Automatic evaluation of prosodic features of tracheoesophageal substitute voice

Tino Haderlein · Elmar Nöth · Hikmet Toy · Anton Batliner · Maria Schuster · Ulrich Eysholdt · Joachim Hornegger · Frank Rosanowski

Abstract In comparison with laryngeal voice, substitute voice after laryngectomy is characterized by restricted aero-acoustic properties. Until now, an objective means of prosodic differences between substitute and normal voices does not exist. In a pilot study, we applied an automatic prosody analysis module to 18 speech samples of laryngectomees (age: 64.2 ± 8.3 years) and 18 recordings of normal speakers of the same age (65.4 ± 7.6 years). Ninety-five different features per word based upon the speech energy, fundamental frequency F_0 and duration measures on words, pauses and voiced/voiceless sections were measured. These reflect aspects of loudness, pitch and articulation rate. Subjective evaluation of the 18 patients' voices was performed by a panel of five experts on the criteria "noise", "speech effort", "roughness", "intelligibility", "match of breath and sense units" and "overall quality". These ratings were compared to the automatically computed features. Several of them could be identified being twice as high for the laryngectomees compared to the normal speakers, and vice versa. Comparing the evaluation data of the human experts and the automatic rating, correlation coefficients of up to 0.84 were measured. The automatic analysis serves as a good means to objectify and quantify the global speech outcome of laryngectomees. Even better results are expected when both the computation of the features and the comparison method to the human ratings will have been revised and adapted to the special properties of the substitute voices.

Keywords Laryngectomy · Substitute speech · Automatic speech recognition

T. Haderlein · H. Toy · M. Schuster · U. Eysholdt · F. Rosanowski
Department of Phoniatrics and Pedaudiology, University of Erlangen-Nuremberg, Bohlenplatz 21, 91054 Erlangen, Germany
e-mail: Tino.Haderlein@informatik.uni-erlangen.de

E. Nöth · A. Batliner · J. Hornegger
Chair for Pattern Recognition (Computer Science 5), University of Erlangen-Nuremberg, Martensstraße 3, 91058 Erlangen, Germany

Introduction

Laryngectomy for laryngeal or hypopharyngeal cancer affects many aspects of life [26] with loss of ability for vocal communication being an outstanding stigma for the affected persons [10]. Today, voice rehabilitation with shunt valves is regarded state-of-the-art [7]. These valves allow for deviation of the air stream into the upper esophagus during expiration. Tissue vibrations of the pharyngoesophageal (PE) segment modulate the streaming air and generate a substitute voice signal used as source for the tracheoesophageal (TE) voice [20]. Of course, in comparison to normal voices, the quality of substitute voices is "low" [6, 25] with a loss of prosodic features being a particular characteristic: The voice sounds monotonous due to the limited change of pitch and intensity; inter-cycle frequency perturbations let the voice sound hoarse [28]. This leads to reduced intonation and voiced-voiceless distinction [13, 29]. Another source of distortion may result from incomplete closure of the tracheostoma. If the patient is not able to do this properly, loud "whistling" noise from the eluding air may occur.

Both for clinical and scientific purposes, a patient's voice has to be evaluated in a quantitative manner. However, as the evaluation is done by human raters today, it may be biased and time-consuming, i.e. expensive. An automatically computed, objective measure would be helpful since it provides a solution for these two problems:

Costs would be reduced and the problem of inter- and intra-rater variability would be eliminated completely, because an automated evaluation algorithm always yields the same result for one specific speech recording. Thus, the global speech restoration outcome after laryngectomy could be evaluated independently of the rater or the rater's experience with substitute voices. Until now, only little data exist on how to objectively evaluate substitute voice.

In earlier experiments we concentrated on the criterion of speech intelligibility [27]. The speakers' recordings were processed by an automatic speech recognition system. The correlation between the word accuracy, which is a measure for the number of correctly recognized words, and the raters' evaluation of intelligibility was computed to $r = -0.84$.

In order to find automatically computable counterparts for subjective rating criteria, we use a "prosody module" to compute features based upon frequency, duration and speech energy (intensity) measures. This is state-of-the-art in automatic speech analysis on normal voices [1, 4, 8, 12, 23, 30].

Prosodic information is modulated onto speech segments, like syllables, words, phrases, and whole utterances. These segments we assign perceptual properties like pitch, loudness, articulation rate, voice quality, duration, pause or rhythm. In general, there is no unique feature in the speech signal corresponding to them exactly, but there are features which highly correlate with them; examples are the fundamental frequency (F_0) which correlates to pitch, and the signal energy correlating to loudness.

