abteilung für **phoniatrie** und **pädaudiologie,** universitätsklinkum erlangen
lehrstuhl für **mustererkennung,** universität erlangen-nürnberg
tino haderlein · andreas maier · maria schuster · elmar nöth · ulrich eysholdt · frank rosanowski

Automatic Quantitative Assessment of Tracheoesophageal Speech

introduction

Tracheoesophageal (TE) voice is state-of-the-art in voice rehabilitation after laryngectomy [1]. Objective methods of voice and speech assessment are desirable in the field of post-laryngectomy speech therapy and its evaluation. Such methods were in the focus of this study [2]. The main difference to established objective speech and voice assessments is that text recordings were used instead of sustained vowels in order to evaluate not only voice but also speech properties.

material and methods

41 laryngectomees with TE substitute voice read the German text “Der Nordwind und die Sonne”. It is a phonetically balanced text; the English version is known as “The North Wind and the Sun”. The technical procedure of data acquisition was previously [1] described. The speech samples were evaluated on 5-point scales by 5 medical experts with intelligibility, speech effort, match of breath and sense units, and vocal tone being the focus of this poster.

Recordings of the Post-Laryngectomy Telephone Intelligibility Test (PLTT, [3]) were available from 31 test persons with TE speech. Data were recorded via telephone with a dialogue system provided by Sympalog Voice Solutions (www.sympalog.com). Naïve listeners evaluated these data (8 male and 3 female inexperienced non-medical students). For recording the PLTT, each patient got a unique sheet of paper with instructions and 22 words and 6 sentences that were randomly chosen. The first two words and the first sentence were neither used for human nor for automatic evaluation.

To the text recordings, an automatic speech recognition (ASR) system and a module for prosodic analysis [4] were applied. The system was trained with normal speech. This simulates a listener who is not influenced by prior experience with pathologic voices. The recognition system for the PLTT recordings was basically the same as for the text recordings. It was, however, trained with telephone speech and could recognize the PLTT vocabulary instead of the words of “Der Nordwind und die Sonne”.

The “prosody module” for the prosodic analysis of the text derives 95 “local” features for each processed word and 15 “global” features per recording, i.e. on the entire text. The local features are derived from information about intensity (speech energy), word and pause durations, and fundamental frequency (F0). The global features are based on jitter, shimmer, and the number and duration of voiced/unvoiced sections in the speech signal.

results

The word accuracy of the ASR system, which resembles the percentage of correctly recognized words, showed a high correlation to the average human intelligibility rating on the text samples (r=-.88). It was higher than the inter-rater correlation among the humans (r=.82) which was computed as the average of all correlations between one listener and the average of the remaining persons. Some automatically computed prosodic features were found to correlate with other human rating criteria:

feature	rating criterion	inter-rater correlation |r|	human-machine correlation |r|
average word duration	speech effort	.80	.76
	match of breath and sense units	.73	.75
error between speech energy trajectory and its regression line within one word	vocal tone	.83	.71
word accuracy	intelligibility	.82	.88

For the PLTT recordings, the inter-rater correlation among the human listeners was r=.90. The naïve listeners’ intelligibility ratings and the word accuracy reached r=.89. For the word recognition rate, which does not count words that were erroneously inserted into the recognized word sequence, the correlation was even r=.93.

conclusion

Automatic methods can be used for objective rating of substitute voice. In addition to established methods, speech properties can be evaluated.

references

[1] Brown DH, Hilgers FJM, Irish JC, Balm AJM. Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium. World J Surg 2003; 27:824-831.
[2] Haderlein T. Automatic Evaluation of Tracheoesophageal Substitute Voices. Volume 25 of Studien zur Mustererkennung. Logos Verlag, Berlin, 2007.
[3] Zenner HP. The postlaryngectomy telephone intelligibility test (PLTT). In: Herrmann IF (ed). Speech Restoration via Voice Prosthesis. Springer, Berlin, 1986, pp 148-152.
[4] Batliner A, Buckow J, Niemann H, Nöth E, Warnke V. The Prosody Module. In: Wahlster W (ed). Verbmobil: Foundations of Speech-to-Speech Translation. Springer, Berlin Heidelberg New York, 2000, pp 106-121.

acknowledgments

This work was partially funded by the German Cancer Aid (Deutsche Krebshilfe, grant 106266).