Evaluation of Tracheoesophageal Substitute Voices Using Prosodic Features

Tino Haderlein[1], *Elmar Nöth*[2], *Maria Schuster*[1], *Ulrich Eysholdt*[1] *& Frank Rosanowski*[1]

[1] Department of Phoniatrics and Pedaudiology
University of Erlangen-Nuremberg, Erlangen, Germany
[2]Chair for Pattern Recognition (Computer Science 5)
University of Erlangen-Nuremberg, Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de

Abstract

Tracheoesophageal (TE) speech is a possibility to restore the ability to speak after laryngectomy, i.e. after the removal of the larynx. TE speech often shows low audibility and intelligibility which makes it a challenge for the patients to communicate. In speech rehabilitation the patient's voice quality has to be evaluated. As no objective classification means exists until now and an automation of this procedure is desirable, we performed initial experiments for automatic evaluation using prosodic features. Our reference were scoring results for several evaluation criteria for TE speech from five experienced raters. Correlation coefficients of up to 0.84 between human and automatic rating are promising for future work.

1. Introduction

The removal of the larynx (the laryngectomy) is an operation which is mostly necessary in severe cases of cancer of the larynx. This means that the patient's voice and thus his or her main means of communication is basically destroyed. Although there were several attempts for voice restoration before, the development of the shunt valves (or "voice prostheses") by Singer and Blom [1] might have been the most important step in voice restoration surgery. This paper focuses on the automatic evaluation of tracheoesophageal (TE) substitute voices which can be established by a shunt valve.

In tracheoesophageal speech, the upper esophagus, the pharyngoesophageal (PE) segment, serves as a sound generator (see Fig. 1). The air stream from the lungs is deviated into the esophagus during expiration via a shunt between the trachea and the esophagus. In order to force the air to take its way through the shunt into the esophagus and allow voicing, the patient usually closes the tracheostoma, which is the upper end of the trachea, with a finger. Tissue vibrations of the PE segment modulate the streaming air and generate a substitute voice signal. In comparison to normal voices the quality of substitute voices is "low" [2, 3]. The change of pitch and volume is limited which causes monotone voice. Inter-cycle frequency perturbations let the voice sound hoarse [4]. This causes reduced ability of intonation or voiced-voiceless distinction [5, 6]. Another source of distortion is the incomplete closure of the tracheostoma. If the patient is not able to do this properly, loud "whistling" noises causes by eluding air may occur.

In speech therapy and rehabilitation a patient's voice has to be evaluated. An automatically computed, objective measure would be a very helpful support for this task. In our work we examine how well TE speech is processed by a speech recognition system, how the recognizer can be adapted to TE voices [7] and whether the results can be used for evaluating the quality of a TE voice automatically, i.e. whether they correlate with experts' ratings on criteria like "intelligibility" [8].

In this paper we present initial results on the use of prosodic features in order to quantify the properties of TE voices. In the VERBMOBIL project we developed prosodic features for the use of prosodic phenomena during the linguistic analysis [9]. The "prosody module" originating from this project was now applied to pathologic voices for the first time.

This paper is organized as follows: In Section 2 the test data and the human evaluation criteria are introduced. Section 3 gives an overview of the prosodic features used for the experiments. Section 4 shows which features were found to be suitable for distinguishing normal and TE voices and which of them correlate with the human ratings. Finally Section 5 gives a short outlook to future work.

2. Test data

The test files were recorded from 18 male laryngectomees (denoted as group *laryng18*) with tracheoesophageal substitute speech. Their average age was 64.2 years (standard dev. 8.3 years). They had undergone total laryngectomy because of laryngeal or hypopharyngeal cancer at least one year prior to the investigation and were provided with a Provox® shunt valve. Each person read the story of "North Wind and Sun", a phonetically rich text with 108 words (71 disjunctive) often used in speech therapy in German-speaking countries. The duration of all 18 audio files together was 21 minutes, the test persons spoke 1980 words. In addition to the words of the text 32 different additional words were produced as reading errors. The data are close-talking speech, quantized with 16 bit at 16 kHz sampling frequency. Five experienced phoniatricians and speech scientists evaluated the voices of the 18 test persons on criteria such as roughness (*"rough"*)[1], match of breath and sense units (*"breath-sense"*), distortions by insufficient occlusion of tracheostoma (*"noise"*), speech effort (*"effort"*) and intelligibility (*"intell"*). The scores given by the experts were represented by integer numbers between 1 ("very high/good") and 5 ("very low/bad"), i.e. a Likert scale [10].