The topics of our new experiments were two related questions: find features that

1. separate normal voices from pathologic voices, i.e. TE voices,
2. correlate with the human rating criteria.

These topics form the basis for the development of an objective and automatic speech evaluation procedure.

Material and methods

Patients and control group

Test files were recorded from 18 male laryngectomees with tracheoesophageal substitute speech (the "TE group"). Their average age was 64.2 ± 8.3 years. They had undergone total laryngectomy because of laryngeal or hypopharyngeal cancer at least 1 year prior to the investigation and had been provided with a Provox® shunt valve. At the time of investigation, none of the test persons suffered from recurrent tumor growth or metastases. Each person read the text "Nordwind und Sonne", a phonetically rich text with 108 words (71 disjunctive) and 172 syllables used in medical speech evaluation in German-speaking countries. It is known as "The North Wind and the Sun" in the Anglo-American language area and is also used for speech evaluation in other languages [16, 22]. The duration of all 18 audio files together was 21 min, the test persons spoke 1980 words. In addition to the words of the text, 32 different additional words were produced as reading errors. The control group consisted of 18 healthy men (abbreviated as "C group") forming an age-matched group with respect to the TE speakers. On average they were 65.4 ± 7.6 years old. The 18 "Nordwind und Sonne" recordings from this group contained 1964 words with a total duration of 15 min. Twenty-two different words were uttered as reading errors. All data were recorded with a close-talk microphone (dnt Call 4U Comfort headset; DNT GmbH, 63128 Dietzenbach, Germany), digitized with 16 bit at 16 kHz sampling frequency.

Subjective evaluation of substitute voices

A panel of five experienced voice professionals subjectively evaluated the substitute speech of each patient from the TE group while listening to a play-back of the recordings of the "Nordwind und Sonne" text. Rating criteria relevant for the purpose of this study were roughness (in Table 3 denoted by "rough"), distortions by insufficient occlusion of tracheostoma (noise), speech effort (effort), intelligibility (intell), the overall quality (overall), and finally the match of breath and sense units (breath-sense), i.e. whether the patient had to breath within a sentence. The speech samples were played to the experts once via loudspeakers in a quiet seminar room without disturbing noise or echoes. After each recording the raters had time to mark their impressions on a preprinted evaluation sheet.

Rating was performed on a five-point Likert scale (1 = very high, 2 = rather high, 3 = medium, 4 = rather low, 5 = very low) for each of the criteria except for the overall quality which was rated on a visual analog scale. This scale had a width of 10 cm; the label at its left end said "very bad", the label at its right end was "very good". The distance of the expert's marking to the left end served as numerical input data for the further experiments. Possible values were between 0.0 and 10.0 cm. The experts were asked not to take normal laryngeal speech into consideration when judging the substitute speech, but TE voices in general in order to use the total range of the Likert scales and the analog scale, respectively.

Prosodic features

The prosody module requires two files as input. The first one is the speech recording itself. The second one is a so-called word hypotheses graph (WHG) which is the output of a speech recognition module and contains the information where each spoken words begins and ends in the recording. The latter, however, can only be computed when a word-wise

transliteration of the recording is available, i.e. a text file containing the word sequence that was spoken by the test person. In this study, it was provided for each speaker by a computer scientist experienced in speech recognition (TH). The speech recognition module is based upon Hidden Markov Models (HMM), an approach that defines a statistical model for each phoneme to be recognized [2, 17, 18]. These models contain information about which frequencies usually occur during the production of the respective phoneme. Coarticulatory effects can be considered in the models. In previous experiments, however, we found out that this may have a negative effect in the case of the highly pathological substitute voices [15]. Therefore, we defined monophone models only. The recognition of phonemes is done by cutting the speech file into frames of 16 ms length. The frequency portions in such a section are summed up in intervals equally spaced on an auditory-based mel scale. The final features are then achieved by a discrete cosine transform; these measures are known as mel-frequency cepstrum coefficients (MFCC, [9]). Twelve of them and their respective first derivative form a 24-dimensional feature vector that is used for the phoneme recognition [27, 31]. The recognized phonemes are connected to form words according to a given vocabulary list. The vocabulary of the recognition system for the generation of the WHGs consisted of the 71 words of the "North Wind and the Sun" text. The speech recognition system and the prosody module could be also applied to any other text in any other language.

For the computation of the prosodic features, a fixed reference point is chosen after each word provided by the word recognizer (see Fig. 1). For each reference point, we extract 95 prosodic features in intervals of different size: 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. For instance, 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 in which they are calculated for a total of 95 prosodic features. Figure 1 shows examples of features computed from the fundamental frequency F_0. The absolute F_0 values, however, are not applicable for prosodic analysis. In order to take into account the logarithmic scale of human perception, a logarithmic normalization is done and the overall mean value is subtracted. The F_0 values that are further processed are thus small numbers close to 0. For similar reasons, also the energy features are normalized. The word duration measures are normalized with respect to articulation rate.

