嗨！Siri: 常見的語音辨識系統在識別異常嗓音的有效率

Hey Siri: How Effective are Common Voice Recognition Systems at Recognizing Dysphonic Voices

Matthew L. Rohlfing, MD ; Daniel P. Buckley, MS, CCC-SLP; Jacquelyn Piraquive, MD;Cara E. Stepp, PhD; Lauren F. Tracy, MD

臺中榮民總醫院 王仲祺醫師

Commentary

近年來嗓音異常的罹病率越來越高，2012 年美國健康訪談調查顯示每 13 名成年人中就有 1 人有嗓音異常。而嗓音異常會影響人際間的溝通，並被證明會對健康的功能、社交、情感和生理機能等面向產生負面影響。Benninger 等人也發現有嗓音異常者，相對患有坐骨神經痛、背痛和心絞痛等其他慢性疾病患者的社會功能差。原因可能來自嗓音異常患者的聲音強度降低、聲門閉合不佳，導致說高頻音或子音時氣流混亂不清，因此語音可被理解性降低所致。然而在過去十年，雲計算及機器學習等電腦科技不斷進步，電腦語音辨識系統功能也日漸強大，智能手機更成為大多數人日常生活最主要使用的電腦設備。再者，各種智能揚聲器和語音辨識助手如 Apple 的 Siri™、Google Assistant™ 和 Amazon 的 Alexa™ 等的出現也進一步擴展了各種語音使用體驗。根據 Google 的報告，全球約有 29% 的網路使用者會使用語音在各種移動電腦設備上進行線上搜尋。而本研究主旨在評估常見的語音辨識系統在識別異常嗓音時的正確性。其次，作者也想衡量哪些聽覺感知面向或聲學參數對語音識別系統準確性的影響較大。

作者錄音記錄 30 位嗓音異常患者和 23 位嗓音正常者朗誦英文短篇 “Rainbow Passage” 最前段的 98 字，然後用 65 到 70 分貝的音量以 Bose 揚聲器在距離 Apple iPhone 6S™、Apple iPhone11 Pro™ 和蘋果電腦上的 Google Voice™ 等電腦設備 12 英吋處播放，並計算語音辨識的正確性；正確性以 ” 單字識別率 ” 呈現，計算方法為正確轉錄朗誦單字的比例。這些朗誦錄音除了有聲學評估如嗓音基本頻率及其標準差等實驗室分析數據，也由兩位訓練中的喉科醫師判讀，並根據語音聽覺評估量表 (CAPEV) 對嗓音整體、嗓音粗糙度、嗓音氣息度、嗓音緊張度、嗓音音高、嗓音聲量等各面向予以評分。結果嗓音異常者和嗓音正常者的單字識別率在 Apple iPhone 6S™ 分別為 68.6% 和 91.9%(p<0.01, 有顯著差異); 在 Apple iPhone 11 Pro™ 分別為 71.2% 和 93.7% (P<.001); 在 Google Voice™ 分別為 68.7% 和 93.8%(P <.001); 而 ” 單字識別率 ” 和語音聽覺評估量表 CAPE-V5 中的嗓音整體嚴重程度之間存在很強的近似線性相關。相關係數 (R^2) 在 iPhone 6S™ 為 0.609、在 iPhone 11 Pro™ 為 0.670，在 Google Voice™ 為 0.619。他們的結論是，常見的語音辨識系統在處理沒有嗓音障礙時的語音效能良好，但有嗓音障礙的患者在使用時表現不佳。而語音聽覺評估的好壞和 ” 單字識別率 ” 有顯著相關。當我們的社會越來越頻繁使用語音辨識技術於日常生活時，嗓音異常患者在此方面的需求值得被進一步關注。

此篇文獻針對嗓音錄音後用於語音辨識的效能做了完整的分析，但因為 COVID-19 疫情，無法請聲量大小不同的受試者直接進行測試。所以結果並無法呈現聲量大小對 ” 單字識別率 ” 的影響。此外有構音問題的病人或非以英文為母語的人之錄音也被排除在研究樣本之外，因此這兩類情況的影響也無法以此文呈現。但如同肢體障礙患者的行動需要一些無障礙設施輔助；未來對於嗓音異常患者，社會也應該考慮提高語音辨識系統的效能，以協助排除嗓音患者在日常生活可能遭遇的障礙。

關鍵詞：嗓音異常，語音辨識，聲音沙啞，手機，科技

The Laryngoscope

Hey Siri: How Effective are Common Voice Recognition Systems at Recognizing Dysphonic Voices?

Matthew L. Rohlfing, MD; Daniel P. Buckley, MS, CCC-SLP; Jacquelyn Piraquive, MD; Cara E. Stepp, PhD; Lauren F. Tracy, MD

Objectives/Hypothesis: Interaction with voice recognition systems, such as Siri™ and Alexa™, is an increasingly important part of everyday life. Patients with voice disorders may have difficulty with this technology, leading to frustration and reduction in quality of life. This study evaluates the ability of common voice recognition systems to transcribe dysphonic voices.

Study Design: Retrospective evaluation of "Rainbow Passage" voice samples from patients with and without voice disorders.

Methods: Participants with (n = 30) and without (n = 23) voice disorders were recorded reading the "Rainbow Passage". Recordings were played at standardized intensity and distance-to-dictation programs on Apple iPhone 6S™, Apple iPhone 11 Pro™, and Google Voice™. Word recognition scores were calculated as the proportion of correctly transcribed words. Word recognition scores were compared to auditory–perceptual and acoustic measures.

Results: Mean word recognition scores for participants with and without voice disorders were, respectively, 68.6% and 91.9% for Apple iPhone 6S™ ($P < .001$), 71.2% and 93.7% for Apple iPhone 11 Pro™ ($P < .001$), and 68.7% and 93.8% for Google Voice™ ($P < .001$). There were strong, approximately linear associations between CAPE-V ratings of overall severity of dysphonia and word recognition score, with correlation coefficients (R^2) of 0.609 (iPhone 6S™), 0.670 (iPhone 11 Pro™), and 0.619 (Google Voice™). These relationships persisted when controlling for diagnosis, age, gender, fundamental frequency, and speech rate ($P < .001$ for all systems).

Conclusion: Common voice recognition systems function well with nondysphonic voices but are poor at accurately transcribing dysphonic voices. There was a strong negative correlation with word recognition scores and perceptual voice evaluation. As our society increasingly interfaces with automated voice recognition technology, the needs of patients with voice disorders should be considered.

