Proceedings of the Institute of Acoustics

SPEECH RECOGNITION BASED HUMAN COMPUTER INTERFACING AND ITS POSSIBLE CONSEQUENCES ON VOICE QUALITY: AN ACOUSTIC STUDY

C. G. de Bruijn (1), S. P. Whiteside (1), P. A. Cudd (1), D. Syder (1), K. M. Rosen (2), L. Nord (2)

(1) University of Sheffield, UK
(2) KTH, Stockholm, Sweden

1. INTRODUCTION

Now that sufficient computing power for automatic speech recognition (ASR) has become affordable for the general public, ASR software packages have become off-the-shelf products. These recognition packages are more widely used by people, both at home and in the office, as an alternative input method to the keyboard. However, recent research and individual reports have shown evidence of vocal fatigue and symptoms of dysphonia, as a result of the use of speech recognition based human computer interfaces. These reports make it clear that it is necessary to qualify the potential risks of voice damage.

The present study is part of ongoing research that is investigating acoustic changes in the voice, after use of a discrete speech recognition system. The study reports on two Swedish subjects who used such a system. These subjects function as demonstrators for the ENABL project, of which this study is part. The ENABL project, in which KTH, Sweden is one of the partners, aims to couple a speech driven user interface together with vocational generative modeling software. The two Swedish subjects will evaluate the use of the entire system. The Sheffield group will provide the demonstrators with voice care and voice monitoring. One of the final aims is to determine at risk populations of users of ASR systems.

Studies that have reported on the topic of ASR related voice problems include Cudd et al. [1] and Kambeyanda et al. [2]. The research carried out by Kambeyanda et al. [2] consisted of a clinical study and a survey with questions regarding subjects' use of a discrete speech recognition system. Four subjects were selected out of the seventy valid responses to the survey, for further clinical examinations. These subjects were reported to have severe voice problems, though none of them had had previous vocal disorders before the use of speech recognition systems. Within a year after the start of using these recognition systems they were said to have developed various throat and voice problems, eventually leading to loss of voice control and almost complete voice loss. The findings of their survey can be summarised as follows:

1. The length of a typical work period T (where T <= 1/2 hour or T > 1/2 hour) did not seem to have any influence on the occurrence of voice problems, as no significant relationship was found between the two. However, cases of severe periods of voice loss were reported in clinical studies, after 4 hours of continuous use of the system.
2. The way in which the recognition system is used appeared to be of influence on the development of voice problems. The percentage of use of speech recognition as a computer access method S (i.e. computer control and navigation as opposed to dictation), where S < 50% or S >= 50%), and the occurrence of voice problems were highly and significantly related.
3. The occurrence of voice problems and the presence of CTD (Cumulative Trauma Disorder) appeared to be positively correlated. This correlation was highly significant.

In the clinical results the authors observed a continuum of voice stress symptoms which could be categorised into three progressive phases. The first phase, or Early Onset, was characterised by a dry and/or tickly throat, coughing bouts, slowly lowering pitch and a chronic turning hoarseness. These symptoms could be partially relieved by drinking liquids and periods of rest. Typical symptoms of phase

II, the Progressive Phase, were strap and neck muscle ache, a sore throat, a voice that was always hoarse and breathy, with the normal pitch lowered, the inability to increase loudness and voice breaks at progressively shorter intervals. Because the vocal cords are bowing, the subject experiences extreme fatigue in speaking and has difficulty in talking or communicating. Recovery is possible, but takes more and more time. The third phase was described as the Cumulative Phase and was characterised by the same symptoms as phase II. However, the symptoms had become chronic and spontaneous recovery without the interference of a voice expert such as a speech and language therapist, was no longer possible.

The authors believe that users tend to maintain a constant pitch, volume and inflection in order to avoid recognition errors. This could result in the fixation of the vocal musculature, which in turn can cause muscle fatigue and eventually laryngeal damage [3, 4]. The authors therefore hypothesise that the development of the voice disorders in their subjects may have been caused by their attempts to avoid these recognition errors.