The second speaker group consisted of 18 healthy men ("control group men", *kom18*) forming an age-matched group with respect to the tracheoesophageal speakers. On average they were 65.4 years (± 7.6 years) old. The 18 recordings of the "North Wind and Sun" text from this group contained 1964 words with a total duration of 15 minutes.

[1] In medical sciences a harsh voice is often called "rough", therefore we follow this convention.

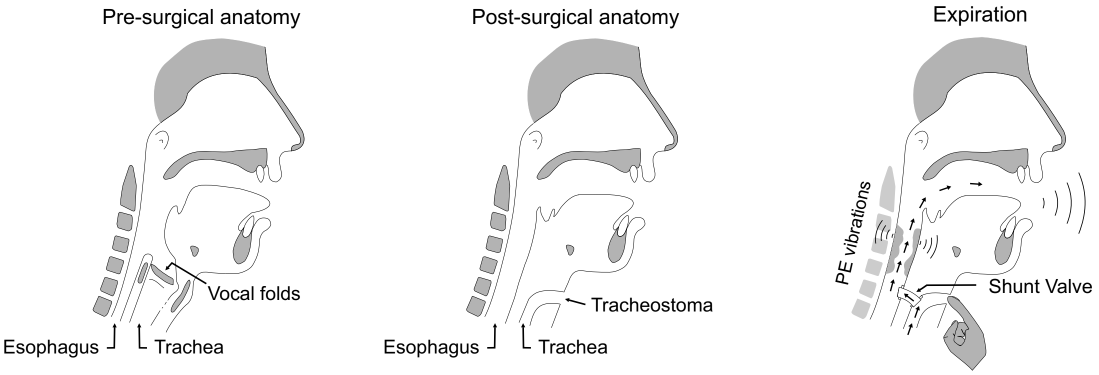

Figure 1: *Anatomy of a person with intact larynx* (left)*, anatomy after total laryngectomy* (middle)*, and the substitute voice* (right) *caused by vibration of the pharyngoesophageal segment (pictures from [11])*

3. Prosodic features

The prosody module takes the output of our word recognition module in addition to the speech signal as input. In this case the time-alignment of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody module. The speech recognizer uses semi-continuous Hidden Markov Models (HMM), monophone models and 24 MFCC-based features per 16 ms frame at a frame shift rate of 10 ms. For details please refer to [12] or [7]. The vocabulary of the recognizer for the generation of the word hypotheses graphs (WHGs) consisted of the 71 words of the "North Wind and Sun" text only.
A fixed reference point has to be chosen for the computation of the prosodic features. We decided in favor of the end of a word because the word is a well–defined unit in word recognition, it can be provided by any standard word recognizer, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point we extract 95 prosodic features over intervals of different sizes: The current word, i.e. after which the reference point is set, gets the number *0*. The interval containing only this word is denoted by "*0,0*". The interval containing the two words before word *0* is called "*-2,-1*", because it begins at word *-2* and ends at the end of word *-1*. In the same way the words after the reference point get positive numbers. The interval code is added to the name of the feature. So the feature *En:Max1,2* denotes the maximum energy value in the two words after the reference point. Table 1 shows the 28 different features and the contexts over which they are calculated for a total of 95 prosodic features. The abbreviations can be explained as follows:

- **duration features '*Dur*'**: absolute (*Abs*) and normalized (*Norm*); the normalization is described in [9]; the global value *DurTauLoc* is used to scale the mean duration values; *AbsSyl* is the absolute duration divided by the number of syllables and represents another sort of normalization;
- **energy features '*En*'**: regression coefficient (*RegCoeff*) and mean square error (*MseReg*) of the energy curve w.r.t. the regression curve; mean (*Mean*), maximum (*Max*) with its position on the time axis (*MaxPos*), absolute (*Abs*) and normalized (*Norm*) values; for the normalization see [9]; the global value *EnTauLoc* is used to scale the mean energy;
- **F_0 features '*F0*'**: regression coefficient (*RegCoeff*) and the mean square error (*MseReg*) of the F_0 curve w.r.t. the regression curve; mean (*Mean*), maximum (*Max*), minimum (*Min*), onset (*On*), and offset (*Off*) values as well as the position of *Max* (*MaxPos*), *Min* (*MinPos*), *On* (*OnPos*), and *Off* (*OffPos*) on the time axis; all F_0 values are not stored as absolute values, but as their logarithm, normalized as to the mean value *F0MeanG*;
- **length of pauses '*Pause*'**: length of silent pause before (*Pause–before*) and after (*Pause–after*) and filled pause before (*PauseFill–before*) and after (*PauseFill–after*) the respective word in context.