Key Words: Dysphonia, voice recognition, hoarseness, mobile phone, technology.

Level of Evidence: 4

***Laryngoscope*, 131:1599–1607, 2021**

INTRODUCTION

There is increasing prevalence of voice disorders, with a 2012 National Health Interview Survey revealing that 1 in 13 adults reported voice problems.[1, 2] Voice disorders impair communication and contribute substantially to decreased quality of life and increased healthcare costs. Dysphonia has been shown to negatively impact many aspects of health, including functional, social, emotional, and physical well-being.[3–5] In addition to decreased quality of life, patients with voice disorders often report poor overall health. Benninger et al compared dysphonia to other chronic diseases and found that dysphonic individuals reported lower social functioning relative to those with sciatica, back pain, and angina.[6] Another study found that patients rated moderate dysphonia equivalent to monocular blindness using health state utility values.[7]

The decreased quality of life experienced by dysphonic speakers may be attributed to difficulty being understood, in addition to difficulty with voice production. Speakers with voice disorders have been repeatedly shown to be more difficult to understand relative to speakers without voice disorders.[8–14] Disordered voice production is associated with a number of deviations that may contribute to decreased intelligibility, including decreased vocal intensity,[8] and noisy turbulent airflow.[9] In the setting of incomplete glottic closure, speakers have more errors in stop consonant production, as well as increased turbulence during high-frequency sounds,[11] which can impact the ability of listeners to differentiate between sounds.[15]

From the Department of Otolaryngology–Head and Neck Surgery (M.L.R., D.P.B., J.P., L.F.T.), Boston Medical Center Boston University School of Medicine, Boston, Massachusetts, U.S.A.; Department of Speech (D.P.B., C.E.S.), Language, and Hearing Sciences, Boston University, Boston, Massachusetts, U.S.A.

Editor's Note: This Manuscript was accepted for publication on August 16, 2020.

This work was accepted for podium presentation at American Laryngological Association Annual Meeting, Atlanta, GA, April 22, 2020.

This work was supported in part by the National Institutes of Health through grants R01DC015570 (C.E.S.)

The authors have no other funding, financial relationships, or conflicts of interest to disclose.

Send correspondence to Lauren F. Tracy, MD, Boston Medical Center, Boston University School of Medicine, Department of Otolaryngology–Head and Neck Surgery, 830 Harrison Avenue, Boston, MA 02118. E-mail: lauren.tracy@bmc.org

DOI: 10.1002/lary.29082

Voice recognition technology enables the translation of voice input to text or to programed actions. This technology is intended to streamline communication by enhancing mobility, multitasking, and information gathering. Over the past decade, substantial advances in computational power via cloud computing and machine learning have enabled voice recognition technology to make unprecedented strides in accuracy and utility. Smartphones have become the primary everyday computational device for the vast majority of people. Furthermore, smart speakers and voice recognition assistants such as Apple's Siri™, Google Assistant™, and Amazon's Alexa™ have emerged to further integrate the experience of data acquisition and programmable tasks. According to Google, 29% of the global online community uses their voice to search on mobile devices, regardless of device type.[16] Voice recognition technology has many applications beyond accessing information, including driving with hands-free mobile access, routing customer service queries, and refilling pharmacy prescriptions. Voice recognition technology grants additional opportunities for individuals with physical impairments or poor literacy to have access to the internet and other assistive tasks.

Given that disordered voices result in impaired intelligibility among human listeners, and with the increased utilization of voice recognition technology, it is important to investigate the impact of dysphonia on voice recognition technology. We hypothesize that the accuracy of voice recognition technology will be reduced for individuals with voice disorders relative to individuals without voice disorders and that voice recognition technology accuracy will be decreased relative to the overall severity of dysphonia. Therefore, the objective of this study is to assess the accuracy of common voice recognition systems in transcribing dysphonic voices. Secondarily, we aim to evaluate which auditory–perceptual or acoustic measures impact the accuracy of voice recognition technology.

MATERIALS AND METHODS

Study Design, Setting, and Patient Selection

After receiving approval from the Institutional Review Board at Boston Medical Center, the Otolaryngology-Head and Neck Surgery department voice recording catalog was searched for English-speaking patients with dysphonic voices who had undergone comprehensive voice assessment and were recorded reading the "Rainbow Passage". Additional recordings of both dysphonic and nondysphonic voices were sourced from existing recordings of study participants at the Stepp Lab for Sensorimotor Rehabilitation Engineering at Boston University, all of whom completed informed consent in compliance with the Boston University Institutional Review Board. All available voice samples were considered for inclusion. Samples were excluded if they did not include a "Rainbow Passage" reading, if they did not complete the first 99 words of the "Rainbow Passage", if the speaker had impaired articulation, or if the speaker spoke English with a nonnative accent. One subject was excluded for mild ataxic dysarthria in the setting of Parkinson's disease. The voice recordings utilized in the study represent a variety of dysphonia severity, diagnoses, and patient demographic characteristics. All voice disorder diagnoses were made by fellowship-trained laryngologists.

All microphone signals were digitally recorded, and analysis occurred offline. Participants were recorded in one of two settings: the first was a quiet room at Boston Medical Center with an omnidirectional headset condenser microphone (model Beta 53; Shure, Niles, IL). These recordings took place between December 19, 2018, and September 4, 2019. The second setting was a quiet room at the Stepp Lab using a unidirectional headset condenser microphone (model SM35XLR; Shure). These recordings were conducted between July 5, 2016, and August 20, 2019. All participants were recorded with the microphone at 7–10 cm from the lips and at a 45-degree angle from midline. Microphone signals were sampled at 44.1 kHz with 16-bit resolution at both locations. In each setting, participants were instructed to read the "Rainbow Passage" in a comfortable speaking voice. As the recordings were accessed retrospectively, the speakers were not aware that their samples would be transcribed with voice recognition technology.

Demographic, Treatment, and Outcome Variables

Demographic variables and potential contributors to intelligibility were documented for each case. These included age, gender, diagnosis, fundamental frequency (f_o), standard deviation of f_o, and cepstral peak prominence, smoothed (CPPS). The CPPS has been demonstrated as a marker of dysphonia, with lower CPPS values corresponding to more severe dysphonia.[17] The f_o, standard deviation of f_o, and CPPS were calculated with Praat software.[18] Speech rate, in words per minute, inclusive of pauses in speech, was also calculated.

