Perceptual Error Identification of Human and Synthesized Voices

***,†Marina Englert, †Glaucya Madazio, †Ingrid Gielow, ‡Jorge Lucero, and *,†Mara Behlau,** **†São Paulo, ‡Brasília, Brazil*

Summary: Objectives/Hypothesis. To verify the discriminatory ability of human and synthesized voice samples.
Study Design. This is a prospective study.
Methods. A total of 70 subjects, 20 voice specialist speech-language pathologists (V-SLPs), 20 general SLPs (G-SLPs), and 30 naive listeners (NLs) participated of a listening task that was simply to classify the stimuli as human or synthesized. Samples of 36 voices, 18 human and 18 synthesized vowels, male and female (9 each), with different type and degree of deviation, were presented with 50% of repetition to verify intrarater consistency. Human voices were collected from a vocal clinic database. Voice disorders were simulated by perturbations of vocal frequency, jitter (roughness), additive noise (breathiness) and by increasing tension and decreasing separation of the vocal folds (strain).
Results. The average amount of error considering all groups was 37.8%, 31.9% for V-SLP, 39.3% for G-SLP, and 40.8% for NL. V-SLP had smaller mean percentage error for synthesized (24.7%), breathy (36.7%), synthesized breathy (30.8%), and tense (25%) and female (27.5%) voices. G-SLP and NL presented equal mean percentage error for all voices classification. All groups together presented no difference on the mean percentage error between human and synthesized voices (P value = 0.452).
Conclusions. The quality of synthesized samples was very high. V-SLP presented a lower amount of error, which allows us to infer that auditory training assists on vocal analysis tasks.
Key Words: Voice–Dysphonia–Auditory perception–Evaluation–Judgment.

INTRODUCTION

The auditory-perceptual evaluation is an essential tool for dysphonia assessment, as it is the basis of vocal clinic. Auditory-perceptual analysis is used as a diagnostic tool, for outcomes measurement, follow-up, and dismissal.[1–3]

Although it is widely used, it has a variable reliability, intra-rater and interrater,[4] due to the multidimensional characteristics of the human voice and also probably due to the human nature of the auditory processing[5] which makes it a difficult task.[4,6–8] Voice auditory-perceptual evaluation depends on various internal standards and, although there are efforts to reduce interferences in this process, many factors contribute to its misidentification, low reliability, and high variability such as presentation context, personal and professional experiences.[9–11]

Studies highlight that the lack of standards and assessment protocols for the auditory-perceptual evaluation contributes to its high variability and seeks to find a way to standardize them.[12,13] However, the complex nature of the human voice itself makes this analysis complicated. Patients' voices are not always stable and are often characterized with mixed components such as roughness and breathiness, breathiness and strain, or these three main deviations together.[6,9,14,15]

Some researchers suggest the use of controlled anchors stimuli to increase the reliability of the auditory-perceptual evaluation.[1,6,12,16–19]

The anchors stimuli are predefined and selected as representative of a particular type and/or degree of deviation and may be human or synthesized. One of the major advantages of the synthesized stimuli is the exact knowledge of their acoustic properties and the possibility of manipulating its acoustic parameters according to ones desire and/or need, enabling the creation of many samples.[1,6]

Researchers have shown that auditory training also increases reliability of the auditory-perceptual evaluation by decreasing the variability and the subjectivity of this task.[12,16,20] The known characteristic of the synthesized stimuli seems to be practical for its use as anchor or for young clinicians' auditory training, yet, for these purposes, the stimuli must sound natural.

Synthesizers are developed considering acoustic models that make the voice sound even more human and natural. These stimuli allow acoustic parameters control and therefore make it possible to be used in clinical practice and scientific research.[6,14,21–28]

Although the promising use of synthesized voices, whether for research or auditory training, they are not yet a common practice, both due to the difficulty of producing them and being considered unnatural or unpleasant by the listener.[22,26,28–32]

The aim of this study was to verify the discriminatory ability of a synthesized vowel produced by the physics-based

Accepted for publication July 30, 2015.
Presented at The Voice Foundation: May 26–31, 2015, Philadelphia, USA.
From the *Department of Speech Language Pathology and Audiology, Universidade Federal de São Paulo, São Paulo, Brazil; †Voice Department, Centro de Estudos da Voz—CEV, São Paulo, Brazil; and the ‡Universidade de Brasília, Brasília, Brazil.
Address correspondence and reprint requests to Marina Englert, Centro de Estudos da Voz—CEV, R. Machado Bittencourt, 361/1001, Vila Mariana, São Paulo, SP 04044-001, Brazil. E-mail: marinaenglert@gmail.com
Journal of Voice, Vol. 30, No. 5, pp. 639.e17-639.e23
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2015.07.017

synthesizer (VoiceSim) according to its nature of production and to check errors in this classification.

