USE OF SPEECH RECOGNITION AND VOICE FATIGUE: MEASURES OF F0 AND SPECTRAL SLOPE

C. de Bruijn[1] and S. Whiteside[2]

1. University of Central England, Birmingham, UK; 2. University of Sheffield, UK
1. christel.debruijn@uce.ac.uk 2. s.whiteside@sheffield.ac.uk

ABSTRACT

This study investigates the effect of a speech recognition task on acoustic measures of voice quality. Type of speech recognition (discrete and continuous) and vocal load of a speaker receive particular attention. A rise in F0, a common finding in voice fatigue studies, appears as the most consistent finding. It is interpreted as part of a hyperfunctional mechanism countering early signs of voice fatigue.

Keywords: voice fatigue, speech recognition, F0, spectral slope, hyperfunction.

1. INTRODUCTION

After decades of research, automatic speech recognition (ASR) came finally within reach of the general public in the early 1990's. The first large-vocabulary discrete speech-to-text systems for general purpose dictation became off-the-shelf products. However, soon after the introduction of speech recognition software to the general consumer market, reports started appearing of people claiming to have developed voice problems as a result of using such software. These reports came from various sources, such as computer magazines, mailing lists for users of speech recognition software, independent web pages, the Bulletin of the Royal College for Speech and Language Therapists (UK) and personal communication with speech therapists as well as ASR users. The problems and symptoms reported varied in severity from a dry throat to being unable to speak for more than half an hour a day.

So far, only two studies have been published in an attempt to examine the influence of speech recognition software on voice [1, 2]. Although the studies reported results from perceptual, aerodynamic and videostroboscopic examinations, relatively little attention was paid to acoustic measures. The aim of the current study is to measure acoustically if and how voice quality is affected by the use of ASR.

Two main variables have been taken into consideration in this study: type of ASR software and vocal load. It could be hypothesised that the use of discrete recognition software may have a stronger effect on voice quality than continuous software, due to the word-by-word speaking style it requires. This requires vocal fold ab- and adduction to a far greater extent than when dictating in entire phrases. This therefore could have a fatiguing effect on the muscles involved. The effect of vocal load is investigated because by using ASR, the vocal load of the speaker may increase substantially. It is well known that people in professions that are heavily reliant on voice are prone to developing voice problems [3].

2. METHODOLOGY

A group of 25 subjects carried out a 2 hour ASR task. All speakers were native British English speakers, had no strong regional accents, were non-smoking, reported no voice problems, considered themselves computer literate and had not used ASR software for any prolonged period of time. Subjects with high and low vocal loads were recruited and this variable was crossed over with type of recognition (discrete or continuous). This resulted in 4 subject groups: one group of low load speakers carrying out discrete recognition (N=8), another group of low load speakers carrying out continuous recognition (N=7), a high load – continuous recognition group (N=6) and a high load – discrete group (N=4). Groups could not be matched for sex or age, but this issue was addressed in the statistical analysis of the results.

Single tokens of 3 sustained vowels (/ɑ/,/i/,/u/) were recorded and measured before and after the dictation task. Acoustic analyses were carried out using Praat software [4] and consisted of F0 related measures (average F0, F0 min, F0 max), measures of spectral slope (Lh1-Lh2, Lh1- LF1, Lh1-LF3, LF1-LF3, CoG)[i] and power in spectral regions (0-1, 1-2, 2-3.5, 3.5-5, 5-6.5, 6.5-8, 8-10, 10-12.5, 12.5-15 kHz). Based on findings by De Krom [5]

the entire vowel was analysed rather than steady state only.

For the statistical analysis, for each acoustic parameter, univariate ANOVAs were carried out on the differences in measures before and after dictation to investigate main effects and interactions between the factors vocal load and recognition type. Because the speaker groups were not matched for age and sex, sex was also entered as a between-subjects factor, and age as a covariate. ANOVA results were followed up with parametric independent-samples tests, or with non-parametric exact tests if cell sizes were smaller than five or if variance assumptions were not met (checked for by Levene's test). Normal distribution of the data was checked with the Kolmogorov-Smirnov test. No significant deviations from normality were found.

