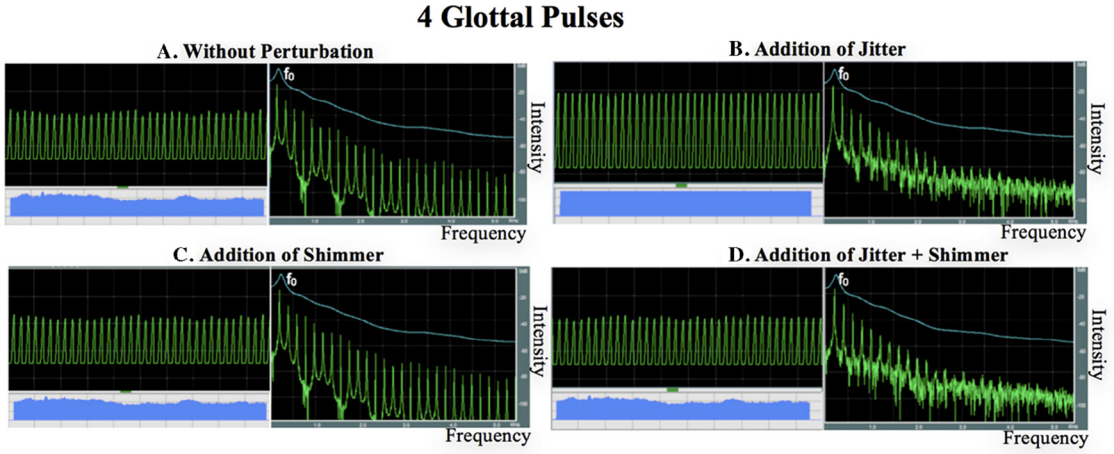

FIGURE 2. Example of four glottal pulses from the same human voice. To the left, sample of the glottal pulses (GP) and the audio signal below. To the right, the spectral envelope with the fundamental frequency peak and no formants. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

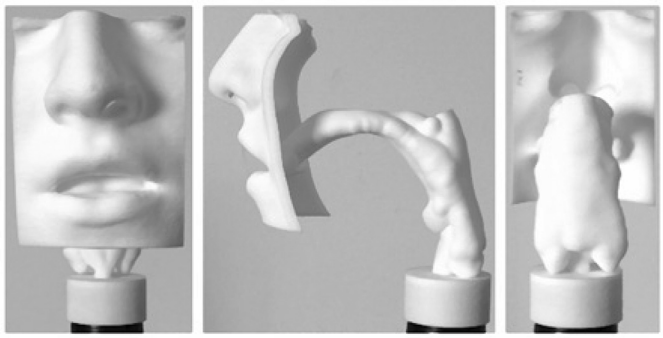

FIGURE 3. Solid model of the vocal tract—vowel /ɛ/.

(nylon) with 2 mm thickness. The lower end of the solid model, near the region where the vocal folds are located, was produced with a circular shape to allow direct connection to the driver, thus allowing transfer of the glottal pulses to the solid model. Figure 3 shows the solid model of the vocal tract with the configuration of the vowel /ɛ/.

Acoustic simulations

The Atlas Sound PD60A driver series with a frequency response of 100–3700 Hz and the audio amplifier TDA7294 100W RMS were used for the acoustic simulations.

Figure 4 shows the instruments and the procedures used for the amplification of the glottal pulses and for the production of the synthesized vowels. When synthesized glottal pulses were amplified and transferred to the driver, only the amplification of the glottal pulses was obtained. The sound was similar to that of an electric shaver. On the other hand, when the solid model of the vocal tract was connected to the driver, synthesized vowel /ɛ/ was obtained. Thus, the driver was a sound generator that excited the solid model of the vocal tract.

Synthesized voices were recorded in a quiet environment using the same notebook computer and software. The microphone was located at a 45-degree and 3-cm distance from the solid model's mouth. Using this procedure, 80 synthesized phonations of vowels with a duration of 2 seconds each were developed: 20 without perturbation, 20 with the addition of jitter, 20 with the addition of shimmer, and 20 with the addition of jitter and shimmer. Supplementary audio signals of the four synthesized vowels are presented in MP3 format.