Each recording of the "Rainbow Passage" was played at standardized intensity for each voice recognition software system. The recordings were played at an intensity between 65 and 70 decibels through a Bose SoundLink Wireless speaker (Bose Corporation, Framingham, MA) to a device 12 inches from the source. The passages were played and processed one sentence or phrase at a time. After completion of the first paragraph (98 words) of the "Rainbow Passage", the number of correct words was tallied and divided by the total to calculate the word recognition score. This process was repeated with Apple iPhone 6S speech dictation (Apple, Cupertino, CA), Apple iPhone 11 Pro speech dictation (Apple), and Google Voice online dictation software (Google, Mountain View, CA) using an Apple MacBook Air laptop computer (Apple). No software updates occurred during the study, and all devices were cleared of data prior to study initiation. To control for the potential of machine learning, the accuracy of word recognition was compared throughout time for all three studied devices. There was no difference in scores throughout the duration of the study period for any device. A fourth system, the Microsoft Windows 10 talk-to-type dictation (Microsoft, Redmond WA), was also attempted. The system was unable to recognize and transcribe nondysphonic "Rainbow Passage" recordings, so further testing was not pursued.

Two trained listeners (authors M.L.R. and D.P.B.) conducted independent perceptual evaluation of voice for each voice sample using the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V).[19] The CAPE-V provides a standardized and validated approach to the evaluation of vocal quality, including six specific vocal attributes: overall severity, roughness, breathiness, strain, pitch, and loudness.[19] For the purpose of this study, loudness was not included as volume was standardized. The listeners were blinded to the diagnosis and to the word recognition scores. Per described technique, a visual analog scale was used to rate the overall severity, roughness, strain, pitch, and loudness.[19] The scale was converted to a numeric value from 0 to 100 by measuring the relative position

TABLE I.
Word Recognition Scores (WRS) using iPhone and Google Voice Dictation Systems to Transcribe the "Rainbow Passage", According to Categorical Patient Factors. Q1–4 = Quartiles.

		Apple iPhone 6s, WRS (%)		Google Voice, WRS (%)		Apple iPhone 11 Pro, WRS (%)	
Patient Factor	N (%)	Mean (SD) 80.3 (19.6)	*P* Value	Mean (SD) 81.3 (19.8)	*P* Value	Mean (SD) 82.2% (19.1)	*P* Value
Gender							
Male	13 (25.0)	79.7 (22.8)	.904	83.8 (17.6)	.578	81.5 (22.0)	.892
Female	39 (75.0)	80.5 (18.7)		80.5 (20.6)		82.4 (18.4)	
Age							
Q1 (18–32)	13 (25.0)	84.8 (4.4)	.689	86.7 (5.9)	.570	86.5 (5.8)	.658
Q2 (33–50)	13 (25.0)	81.1 (21.0)		80.5 (25.0)		82.8 (21.4)	
Q3 (54–67)	12 (23.1)	80.0 (18.7)		82.8 (17.5)		82.8 (17.9)	
Q4 (68+)	14 (26.9)	75.6 (27.1)		75.9 (24.6)		77.1 (25.7)	

SD = standard deviation.

of the mark (i.e., a mark directly in the center of the line would be 50/100). The voice samples were reviewed as ordered by their study identifiers, which allowed a consistent mix of dysphonic and nondysphonic speakers.

Statistical Analysis

Statistical analyses were conducted using SPSS Version 23 (IBM, Armonk, NY). Descriptive statistics are shown in Table I (categorical variables) and Tables II–IV (continuous variables). For the CAPE-V evaluations, inter-rater reliability (Cohen's kappa for two raters) was calculated for ratings of the overall severity of dysphonia and for the ratings of roughness, breathiness, strain, and pitch. Given the high reliability with correlation coefficient 0.913 for CAPE-V (overall), the ratings were combined. The final combined rating is the average of the individual ratings. As recordings were made in two settings under differing conditions, a pairwise analysis was conducted to assess for difference in voice recognition score outcomes. Ten cases from each site were matched based on overall CAPE-V ratings. There was high correlation (correlation coefficient 0.879, 95% confidence interval [CI] 0.686–0.996, $P = .001$) and no significant difference between means ($P = .227$) in paired comparisons of iPhone 6s word recognition score. This was true for the Google system (correlation coefficient 0.965, 95% CI 0.883–0.996, $P < .001$; no difference between means, $P = .294$). This was also true for the iPhone 11 Pro (correlation coefficient 0.934, 95% CI 0.361–0.990, $P < .001$; no difference between means, $P = .145$). Given this, the recordings were pooled for all further analyses.

Comparative analysis was performed between binary patient groupings (i.e., voice disorder or no voice disorder) and also by considering degree of perceptual dysphonia (as represented by CAPE-V ratings of overall severity) as a continuous variable. The analysis treating dysphonia as a continuous variable is emphasized as the authors believe it provides a more comprehensive assessment across varying degrees of dysphonia. The comparative analysis was performed to assess the association between potential covariates and the outcome of interest, word recognition score. This was repeated for each voice recognition system. One-way analysis of variance tests were used to compare word recognition score means for categorical variables, and linear regression was used to assess correlation between continuous variables and word recognition score. From these analyses, variables with *P*-values of .05 or lower were considered potential contributors and selected to be included in multivariate analysis.

Multivariate stepwise linear regression was used to assess the relationships between the selected variables and the word recognition score. Covariates selected from the univariate analyses were considered in the stepwise regression model. This was repeated for each voice recognition system. The final model is

TABLE II.
Word Recognition Scores (WRS) using iPhone 6s Dictation System to Transcribe the "Rainbow Passage", with Associations to Continuous Patient Factors Calculated via Simple Linear Regression Analysis.

Patient Factor	Coefficient (β)	95% CI for β, Low	95% CI for β, High	R^2	*P* Value
Age	−0.203	−0.476	0.070	0.043	.142
CAPE-V (overall)	−0.648	−0.794	−0.501	0.612	**<.001**
CAPE-V (roughness)	−0.562	−0.810	−0.314	0.293	**<.001**
CAPE-V (breathiness)	−0.752	−0.960	−0.545	0.515	**<.001**
CAPE-V (strain)	−0.471	−0.709	−0.232	0.239	**<.001**
CAPE-V (pitch)	−0.433	−0.813	−0.054	0.095	**.026**
f_o	−0.148	−0.289	0.007	0.081	.054
Standard deviation of f_o	−0.344	−0.93	−0.095	0.133	**.008**
CPPS	6.46	4.59	8.33	0.491	**<.001**
Speech rate	0.252	0.090	0.415	0.162	**.003**

Bold values signifies $p < 0.05$.
CAPE-V = Consensus Auditory-Perceptual Evaluation of Voice; CI = confidence interval; CPPS = cepstral peak prominence, smoothed; f_o = fundamental frequency.

