Eur Arch Otorhinolaryngol (2006) 263: 188–193
DOI 10.1007/s00405-005-0974-6

Maria Schuster · Tino Haderlein · Elmar Nöth
Jörg Lohscheller · Ulrich Eysholdt · Frank Rosanowski

Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating

Received: 27 December 2004 / Accepted: 4 April 2005 / Published online: 7 July 2005

Abstract Substitute speech after laryngectomy is characterized by restricted aero-acoustic properties in comparison with laryngeal speech and has therefore lower intelligibility. Until now, an objective means to determine and quantify the intelligibility has not existed, although the intelligibility can serve as a global outcome parameter of voice restoration after laryngectomy. An automatic speech recognition system was applied on recordings of a standard text read by 18 German male laryngectomees with tracheoesophageal substitute speech. The system was trained with normal laryngeal speakers and not adapted to severely disturbed voices. Substitute speech was compared to laryngeal speech of a control group. Subjective evaluation of intelligibility was performed by a panel of five experts and compared to automatic speech evaluation. Substitute speech showed lower syllables/s and lower word accuracy than laryngeal speech. Automatic speech recognition for substitute speech yielded word accuracy between 10.0 and 50% (28.7 ± 12.1%) with sufficient discrimination. It complied with experts' subjective evaluations of intelligibility. The multi-rater kappa of the experts alone did not differ from the multi-rater kappa of experts and the recognizer. Automatic speech recognition serves as a good means to objectify and quantify global speech outcome of laryngectomees. For clinical use, the speech recognition system will be adapted to disturbed voices and can also be applied in other languages.

M. Schuster (✉) · J. Lohscheller · U. Eysholdt · F. Rosanowski
Department of Phoniatrics and Pedaudiology,
University of Erlangen, Bohlenplatz 21,
91054 Erlangen,
Germany
E-mail: maria.schuster@phoni.imed.uni-erlangen.de
Tel.: +49-9131-8533146
Fax: +49-9131-8539272

T. Haderlein · E. Nöth
Department of Pattern Recognition, University of Erlangen,
Erlangen, Germany

Keywords Laryngectomy · Substitute speech · Automatic speech recognition · Speech intelligibility

Introduction

After laryngectomy, patients suffer from several impairments, the loss of laryngeal speech being of outstanding importance for affected patients and their social functioning. In these patients, speech can be restored by different methods, the tracheoesophageal technique being increasingly popular because it resembles most closely laryngeal voice production [9]: A silicone one-way valve is placed into a shunt between the trachea and the esophagus. On the one hand the valve prevents aspiration and on the other hand it deviates the air stream into the upper esophagus during expiration. This air causes vibrations of the mucosa of the pharyngo-esophageal segment (PE segment), which serves as a substitute sound generator.

Alaryngeal substitute speech is characterized by high perturbation and low prosody and modulation. It presents high jitter and shimmer, low fundamental frequency, short maximum phonation time, bad availability and a different ratio of voiced to voiceless phonation in comparison with normal speech [2, 5, 6, 7, 8, 11, 14, 16]. Sometimes it is deteriorated by additional noise in consequence of insufficient closure of the tracheostoma. All these aspects cause a decreased intelligibility, which is suspected to be an important factor for psychosocial and communicative restrictions of the laryngectomee [10].

In clinical practice, subjective and objective diagnostic tools for the description and evaluation of laryngeal speech, e.g., stroboscopy, rating instruments such as the RBH or the GRBAS scale, and technical perturbation measurements proved to be inappropriate for the examination of alaryngeal speech because of its highly pathologic acoustic properties. Presently, there is no consensus on which measures to use for the description and evaluation of alaryngeal voices in laryngectomized patients.

In this paper, we present a new technical procedure for the measurement and evaluation of alaryngeal speech and compare the results obtained with subjective ratings of a panel of expert listeners.

Material and methods

Automatic speech recognition system

For objective measurement, an automatic speech recognition system was applied, a state of the art word recognition system developed at the Chair for Pattern Recognition at the University of Erlangen. In this study, the latest version as described in detail by Stemmer [13] was used. A commercial version of this recognizer is used for conversational dialogue systems (www.sympalog.com). In a commercial environment, the recognizer can handle spontaneous speech with mid-sized vocabularies of up to 10,000 words.