Besides the 95 local features per word, 15 global features are computed for the entire utterance derived from jitter (fluctuations of F_0), shimmer (fluctuations of intensity)

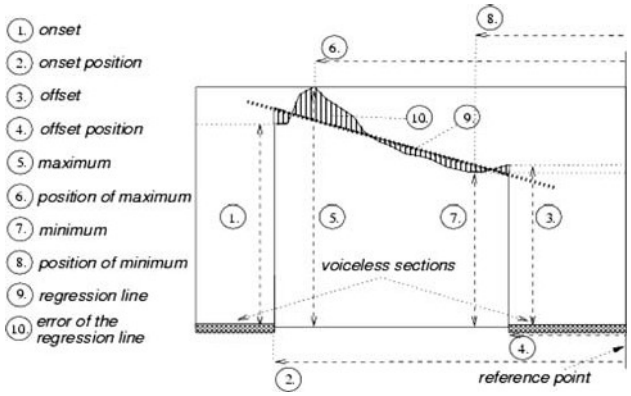

Fig. 1 Computation of prosodic features from the F_0 trajectory within one word (according to [19]); the abscissa shows the time, the ordinate represents the detected frequency value. The word begins and ends with an unvoiced section where no F_0 was detected. All durations are measured from the reference point at the end of the word

and the number of detected voiced and unvoiced (V/UV) sections in the speech signal. Among them are 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 F_0. Details and further references of the features are given in [4].

The 95 local features are computed for each word of the spoken text. The human experts, however, gave ratings not for each word, but for the entire text. In order to receive one single value for each feature that can be compared to the human ratings, the average of each word-based feature served as final feature value. The overall average of the respective feature across all speakers of the TE group was then compared to the respective value of the C group in order to find significant differences between the groups.

For the computation of the correlation between a prosodic feature and a human rating criterion, the average prosodic feature value of the TE group was compared to the average score given by the raters (cf. [15]). The experiment was then repeated twice on all the features that had reached a correlation of $|r| \geq 0.7$. In the first case, the experts' original ratings were replaced by random numbers between 1 and 5. In the second case, the original ratings were replaced by a score of 3 for each criterion in order to simulate undecided raters.

Results

Prosodic features on TE and laryngeal speakers

Table 2 displays the average values $\mu_{\text{TE group}}$ and $\mu_{\text{C group}}$ for the prosodic features with the largest differences between

Table 1 Ninety-five word-based prosodic features, based upon duration (Dur), energy (En) and fundamental frequency (F_0) measures

Features	Context size				
	−2	−1	0	1	2
DurTauLoc; EnTauLoc; F0MeanG			•		
Dur: Abs, Norm, AbsSyl		•	•	•	
En: RegCoeff, MseReg, Mean, Max, MaxPos, Abs, Norm		•	•	•	
F0: RegCoeff, MseReg, Mean, Max, MaxPos, Min, MinPos		•	•	•	
Pause-before, PauseFill-before		•	•		
F0: Off, OffPos		•	•		
Pause-after, PauseFill-after			•	•	
F0: On, OnPos			•	•	
Dur: Abs, Norm, AbsSyl	•				•
En: RegCoeff, MseReg, Mean, Abs, Norm	•				•
F0: RegCoeff, MseReg	•				•
Dur: Norm		•			
En: RegCoeff, MseReg		•			
F0: RegCoeff, MseReg		•			

The context size denotes the interval of words on which the features are computed, e.g. a circle in column '0' means "computed on current word"; a bullet between column '−2' and '−1' means "computed in the interval that contains the second and first word before the current word and the pause between them" (cf. [4]). The features are abbreviated as follows:

Global normalizing factors: DurTauLoc is used to scale the duration values, EnTauLoc scales the energy values, and F0MeanG is the F_0 mean value for the entire file. For the details of the normalization see [4]

Duration features "Dur": absolute (Abs) and normalized (Norm) word duration; the global value 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 within a word w.r.t. the regression curve; mean (Mean) and maximum energy (Max) with its position on the time axis (MaxPos); absolute (Abs) and normalized (Norm) energy values

F_0 features "F0": regression coefficient (RegCoeff) and the mean square error (MseReg) of the F_0 curve w.r.t. its regression curve; mean (Mean), maximum (Max), minimum (Min), voice 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 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

both speaker groups. The average of the pause duration before the current word (Pause-before0,0) is much higher for TE speakers than for normal speakers (318 vs. 147 ms). The normalized word duration Dur:Norm-2,-1 is about four times as high for the TE group as for the control group. The normalized F_0 measures and the information on number and duration of unvoiced sections also show large differences between the speaker groups.

