Voice fatigue and use of speech recognition: A study of voice quality ratings

Christel de Bruijn [1], *Sandra Whiteside* [2]

[1] Division of Speech and Language Therapy, University of Central England, Birmingham, UK
[2] Department of Human Communication Sciences, University of Sheffield, UK

christel.debruijn@uce.ac.uk, s.whiteside@sheffield.ac.uk

10.21437/Interspeech.2007-384

Abstract

Previous studies have suggested the use of speech recognition software may be related to the development of voice problems. The aim of this study is to investigate the effects of using such software on perceptual voice quality. In particular, the variables type of speech recognition (discrete and continuous) and vocal load of a speaker are considered. One of the most consistent results was a rise in pitch, a common finding in voice fatigue studies. It is interpreted as part of a hyperfunctional mechanism countering early signs of voice fatigue.

Index Terms: voice fatigue, speech recognition, pitch, hyperfunction

1. Introduction

During the early 1990s, speech recognition software made its way from the confinement of the research laboratories and purpose-built systems of large companies, to the general consumer market. However, as is the case with the introduction of many new technologies, reports started to appear soon after about potentially adversary effects the use of this software might have on voice. Reports appeared in various media, ranging from computer magazines, mailing lists for users of speech recognition software to the Bulletin of the Royal College for Speech and Language Therapists (UK). The problems and symptoms reported varied from a dry throat to being unable to speak more than half an hour a day.

Until now, only two studies have been published in an attempt to examine the influence of speech recognition software on voice. One study reported on a single case study [1], whereas the other study reported on a survey and clinical examination of speech recognition users [2]. The aim of the study presented here is to investigate the effects of speech recognition on voice in an experimental setting. 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 which could potentially 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

Four different subject groups carried out a speech recognition task for 2 hours. Voice samples were recorded before and after dictation and rated by a listeners panel on several parameters. Results were statistically analysed with ANOVAs.

2.1. Speakers and voice recordings

A group of 25 subjects undertook a 2-hour dictation task using ASR software. Cross over of the two variables, recognition type (discrete vs. continuous) and vocal load (high vs. low), 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). Categorisation of a speaker into the high or low vocal load group was dependent on their typical daily voice use. All participants were native speakers of British English, non-smoking, did not suffer from voice problems and were computer literate. None of the participants had used ASR software for any prolonged period of time. In addition, candidates with any strong regional accent or dialect were excluded from the study, as different accents may affect recognition performance. Groups could not be matched for sex or age, but this issue was addressed in the statistical analysis of the results. Single tokens of sustained vowels /ɑ/,/i/ and /u/ were recorded before and after dictation.

2.2. Listeners panel and ratings

The voice recordings were evaluated by a panel of 11 listeners, consisting of speech and language therapists and final year speech and language therapy students. All listeners were native speakers of British English. None of the listeners reported any hearing problems.

The voice fragments were rated on a number of parameters, as shown in table 1. The choice of parameters was based on the symptoms reported in the two published studies on use of speech recognition software and voice quality [1, 2]. Additional parameters were chosen from the GRBAS scale [4], and on the basis of other studies on vocal fatigue [5-11].

Each parameter was rated either on a 5-point scale or a 9 point scale if the parameter was bipolar. Listeners were trained on a random selection of samples taken from the entire set of voice recordings. All ratings were carried out relative to this training set. For example, a rating of 0 on the breathiness scale means “no breathiness at all”, where 4 means “most breathy compared to the samples of the training set”. For a bipolar parameter such as pitch, a rating of 0 indicated “lowest pitch”, 4 indicated “neutral” and 8 “highest pitch”. Listeners were encouraged to use the full range of the scale. The approach of “normalising” the ratings to the range of qualities in the entire set was chosen because changes in voice quality were expected to be relatively small and might not be revealed if samples were rated against clinical standards.

Table 1**:** *Perceptual parameters and rating scales.*

Parameter	Scale range
Breathiness	0-4
Roughness	0-4
Creak/ glottal fry	0-4
Strain/ vocal effort	0-4
Hard glottal attack	0-4
Asthenicity/ voice weakness	0-4
Lacking sonority	0-4
Overall instability	0-4
Overall vocal deviation	0-4
Pitch	0-8
Hypo/ hyperfunctionality	0-8

2.3. Statistical analysis

For each parameter (and for each speaker) differences were calculated between before and after ratings. Univariate ANOVAs were carried out for each parameter on these differences 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. When cell sizes were smaller than five or when variance assumptions were not met (checked for by Levene's test) ANOVA results were followed up with parametric independent-samples tests, or with non-parametric exact tests. 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.