In order to optimize the statistical analysis, and filter out spurious findings potentially resulting from the small sample sizes, the following approach was taken. First, ANOVAs were carried out collapsing across factor vocal load, but including the factors recognition type and sex. Then, ANOVAs were carried out this time collapsing across recognition type. Finally, ANOVAs were carried out with all between-subject factors. The main effects and interactions from the latter analysis are reported here, but only if they were also significant in the earlier ANOVAs. Further details about the methodology can be found in [6].

3. RESULTS

ANOVA results for measures of F0 are shown in table 1. Only significant results ($p<0.05$) or those approaching significance ($p<0.10$) are reported. For parameter F0, interactions of vocal load by recognition type were found for all three vowels.

Table 1: Between-subject effects for measures of F0.

para meter	stim	source	df	F	sig	obs. power
F0	/ɑ/	load * recog	1	16.565	.002	.961
	/i/	load * recog	1	3.926	.071	.445
	/u/	load * recog	1	12.652	.003	.911
F0 min	/ɑ/	load * recog	1	6.356	.027	.639
	/i/	load * recog	1	5.328	.040	.564
	/u/	load * recog	1	10.152	.007	.842
F0 max	/u/	recog	1	4.321	.057	.490

Table 2: Between-subject effects for measures of spectral slope.

para meter	stim	source	df	F	sig	obs. power
Lh1-Lh2	/i/	sex	1	3.655	.080	.420
	/u/	sex	1	8.785	.010	.787
Lh1-LF1	/ɑ/	intercept	1	3.986	.071	.445
	/i/	intercept	1	6.031	.030	.617
		age	1	8.772	.012	.776
		load	1	4.987	.045	.537
	/u/	sex	1	6.384	.024	.652
Lh1-LF2	/ɑ/	age	1	3.821	.077	.430
		load	1	9.870	.009	.816
LF1-LF3	/ɑ/	intercept	1	5.855	.034	.597
		age	1	5.981	.032	.606
		load	1	10.721	.007	.846
	/i/	load * sex	1	3.557	.086	.406
cog	/ɑ/	sex	1	18.691	.001	.975
		load * recog	1	5.643	.037	.582
	/u/	recog	1	18.814	.001	.981
		load * recog	1	21.313	.000	.990

Table 3: Between-subject effects for measures of spectral power.

para meter	stim	source	df	F	sig	obs. power
p 0-1	/ɑ/	load * recog	1	4.247	.064	.468
p 1-2	/ɑ/	sex	1	8.165	.016	.740
p 2-3.5		age	1	8.884	.013	.774
		load	1	10.544	.008	.840
	/u/	load * recog	1	6.164	.026	.637
p 3.5-5	/u/	load * recog	1	3.919	.068	.454
p 5-6.5	/i/	sex	1	3.250	.097	.382
p 6.5-8	/ɑ/	sex	1	5.637	.037	.581
	/i/	Intercept	1	3.376	.091	.394
		load	1	7.593	.017	.716
		sex	1	6.859	.022	.672
p 8-10	/ɑ/	load * recog	1	4.041	.070	.450
p10-12.5	/u/	Intercept	1	3.596	.079	.423
		age	1	5.744	.031	.607
		load	1	7.727	.015	.734
		sex	1	4.927	.043	.542
		load * sex	1	9.592	.008	.821

The estimated marginal means for vocal load by recognition type showed that in nearly all conditions the subjects experienced a rise in F0. Following up the interactions by exact Mann-Whitney tests, for all three vowels a difference was found between high vocal load speakers who carried out the discrete task, and those who carried out the continuous recognition task. The high load - discrete group revealed increases in F0 for all three vowels. The high-load continuous group however, saw a decrease in F0 for /ɑ/ and /u/,

though a relatively small increase in F0 for /i/. Another significant difference was found between the high and low vocal load speakers in the discrete recognition group for /ɑ/ and /u/. Both revealed an F0 increase, though the increase for the high vocal load speakers was significantly larger.

