USING AUTOMATIC SPEECH RECOGNITION AND ITS POSSIBLE EFFECTS ON THE VOICE

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

[1] University of Sheffield, UK; [2] KTH, Stockholm, Sweden

ABSTRACT

Literature and individual reports contain indications that the use of speech recognition based human computer interfaces could potentially lead to vocal fatigue, or even to symptoms associated with dysphonia. While more and more people opt for a speech driven computer interface as an alternative input method to the keyboard, and these speech recognition systems become more and widely used, both in the home and office environment, it has become necessary to qualify any potential risks of voice damage. This study reports about ongoing research that investigates acoustic changes in the voice, after use of a discrete speech recognition system. Acoustic analyses were carried out on two Swedish users of such a system. So far, for one of the users, two of the acoustic parameters under investigation that could be an indicator of vocal fatigue, show a significant difference directly before and after use of a speech recognition system.

1. INTRODUCTION

Literature and individual reports contain indications that the use of speech recognition based human computer interfaces could potentially lead to vocal fatigue, or even to symptoms associated with dysphonia. As increasing numbers of people opt for a speech driven computer interface, and these speech recognition systems become more and widely used, both in the home and office environment, it has become necessary to qualify any potential risks of voice damage.

Studies reporting on this topic include Cudd et al. [1] and Kambeyanda et al. [2]. The study by Kambeyanda et al. [2] consisted of two different parts: a clinical study and a survey. Respondents to the survey were asked to answer questions with regard to their use of speech recognition systems. Out of the seventy valid responses, four individuals were chosen to undergo clinical testing. These four subjects were reported to have severe voice problems. None of them had reported a previous history of voice disorders, but within less than a year they were said to suffer from a series of throat and voice problems, which eventually lead to loss of voice control and almost complete voice loss. The survey results revealed several findings:

1. A highly significant positive relationship was found between the occurrence of voice problems and the presence of CTD (Cumulative Trauma Disorder).

2. No significant relationship was found between the length of a typical work period T (where T <= 1/2 hour or T > 1/2 hour) and the occurrence of voice problems. However, they reported cases of severe periods of voice loss in clinical studies, after 4 hours of continuous use of the system.

3. A highly significant relationship was found between 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.

The results of the clinical study showed a continuum of voice stress symptoms in progressive phases. They described Phase I as the Early Onset which could be relieved partially by rest periods and drinking fluids. Typical symptoms are a dry or tickly and aching throat, coughing bouts, slowly lowering pitch and hoarseness turning chronic. Phase II or the Progressive Phase, is characterised by strap and neck muscles ache, sore throat, always hoarse and breathy voice, lower normal pitch, inability to increase loudness, voice cuts at progressively shorter intervals, vocal cords bowing, extreme fatigue in speaking and difficulty in talking.

The authors hypothesised that these symptoms might be due to a tendency of the users to maintain constant pitch, volume and inflection in order to avoid recognition errors. This in turn could lead to a fixation of vocal musculature, which could result in muscle fatigue and eventually laryngeal damage [3, 4].

The current study reports on some ongoing research on 2 Swedish subjects using a speech recognition system. The research is being carried out in the framework of the ENABL project. The aim of the project is to couple a speech user interface together with a vocational generative modeling software package. The task of the Sheffield group is to provide voice care and monitoring for the demonstrators (users U and R) of this project, and determine at risk populations of users of these systems. We present some acoustic analyses that have been carried out 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. METHOD

2.1. Speech Material and Subjects

In order to get an insight in the vocal history of the two Swedish subjects, they were asked to fill in the Victoria

Infirmary Voice Questionnaire [5]. Both users reported no current vocal problems, though user R was assessed by a speech and language therapist as having some breath control problems. Both users had had a tracheotomy in the past, but this did not have long term effects on their voices. Both speakers are male and in their mid-thirties.

As speech material, sustained vowels were chosen. The subjects were asked to produce the vowels [ɑ], [ɪ], [u], [æ], and [schwa] at a comfortable pitch and loudness level and hold the vowels for three seconds. For acoustic analysis, the steady states of the vowels were used.

Recordings were made onto a DAT recorder (SONY TCD-D10) with a SONY ECM-959DT microphone. Head to microphone distance was approximately 20 cm. All recordings were made in a sound treated room. The speech was later recorded from DAT tape onto computer hard disk with a sample rate of 44.1 kHz and a resolution of 16 bits. The original relative intensity levels were maintained during this process. The acoustic analyses of the speech data were carried out using the software package Multi-Speech, Model 3700 from Kay Elemetrics Corp.

2.2. Dictation Task

The subjects were asked to carry out a dictation task of their choice for twenty minutes. The speech recognition software used for the dictation task was Dragon Dictate, which is a discrete recognizer (i.e. one has to speak word by word with slight pauses between words). During this task they were provided with a glass of water. Audio recordings of their voices were made before and after this dictation task.

2.3. Acoustic Analysis

The following measures were taken during the acoustic analysis:

- Fundamental frequency F0 (Hz). In the study of Kambeyanda et al. [2], lowering of pitch was listed in each of the progressive phases in the observed continuum of voice stress symptoms.
- Overall energy (dB). Though no absolute values of sound pressure level were used, the relative energy values before and after dictation can be used as an indicator of the level of effort being used.
- Harmonic-to-Noise Ratio HNR (dB). Hoarseness is characterised by noise energy which replaces the harmonic structures in a speech signal. The harmonic-to-noise ratio can be used as an objective and quantitative measure to evaluate the degree of hoarseness, as described by Yumoto, Gould and Baer [6] and Yumoto, Sasaki and Okamura [7].
- Jitter (%). In a study by Deal and Emanuel [8], increases in jitter or frequency perturbation have been associated with increases in spectral noise levels and perceived roughness. In a study by Klingholz and Martin, jitter has also been associated with differentiation between hype- and hyperfunctional voice disorders [9].
- Shimmer (dB). Though shimmer has not been as extensively studied as jitter, it has been reported to contribute to the perception of hoarseness [10, 11].
- Energy (dB) under 6 kHz. De Jonckere [12] has found that the best distinction between pathological and normal voices could be made by comparing the average spectral energy above and below the reference frequency of 6 kHz. Pathological voices would show higher energy levels above 6 kHz.
- Energy (dB) over 6 kHz. See above.