In the current study we present some acoustic analyses that have been carried out on recordings which were made before and after a dictation task, in order to determine whether there are any detectable acoustic changes after using a dictation system. The acoustic parameters that were investigated included fundamental frequency, overall energy, harmonic-to-noise ratio, jitter, energy under 6 kHz, energy above 6 kHz, and shimmer. The results are presented and discussed.

2. MATERIAL AND SUBJECTS

2.1 Subjects

The subjects were two male Swedish users, aged in their mid-thirties. They were asked to fill in the Victoria Infirmary Voice Questionnaire [5], which provided us with information about their vocal histories. It showed that in the past both users had had a tracheotomy. This however did not influence their voices in the long term. Both users reported not to have any current voice problems, though after examination by a speech and language therapist, one user appeared to have some breath control problems.

2.2 Speech material

The steady states of 5 different sustained vowels were used as speech material for the acoustic analysis. The vowels that were chosen, were the four vowels [ɑ], [ɪ], [u], [æ] of the vowel quadrangle and [ə]. The subjects were asked to produce and hold the vowels for three seconds at a comfortable pitch and loudness level. In addition they were asked to read out the Rainbow passage, which is a phonetically balanced text [6]. The recordings of this text were also used in the acoustic analyses.

2.3 Recording

The speech was recorded in a sound treated room, using a DAT recorder (SONY TCD-D10) with a SONY ECM-959DT microphone. The distance from head to microphone was approximately 20 cm. In order to carry out the acoustic analysis, the speech material was later recorded from the DAT tape onto computer hard disk at a sampling rate of 44.1 kHz and with 16-bit resolution. During this process the original relative intensity levels were maintained. For the acoustic analyses the software package Multi-Speech, Model 3700 from Kay Elemetrics Corp. was used.

3. METHOD

3.1 Dictation Task.

The subjects undertook a task which consisted of twenty minutes of dictation. Before and after this dictation task, audio recordings were made of the subjects' voices. The topic of dictation was up to the choice of the subjects. During the task they had a glass of water at their disposal. Dragon Dictate, a

software program for recognition of discrete speech, was used for the dictation task. This style of dictating requires a slight pause between words. This means that after each word, the vocal cords are in a state of abduction and have to adduct with the start of each new word. This way of speaking places a heavier load on the voice than for instance normal conversation. Therefore, it is the hypothesis of the authors of this paper that this style of dictation may contribute to the development of vocal fatigue.

3.2 Acoustic Analysis

The acoustic analysis consisted of six different measures, as listed below. The measures were taken before and after the dictation task.

- Overall energy (dB). During the audio recordings, no absolute SPL values were measured. However, since the relative energy values were preserved, these could be used as valid indicators for the amount of effort being used.

- Jitter (%). Increases of spectral noise levels and perceived roughness have been associated with higher levels of jitter, as was described by Deal and Emanuel [8]. Jitter has also been reported by Klingholz and Martin [9], to play a role in the differentiation between hypo- and hyperfunctional voice disorders.

- Shimmer (dB). Not as much research has been carried out on shimmer as it has been on jitter. However, shimmer has been reported to contribute to the perception of hoarseness [10, 11].

- Harmonic-to-Noise Ratio HNR (dB). The replacement of harmonics in a speech signal by noise is a typical characteristic of hoarseness. According to studies by Yumoto, Gould and Baer [6] and Yumoto, Sasaki and Okamura [7], the use of the harmonics-to-noise ratios is an objective and quantitative way of evaluating the degree of hoarseness.

- Fundamental frequency F0 (Hz). Lowering pitch was one of the features that characterised the progressive stages in the observed continuum of voice stress symptoms, as described in the study by Kambeyanda et al. [2].

- Energy (dB) under 6 kHz. Comparing energy levels in a spectrum below and above 6 kHz is the best way of differentiating between normal and pathological voices, according to De Jonckere [12]. He found that the spectra of pathological voices contain higher energy levels above 6 kHz.

- Energy (dB) over 6 kHz. See above.