For F0 min, load by recognition interactions were also found for all three vowels. Vowels /u/ and /i/ revealed differences between the high vocal load speakers in the continuous and in the recognition group. Both groups revealed an increase in F0 min, but the increase for the high load continuous speakers was substantially lower. In addition, a significant difference was found between high and low vocal load speakers in the discrete recognition group for /ɑ/. Whereas those in the high-discrete group revealed an increase in F0 min, those in the low-discrete group revealed a decrease. Follow up of the main effect for F0 max did not produce significant results.

ANOVA results for measures of spectral slope are reported in table 2. Main effects for sex were found for /i/ and /u/ for the level difference between the first and second harmonic Lh1-Lh2. For both vowels female speakers showed an increase in the difference between the two harmonics, versus a decrease in the male speakers .

The intercept for Lh1-LF1 for vowels /ɑ/ and /i/ was caused by an overall decrease in the difference between Lh1 and LF1 after the dictation task. The main effect of vocal load for vowel /i/ resulted from an increase in the difference between Lh1 and LF1 for high vocal load speakers, in contrast with a decrease for low vocal load speakers. Follow up of the main effect of sex for vowel /u/ revealed an increase in the difference between Lh1 and LF1 for female speakers, and a decrease for male speakers.

For LF1-LF3 (/ɑ/) a significant intercept was found, resulting from an overall decrease between the levels of F1 and F3. Main effects of vocal load were found for Lh1-LF2 and LF1-LF3 for /ɑ/. These effects required non-parametric follow up due to variance violation. However, because age was a significant covariant, it was not possible to confirm these effects. For LF1-LF3 an interaction of vocal load by sex approaching significance was found for /i/. All groups experienced a decrease in LF1-LF3, except for low vocal load females.

For CoG a main effect of sex for vowel /ɑ/ resulted from an increase in male speakers as opposed to a decrease in female speakers. Interactions of vocal load by recognition for /ɑ/ and /u/ were caused by differences between high vocal load speakers in the discrete and the continuous group. For /ɑ/ both saw a decrease in CoG although the decrease was larger for the speakers in the continuous recognition group. For /u/ however, high vocal load speakers in the discrete group revealed an increase in C0G, in contrast with those in the continuous group who experienced a decrease.

ANOVA results for measures of spectral power are displayed in table 3. The vocal load by recognition interaction for power in 0-1 kHz (/ɑ/) resulted from a decrease in the low vocal load speakers in the discrete group, versus an increase in the low load - continuous group. In addition, the continuous recognition group revealed an increase for low load speakers, versus an increase for high load speakers.

In the region 1-2 kHz, a main effect of sex for vowel /ɑ/ was caused by an increase in power for male speakers, as opposed to a decrease for female speakers. A main effect of vocal load for 2-3.5 kHz (vowel /ɑ/) could not be followed up due to the significance of the covariate age. An interaction of vocal load by recognition for the same parameter for vowel /u/ resulted from an increase in high vocal load speakers versus a decrease in low load speakers in the discrete recognition group. Another difference resulted from a decrease in high load speakers in the continuous group versus an increase in those in the discrete group.

For 3.5-5 kHz, the interaction of vocal load by recognition for vowel /u/ resulted from an increase in high load speakers in the discrete recognition group, in contrast with a decrease in low load speakers. Low load speakers in the continuous group also revealed an increase as opposed to those in the discrete group.

In the region 5-6.5 kHz, the main effect of sex for vowel /i/ was caused by an increase for both female and male speakers, though the increase was significantly larger for the male speakers.

Follow up of the main effect for sex for /ɑ/ for the region 6.5-8 kHz, as well as the intercept and main effect for vocal load for /i/ did not produce

significant results. For the latter vowel, the main effect of sex resulted from an increase for male, versus a decrease for female speakers.

The vocal load by recognition interaction for /ɑ/ for the region 8-10 kHz resulted from an increase in high and low vocal load speakers in the discrete group, though the increase was larger for the high load speakers. Low load speakers in the discrete as well as the continuous group also revealed an increase, though the increase was largest for the continuous group.