TABLE III.
Word Recognition Scores (WRS) using Google Voice Dictation System to Transcribe the "Rainbow Passage", with Associations to Continuous Patient Factors Calculated via Simple Linear Regression Analysis.

Patient Factor	Coefficient (β)	95% CI for β, Low	95% CI for β, High	R^2	*P* Value
Age	−0.214	−0.490	0.062	0.046	.126
CAPE-V (overall)	−0.665	−0.810	−0.520	0.630	**<.001**
CAPE-V (roughness)	−0.549	−0.804	−0.295	0.273	**<.001**
CAPE-V (breathiness)	−0.643	−0.883	−0.404	0.368	**<.001**
CAPE-V (strain)	−0.521	−0.754	−0.287	0.286	**<.001**
CAPE-V (pitch)	−0.167	−0.568	0.234	0.014	.407
f_o	−0.165	−0.307	−0.024	0.100	**.023**
Standard deviation of f_o	−0.337	−0.626	−0.128	0.156	**.004**
CPPS	6.23	4.26	8.20	0.446	**<.001**
Speech rate	0.227	0.059	0.395	0.128	**.009**

Bold values signifies $p < 0.05$.
CAPE-V = Consensus Auditory-Perceptual Evaluation of Voice; CI = confidence interval; CPPS = cepstral peak prominence, smoothed; f_o = fundamental frequency.

described in Table V. Findings were considered statistically significant at $P < .05$.

RESULTS

There were 52 voice recordings of "Rainbow Passage" analyzed. Of these participants, 29 (55.8%) had a diagnosis of voice disorder, and 23 (44.2%) were without voice disorder. All 15 patients recorded in the Boston Medical Center setting had a diagnosis of voice disorder. Of the patients recorded at Boston University, 14 (37.8%) had a voice disorder. For the 23 participants without voice disorder, the mean age was 52.2 years (standard deviation 20.4 years, range 19–83 years). For the 29 participants with a voice disorder, the mean age was 48.9 years (standard deviation 19.8 years, range 19–82 years). There was no significant difference in mean age ($P = .561$). The cohort without voice disorders was 21.7% male and 78.3% female; the cohort with voice disorders was 20.7% male and 79.3% female. There was no significant difference in gender distribution by Fisher's exact test ($P = .595$). The most common voice disorder diagnoses were muscle tension dysphonia (N = 11, 21.2%), benign vocal fold lesion (N = 5, 9.6%), and chronic laryngitis (N = 4, 7.7%). Other diagnoses included paresis or paralysis (N = 3, 5.8%), spasmodic dysphonia (N = 3, 5.8%), vocal fold atrophy (N = 2, 3.8%), and laryngeal trauma (N = 1, 1.9%). Mean word recognition scores for participants with and without voice disorders were, respectively, 71.1% and 91.9% for Apple iPhone 6S™ ($P < .001$), 71.5% and 93.7% for Apple iPhone 11 Pro™ ($P < .001$), and 73.0% and 93.8% Google Voice™ ($P < .001$).

For the iPhone 6s, there was significant negative correlation between all attributes of CAPE-V and word recognition scores, with the strongest association for the CAPE-V rating of overall severity of dysphonia ($R^2 = 0.612$, $P < .001$) (Table II, Figure 1). There was also a significant negative correlation for standard deviation of f_o ($R^2 = 0.133$, $P = .008$) and significant positive correlation for CPPS ($R^2 = 0.491$,

TABLE IV.
Word Recognition Scores (WRS) using iPhone 11 Pro Dictation System to Transcribe the "Rainbow Passage", with Associations to Continuous Patient Factors Calculated via Simple Linear Regression Analysis.

Patient Factor	Coefficient (β)	95% CI for β, Low	95% CI for β, High	R^2	*P* Value
Age	−0.183	−0.451	0.085	0.036	.176
CAPE-V (overall)	−0.664	−0.795	−0.553	0.674	**<.001**
CAPE-V (roughness)	−0.559	−0.800	−0.319	0.304	**<.001**
CAPE-V (breathiness)	−0.723	−0.919	−0.526	0.498	**<.001**
CAPE-V (strain)	−0.506	−0.731	−0.281	0.290	**<.001**
CAPE-V (pitch)	−0.437	−0.807	−0.068	0.101	**.021**
f_o	−0.143	−0.281	−0.006	0.080	**.042**
Standard deviation of f_o	−0.300	−0.547	−0.053	0.106	**.018**
CPPS	6.22	4.36	8.07	0.476	**<.001**
Speech rate	0.272	0.116	0.428	0.198	**.001**

Bold values signifies $p < 0.05$.
CAPE-V = Consensus Auditory-Perceptual Evaluation of Voice; CI = confidence interval; CPPS = cepstral peak prominence, smoothed; f_o = fundamental frequency.

TABLE V.
Stepwise Regression Analysis to Predict the Word Recognition Scores (WRS) using the Apple iPhone and Google Voice Systems to Transcribe the "Rainbow Passage", According to Patient Factors Identified in Univariate Analysis.

Patient Factor	Coefficient (β)	95% CI for β, Low	95% CI for β, High	*P* Value	R^2
	Apple iPhone 6s, WRS (%)				
CAPE-V (overall)	−0.648	−0.794	−0.501	**<.001**	0.612
Constant	98.7	93.3	104.1	—	
	Google Voice, WRS (%)				
CAPE-V (overall)	−0.665	−0.810	−0.520	**<.001**	0.630
Constant	100.2	94.9	105.5	—	
	Apple iPhone 11 Pro, WRS (%)				
CAPE-V (overall)	−0.664	−0.795	−0.533	**<.001**	0.674
Constant	101.0	96.2	105.8	—	

Bold values signifies $p < 0.05$.
CAPE-V = Consensus Auditory-Perceptual Evaluation of Voice; CI = confidence interval.