Intrarater reliability was calculated for each speaker using Pearson's correlation coefficients. Ratings for a parameter (by a particular listener) were not entered for further statistical analysis if that parameter reached a correlation of less than 0.60. Interrater reliability was calculated using intra-class correlation coefficients. ANOVA results for voice fragments that did not reach a correlation of 0.60 were ignored. All results were obtained with SPSS statistical software. More details about the methodology can be found in [12].

3. Results

ANOVA results for the various parameters are shown in table 2. Only significant results ($p<0.05$) or those approaching significance ($p<0.10$) are reported.

For the parameter roughness, interactions of vocal load by recognition and vocal load by sex approaching significance were found for /i/. Follow up of the recognition by load interaction by a Mann-Whitney test revealed a significant difference between speakers in the discrete recognition group with a high and a low vocal load. Those with a low vocal load saw an increase in roughness, whereas those with a higher vocal load showed a decrease. Follow up of the vocal load by sex interaction did not produce any further significant results.

For creak, again for vowel /i/, inspection of the grand mean showed that, when the difference subject groups were ignored, there was an overall increase after dictation. The marginal means showed that the main effect of recognition was attributable to an increase in creak in the continuous recognition group and a decrease for the discrete group. The vocal load by sex and the recognition type by sex interactions could not be further explored due to significance of the covariate age.

Table 2**:** *ANOVA results with recognition type, vocal load and sex as between-subjects factors.*

parm	stim	source	df	F	sig	obs power
roughness	/i/	load * recog	1	4.607	.053	.506
		load * sex	1	3.328	.093	.389
creak	/i/	intercept	1	11.373	.006	.871
		age	1	11.299	.006	.869
		recog	1	7.149	.020	.690
		load * sex	1	14.770	.002	.941
		recog * sex	1	4.291	.061	.478
strain	/i/	load * recog	1	3.856	.073	.439
		recog * sex	1	6.749	.023	.665
glottal attack	/u/	load	1	5.139	.040	.560
asthenicity	/u/	recog * sex	1	5.157	.039	.561
lack of sonority	/α/	intercept	1	3.883	.072	.441
		age	1	3.289	.095	.386
	/i/	recog * sex	1	8.451	.013	.761
overall instability	/α/	recog	1	6.514	.025	.650
		load * sex	1	3.844	.074	.438
	/i/	age	1	3.420	.089	.398
		recog * sex	1	23.297	.000	.993
overall deviation	/α/	intercept	1	4.098	.066	.461
	/i/	load * recog	1	3.921	.071	.445
		recog * sex	1	7.834	.016	.729
	/u/	load	1	3.789	.072	.442
		recog	1	6.613	.022	.667
		sex	1	6.203	.026	.640
pitch	/α/	load * recog	1	4.586	.053	.504
	/i/	load	1	4.022	.068	.454
		load * sex	1	3.303	.094	.387
	/u/	intercept	1	11.411	.005	.881
		age	1	6.243	.026	.642
		load	1	20.523	.000	.988
		recog	1	3.507	.082	.415
		sex	1	17.438	.001	.972
		load * recog	1	7.276	.017	.709
Hypo/hyper function	/i/	intercept	1	5.129	.043	.549
		age	1	3.615	.082	.417
		load * sex	1	3.755	.077	.430

For strain, the interaction of vocal load by recognition for /i/ resulted from a difference between speakers with a high and those with a low vocal load in the continuous recognition group. Those with a high load experienced an increase in

strain whereas those with a lower load revealed a decrease. The interaction of recognition type by sex resulted from differences between male and female speakers in the continuous group and differences between female speakers in the continuous and discrete group. The marginal means showed that female speakers in the continuous group saw a decrease in strain, whereas the male speakers in the same group saw an increase. In addition, the female speakers in the discrete group saw an increase in strain as well, in contrast with the female speakers in the continuous group.

The significant main effect of vocal load for glottal attack for vowel /u/ was caused by an increase after dictation for those with a high vocal load and a decrease for those with a lower load.

For asthenicity, follow up of the recognition type by sex interaction for vowel /u/ did not reveal any significant differences between the groups.

Lack of sonority revealed an intercept approaching significance for vowel /ɑ/. The grand mean revealed a decrease in lack of sonority after dictation, thus reflecting an overall increase in sonority. For /i/, the recognition by sex interaction was caused by an increase in sonority for the male speakers in the discrete recognition group versus a decrease for the female speakers in that same group. Another difference was found between the female speakers in the continuous and discrete recognition groups. In the continuous group they revealed an increase in sonority and in the discrete group a decrease. Finally, a difference was found between the two diametrically opposite groups male speakers-continuous recognition and female speakers-discrete recognition. The former group revealed an increase in sonority, whereas the latter showed a decrease.