Fig. 2 shows examples of the F_0 features. It is obvious that there is a strong correlation between some of the 95 features, e.g., between *Dur:Norm* for context *0,0* and for context *-1,0*. In our previous work on the use of prosodic information for linguistic analysis (for instance for finding phrase boundaries) the neural net classifiers were trained with some 13,000 events. Therefore we decided to be as exhaustive as possible and use a large feature vector. Our experiments showed that it was always the best to use all features if there are enough training data available [13]. A full description of the features used is beyond the scope of this paper; details and further references are given in [9]. The features proved to be effective for linguistic and emotion analysis (see [14]), so we expected them to be sufficient for the analysis of the rating criteria used in this study.

Besides the 95 local features per word, 15 global features were computed per utterance from jitter, shimmer and the number of voiced/unvoiced (V/UV) decisions. They cover each of mean and standard deviation for jitter and shimmer, the number, length and maximum length each for voiced and unvoiced sections, the ratio of the numbers of voiced and unvoiced sections, the ratio of length of voiced sections to the length of the signal and the same for unvoiced sections. The last global feature is the standard deviation of the fundamental frequency F_0.

4. Experiments and results

In our experiments we addressed two related questions. The first one was to find features that can separate normal voices from pathologic voices, in our case TE voices. The second one was to find features that correlate with the human rating criteria

<table>
<tr><th rowspan="2">features</th><th colspan="5">context size</th></tr>
<tr><th>-2</th><th>-1</th><th>0</th><th>1</th><th>2</th></tr>
<tr><td><i>DurTauLoc; EnTauLoc; F0MeanG</i></td><td></td><td></td><td>●</td><td></td><td></td></tr>
<tr><td><i>Dur: Abs, Norm, AbsSyl</i></td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>En: RegCoeff, MseReg, Mean,</i></td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>Max, MaxPos, Abs, Norm</i></td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>F0: RegCoeff, MseReg, Mean,</i></td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>Max, MaxPos, Min, MinPos</i></td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>Pause–before, PauseFill–before</i></td><td></td><td>●</td><td>●</td><td></td><td></td></tr>
<tr><td><i>F0: Off, OffPos</i></td><td></td><td>●</td><td>●</td><td></td><td></td></tr>
<tr><td><i>Pause–after, PauseFill–after</i></td><td></td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>F0: On, OnPos</i></td><td></td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td><i>Dur: Abs, Norm, AbsSyl</i></td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td><i>En: RegCoeff, MseReg, Mean,</i></td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td><i>Abs, Norm</i></td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td><i>F0: RegCoeff, MseReg</i></td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td><i>Dur: Norm</i></td><td></td><td colspan="2">●</td><td></td><td></td></tr>
<tr><td><i>En: RegCoeff, MseReg</i></td><td></td><td colspan="2">●</td><td></td><td></td></tr>
<tr><td><i>F0: RegCoeff, MseReg</i></td><td></td><td colspan="2">●</td><td></td><td></td></tr>
</table>

Table 1: *95 local prosodic features and their context [9]*

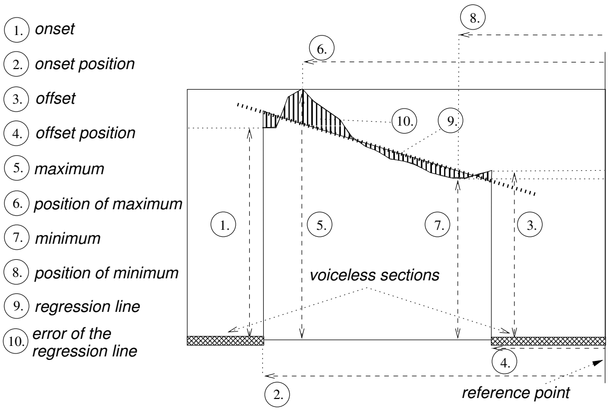

Figure 2: *Computation of prosodic features within one word (after [15])*

mentioned in Section 2.