A speech recognizer converts spoken speech into a set of features that is representative for the language in a first acoustic analysis. It compares spectral and temporal characteristics of the speech signal with an internal dictionary that is given by acoustic speech samples with according transliteration. A recognizer "knows" all common phonemes of the language and also morphosyntactic rules. The recognition of phonemes is supported by semi-continuous Hidden Markov Models (HMM). They describe the likelihood of an analyzed acoustic signal to be identical with a phoneme. The current recognizer works in a monophone-based manner, which means that the acoustic characteristics of one phoneme are only represented by one HMM independently from co-articulatory modulations of the phoneme. Usually, there is more than one basic model for each phoneme as we also register coarticulatory effects. In most recognizers, the phoneme's predecessor and successor are taken into account, which leads to the so-called triphone models. However, we used a monophone-based approach and favored a context-independent monophone model over a context-dependent triphone model because context-independent models are more robust towards the strong deviation of tracheoesophageal speech in comparison with normal laryngeal speech, as shown in preliminary experiments. But this also means that a monophone-based recognizer shows worse results for good voices.

Due to the strong deviation of substitute speech quality to normal speech, we used a so-called unigram language model to weight the outcome of each word model [4]. Thus, the frequency of occurrence for each word in the used text was known to the recognizer. Most speech recognizers use a bigram or trigram language model containing information about the occurrence of word pairs or sequences of three words, respectively. This helps to enhance recognition results by including linguistic information. However, for this purpose it was necessary to put more weight on the recognition on acoustic features as we wanted to evaluate substitute voices. Therefore, we restricted the linguistic information to the unigram language model.

The system had been trained with acoustic information from dialogues of the VERBMOBIL project [15]; the ISADORA system was used, and the recognition was done by the program lr_beam. Normal adult speakers from all over Germany served as the training population and thus covered most dialectal regions. All speakers were, however, asked to speak "standard" German. Ninety percent of the training population (304 males, 274 females) were younger than 40 years. For the VERBMOBIL German data, we used 27 h of speech (11,714 utterances, 257,810 words) for training. Although the system had been trained with adult speakers, it has also been successfully applied in speech recognition for children [12].

With the recognizer we calculated the so called "word accuracy" (WA) of the tracheoesophageal recordings. WA is a standard measurement to evaluate recognizers and shows how much a recognized word chain deviates from the spoken utterance. It is calculated with the following formula (formula 1):

$$\mathrm{WA}\ [\%] = 100^*(\mathrm{NC} - \mathrm{NW})/\mathrm{N} \qquad (1)$$

(where NC is the number of correctly recognized words, NW the number of wrongly inserted words and N the number of all spoken words).

Thus, if the speaker said "Now it was the sun's turn" and the output of the recognizer is "New its was her the sun's turn," then the WA is 50%. The sentence consists of seven words (N = 7), four words were correctly recognized (NC = 4), while one word was recognized as wrongly inserted (NW = 1). Using formula 1, WA is calculated as:

$$\mathrm{WA}\ [\%\ ] = 100^*(4 - 1)/6 = 50.$$

Patients

Acoustic files were recorded from 18 male laryngectomees aged 64.2 ± 8.3 years with tracheoesophageal substitute speech. Informed consent had been obtained by all participants prior to the examination. At the time of the examination, the patients had been using a Provox voice prosthesis device for between 5 and 136 months (63.2 ± 35.7 months). Fourteen had undergone total laryngectomy because of laryngeal cancer, and four because of hypopharyngeal cancer. All patients were native German speakers using a local dialect.

Speech samples

The participants read out a standard German text "Nordwind und Sonne," a fable from Aesop that is known as "north wind and the sun" in the Anglo-American language area. The German text consists of

108 words (71 disjunctive) and 172 syllables. It is phonetically balanced and includes all possible phonemes of the German language. For "normal" speakers it takes 43 s on average to read the text loudly, i.e., four syllables per second. The speech samples were recorded with a close-talk microphone (dnt Call 4U Comfort-Headset) at a sampling frequency of 16 kHz and quantized with 16 bit. Eighteen male laryngeal speakers (65.4 ± 7.6 years old) without laryngeal diseases and with normal voice served as a control group.