For parameter overall instability, a significant main effect was found for recognition type for vowel /ɑ/, as well as an interaction of vocal load by sex which approached significance. Because Levene's test for homogeneity of variance was significant for vowel /ɑ/, the main effect and interaction were followed up by non-parametric exact Mann-Whitney tests. Neither of these produced any significant results. A significant interaction of recognition type by sex was found for vowel /i/. Because one of the cell sizes was smaller than five (N=3) the result should have been followed up by an exact test. This was however, not possible due to a significant covariance with age.

For parameter overall deviation, an intercept approaching significance for vowel /ɑ/ resulted from an overall decrease after dictation. Follow up of the interactions of recognition type by sex and vocal load by recognition for vowel /i/ did not reveal any significant group differences. Finally, for vowel /u/ significant main effects were found for recognition type and sex, and a main effect for vocal load approaching significance. Although cell sizes were sufficient in each case, Levene's test was significant, and therefore these main effects were followed up by non-parametric statistics. Follow up of the main effect for sex was found to be significant, with female speakers showing an increase in deviation after dictation whereas male speakers showed a decrease. Follow up of the main effect for recognition type was also significant with those in the discrete recognition group showing an increase in overall deviation, and those in the continuous group showing a decrease. Follow up of the main effect vocal load did not produce a significant result.

For pitch, an interaction of vocal load by recognition approached significance for vowel /ɑ/. Because of small cell sizes, this was followed up by non-parametric exact statistics. A difference approaching significance was found between the two diametrically opposite groups discrete-high load and continuous-low load. Both groups saw an increase in pitch but the increase was largest for the discrete recognition-high load group.

A vocal load by sex interaction approaching significance was found for vowel /i/. Follow up by non-parametric exact statistics (due to small cell sizes) revealed a significant difference between male speakers with a high and a low vocal load. Although speakers in both groups experienced an increase in pitch after dictation, the increase was significantly larger for the male speakers in the high vocal load group. A significant difference was also found between the two diametrically opposite groups male-high load and female-low load. Both revealed an increase in pitch after dictation, although the increase was significantly larger for the male speakers in the high vocal load group.

For vowel /u/ a significant intercept for pitch was found, as well as significant main effects for vocal load and sex and a main effect for recognition type approaching significance. A significant interaction between vocal load and recognition was also found, but could not be further explored due to significant covariance with age. The intercept reflected an overall increase (when all groups are pooled together) of pitch after dictation. The main effect for sex revealed that both male and female speakers showed an increase in pitch after dictation, although the increase was larger for male than for female speakers. Investigating the means for the main effect of vocal load revealed that speakers with a high vocal load revealed an increase in pitch whereas those with a low vocal load revealed a decrease. Finally, the main effect for recognition type showed that speakers in both the continuous and discrete recognition group revealed an increase in pitch, although those in the discrete group revealed a larger increase. It should however, be borne in mind that a high-order interaction of load by recognition type could not be interpreted due to significance of the covariate age, therefore impacting on the interpretation of the main effects of vocal load and recognition type.

The final parameter that revealed significant results was hypo/hyperfunction. For vowel /i/ a significant intercept was found. The overall means showed that this intercept revealed an overall shift towards hyperfunction after the dictation task. A vocal load by sex interaction approaching significance was also found for this vowel, but could not be further explored due to the covariate age approaching significance.

4. Discussion

A number of perceptual parameters appear to be affected by the speech recognition task. Roughness became worse for low load speakers, but less for high load speakers in the discrete recognition group. Creak increased in the continuous recognition group, but decreased in the discrete group. A higher-order interaction was present for this parameter, but could not be interpreted hence limiting the interpretation of the main effect for recognition. Strain decreased in female speakers in the continuous-recognition group – in contrast with female speakers in the discrete group who experienced an increase – but increased for male speakers in the same continuous group. In the continuous recognition group strain

also increased for speakers with a high vocal load and decreased for those with a low vocal load. Hypo/hyperfunction revealed an overall shift towards hyperfunction. Finally, every interpretable main effect or interaction for pitch was caused by an increase in pitch.

The above findings paint a complicated picture. Although a number of significant results were found, these occurred in most cases in very specific circumstances, i.e. for one particular combination of vowel, vocal load group and speech recognition type, rather than across the board. This means that although a number of voice quality parameters appear to be affected by the dictation task, it could not be established if these results were systematically influenced by the variables vocal load and type of recognition.