4. RESULTS

Acoustic analyses were carried out on the steady states of sustained vowels, which were recorded before and after dictation. The results for user U are shown in tables 1 and 2. Mean values, standard deviation and minimum and maximum values were measured on the basis of 15 data points, namely three repetitions for each of the five vowels. The results for user R are shown in tables 3 and 4. For the calculations of these results only one repetition was available per vowel. Tables 5 and 6 show the results of the acoustic analyses (F0 and Energy) on the Rainbow passage, for users R and U respectively.

	Mean	S.d.	Min.	Max.
Energy (dB) under 6 kHz before	-6.01	2.36	-9.66	-2.04
Energy (dB) under 6 kHz after	-7.07	5.20	-18.15	-0.85
Energy (dB) over 6 kHz before	-18.59	3.78	-25.32	-15.05
Energy (dB) over 6 kHz after	-18.36	5.20	-24.72	-13.84
Shimmer (dB) before	0.30	0.29	0.07	0.58
Shimmer (dB) after	0.26	0.22	0.06	0.87

Table 1: Results of acoustic analysis for user U. Calculated parameters are energy (dB) under and over 6 kHz and shimmer (%), measured before and after the dictation task.

	Mean	S.d.	Min.	Max.
F0 (Hz) before	167.88	12.50	149.24	188.52
F0 (Hz) after	164.97	10.67	150.80	191.63
Energy (dB) before	70.17	3.35	62.27	74.92
Energy (dB) after	69.52	3.00	66.12	74.22
HNR (dB) before	6.14	4.54	0.30	14.15
HNR (dB) after	7.87	3.52	-1.16	13.43
Jitter (%) before	1.29	1.48	0.190	5.63
Jitter (%) after	2.23	3.34	0.24	10.55

Table 2: Results of acoustic analysis for user U. Parameters are fundamental frequency F0 (Hz), overall energy (dB), harmonic-to-noise ratio (dB) and jitter (Hz), measured before and after the dictation task.

	Mean	S.d.	Min.	Max.
Energy (dB) under 6 kHz before	-16.39	4.21	-20.71	-11.25
Energy (dB) under 6 kHz after	-14.02	2.76	-18.65	-11.25
Energy (dB) over 6 kHz before	-17.86	4.73	-19.06	-10.79
Energy (dB) over 6 kHz after	-21.12	2.38	-23.96	-17.98
Shimmer (dB) before	0.22	0.06	0.15	0.28
Shimmer (dB) after	0.24	0.12	0.17	0.44

Table 3: Results of acoustic analysis for user R. Calculated parameters are energy (dB) under and over 6 kHz and shimmer (%), measured before and after the dictation task.

	Mean	S.d.	Min.	Max.
F0 (Hz) before	123.53	10.45	106.01	132.94
F0 (Hz) after	138.03	23.27	108.63	173.74
Energy (dB) before	67.42	1.83	64.62	69.51
Energy (dB) after	64.23	0.70	63.50	64.99
HNR (dB) before	6.88	2.79	-2.88	10.59
HNR (dB) after	3.89	3.29	-0.81	9.04
Jitter (%) before	2.38	2.10	0.53	5.39
Jitter (%) after	1.11	1.24	0.29	3.28

Table 4: Results of acoustic analysis for user R. Parameters are fundamental frequency F0 (Hz), overall energy (dB), harmonic-to-noise ratio HNR (dB) and jitter (Hz), measured before and after dictation.

	Before	After
F0 (Hz)	121.155	118.564
Energy (dB)	67.447	61.19

Table 5: measurements of F0 and energy in the Rainbow passage, before and after the dictation task, for user R.