Figure 5 presents an example of the voice signal and the spectral envelope with the f_0 and the two first formant peaks of the synthesized vowel /ɛ/ for the four acoustic variations.

Auditory-perceptual evaluation

Experiment 1: Human Voices Vs. *Synthesized Voices*

The auditory-perceptual evaluation was performed by three speech-language pathologists specializing in voice, aged from 35 to 50 years. The judges, two women and one man, all had more than 10 years of clinical experience in voice assessment. The judges were asked to classify the vocal sample as human or synthesized without receiving any additional information about the stimuli. The sample consisted of 120 sustained vowel /ɛ/: 20 human, 80 synthesized, and 20 repetitions to test intra-rater reliability. The 20 repetitions comprise both synthesized and human samples that were randomized when selected.

The 20 human vowels were the same as that used for the extraction of the acoustic parameters and production of the synthesized glottal pulses. The vowels were presented in randomized order. The auditory-perceptual analysis was performed individually. The listeners wore a headset that allowed the total

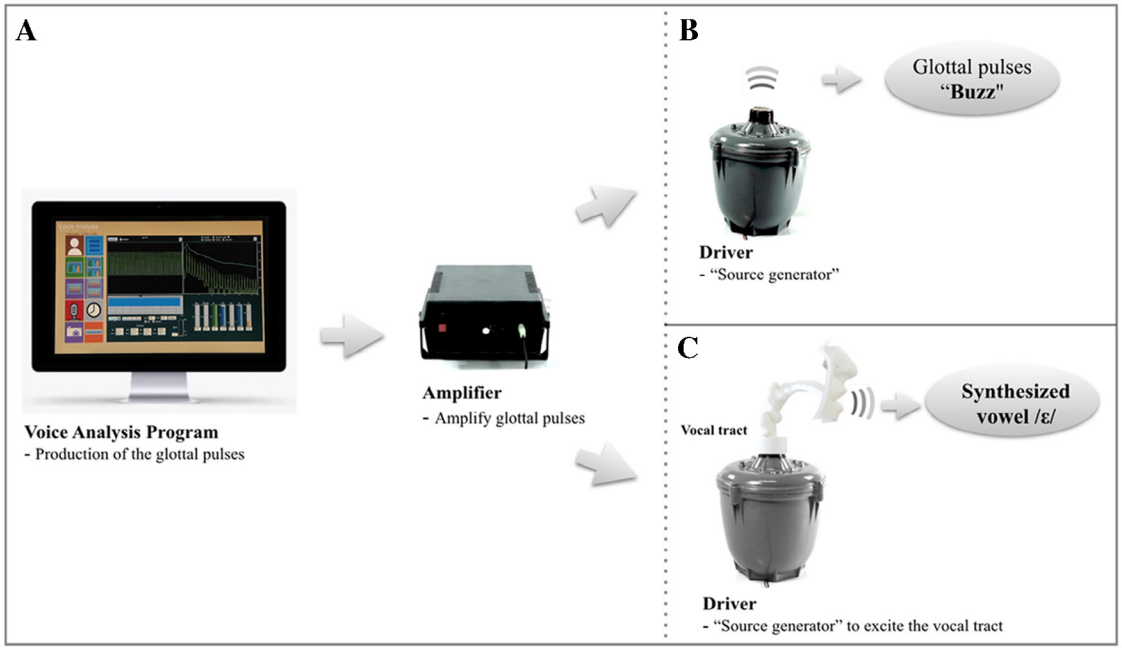

FIGURE 4. Instruments and procedures used for amplification of the glottal pulses and for production of synthesized vowel: **(A + B)** = amplification of the glottal pulses; **(A + C)** = production of synthesized vowel /ɛ/.