Pitch was the only parameter for which consistent results were found across the different vowels, showing an increase in all cases. Few studies have been reported that systematically evaluated perceptual parameters of voice quality in relation to voice fatigue. Most studies involved a vocally fatiguing task, such as prolonged reading, and evaluated voice quality before and after. However, considerable variability exists in the nature and length of the experimental tasks in these studies. Moreover, different perceptual parameters were evaluated by either the subjects and patients themselves or by trained listeners (e.g. therapists) in different studies. This variability spreads the water thin and leads to complications in interpreting and comparing results across studies.

One salient finding across several fatigue studies has been the increase in pitch or fundamental frequency [e.g. 8, 10]. The core finding of this study, an increase in (perceived) pitch, seems to correspond to a majority of published results for this parameter.

Stemple et al. [7] proposed, based on Greene [13], that vocal fatigue is caused by weakness of the thyroarytenoid muscle, which they claim is responsible for low pitch attainment. In their study, they found a higher fundamental frequency as well as anterior glottal chinks for 6 out of 10 subjects after a 2-hour reading task. They hypothesised that because low pitch attainment is accomplished by contraction of the TA muscle, strain or weakness of the TA muscle may lead to an increased fundamental frequency.

There are however, at least two arguments against the theory proposed by Stemple et al. [7]. First of all, it has been pointed out by Welham and Maclagan [14] that the human thyroarytenoid muscle (which is responsible for adduction of the vocal folds) contains a relatively high number of type 1 (slow-twitch) motor unit fibre types, which are more fatigue resistant than type 2 (fast-twitch) units. They have therefore raised the issue that the human TA muscle may in fact be high resistant to muscular fatigue. Moreover, findings by Titze et al. [15] suggest that at low (habitual speaking) frequencies (i.e. when the cricothyroid muscle is not or little contracted), contraction of the TA muscle leads to an increase of F0. Weakening of the TA muscle should therefore lead to a drop, rather than an increase, in F0.

A different explanation for the increase of F0 has been offered by Vilkman et al. [16, see also 3]. They proposed that changes in some parameters, including fundamental frequency, are caused by the speakers' compensatory reactions to changes in their voice. They hypothesised that physiologic changes such as alterations of the mucosa, leads the speaker to increase the glottal adductory forces (i.e. hyperfunction). As a result, the subglottal pressure supposedly increases and raises the fundamental frequency. Indeed, a shift towards hyperfunctionality, as well as an increase in glottal attack and strain/ vocal effort was found in some conditions in the current study, supporting the hypothesis put forward by Vilkman et al.

To summarise then, a rise in pitch, a common finding in voice fatigue studies, was found across the different experimental conditions following the speech recognition task. Although several other parameters also appeared to be affected, it could not be established if these results were systematically influenced by the variables vocal load and type of recognition.

5. References

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

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

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

[4] Hirano, M. Clinical examination of voice. Springer-Verlag, Vienna and New York: 1981.

[5] Sander, E. K. and Ripich, D. E., "Vocal fatigue", Ann. Otol. Rhinol. Laryngol. 92: 141-145, 1983.

[6] Gotaas, C. and Starr, C. D., "Vocal fatigue among teachers", Folia Phoniatrica 45: 120-129, 1993.

[7] Stemple, J. C., Stanley, J. and Lee, L., "Objective measures of voice production in normal subjects following prolonged voice use", Journal of Voice 9(2): 127-133, 1995.

[8] Rantala, L., Paavola, L., Körkkö, P. and Vilkman, E., "Working-day effects on the spectral characteristics of teaching voice", Folia Phoniatrica et Logopaedica 50: 205-211, 1998.

[9] Solomon, N. P. and DiMattia, M. S., "Effects of a vocally fatiguing task and systemic hydration on phonation threshold pressure", Journal of Voice 14(3): 341-362, 2000.

[10] Jónsdottir, V., Laukkanen, A. and Vilkman, E., "Changes in teachers' speech during a working day with and without electric sound amplification", Folia Phoniatrica et Logopaedica 54: 282-287, 2002.

[11] Jónsdottir, V., Laukkanen, A. and Siiki, I., "Changes in teachers' voice quality during a working day with and without electric sound amplification", Folia Phoniatrica et Logopaedica 55: 267-280, 2003.

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

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

[14] Welham, N. V. and Maclagan, M. A., "Vocal fatigue: Current knowledge and future directions", Journal of Voice 17(1): 21-30, 2003.

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

[16] Vilkman, E., Lauri, E., Alku, P., Sala, E., Sihvo, M., "Ergonomic conditions and voice", Logopedics Phoniatrics Vocology 23: 11-19, 1989.