$P < .001$) and speech rate ($R^2 = 0.162$, $P = .003$). In stepwise regression analysis, only the CAPE-V overall severity demonstrated significant correlation ($P < .001$). All other potential covariates were excluded from the stepwise model as they did not generate a significant change in correlation.

For the Google Voice system, there was significant negative correlation between all CAPE-V scores except pitch and word recognition scores, with strongest association for the overall CAPE-V ($R^2 = 0.630$, $P < .001$) (Table II, Figure 2). There was also significant negative

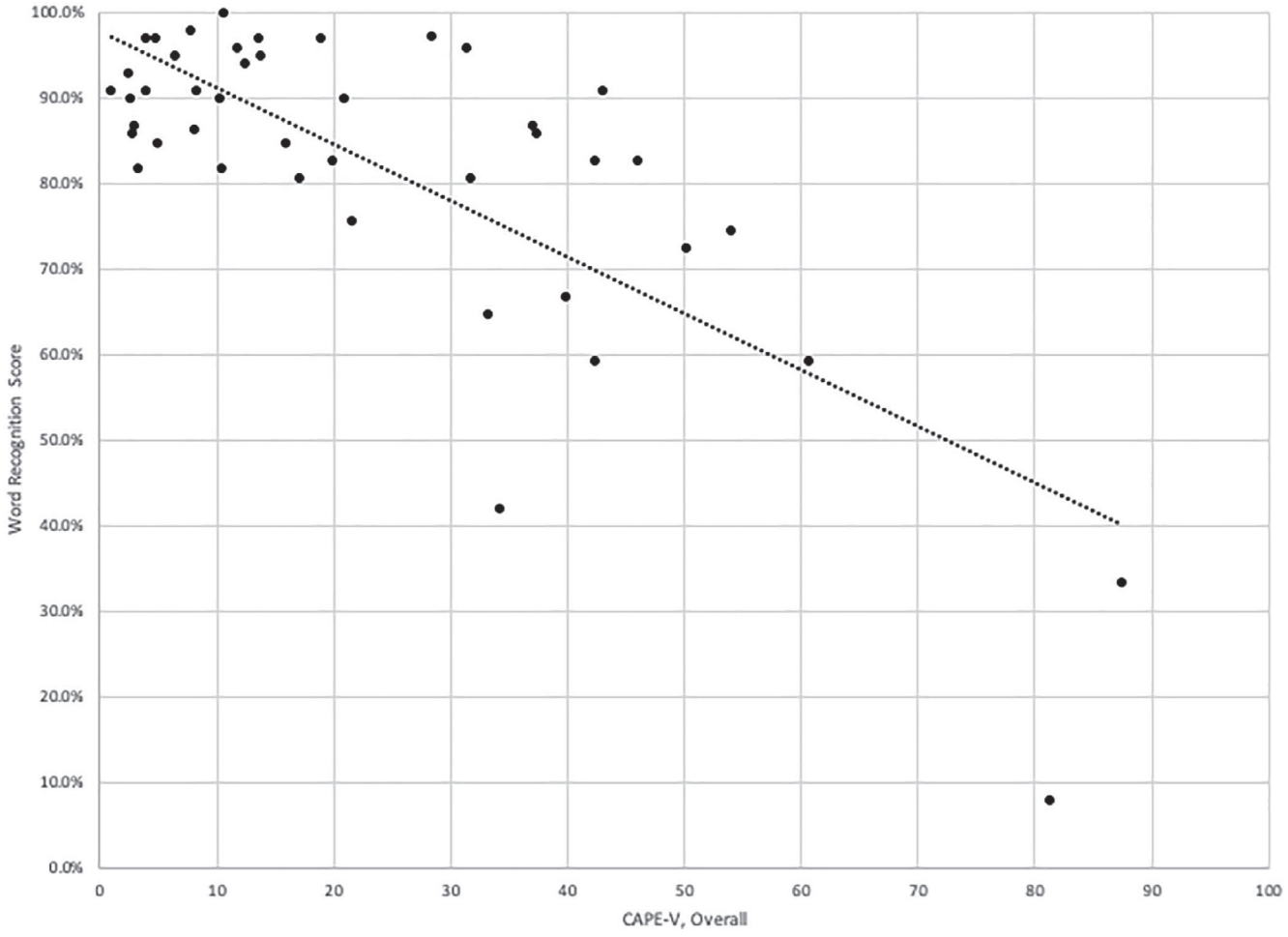

Fig. 1. Correlation of word recognition score and CAPE-V ratings of overall severity of dysphonia for Apple iPhone 6s voice recognition software.