	Before	After
F0 (Hz)	181.868	120.356
Energy (dB)	62.081	62.057

Table 6: measurements of F0 and energy in the Rainbow passage, before and after the dictation task, for user U

5. DISCUSSION

Despite the fact that a substantial amount of voice patients suffer from vocal fatigue, this disorder has been an underexplored area in speech research and has not obtained the amount of attention it deserves. It is often assumed that individuals with jobs that place a high load on the voice (such as teachers), form a group of people with an increased risk of developing voice problems, of which vocal fatigue is one of the most frequent. Gotaas and Starr [14] defined vocal fatigue as a problem that starts developing as the length of speaking time progresses during the day, and which is most evident at the end of the day. Usually the symptoms disappear by the following day. They characterise the disorder by changes in vocal quality, loudness, pitch or effort, some of which possibly occurring together. These changes may be associated with vocalisation over a prolonged period of time, with high intensity and pitch levels. The vocalisation may be excessively tense and accompanied by unhealthy vocal cords. However, although extended speaking time at high intensity levels may account for vocal fatigue in some persons, other individuals develop vocal fatigue without the presence of these factors, as was reported by Gotaas and Starr [14]. They posed the question whether there are any other variables which may interact with speaking time and intensity levels and suggested that psychological tension might be at least one other factor.

The tension factor in relation to vocal fatigue is a topic of investigation in a research project currently being carried out at the University of Sheffield. The project involves four groups of individuals, who carry out a dictation task for a certain period of time using discrete speech recognition software, before and after which recordings are made of their voices. The volunteers are divided into a control group, a group of persons with a daily high vocal load, individuals with a strong regional British accent or foreign accent, and a group with individuals who have a certain anxiety to use computers, or who are less computer literate. The latter group was specially designed to investigate the tension factor in the development of vocal fatigue.

The findings presented in this paper are results from a pilot study carried out for the research project mentioned above. The paper focuses mainly on the acoustic aspects, and the parameters which were investigated included fundamental frequency, energy, jitter, shimmer, harmonic-to-noise ratio and power under and above 6 kHz.

	User R	User U
F0 (Hz)	-1.00	0.82
Energy (dB)	*5.62* *	0.66
HNR (dB)	1.30	-1.33
Jitter (%)	1.06	-0.94
Shimmer (dB)	-0.77	0.42
Mean power (dB) under 6 kHz	-0.85	0.90
Mean power (dB) above 6 kHz	*2.75* *	-0.44

Table 7: Results of statistical analysis (t-values) of acoustic results for user U and user R. Results followed by * are significant at the level $p < 0.05$.

A two-tailed t-test for paired samples of vowels was carried out in order to find significant differences in measurements before and after the dictation task. The results of the statistical analysis are shown in table 7.

For user R, the results revealed a significant difference at the 5% level for two out of the seven parameters under investigation. The overall energy parameter showed for user R a significant decrease after dictation (see figure 1). This could be interpreted as a symptom of vocal fatigue, and even preliminary stages of voice damage. The results for user U also showed a decrease in energy, but this loss of energy was not significant at the 5% level. A significant decrease in energy was also found for user R in the spectral region above 6 kHz (see figure 2).