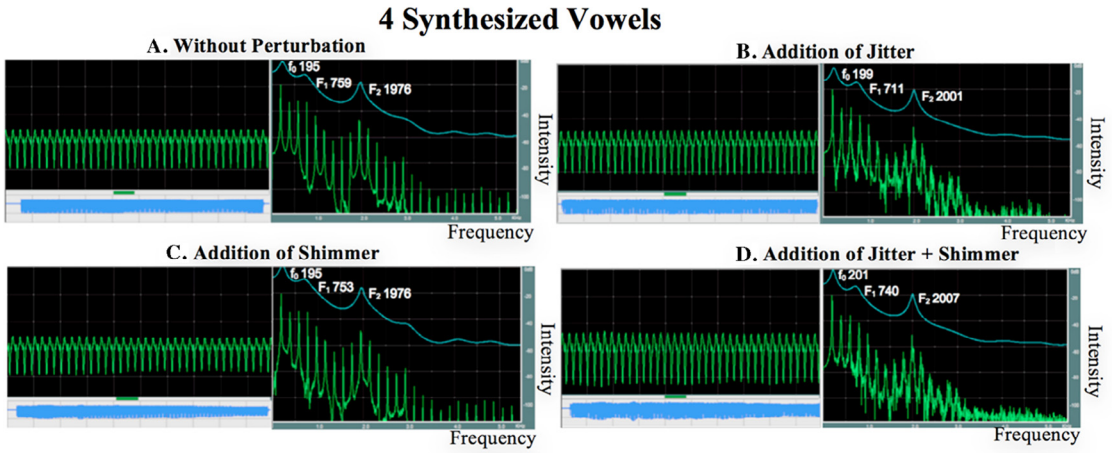

FIGURE 5. Example of four synthesized vowels with glottal pulses obtained from the same vocal sample. Vocal signal on the left, spectral envelope on the right. The spectral envelope shows the fundamental frequency and the first two formants, F_1 and F_2.

occlusion of both ears. Repetition of signals was provided on request.

The final classification of the voices as human or synthesized was based on the consensus response among at least two judges.

The level of inter-rater agreement was verified using the Cronbach test. The intra-rater reliability was tested through McNemar test.

Experiment 2: Synthesized Voices Vs. *Degree of Naturalness*

The same raters of experiment 1 carried out the second experiment following a 15-day interval. The vocal sample consisted of 80 synthesized vowels using 20 PowerPoint slide sequences. Each slide had four synthesized phonations of vowel /ɛ/: without perturbation, with addition of jitter, with addition of shimmer, and with addition of jitter and shimmer. The evaluators were asked to rate the synthesized vowels from the most natural to the most artificial, classifying the most natural vowel as 0 and the most artificial vowel as 3. Vowels were placed randomly on each slide.

The level of inter-rater agreement was verified using the Cronbach test, and the intra-rater reliability was verified using Wilcoxon test.

RESULTS

Experiment 1: Human Voices *Vs.* Synthesized Voices

The inter-rater reliability was 0.793 and the intra-rater reliability varied from 70% to 90%. All human vowels were classified as human (100%). Of the 80 synthesized vowels, 53 (66%) were classified as synthesized and 27 (34%) were classified as human according to the answer key.

Figure 6 shows the distribution of the 27 synthesized voices classified as human, according to glottal pulse characteristics. Of these 27 voices, 15 were produced with the combined use of jitter and shimmer, 10 with the addition of jitter, and 2 without any perturbation measures. No vowel produced only with addition of shimmer alone was classified as human.

Experiment 2: Synthesized Voices *Vs.* Degree of Naturalness

Inter-rater reliability ranged from 0.89 to 0.96. The results of intra-rater agreement were considered acceptable by statistical analysis. Two evaluators, judge 1 and judge 3, considered the vowels produced with jitter and shimmer as more natural, followed by the vowels produced only with addition of jitter. Judge

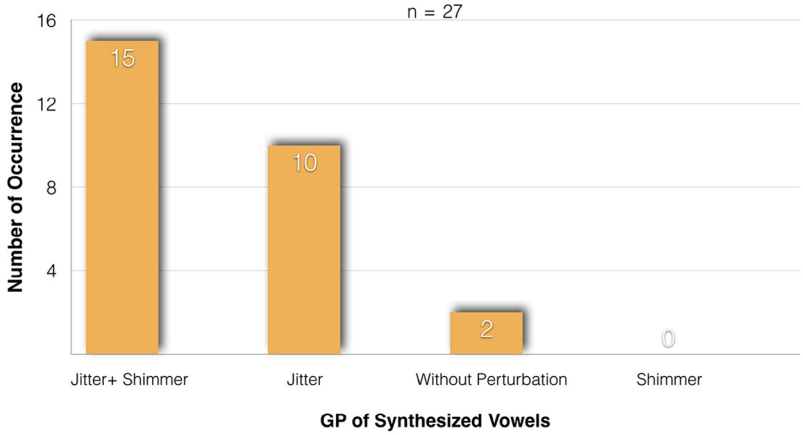

FIGURE 6. Distribution of the 27 synthesized vowels classified as human voices, according to glottal pulses characteristics.