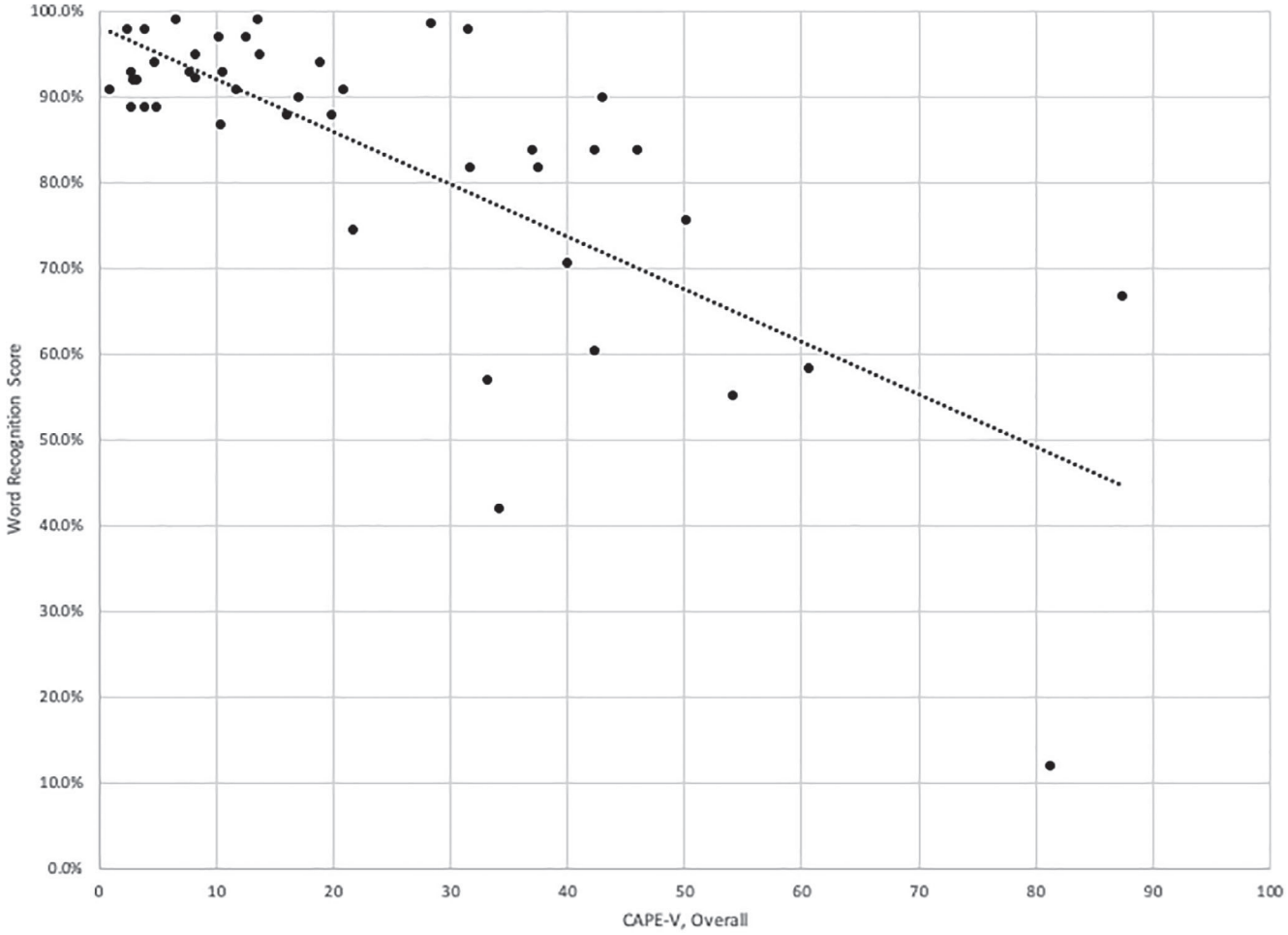

Fig. 2. Correlation of word recognition score and CAPE-V ratings of overall severity of dysphonia rating for Google Voice recognition software.

correlation for f_o ($R^2 = 0.100$, $P = .023$), standard deviation of f_o ($R^2 = 0.156$, $P = .004$), and significant positive correlation for CPPS ($R^2 = 0.446$, $P < .001$) and speech rate ($R^2 = 0.128$, $P = .009$). In stepwise regression analysis, CAPE-V overall severity ($P < .001$) was the only factor that demonstrated significant correlation. All other potential covariates were excluded from the stepwise model as they did not generate a significant change in correlation.

For the iPhone 11 Pro, there was a significant negative correlation between CAPE-V scores and word recognition scores, with the strongest association for the CAPE-V overall severity of dysphonia ($R^2 = 0.674$, $P < .001$) (Figure 3). There was also a significant negative correlation for f_o ($R^2 = 0.080$, $P = .042$) and standard deviation of f_o ($R^2 = 0.106$, $P = .018$) and significant positive correlation for CPPS ($R^2 = 0.476$, $P < .001$) and speech rate ($R^2 = 0.198$, $P = .001$). In stepwise regression analysis, the only factor that retained significant correlation was CAPE-V overall rating of severity ($P = .001$). All other potential covariates were excluded from the stepwise model as they did not generate a significant change in correlation.

DISCUSSION

The purpose of this study was to examine the impact of dysphonia on voice recognition technology accuracy as measured by word recognition scores. The results indicated that word recognition scores were significantly lower for speakers with voice disorders in comparison to those without voice disorders, and this finding was consistent across all three studied systems.

A secondary interest was investigating which measures of dysphonia had the highest correlation with voice recognition technology intelligibility. Several factors were considered potential contributors to word recognition scores. Participants' f_o demonstrated a weak but significant association with word recognition score in univariate analysis, with lower f_o associated with higher word recognition scores. This suggests that the voice recognition technology may be better equipped to understand voices with lower f_o. This finding is similar to previous work demonstrating improved listener accuracy in speech recognition for lower f_o.[20] This may be due to the more narrow spacing of harmonics in a lower-frequency domain, which allows for better resolution. This then allows for improved resolution in resolving vocal tract resonances at lower