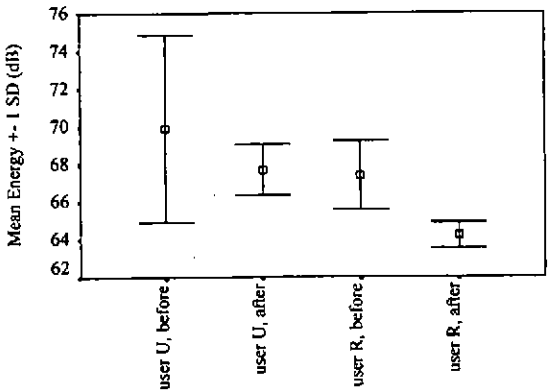

Figure 1: Mean energy (dB) before and after the dictation task

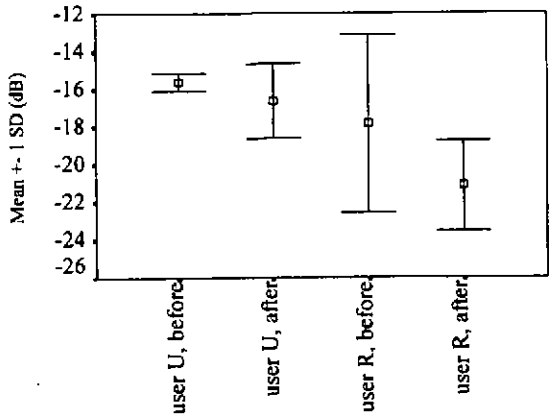

Figure 2: Energy (dB) above 6 kHz before and after the dictation task

The tendency of loss of energy is also evident from the results of the acoustic analyses on the Rainbow passage (tables 5 and 6), in which the energy level before the dictation task drops for user R from 67.4 dB to 61.2 dB. The results also show a decrease in fundamental frequency for both speakers, of which the decrease for user U is rather substantial, dropping from 182 Hz to 120 Hz. These changes can be interpreted as symptoms of vocal fatigue, as described by Gotaas and Starr [14].

Our results show that for at least two out of the seven parameters under investigation for one user, a significant voice change has occurred. One reason for the absence of any other significant differences before and after the dictation task, may be the short duration of the task, which was twenty minutes. It

could be reasoned that during this period the voice is still in a "warming up phase". Because user R had slight breath control problems, symptoms would probably become earlier evident, which explains why we only found significant results for him. In further studies we will extend the dictation task to two hours. These studies are also carried out on a larger number of subjects and under several different conditions, as outlined above.

6. ACKNOWLEDGMENTS

The authors would like to thank the Department of Speech, Music and Hearing of KTH, Stockholm, Sweden for making the audio recordings available. We would also like to thank Stephanie Martin to allow us to use the Victoria Infirmary Voice Questionnaire and make some additions to it for future research.

This research has been carried out on the ENABL project (DE 3206) and was funded by the Fourth Framework Programme of European Community activities in the field of Research and Technological Development "Telematics Applications Programme".

7. REFERENCES

1. P A CUDD, S P WHITESIDE, H STONEHAM, D SYDER & C DE BRUIJN "Using dictation systems: a contributory cause of dysphonia?", *Proceedings of VoiceData98*, p 98-103 (1998)
2. D KAMBEYANDA, L SINGER & S CRONK "Potential Problems Associated with Use of Speech Recognition Products", *Asst. Technol.*, 9 p 95-101 (1997)
3. E K SANDER & D E RIPICH, "Vocal fatigue", *Annals of Otology, Rhinology, and Laryngology*, 92 p 141-145 (1983)
4. J STEMPLE, J STANLEY,& L LEE, "Objective measures of voice production in normal subjects following prolonged voice use", *Journal of Voice*, 9 (2) p 127-133
5. S MARTIN, *Working with dysphonics: a practical guide to therapy for dysphonia*, Winslow Press, Bicester, Oxon, 1987.
6. G FAIRBANKS, *Voice and articulation drill book*, Harper & Row, New York, 1960.
7. E YUMOTO, W J GOULD & T BAER, "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am.* 71 (6) p 1544-1550 (1982)
8. E YUMOTO, Y SASAKI & H OKAMURA, H., "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness", *Journal of Speech and Hearing Research*, 27 p 2-6 (1984)
9. R E DEAL & F W EMANUEL "Some waveform and spectral features of vowel roughness", *Journal of Speech and Hearing Research*, 21 p 250-264 (1978)
10. F KLINGHOLZ & F MARTIN, "Quantitative spectral evaluation of shimmer and jitter", *Journal of Speech and Hearing Research*, 28 p 169-174 (1985)
11. R W WENDAHL, "Some parameters of auditory roughness", *Folia Phoniatrica*, 18 p 26-32 (1966)
12. R W WENDAHL, "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness", *Folia Phoniatrica*, 18 p 98-108, (1966)
13. P H DEJONCKERE, "Recognition of hoarseness by means of LTAS", *International Journal of Rehabilitation Research*, 6 p 343-345, (1983)
14. C GOTAAS & C D STARR, "Vocal Fatigue Among Teachers", *Folia Phoniatrica*, 45 p 120-129, (1993)