Prosodic features in correlation with human rating

Table 3 shows the correlation between the human rating criteria and the automatically computed features for features that had reached $|r| \geq 0.7$. The absolute average correlation for these features was $|r| = 0.74$.

The information in the score for the criterion "match of breath and sense units" (breath-sense) corresponds best with some pause and duration features. The highest correlation of $r = 0.84$ is achieved with the voice onset position in the word after the reference point, F0:OnPos1,1. Speech effort is indicated best by duration values.

When the experiment was repeated with random scores between 1 and 5 instead of the experts' ratings, the average correlation of the prosodic features to these random scores was only 0.22. When the original ratings were replaced by a constant score of 3 for each criterion, the average absolute correlation was 0.20.

Discussion

Until now, there is no generally accepted objective method for the evaluation of speech restoration outcome after laryngectomy. Here, we present an automatic objective measurement of clinically valid speech quality criteria based upon prosodic features. It is achieved by the analysis of running speech rather than sustained vowels like other approaches for measuring laryngeal voice quality [11, 14,

Table 2 Selection of prosodic features with largest differences between laryngeal (C group) and TE speakers (TE group); values marked with an asterisk are normalized

Feature name	$\mu_{\text{TE group}}$	$\mu_{\text{C group}}$
Pause-before0,0	318 ms	147 ms
En:RegCoeff0,0	−12.90	−5.64
En:Norm-2,-1*	−0.29	−0.55
Dur:Norm-2,-1*	0.92	0.23
F0:Max0,0*	0.33	0.15
F0:Min0,0*	−0.37	−0.14
F0:OnPos1,1	326 ms	184 ms
Number_UV_Sections	1.71	0.74
Length_UV_Sections	90 ms	46 ms
Max_Length_UV_Section	66 ms	40 ms
StandardDeviation_F0*	0.40	0.15

In order to take into account the logarithmic scale of human perception, a logarithmic normalization is done and the overall mean value is subtracted from the F_0 values. The marked duration measures are normalized w.r.t. articulation rate. The measures for unvoiced (UV) sections are given per word

Table 3 Correlation between selected prosodic features and human ratings for TE speakers (TE group); the correlation was measured using the mean value of all words per file for the respective feature

Feature name	Criterion and correlation
Pause-after0,0	effort −0.71; breath-sense +0.79; overall +0.76
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; overall +0.72
Dur:Abs-2,-1	effort −0.76; breath-sense +0.82; overall +0.78
F0:OnPos1,1	effort −0.75; breath-sense +0.84; overall +0.77

Presented are criteria with a correlation of $|r| \geq 0.7$

24, 32]. Thus it is possible to evaluate the patient's voice and speech at the same time. A similar approach was introduced in [21], but the text used there was shorter: it contained 18 words only. Correlations to human ratings are only given for the "overall impression" of the substitute voice, and they do not exceed $r = 0.49$. The features were based upon measures like jitter, voicing duration or pause durations similar to our study.

The German version of the text "North Wind and the Sun" contains 108 words with 172 syllables. The total duration of the entire text was 70.9 ± 23.1 s for the average TE speaker and 52.2 ± 8.1 s for a normal speaker. The average articulation rate for the TE group was 2.81 ± 0.76 syllables per second, for the control group it was 3.54 ± 0.55 syllables per second. Due to the high standard deviation, these measures alone cannot be reliable distinctive features.

Our "prosody module", which had been applied to normal speakers before, was now applied to pathologic, alaryngeal voices for the first time. The goal was to identify characteristics of substitute voices related to duration measures, signal energy or F_0 that can be described by single and easily extractable features. Currently, we are using a set of 95 prosodic features per word. The features proved to be effective for linguistic and emotion analysis [4, 5, 23], so we expected them to be sufficient for the analysis of the rating criteria used in this study. It is obvious that there is a strong correlation between some of the features, e.g., between the "duration of the current word" (Dur:Norm0,0) and the "duration of the two previous words" (Dur:Norm-2,-1). Due to this redundancy in different intervals for a respective feature, only one interval is presented for each feature in Table 2. In earlier experiments on normal speech, we applied automatic feature selection methods to choose the best features for the respective problem [3]. In the current experiments, this selection is done on the basis of the correlation values between automatically computed features and the human ratings.