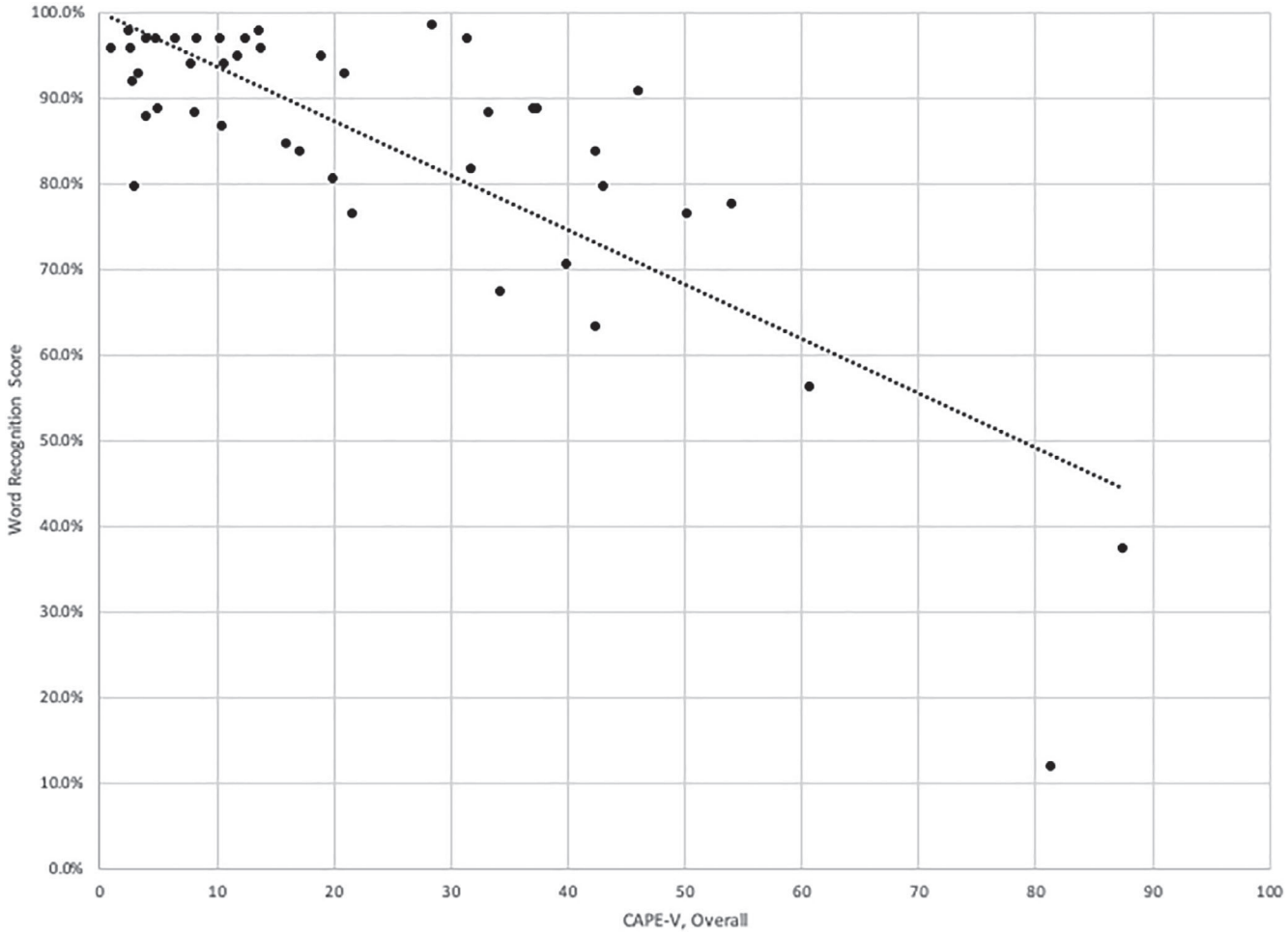

Fig. 3. Correlation of word recognition score and CAPE-V ratings of overall severity of dysphonia for Apple iPhone 11 Pro voice recognition software.

frequency. In contrast, higher f_o values have decreased resolution, and formants become poorer estimates of underlying resonances. The lack of being able to track formants in higher f_o, which relate to vocal tract resonances could be related to decreased word recognition.

A higher standard deviation of f_o was also associated with lower word recognition scores. A higher f_o in this studied cohort correlated with higher standard deviation of f_o, which has been demonstrated previously.[21] Because of the close relationship between f_o and standard deviation of f_o, it is difficult to conclude which variable had the more significant effect on the outcome. The authors speculate that perhaps a voice with more frequency variation could be more difficult for the technology to follow and recognize. The correlation coefficients (R^2) were quite low for these variables; for f_o, R^2 was 0.100 (Google Voice) and 0.080 (iPhone 11), and iPhone 6s demonstrated nonsignificant correlation. For standard deviation of f_o, R^2 was 0.133 (iPhone 6s), 0.156 (Google Voice), and 0.106 (iPhone 11). So although there were significant correlations, f_o and standard deviation of f_o have limited impact on the variability in the outcome.

Speech rate was considered a potential confounder, and it was hypothesized that a slower speech rate would result in improved voice recognition technology accuracy. However, faster speech rate demonstrated a positive association with word recognition scores. The authors' hypothesize that this discrepancy occurred because participants with more severe dysphonia spoke at a slower rate. The mean speech rate between individuals with a voice disorder was 152.6 words per minute (95% CI 141.1–163.9) in comparison to 175.6 words per minute for individuals without a voice disorder (95% CI 164.7–186.0; P = .006). Thus, speech rate in this study is likely another proxy for dysphonia.

CPPS showed a significant correlation with the voice recognition technology performance in univariate analysis. As CPPS is a correlate of the overall severity of dysphonia,[22] the authors hypothesized that lower CPPS values (which designate more severe dysphonia) would correlate with decreased word recognition. In this study, CPPS correlated significantly with voice recognition performance with coefficients of 0.491 (iPhone 6s), 0.446 (Google Voice), and 0.476 (iPhone 11), meaning that

changes in CPPS can explain a moderate amount of the change in outcome. In the stepwise regression analysis, they did not generate significant improvement of the model and therefore were not included in the final model. This is likely a representation of the covariance between CPPS and CAPE-V. In linear regression, these two variables have a moderate, negative correlation ($R^2 = 0.605$, $P < .001$). Therefore, although CPPS did not change the regression model, it is a useful marker of dysphonia that did correlate with the performance of the technologies.

In multivariate analysis, the CAPE-V rating of the overall severity of dysphonia was the most consistent predictor of word recognition scores across all three systems. Therefore, increases in degree of perceived dysphonia – even when controlling for other factors – was associated with a significant decrease in accuracy of the voice recognition technology. Voice recognition technology accuracy was not impacted by any clinically perceived or patient-reported articulation disorder in the studied cohort; therefore, it is possible that compensatory laryngeal actions and posturing in the setting of dysphonia decreased the clarity of the speech output. As dysphonic voices have been shown to have decreased intensity, this potential confounder was controlled by equalizing vocal intensity for each recording.[17] Therefore, individuals with voice disorders associated with lower vocal intensity may have worse intelligibility than demonstrated in this study. The recordings were obtained in quiet rooms without background voice, given that background noise has been shown to limit listener intelligibility of dysphonic voices.[8] However, voice recognition technology is often used in situations with surrounding noise, and functionality may be additionally limited in this setting.

The accuracy of voice recognition systems may be improving with advances in technology. Improved software, independent of learning, may contribute to improved accuracy as demonstrated by the improved word recognition scores for iPhone 11 Pro in comparison to iPhone 6s. However, the absolute difference for dysphonic voices was small, with 69.1% for Apple iPhone 6S™ versus 71.1% for Apple iPhone 11 Pro™, and there remains substantial opportunity for improvement. Furthermore, similar to listeners using cognitive–perceptual techniques to improve comprehension,[23] voice recognition technology may have the capacity to "learn" accuracy through increased experience with voice. It is possible that voice recognition technology can improve word recognition score with increased familiarity with a speaker's vocal profile; additional investigation is warranted to evaluate this hypothesis. The adaptability of voice recognition technology to "learn" how to transcribe dysphonic voices through repetition was not investigated, and this would benefit from additional study.