TABLE 1.
Distribution of the Classification of Synthesized Vowels Produced With Four Perturbation Conditions From the Most Natural (0) to the Most Artificial (3), According to Auditory-Perceptual Evaluation Performed by Three Judges

	Vowels	0	1	2	3	Total
Judge 1	Jitter + Shimmer	12	8	0	0	80
	Jitter	8	12	0	0	
	Shimmer	0	0	11	9	
	No perturbation	0	0	9	11	
Judge 2	Jitter + Shimmer	7	13	0	0	80
	Jitter	13	7	0	0	
	Shimmer	0	0	11	9	
	No perturbation	0	0	9	11	
Judge 3	Jitter + Shimmer	11	9	0	0	80
	Jitter	9	11	0	0	
	Shimmer	0	0	12	8	
	No perturbation	0	0	8	12	

2 considered the vowels produced with jitter alone as more natural, followed by vowels produced with jitter and shimmer. The three evaluators classified the vowels produced without perturbation as more artificial, followed by the vowels produced with the addition of shimmer.

Table 1 shows the distribution of synthesized vowels produced with different perturbation conditions, according to the degree of naturalness classification.

DISCUSSION

The naturalness of speech is probably one of the most important aspects in synthesized voice production.[2] The perceived naturalness of the synthesized signal may be obtained by using certain acoustic characteristics of the human voice. Vocal quality is multidimensional, that is, a single voice has different combinations of vocal characteristics.[31] Dysphonic voices are acoustically more complex than normal voices, and some studies have investigated which acoustic properties affect the perception of naturalness.[16,20,32] An important advantage of voice synthesis is the ability to control and systematically vary specific acoustic dimensions that often cannot be achieved with the use of human voices.[20] Also, vocal synthesis has enabled the production of voices with specific qualities and predetermined deviations. These voices can be used for auditory-perceptual training of inexperienced listeners and as anchor stimuli.[21,22]

This study aimed to determine the impact of short-term perturbation measures—jitter and shimmer, both traditionally used in vocal clinical and scientific research—on the degree of naturalness perception of the synthesized vowels. To do so, synthesized vowels were developed using the Rosenberg Mathematical Model[27] to produce the glottal pulses, and a solid model of the vocal tract of a young woman was built to generate the vowel /ɛ/. The acoustic analysis of the synthesized vowels showed small variations in the f_0, F_1, and F_2 values. Usually, the f_0 values of the glottal pulses originated with the addition of jitter alone or with jitter and shimmer were slightly higher than those produced without jitter. F_1 value was lower in vowels produced with addition of jitter alone, and F_2 values were higher in vowels produced with the addition of jitter and shimmer, and with addition of jitter alone (Figure 5).

The first experiment had a vocal sample of 20 human and 80 synthesized vowels. The synthesized stimuli were controlled considering the addition of jitter and/or shimmer to the glottal pulses. The three evaluators were asked to classify the voices as human or synthesized. All human voices and 53 synthesized voice samples (66%) were correctly identified. Evaluators reported greater difficulty classifying the synthesized voice samples than classifying the human voices. The greater amount of perceptual errors occurred in vowels produced with the addition of jitter and shimmer; 15 of the 20 voices produced with this combination were identified as human. Ten of 20 voice samples produced with the addition of jitter were classified as human, only two voice samples produced without perturbation were identified as human, and none with the addition of shimmer alone was considered human. These data suggest that jitter, added alone or in combination with shimmer, has a significant positive impact on the perception of the stimuli naturalness, giving it a potential human identity. The positive effect of perceived naturalness on the variation in pitch perturbation has been described in previous studies.[33,34] However, vowels without perturbation or with the addition of shimmer alone had a negative impact on the perception of naturalness. Vowels produced without jitter were perceived as metallic voices by the three evaluators.