METHODS

This prospective research was approved by the Ethics Committee of the Federal University of Sao Paulo (UNIFESP) under the protocol number 897.232.

Stimuli

A set of human voices, male and female, was selected from a voice bank of a vocal clinic, CLINCEV. The vocal samples represented different types of voices (rough, breathy, and strain) and different degrees of deviation (mild, moderate, and severe). The voices were selected by three voice specialist speech-language pathologists (V-SLPs). The voices' selection was performed by convenience, in order for each gender to contain a rough, a breathy, and a strain voice, all with mild, moderate, and severe degree of deviation, totalizing 18 stimuli, nine male and nine female. The speech material was the Brazilian vowel /æ/, sustained for 1 second.

The set of synthesized voices was developed by a physics-based synthesizer (VoiceSim), produced in the Department of Computer Science at *Universidade de Brasília* (UnB), in support with *Laboratoires d 'Images, Signaux Dispositifs et des Télécommunications* (LIST) of the *Université Libre de Bruxelles* (ULB), in partnership with the researchers Prof. Jorge Lucero (UnB) and Prof. Jean Schoentgen (ULB). The synthesizer contains a vocal fold model and a representation of the vocal tract in the form of concatenated tubes through which an acoustic wave propagates.

Vocal deviations were produced using three parameters manipulation: for the roughness, the length of the glottal excitation cycle, jitter, was generated by introduction of a stochastic disturbance in the stiffness of the vocal fold tissue, in the form:

$$\Delta K = a\varepsilon K,$$

where a is a scale parameter, ε is a random variable, and K is a vocal fold stiffness coefficient; for the breathiness, additive noise was added, in the form:

$$\Delta u = b\varepsilon u,$$

where u is the glottal flow rate, b is a scale parameter, and ε is a random variable similar to jitter; in the strain voice, increasing of tension, K, and subglottal pressure and decreasing of vocal fold separation were performed. For further details of the synthesizer implementation, see Lucero et al (2013).[28]

The speech material of the synthesized stimuli was also the Brazilian vowel /æ/ sustained for 1 second; same that was used for the human stimuli.

The same three V-SLPs who selected the human voices also selected the synthesized voices in order that they were in accordance and paired with the type and degree of deviation previously selected for the human voices; the voices' selection had to respect the consensus of the three voice specialists. Subsequently, 18 synthesized voices were selected, nine female and nine male, with roughness, breathiness, and strain with mild, moderate, and severe degree of deviation.

Finally, there were a total of 36 stimuli, 18 human and 18 synthesized with different type and degree of deviation.

Listening session

The study included 70 subjects for the listening task, 20 V-SLPs forming the V-SLP group, 20 general SLPs (G-SLPs) with at least one from graduation, on the G-SLP, and 30 naive listeners (NLs) forming the NL group. The SLPs were recruited by request via e-mail sent by the researchers and NLs by indication. The average number of years in the profession for the V-SLP group was 5.75 years and 5.42 years for the G-SLP. All study participants signed an informed consent form. All participants reported normal hearing and no hearing complaints in the past.

The subject underwent a listening session of approximately 15 minutes in a quiet room, using loudspeakers. Several listening test groups were formed with an average of six participants each. A total of 54 stimuli were presented; the 18 human and 18 synthesized predefined stimuli and 18 (50% of random selection) repetition to verify intrarater consistency. The task was to classify these stimuli as human or as synthesized voices. Repetition was provided on request. This research considered only the responses of subjects with intrarater consistency above 72.2%; in other words, of the 18 repeated voices, at least 13 should have been equally classified.

Statistical analysis

Data were analyzed using the software: *SPSS V17 (SSPS Inc, Chicago, IL), Minitab 16 (Minitab, Inc, State College, PA), and Office Excel 2010 (Microsoft corporation, Redmond, Washington, USA)*. Significance level of 0.05 (5%) was considered, and all confidence intervals were stated at 95% statistical confidence. Statistical analysis used Analysis of Variance Test to compare groups and Multiple Range Test (Tukey's HSD) when necessary to detect differences between groups.

RESULTS

Human and synthesized samples produced a certain amount of errors identification.

The error average, regardless of the stimuli nature, considering all groups was 37.8%. The V-SLP group presented lower error percentage than the NL and the G-SLP groups, with statistically significant difference. NL and G-SLP presented statistically similar error percentage as summarized in Table 1.

The error average per group related to the voice nature, human or synthesized, showed that the V-SLP group had less error identification for the synthesized voices than the other groups, with strong statistical significance and that all groups together present equal error for the human voices. The error percentage for the voices' type, rough, breathy, or strain, and gender showed that the V-SLP group had less identification error for the breathy and for the female voices; all groups equally misclassified roughness and strain, and all groups presented greater error percentage for the male voices. These data are presented in Table 2.