To our knowledge, this is the first study to evaluate the ability of common voice recognition technology systems to transcribe English-speaking voices with variable degrees of dysphonia. A study of automatic speech recognition with Arabic digits on dysphonic patients similarly showed decreased speech recognition ranging from 56% to 84%, compared to 100% with normal voices.[24] In this study, there was variable recognition based on pathology, with sulcus vocalis having the least accuracy in voice recognition and the highest accuracy with vocal cord nodules.[8] Other literature concerning voice recognition and dysphonia has focused on the identification of vocal pathology using automatic speech recognition technology or to evaluate intelligibility of speech in patients with head and neck cancer to guide rehabilitation after treatment.[22, 25]

Of note, there is limited publicly available information about the design of these proprietary technologies. This study evaluates the ability to transcribe recordings of dysphonic and nondysphonic voices and is not a comparison of overall quality of these technologies. This area represents an opportunity for additional investigation.

Limitations to this study include the heterogenous diagnoses contributing to the voice disorder cohort. Further investigation is warranted for subgroup analysis of how these diagnoses separately impact efficacy of voice recognition technology. The inclusion of a broader range of voice recognition technology systems would also improve generalizability. In addition, the recordings were conducted at two different clinical sites, with different clinicians or researchers and slightly different protocols and equipment. A paired comparison is included that demonstrates no significant difference between sites as it relates to word recognition, but this should be considered a methodological limitation. Furthermore, the methodology would be strengthened by use of live speakers in lieu of or in addition to recorded voices; however, this was not possible due to restrictions associated with the COVID-19 pandemic. The authors are planning to perform follow-up investigations with live speakers when restrictions are lifted to better understand and optimize the interactions of dysphonic voices with these technologies.

As voice recognition technology becomes increasingly embedded in daily life, dysphonic patients remain potentially excluded from these innovations and interactions. Early models of voice recognition were based on recordings with ideal acoustic qualities and limited diversity. Future advancements in voice recognition technology should be directed toward adding more diverse voice models to enable recognition of non-normative voices and speech patterns.

CONCLUSIONS

Commonly used voice recognition technology systems are deficient in their ability to transcribe moderately and severely dysphonic voices. The results from auditory–perceptual ratings show that increasing severity of dysphonia correlates with worse function of voice recognition technology. Daily interaction with voice recognition technology is increasing; therefore, additional investigation with more diverse voice models is warranted to improve applicability and accessibility of this technology.

BIBLIOGRAPHY

1. Benninger MS, Holy CE, Bryson PC, Milstein CF. Prevalence and occupation of patients presenting with dysphonia in the United States. *J Voice* 2017;31:594–600.

2. Bhattacharyya N. The prevalence of voice problems among adults in the U.S. *Laryngoscope* 2014;124:2359–2362.
3. Spina AL, Crespo AN. Assessment of grade of dysphonia and correlation with quality of life protocol. *J Voice* 2017;31:243.e21–243.e26.
4. Cohen SM, Dupont WD, Courey MS. Quality-of-life impact of non-neoplastic voice disorders: a meta-analysis. *Ann Otol Rhinol Laryngol* 2006;115: 128–134.
5. Wilson JA, Deary IJ, Millar A, Mackenzie K. The quality of life impact of dysphonia. *Clin Otolaryngol Allied Sci* 2002;27:179–182.
6. Benninger MS, Ahuja AS, Gardner G, Grywalski C. Assessing outcomes for dysphonic patients. *J Voice* 1998;12:540–550.
7. Naunheim MR, Goldberg L, Dai JB, Rubinstein BJ, Courey MS. Measuring the impact of dysphonia on quality of life using health state preferences. *Laryngoscope* 2019;00:1–6.
8. Ishikawa K, Boyce S, Kelchner L, et al. The effect of background noise on intelligibility of dysphonic speech. *J Speech Lang Hear Res* 2017;60: 1919–1929.
9. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality dysphonic voices and continuous speech. *J Speech Lang Hear Res* 1996;39: 311–321.
10. Hillenbrand J, Cleveland RA, Erickson RL. Acoustic correlates of breathy vocal quality. *J Speech Lang Hear Res* 1994;37:769–778.
11. Isshiki N, Kitajima K, Kojima H, Harita Y. Turbulent noise in dysphonia. *Folia Phoniatr Logop* 1978;30:214–224.
12. Casper JK, Leonard R. *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. Baltimore, MD: Lippincott, Williams & Wilkins; 2006.
13. Evitts PM, Starmer H, Teets K, et al. The impact of dysphonic voices on healthy listeners: listener reaction times, speech intelligibility, and listener comprehension. *Am J Speech Lang Pathol* 2016;25:561–575.
14. Bender B, Cannito MP, Murry T, et al. Speech intelligibility in severe adductor spasmodic dysphonia. *J Speech Lang Hear Res* 2004;47:21–33.
15. Hartl DM, Crevier-Buchman L, Vaissiere J, Brasnu D. Phonetic effects of paralytic dysphonia. *Ann Otol Rhinol Laryngol* 2005;114:792–798.
16. Voice Search Mobile Use Statistics. Think with Google. https://www.thinkwithgoogle.com/data/voice-search-mobile-use-statistics/. Accessed March 1, 2020.
17. Watts CR, Awan SN, Maryn Y. A comparison of cepstral peak prominence measures from two acoustic analysis programs. *J Voice* 2017;31:387.
18. Boersma P. Praat, a system for doing phonetics by computer. *Glot Int* 2001; 5:341–345.
19. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech Lang Pathol* 2009; 18:124–132.
20. Goldwater S, Jurafsky D, Manning CD. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Commun* 2010;52:181–200.
21. Brown CA, Helms Tillery K, Apoux F, Doyle NM, Bacon SP. Shifting fundamental frequency in simulated electric-acoustic listening: effects of F0 variation. *Ear Hear* 2016;37:e18–e25.
22. Sauder C, Bretl M, Eadie T. Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and analysis of dysphonia in speech and voice (ADSV). *J Voice* 2017;31:557–566. https://doi.org/10.1016/j.jvoice.2017.01.006.
23. Lindblom B. On the communication process: speaker-listener interaction and the development of speech. *Augment Altern Commun* 1990;6:220–230.
24. Muhammad G, Mesallam TA, Malki KH, Farahat M, Alsulaiman M, Bukhari M. Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. *Biomed Eng Online* 2011;10:41.
25. Maier A, Haderlein T, Stelzle F, et al. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP J Audio Speech Music Process* 2009;2010: 926951.