Subjective evaluation of substitute voice

A panel of five voice professionals subjectively estimated the intelligibility of the substitute speech of each patient while listening to a play-back of the recordings of the "Nordwind und Sonne" text. A five-point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low and 5 = very low) was applied to rate the intelligibility of all individual samples. The experts were asked not to take normal laryngeal speech into consideration when judging the intelligibility of substitute speech in order to use the total range from 1 to 5.

Analysis and evaluation of the data

Statistical analysis was performed using Microsoft Excel and scripts written in the Perl programming language. For the agreement computations between different raters on the one hand and raters/recognizer on the other hand, not Cohen's "basic" kappa, but the weighted multi-rater kappa by Davies and Fleiss [1] was used. It allows the comparison of an arbitrary number of raters and weights the difference between the values of intelligibility or WA, respectively.

When comparing the ratings of the human experts and a speech recognition system, several problems occur. First of all, the human ratings were made on a Likert scale while the word accuracy is a continuous measure within a completely different range. As the kappa value can also be applied only on discrete data, a mapping of the word accuracy to the Likert scale had to be defined. We rounded the experts' average intelligibility scores to the next integer and set thresholds on the WA results, so that the difference between the experts' scores and the scores derived from the WA values was minimal (0 in the particular case). The segmentation of the WA range was then made as follows: Word accuracies smaller than zero got a score of 5 (but this case did not occur in the data), results below 15% got a score of 4, and the next interval boundaries were 25 and 40%. Thus, for a score of 1 more than 40% WA was necessary.

Results

The total duration of the laryngectomee's audio files was 21 min, consisting of 1,980 words. In addition to the words of the text, 36 additional words were produced and recognized as reading errors. The vocabulary of the word recognizer contained all words occurring in the test data (71 unique words of the text, and 32 additional words representing the reading errors).

The duration of the reading by the laryngectomees was 2.81 ± 0.76 syllables/s and differed significantly from the control group with 3.54 ± 0.55 syllables/s ($P < 0.01$).

The recordings showed a wide range in intelligibility (see Fig. 1). The recognizer's evaluation of word accuracy WA is shown in Table 1, with a significant difference between laryngectomees and the control group ($P < 0.05$).

Subjective speech evaluation showed good consistency (see Table 2). The lowest correlation value between a rater and the mean of the other four raters was 0.68, the highest 0.85.

The results for the correlations of the WA and the subjective speech evaluation are shown in Table 3. Considering the average of the raters, the WA of the recognizer has a significant correlation of –0.84 ($P < 0.01$), as shown in Fig. 1. The coefficient is negative because high recognition rates came from "good" voices with a low score number and vice versa (note the inverse y-axis in Figs. 1 and 2). The multi-rater kappa (achieved by [1]) for the group of the five raters was 0.44. The kappa for the rater group vs. the monophone-based recognizer is 0.43, i.e., the agreement between the human raters and the machine and the agreement among the humans alone can be regarded as identical (note that a result greater than 0.4 is said to represent fair agreement beyond chance [3]). Figure 2 shows the scores of human raters (average and rounded) and the machine and the applied thresholds: 12 results were identical and 6 results differed only by a grade of 1.

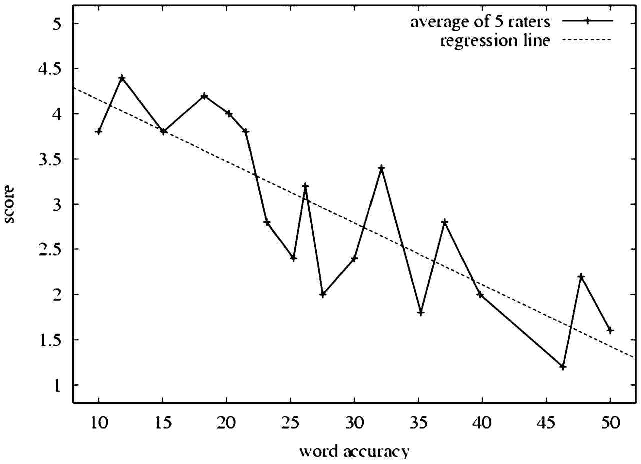

Fig. 1 Word accuracy (*WA*) versus the average of the five experts' estimation of intelligibility of 18 male laryngectomees with TE voice with corresponding *regression line*. The laryngectomees are ordered w.r.t. increasing WA. It is clearly visible that there's a strong correlation (−0.84) between the human and the automatic evaluation results