The region 10-12.5 kHz revealed several results for /u/. The intercept reflected an overall increase in power. The main effect for vocal load was caused by an increase in power for low vocal load speakers and a decrease for high load speakers. The main effect for sex resulted from an increase for male as well as female speakers, although the increase was larger for the female speakers.

4. DISCUSSION

The most salient finding appears to be an increase in F0 and F0 min after the speech recognition task, with only a few isolated exceptions. However, there are no clear patterns in the results that suggest that use of discrete versus continuous recognition software, or the higher or lower vocal load of a speaker has a greater or lesser effect on the measurements.

The rise in F0 corresponds with results found in the vast majority of studies on voice fatigue. Stemple et al. [7] proposed, based on Greene [8], that vocal fatigue is caused by weakness of the thyroarytenoid muscle, which they claim is responsible for low pitch attainment. However, anatomy of the TA muscle suggests it may be highly resistant to muscular fatigue [9]. Moreover, findings by Titze et al. [10] suggest that at low (habitual speaking) frequencies weakening of the TA muscle should lead to a drop, rather than an increase, in F0.

Vilkman et al. [11, 3] propose that vocal loading leads to physiological changes in the vocal folds, e.g. in the mucosa. They suggest that, in order to counter these changes, a speaker increases the glottal adductory forces. This would raise the subglottal pressure, causing a rise in F0.

An increase in glottal adductory forces should translate into more abrupt closure of the vocal folds. This would be reflected in a leveling of the spectral slope. The results in this study for the spectral slope measures, as well as for the measures of power distribution across the spectrum are spurious. A significant result for one parameter or one vowel was most often not repeated in other vowels or similar parameters. One relatively consistent pattern though is the increase in power above 5 kHz, in particular for male speakers, pointing at a leveling of the spectrum. This corresponds with the decrease in Lh1-Lh2 for /i/ and /u/, a decrease in Lh1-LF1for /u/ and an increase in CoG for /ɑ/. These findings thus seem to support the hypothesis by Vilkman et al. [11, 3].

5. REFERENCES

[1] Kambeyanda, D., Singer, L., Cronk, S. 1997. Potential problems associated with use of speech recognition products. *Assistive Technology* 9, 95-101.

[2] Haxer, M.J., Guinn, L.W., Hogikyan, N.D. 2001. Use of speech recognition software: a vocal endurance test for the new millennium. *Journal of Voice* 15, 231-236.

[3] Rantala, L., Vilkman, E., Bloigu, R. 2002. Voice changes during work: subjective complaints and objective measurements for female primary and secondary schoolteachers. *Journal of Voice* 16, 344-355.

[4] Boersma, P., Weenink, D.J.M. 1996. Praat, a system for doing phonetics by computer. *Report of the Institute of Phonetic Sciences 132,* University of Amsterdam.

[5] De Krom, G. 1994. *Acoustic correlates of breathiness and roughness: experiments on voice quality*. OTS, Research Institute for Speech and Language, Utrecht University, The Netherlands: Published doctoral dissertation.

[6] De Bruijn, C. 2007. Voice quality after dictation to speech recognition software: a perceptual and acoustic study. University of Sheffield, UK: Doctoral dissertation

[7] Stemple, J.C, Stanley, J., Lee, L. 1995. Objective measures of voice production in normal subjects following prolonged voice use. *Journal of Voice* 9, 127-133.

[8] Greene, M.C.L. 1972. *The voice and its disorders*. Philadelphia: Lippincott.

[9] Welham, N.V., Maclagan, M.A. 2003. Vocal fatigue: Current knowledge and future directions. *Journal of Voice* 17, 21-30.

[10] Titze, I.R., Luschei, E.S., Hirano, M. 1989. Role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice* 3, 213-224.

[11] Vilkman, E., Lauri, E., Alku, P., Sala, E., Sihvo, M. 1998. Ergonomic conditions and voice. *Logopedics Phoniatrics Vocology* 23, 11-19.

[i] CoG stands for centre of gravity, Lh for harmonic level (hand measured in FFT spectrum), LF for formant level (FFT spectrogram with 10 ms Gaussian window). All levels were normalized to RMS.