Measures were taken before and after the dictation task.

3. RESULTS

The results of the acoustic analyses on the sustained vowels for user U are shown in tables 1 and 2. Measures were taken before and after dictation. The tables show the mean, standard deviation and minimum and maximum values. Tables 3 and 4 show the same parameters for user R.

For user U, calculations are based on 15 datapoints, namely three repetitions for each of the five vowels. For user R, only one repetition per vowel was available.

	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 1: Results of acoustic analysis for user U. Parameters are fundamental frequency, overall energy, harmonic-to-noise ratio and jitter, measured before and after the dictation task.

	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 2: Results of acoustic analysis for user U. Calculated parameters are energy 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 3: Results of acoustic analysis for user R. Parameters are fundamental frequency, overall energy, harmonic-to-noise ratio and jitter, measured before and after dictation.

	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 4: Results of acoustic analysis for user R. Calculated parameters are energy under and over 6 kHz and shimmer, measured before and after the dictation task.

4. DISCUSSION

Table 5 displays the results of the statistical analysis on the acoustic data. Because of the nature of the data, repeated measurements before and after the dictation task, a two-tailed t-test for paired samples was carried out. Two of the parameters under investigation showed a significant difference before and after dictation, for user R This difference was significant at the 5% level.

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

The overall energy parameter for users U and R shows a decrease in energy after the dictation task, but this change was only found to be significant for user R (see Table 5 and Figure 1). This loss of energy can be interpreted as a sign of vocal fatigue, which in turn could be interpreted as a preliminary stage of voice damage.

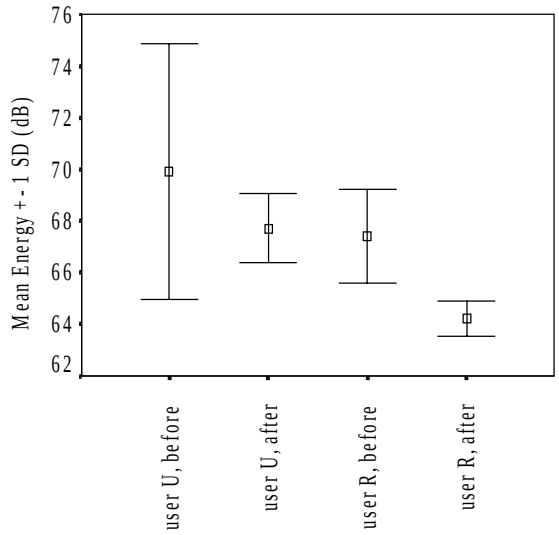

Figure 1: Mean energy (dB) before and after dictation task.

A significant difference for user R was also found in the energy levels of the spectral region above 6 kHz. The results revealed a decrease in energy over 6 kHz after the dictation task.

Our results show that for at least two out of the seven parameters under investigation for one user, a deterioration 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. In further studies we are planning to extend the dictation task to two hours. These studies will also be carried out on a larger number of subjects and under several different conditions. We will also record more data per subject.

5. 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".

6. REFERENCES

1. Cudd, P. A., Whiteside, S. P., Stoneham, H., Syder, D. and De Bruijn, C. "Using dictation systems: a contributory cause of dysphonia?", *Proceedings of VoiceData98: 98-103, 1998.*
2. Kambeyanda, D., Singer, L. and Cronk, S. "Potential Problems Associated with Use of Speech Recognition Products", *Asst. Technol. 9: 95-101, 1997.*
3. Sander, E. K., and Ripich, D. E., "Vocal fatigue", *Annals of Otology, Rhinology, and Laryngology 92: 141-145.*
4. Stemple, J., Stanley, J., and Lee, L., "Objective measures of voice production in normal subjects following prolonged voice use", *Journal of Voice 9 (2): 127-133.*
5. Fairbanks, G., *Voice and articulation drill book*, Harper & Row, New York, 1960.
6. Yumoto, E., Gould, W. J., and Baer, T., "Harmonics-to-noise ratio as an index of the degree of hoarseness", *J. Acoust. Soc. Am. 71 (6): 1544-1550, 1982.*
7. Yumoto, E., Sasaki, Y., and Okamura, H., "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness", *Journal of Speech and Hearing Research 27: 2-6, 1984.*
8. Deal, R. E., and Emanuel, F. W., "Some waveform and spectral features of vowel roughness", *Journal of Speech and Hearing Research 21: 250-264, 1978.*
9. Klingholz, F., and Martin, F., "Quantitative spectral evaluation of shimmer and jitter", *Journal of Speech and Hearing Research 28: 169-174, 1985.*
10. Wendahl, R. W., "Some parameters of auditory roughness", *Folia Phoniatrica 18: 26-32, 1966.*
11. Wendahl, R. W., "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness", *Folia Phoniatrica 18: 98-108, 1966.*
12. DeJonckere, P. H., "Recognition of hoarseness by means of LTAS", *International Journal of Rehabilitation Research 6: 343-345, 1983.*