4.1. Prosodic features on TE and laryngeal speakers

In the following the prosodic features of tracheoesophageal speakers (the *laryng18* group) and laryngeal speakers (the *kom18* group) are compared. Each one of the features computed per word or per file was reduced to its mean value for each speaker group to do a quick elimination of all features which will probably not be suitable for the distinction between laryngeal and TE speakers. Of course this is a rough reduction that does not take into account the trajectory of the features over time.

Table 2 contains the prosodic features that were on average most significantly different for both speaker groups. Due to the redundancy in different feature intervals mentioned above, for each feature only the interval with the largest difference between the speaker groups is presented. As expected, the average of the pause duration before the current word *Pause-before0,0* is much higher for TE speakers than for normal speakers (31 frames vs. 14 frames). The normalized word duration *Dur:Norm-1,0* is about four times as high for *laryng18* in comparison to *kom18*, as it does not only take into account two words, but also the pause between them. Very important are also the different F_0 measures and the global information on number and duration of unvoiced sections. The last column of Table 2 shows which features have a remarkably higher or lower mean for the substitute voices than for the normal voices. For those the quotient is clearly different from 1.

The F_0 features are suffering from the fact that it is very hard to find a periodic signal in TE speech at all. The “F_0” values themselves were not very helpful in this study. Nevertheless the binary decision whether a section is voiced or unvoiced and the number and duration of such sections is still useful for the comparison to normal voices, as the results show.

feature name	μ_{laryng18}	μ_{kom18}	$\frac{\mu_{\text{laryng18}}}{\mu_{\text{kom18}}}$
Pause-before0,0	31.20	14.06	2.22
En:RegCoeff0,0	-12.90	-5.64	2.29
En:Norm-2,-1	-0.29	-0.55	0.52
Dur:Norm-1,0	0.94	0.23	4.04
F0:Max0,0	0.33	0.15	2.27
F0:Min0,0	-0.37	-0.14	2.67
F0:OnPos1,1	32.01	17.77	1.80
Number_UV_Sections	1.71	0.74	2.30
Length_UV_Sections	8.43	4.04	2.09
Max_Length_UV_Section	6.01	3.36	1.79
StandardDeviation_F0	0.40	0.15	2.73

Table 2: *Selection of prosodic features with significant differences between laryngeal and TE speakers; the single values are averaged on all words of all speakers. Some values have been transformed by normalization and/or a logarithmic function (cmp. the feature description in Section 3).*

4.2. Prosody features in correlation with human rating

Finding the prosodic features that correlate to rating criteria of the human raters introduced in Section 2 is a statistical problem due to the high number of measures. If a recorded paragraph contains 108 words, like the “North Wind and Sun” text read by the test persons of this study, then this means that per recording $95 \cdot 108 + 15 = 10275$ features are computed. They have to be compared to one single Likert value per rating criterion given by a human rater. For the initial experiments described here a rather simple method was applied to quickly exclude the feature/score pairs probably least useful for automatic speech evaluation: First all 108 values for each single local feature in a file were averaged. This was done for the 18 signals of the *laryng18* speaker group. Then the correlation between these mean values and the rating criteria was computed. The reference score for a criterion from the entire group of 5 raters was also achieved by averaging the Likert values of the single raters (cmp. [7]).

Table 3 shows a clear correlation between the criterion “match of breath and sense units” (*breath-sense*) and some pause and duration features. A very good indicator with a correlation of $r = 0.84$ is the voice onset position in the word after the reference point *F0:OnPos1,1*. It is very likely caused by artefacts of the word recognition process. Many filled pauses with breathing noise are classified as initial unvoiced sections of the “following” word and thus result in a high onset position value. The speech effort is also well-indicated by duration values. It is intuitive that tracheostoma noise is reflected by energy measures

feature name	criterion and correlation
Pause-after0,0	*effort* -0.71; *breath-sense* +0.79
En:Norm-2,-1	*noise* -0.76
En:Abs-2,-1	*rough* -0.74
Dur:Norm-2,-1	*breath-sense* +0.71; *noise* -0.71
Dur:AbsSyl-2,-1	*effort* -0.75; *breath-sense* +0.81
F0:OnPos1,1	*effort* -0.75; *breath-sense* +0.84

Table 3: *Correlation between selected prosodic features and human ratings for TE speakers; the correlation was measured using the mean value of all words per file. Presented are criteria with a correlation of* $|r| \geq 0.7$.

as it is an additive distortion on the speech signal. This also basically holds for roughness. However, a more detailed look at their trajectory in the recording should give more information than a mean value that was computed from voiced and unvoiced sections together. The connection between *Dur:Norm-2,-1* and the *noise* criterion might have its reason in a lower speaking rate when a lot of air is getting lost through the tracheostoma.