Table 1 Word accuracy (WA) analyzed by a speech recognizer of 18 male laryngectomees and control group of 18 male persons without diseases of the voice

	Mean ± SD	Minimum	Maximum
Laryngectomees $n = 18$	28.7 ± 12.1	10.0	50.0
Control group $n = 18$	57.6 ± 6.1	46.8	71.6

Discussion

Until now, no objective method of determining global speech restoration outcome after laryngectomy has existed. Here, we present a new automatic objective measurement of speech quality: the recognition of spoken words, i.e., word accuracy, by an automatic speech recognizer. First results for severely disturbed speech are shown.

Today, automatic voice recognition is used in many domains: for professional and private use as dictating machines, in call centers when a restricted vocabulary and ‘‘normal’’ voice quality and speech without background noise can be expected and in the support of handicapped persons. Nevertheless, the technique often doesn’t qualify for higher requirements such as dictation of professional reports with low speech quality. Furthermore, background noise has an effect on speech recognition systems: without background noise some dictating systems with a vocabulary of 60,000 words and more recognize about 95% (1 of 20 words is not correctly identified), whereas with background noise such as in a driving car the rate of recognized words could diminish considerably. Further research is done to enhance the possibilities of automatic speech recognition. It can also be of special interest as applied in diagnostics, e.g. for speech disturbances.

Speech recognition depends on five factors: the speaker, the speech (read speech and spontaneous speech), the vocabulary, the grammatical complexity or perplexity (average probability of words possibly following a sequence of others) and the input medium [4]. The influence of most of these factors can be minimized when using a standard text and stable recording setting as practiced in this study for diagnostic application. Thus, the speaker remains the only influencing factor.

Table 2 Inter-rater correlations between five experts (K, L, R, S and U) judging the intelligibility of 18 recordings of laryngectomees with tracheoesophageal speech. ‘‘All’’ means the average of the remaining four raters that is compared to each single rater (lower row)

Rater	K	L	R	S	U
K		+0.60	+0.82	+0.70	+0.23
L			+0.53	+0.77	+0.89
R				+0.66	+0.29
S					+0.46
All (mean)	+0.83	+0.82	+0.77	+0.85	+0.68

Table 3 Correlation coefficient between the objective speech evaluations by the monophone-based recognizer concerning word accuracy (WA) and the subjective evaluations by five experts concerning ‘‘intelligibility.’’ The automatic speech evaluations mostly agree with subjective evaluation by experts. Negative correlation coefficients result of the opposite scales for intelligibility (1 = high intelligibility to 5 = low intelligibility) and word accuracy in %

Rater	Correlation coefficient rater versus recognizer
K	−0.81
L	−0.65
R	−0.81
S	−0.79
U	−0.55
All (mean)	−0.84

We examined speech samples of 18 male laryngectomized speakers. The reading duration of the standard text between laryngeal speakers and laryngectomees showed typical differences. This is consistent with former descriptions of substitute speech characteristics [6]. For this study, we applied a non-adapted speech recognizer for automatic speech evaluation that has previously been proven to be adequate for ‘‘normal’’ speech samples. The automatic speech evaluations were compared to the results of subjective evaluation and to 18 speakers without speech pathology. Increased age has been shown to have a decreasing influence on automatic speech recognition [17]. Therefore, a control group of similar age was chosen. In our study, normal laryngeal speakers at the same age as the laryngectomees reached a word accuracy of up to 71.6%. This value seems relatively low compared to other applications because we used the monophone-based recognizer and unigram language model as described before. But even with