The second experiment included only the analysis of the synthesized vowels. The evaluators were asked to rate each sequence of four synthesized vowels (one with jitter, one with shimmer,

one with jitter and shimmer, and another one without perturbation) from the most natural to the most artificial. It was observed that the voices produced with jitter and shimmer or with jitter alone were judged as more natural. The voices produced without any perturbation measurements or with shimmer alone were considered more artificial. The unnatural quality of the synthesized stimuli varying in amplitude perturbation was described by Hillenbrand.[20]

In a different design, Kreiman and Gerratt[16] performed a study using human and synthetic voices to determine the role of jitter, shimmer, and noise-to-signal ratio in the perception of vocal quality. Listeners were asked to adjust levels of perturbation measures until synthetic voice-matched voice was naturally produced. The results showed that listeners agreed well in their judgments of the noise-to-signal ratio, but they did not agree in their chosen settings for jitter and shimmer. The authors concluded that jitter and shimmer are not useful as independent indices of perceived vocal quality. In the present study, the results of the two auditory-perceptual experiments showed that jitter seems to be important to the perception of naturalness in synthesized vowels, maintaining the quality of human vocal identity. Vowels produced with the addition of shimmer alone were very similar to those with no perturbation and were easily perceived as synthesized. However, the combination of both jitter and shimmer contributed most to the perceived naturalness of the vowels.

LIMITATIONS OF THIS STUDY

Specific limitations suggest caution in the interpretation of the data. First of all, a single solid model of the vocal tract was used, and results need to be analyzed with caution. Second, the MRI images were obtained with the participant in supine position during phonation; therefore, the solid model of the vocal tract was developed with this positional condition. Finally, the noise during MRI acquisition was very high. Although, the participant did not have an ideal feedback of her own voice, she felt that she produced her normal voice.

CONCLUSIONS

The association of jitter and shimmer perturbation measurements increased the degree of naturalness of synthesized vowels. The addition of jitter provided naturalness, whereas the vowels generated with the addition of shimmer alone had a negative impact on the perception of naturalness. This study suggests that the acoustical simulation performed with glottal pulses generated by means of the mathematical model of voice source and using a solid model of the vocal tract is a possible method to test the effect of the perturbation measurements on synthesized voices.

Acknowledgments

The authors thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP: 2012/17390-3) for financial support.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online at doi:10.1016/j.jvoice.2016.09.020.