The low number of sound files evaluated and the Likert scheme with only 5 values to choose raised the question whether the high correlation measured for several features occurred just by coincidence. Therefore the experiment was repeated twice on all the features with $|r| \geq 0.7$. In the first case the file with the experts' original ratings was replaced by random Likert numbers between 1 and 5. The average absolute correlation of the prosodic features to these random scores was 0.22 while it was 0.74 for the raters' judgments. In the second case the original ratings were replaced by a score of 3 for each criterion in order to simulate undecided raters. It is a known characteristic of Likert scales that the use of extreme scales, here 1 or 5, is reduced. The average absolute correlation for these data was 0.20. Those results confirm that the correlation values measured on the original data reveal a real connection between human ratings and machine-computed features and that the high correlation is not just caused by coincidence.

5. Conclusions and outlook

The results show that it is not only possible to distinguish normal and pathologic voices by prosodic features, but the features can also serve as automatic measures for several evaluation criteria. The correlation between human raters and the prosody module is very encouraging for further work. As the main goal of our project is to allow evaluation of substitute voices via telephone, we are currently collecting telephone speech data from laryngectomees. We will examine whether the results described for close-talking speech in this paper also hold for the new database. More work has to be done on the reduction of dimension of the prosodic feature vectors. The simple averaging must be replaced by statistical regression methods in order to keep the loss of information as small as possible.

In future work out-of-vocabulary (OOV) errors during creation of the word hypotheses graphs need a special treatment, as the vocabulary of the speech recognizer knows only the words of the reference text. Unknown words in the utterance result in alignment errors. By using confidence measures and language models the sections with reading errors can be detected in the recording. Then the remaining parts of the file will be used for the computation of the prosodic features only.

6. Acknowledgments

This work was funded by the German Cancer Aid (Deutsche Krebshilfe) under grant 106266. The responsibility for the content of this paper lies with the authors.

7. References

[1] M.I. Singer and E.D. Blom. An Endoscopic Technique for Restoration of the Voice after Laryngectomy. *Ann Otol Rhinol Laryngol*, 89:529–533, 1980.

[2] J. Robbins, H.B. Fisher, E.D. Blom, and M.I. Singer. A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production. *J Speech Hear Disord*, 49:202–210, 1984.

[3] M.H. Bellandese, J.W. Lerman, and H.R. Gilbert. An Acoustic Analysis of Excellent Female Esophageal, Tracheoesophageal, and Laryngeal Speakers. *J Speech Lang Hear Res*, 44:1315–1320, 2001.

[4] H.K. Schutte and G.J. Nieboer. Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. *Folia Phoniatr Logop*, 54:8–18, 2002.

[5] J. Gandour and B. Weinberg. Perception of Intonational Contrasts in Alaryngeal Speech. *J Speech Hear Res*, 26:142–148, 1983.

[6] J.P. Searl and M.A. Carpenter. Acoustic Cues to the Voicing Feature in Tracheoesophageal Speech. *J Speech Lang Hear Res*, 45:282–294, 2002.

[7] T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski, and M. Schuster. Automatic Recognition and Evaluation of Tracheoesophageal Speech. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proc. TSD 2004*, volume 3206 of *Lecture Notes in Artificial Intelligence*, pages 331–338, Springer, Berlin, 2004.

[8] M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski. Can You Understand Him? Let's Look at His Word Accuracy – Automatic Evaluation of Tracheoesophageal Speech. In *Proc. ICASSP*, volume I, pages 61–64, Philadelphia, PA, 2005.

[9] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 106–121. Springer, Berlin, 2000.

[10] R. Likert. A technique for the measurement of attitudes. *Archives in Psychology*, 140:1–55, 1932.

[11] J. Lohscheller. *Dynamics of the Laryngectomee Substitute Voice Production*. Shaker-Verlag, Aachen, Germany, 2003. PhD thesis.

[12] G. Stemmer. *Modeling Variability in Speech Recognition*, volume 19 of *Studien zur Mustererkennung*. Logos Verlag, Berlin, 2005.

[13] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. 14th ICPhS*, volume 3, pages 2315–2318, San Francisco, 1999.

[14] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*, 40:117–143, 2003.

[15] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker, Aachen, 1997.