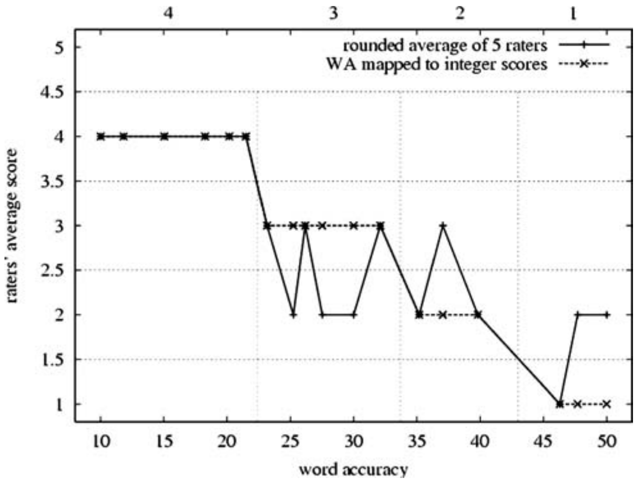

Fig. 2 Scores derived from WA versus the rounded average of the five experts’ scores for 18 laryngectomees with TE speech. The laryngectomees are ordered w.r.t. increasing WA. The recognizer’s results and the experts’ scores are the same for 12 of 18 laryngectomees and differed only by one grade for six laryngectomees

polyphone-based recognizers, results of normal speakers don't reach 100% of WA. Our preliminary experiments showed a mean WA of 84% for normal young speakers in opposition to a mean of 69% with the monophone-based recognizer. As alaryngeal speech has many characteristics that inhibit good word recognition, such as hoarseness and phonematic alterations, the WA of alaryngeal speakers is significantly worse than the WA of laryngeal speakers. With the here applied monophone-based system, however, some laryngectomees with low voice quality achieved a remarkably high quota of recognized words, although they spoke in a dialectally altered way and were by far older than the speakers of the training population. The 18 laryngectomees all used the same substitute speech and were all equipped with the same type of voice prosthesis device. Nevertheless, all examinations showed considerable variation of the results referring to individual speech outcome. So, in spite of low speech quality, the presented system allows for sufficient discrimination between different speech qualities of laryngectomees, albeit the system had only been trained with normal speakers and doesn't dispose of any special information on disturbed voices. However, further improvements of automatic recognition of "special" speech might be reached by interpolating the applied recognizer with data from other recognizers [12, 13].

Until now, only recordings of 18 laryngectomees have been evaluated, but there is every indication that the method can yield valuable information in evaluating substitute voices on an expert level when taking into account the conformity with experts' estimations of intelligibility.

Word accuracy is not similar, but akin to intelligibility. Both are influenced by voice quality, phonematic and morpho-syntactic structure, background noise, amplitudes and speaking velocity. Intelligibility includes also the "human factor." Even if one does not understand every word or syllable of a spoken sequence, the meaning can be understood by extrapolating from contextual, pragmatic and prosodic characteristics. Nevertheless, word accuracy shows good consistency to subjective estimation of intelligibility. Though word accuracy is only one aspect of speech quality, it obviously represents one major aim of speech restoration after loss of the larynx, i.e., the intelligibility.

Although the comparison of automatic evaluation with subjective evaluation by means of Likert scales is practical, it shows some restrictions. The discrepancy between the automatic speech recognizer and the experts' estimations, especially in the scales 2, 3 and 4 (rather high, medium, rather low intelligibility), arises from a reduced discriminatory power of Likert scales. The experts' estimation is demonstrated on a linear ordinate in Fig. 1, although the relation of distances between the values 1 to 5 could not be determined, i.e., the distance between "rather low" and "medium" is commonly not half the distance of "very low" to "medium." We would expect a bigger distance between the outer values (1 or 5) to their adjacent value than between the inner values (2, 3 and 4). Another characteristic of Likert scales is the reduced use of extreme scales, here 1 or 5. Both features of Likert scale estimation might be responsible for restricted compliance of subjective and automatic evaluation of some recordings.

For the German language, the mentioned automatic speech recognition system is shown to be a valuable means of quantifying laryngectomees' global speech quality. This ought to be alike in other languages. For clinical application, we currently replace the models for the reading errors by background models representing all the words (reading errors) outside of the 71 words of the text. Further automatic speech evaluation has to be done in order to get standard values of tracheoesophageal speech.