REFERENCES

1. Smruti S, Sahoo J, Dash M, et al. An approach to design an intelligent parametric synthesizer for emotional speech. In: Satapathy SC, Biswal BN, Udgata SK, et al., eds. *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, Vol. 2. New York: Springer; 2015:367–374.
2. Nusbaum HC, Francis AL, Henly AS. Measuring the naturalness of synthetic speech. *Int J Speech Tech*. 1995;1:7–19.
3. Titze IR. *Principles of Voice Production*. 2nd ed. Iowa City: NCVS; 2000:313–314.
4. Behlau M. *Voice: The Book of the Specialist*. Rio de Janeiro: Revinter; 2001:1–51.
5. Brockmann M, Drinnan MJ, Storck C, et al. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice*. 2011;25:44–53.
6. Brockmann M, Storck C, Carding PN, et al. Voice loudness and gender effects on jitter and shimmer in healthy adults. *J Speech Lang Hear Res*. 2008;51:1152–1160.
7. Zhang Y, Jiang JJ, Biazzo L, et al. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. *J Voice*. 2005;19:519–528.
8. Lopes LW, Cavalcante DP, Costa PO. Severity of voice disorders: integration of perceptual and acoustic data in dysphonic patients. *Codas*. 2014;26:382–388.
9. Petrovic-Lazic M, Jovanovic N, Kulic M, et al. Acoustic and perceptual characteristics of the voice in patients with vocal polyps after surgery and voice therapy. *J Voice*. 2015;29:241–246.
10. Mattioli F, Menichetti M, Bergamini G, et al. Results of early versus intermediate or delayed voice therapy in patients with unilateral vocal fold paralysis: our experience in 171 patients. *J Voice*. 2015;29:455–458.
11. Ziwei Y, Zheng P, Pin D. Multiparameter voice assessment for voice disorder patients: a correlation analysis between objective and subjective parameters. *J Voice*. 2014;28:770–774.
12. Topaloğlu I, Salturk Z, Atar Y, et al. Evaluation of voice quality after supraglottic laryngectomy. *Otolaryngol Head Neck Surg*. 2014;151:1003–1007.
13. Sørensen MK, Durck TT, Bork KH, et al. Normative values and interrelationship of MDVP voice analysis parameters before and after endotracheal intubation. *J Voice*. 2015;30:S0892-1997(15)00146-0 [pii].
14. Kreiman J, Garellek M, Chen G, et al. Perceptual evaluation of voice source models. *J Acoust Soc Am*. 2015;138:1–10.
15. Murphy PJ. Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *J Acoust Soc Am*. 2000;107:978–988.
16. Kreiman J, Gerratt BR. Perception of aperiodicity in pathological voice. *J Acoust Soc Am*. 2005;117(4 pt 1):2201–2211.
17. Sofranko JL, Prosek RA. The effect of levels and types of experience on judgment of synthesized voice quality. *J Voice*. 2014;28:24–35.
18. Kisenwether JS, Sataloff RT. The effect of microphone type on acoustical measures of synthesized vowels. *J Voice*. 2015;29:548–551.
19. Rozsypal AJ, Millar BF. Perceptual dimensionality of jitter and shimmer in synthetic vowels. *J Acoust Soc Am*. 1975;58:S23.
20. Hillenbrand J. Perception of aperiodicities in synthetically generated voices. *J Acoust Soc Am*. 1988;83:2361–2371.
21. Englert M, Madazio G, Gielow I, et al. Perceptual error identification of human and synthesized voices. *J Voice*. 2015;31:S0892-1997(15)00165-4 [pii].
22. Fraj S, Schoentgen J, Grenez F. Development and perceptual assessment of a synthesizer of disordered voices. *J Acoust Soc Am*. 2012;132:2603–2615.
23. Dang J, Honda K. Acoustic characteristics of the piriform fossa in models and humans. *J Acoust Soc Am*. 1997;101:456–465.
24. Honda K, Kitamura T, Takemoto H, et al. Resonance characteristics of the hypopharyngeal cavities. *J Acoust Soc Am*. 2008;123:3731.
25. Fujita S, Honda K. An experimental study of acoustic characteristics of hypopharyngeal cavities using vocal tract solid models. *Acoust Sci Tech*. 2005;26.

26. Oppenheim AV, Schafer RW. *Discrete-Time Signal Processing*. 3rd ed. New Jersey: Pearson; 2010.
27. Rosenberg AE. Effect of glottal pulse shape on the quality of natural vowels. *J Acoust Soc Am*. 1971;49(2 suppl 2):583–590.
28. Montagnoli AN. Voice Analysis Program—1.5. Support system to acoustic analysis of voice: voice analysis; 2015.
29. Gonçalves MI, Pontes PA, Vieira VP, et al. Transfer function of Brazilian Portuguese oral vowels: a comparative acoustic analysis. *Braz J Otorhinolaryngol*. 2009;75:680–684.
30. Amorim P, Moraes T, Silva J, et al. InVesalius: an interactive rendering framework for health care support. In: Bebis G, Boyle R, Parvin B, et al., eds. *Advances in Visual Computing—11th International Symposium—ISVC 2015*. Las Vegas: Springer; 2015:14–16.
31. Gerratt BR, Kreiman J. Measuring vocal quality with speech synthesis. *J Acoust Soc Am*. 2001;110:2560–2566.
32. Yiu EM, Murdoch B, Hird K, et al. Perception of synthesized voice quality in connected speech by Cantonese speakers. *J Acoust Soc Am*. 2002;112:1091–1101.
33. Kersta LG, Bricker PD, David EE Jr. Human or machine? A study of voice naturalness. *J Acoust Soc Am*. 1960;32:1502.
34. Rozsypal AJ, Millar BF. Perception of jitter and shimmer in synthetic vowels. *J Phon*. 1979;343–355.