As substitute speech differs essentially from "normal" laryngeal speech and therefore classifications such as high, moderate or low intelligibility of substitute speech do not confirm with a classification of laryngeal speech, an extra scale should be applied when judging the quality. A classification into high intelligibility/low intelligibility or high quality/low quality of substitute speech could be deviated from experts' estimations, but should take Likert scale characteristics into account.

Conclusion

Automatic speech evaluation after laryngectomy by a speech recognizer is a valuable means for research and clinical purposes in order to determine the global speech outcome. It enables the quantification of the quality of speech, also in severely disturbed voices. It can easily be been transposed into other languages and could probably also be used for the evaluation of other speech and voice disorders.

Acknowledgments This work was partially supported by the EU in the project PF-Star under grant IST-2001–37599 and by the DFG (Deutsche Forschungsgemeinschaft, German Research Council), SFB 603, subproject B5, and the Deutsche Krebshilfe (registration no. 106266). The authors are responsible for the content of this article. We thank PD Dr. A. Pfahlberg, Institute for Medical Informatics, Biometry and Epidemiology, University of Erlangen, for the helpful suggestions, and Prof. Dr. H. Iro, Department of ENT, University of Erlangen, for supplying the data of the control group.

References

1. Davies M, Fleiss JL (1982) Measuring agreement for multinomial data. Biometrics 38:1047–1051
2. Debruyne F, Delaere P, Wouters J, Uwents JP (1994) Acoustic analysis of tracheo-oesophageal versus oesophageal speech. J Laryngol Otol 108:325–328
3. Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. John Wiley & Sons, New York
4. Gallwitz F, Niemann H, Nöth E (1999) Speech recognition—state of the art, applications, and future prospects. Wirtschaftsinformatik 41:538–547

5. Gandour J, Weinberg B (1983) Perception of intonational contrasts in alaryngeal speech. J Speech Hear Res 44:1315–1320
6. Pauloski BR (1998) Acoustic and aerodynamic characteristics of tracheoesophageal voice. In: Blom ED, Singer MI, Hamaker RC (eds) Tracheoesophageal voice restoration following total laryngectomy, PA. Singular Publishing Group Inc, San Diego London, pp 123–141
7. Pindzola RH, Cain BH (1989) Duration and frequency characteristics of tracheoesophageal speech. Ann Otol Rhinol Laryngol 98:960–964
8. Qi Y, Weinberg B (1995) Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers. J Speech Hear Res 38:536–548
9. Robbins J, Fisher HB, Blom ED, Singer MI (1984) A comparative study of normal, esophageal and tracheoesophageal speech production. J Speech Hear Disord 49:202–210
10. Schuster M, Lohscheller J, Kummer P, Hoppe U, Eysholdt U, Rosanowski F (2004) Voice handicap of laryngectomees with tracheoesophageal speech. Folia Phoniatr Logop 56:62–67
11. Searl JP, Carpenter MA (2002) Acoustic cues to the voicing feature in tracheoesophageal speech. J Speech Lang Hear Res 45:282–294
12. Steidl S, Stemmer G, Hacker C, Nöth E, Niemann H (2002) Improving children's speech recognition by HMM Interpolation with adults' speech recognizer. In: Michaelis B, Krell G (eds) Pattern recognition, 25 th DAGM Symposium, vol 2781 of lecture notes in computer science. Springer, Heidelberg New York Berlin, pp 600–607
13. Stemmer G (2005) Modeling variability in speech recognition. PhD Thesis, chair for pattern recognition. University of Erlangen-Nuremberg, Germany
14. Van As CJ, Hilgers FJM, Verdonck-de Leeuw IM, Koopmans-van Beinum FJ (1998) Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. J Voice 12:239–248
15. Wahlster W (ed) (2000) Verbmobil: Foundations of speech-to-speech translation, Springer, Berlin Heidelberg New York
16. Wiliams SE, Scanio TS, Ritterman SI (1989) Temporal and perceptual characteristics of tracheoesophageal voice. Laryngoscope 99:846–850
17. Wilpon JG, Jacobsen CN (1996) A study of speech recognition for children and the elderly. Proc. of ICASSP, pp 349–352