In addition to the 95 features for each one of the 108 words of the "North Wind and Sun" text, 15 global features for the entire text are computed resulting in a total number of 10,275 features per recording. For a quick elimination of all features which will probably not be suitable for the distinction between laryngeal and TE speakers, each one of the local features was averaged across all words in a recording. This is a rough reduction that does not take into account the time-dependent changes. The averaging is only possible if the patients' speech is restricted to a standard text. This means that all patients were supposed to produce the same phonemes and words, so the differences among them concerning the measured duration, pitch and loudness features are mainly caused by their individual degree of the speech impairment. Therefore, we regard this procedure justified for a pilot study. The dimension reduction for identifying the features that correlate best with the human rating criteria for the TE group was done in the same way. The features were compared to one single average value per rating criterion obtained from all raters. Thus the feature-score pairs probably least useful for automatic speech evaluation were quickly excluded.

The F_0 features are influenced by the difficulties in finding a periodic signal in TE speech at all. The algorithm for the detection of the fundamental frequency does a voiced-unvoiced (V/UV) decision first. On all 16 ms speech frames that were classified as voiced, the program performs pitch detection. Because of the high degree of aperiodicity, the algorithm often makes octave errors, i.e. it finds the double, triple or the half of the actual F_0 value. This explains why the extreme values (F0:Max0,0, F0:Min0,0) and the standard deviation differ so much from the normal speakers. Since the detected values are not reliable, they cannot be used for the detailed evaluation of a speaker's voice pathology.

However, restricting to the binary voiced-unvoiced decision, i.e. the information whether a section contains a quasi-periodic signal or not, is very useful for the comparison between normal and TE voices. The number and duration of these sections are thus a valid and important component of the objective speech analysis (Table 2).

For the intelligibility criterion, no correlation to prosodic features above $|r| = 0.7$ was found. For this criterion, the word accuracy of the recognition module is by far better ($r = -0.84$; see [27]). The overall quality score corresponds well to several of the duration measures which shows how important speech fluency is for a human rater's impression of voice quality. This cannot be covered by the analysis of a sustained vowel.

The normalized word duration Dur:Norm-2,-1 shows the effect of the combination of two different measures. Taking into account durations of two words and the pause between them explains the large difference for this feature between the speaker groups. Such a combination is also helpful for the human evaluation criterion "match of breath and sense units" (breath-sense) which correlates very well with several pause and duration features, as Table 3 shows. The best indicator in our experiments with a correlation of $r = 0.84$ is the voice onset position in the word after the reference point, F0:OnPos1,1. It includes the duration of the pause after the current word (Pause-after0,0) and the length of the first unvoiced section in the following word. The longer the pause and the higher the degree of aperiodicity, the higher the voice onset position value. This leads to a good correlation for this particular feature to the breath-sense and the speech effort criteria, but we expect better results for other duration and F_0 features when the speech recognition module will be retrained with a small amount of TE speech [15, 31] for a better distinction of breath and unvoiced speech. However, this will also have a disadvantage. Up to now, we use a recognition system trained on "normal" speech samples which resembles a human listener who has not heard this kind of voices before and can thus evaluate them "objectively". Integrating knowledge about TE speech into the system might mean the loss of a certain degree of this objectivity.

We also plan to detect reading errors and to exclude them from the evaluation automatically. In the current study, the reading error rate was 3.4% for the TE group and 2.8% for the control group which means that the average reader misreads about 3 of 100 words.

It is obvious that tracheostoma noise is reflected by energy measures as it is represented by additive distortions on the speech signal. This also basically holds for roughness. However, a more detailed look at the run of their trajectory in the recording should give more information than a mean value that was computed from voiced and unvoiced sections together. So, this will be subject of future work as well. The connection between the normalized word duration in the two words before the reference point (Dur:Norm-2,-1) and the noise criterion is not completely clear yet; the reason is very likely the lower articulation rate when a lot of air is getting lost through the tracheostoma. Future experiments on a bigger data set will give the answer. The number of sound files available for this study was rather low, so the statistical validity of the results had to be verified. Therefore, the experiment for the man-machine correlation was repeated on all the features that had reached $|r| \geq 0.7$. When we replaced the original experts' ratings by random scores, we achieved a much lower correlation ($r = 0.22$). Thus we showed that the correlation between the objective automatic measures and the real raters' data are not a product of coincidence, but indeed reveal the decisions of the human experts expressed in the rating criteria.

When we replaced the experts' ratings by constant marks of "3" for all criteria to simulate an inexperienced and thus undecided rater, the correlation was very low again at $r = 0.20$. For the correlation on the original data, this means that the raters' experience is expressed by the automatic measures. Therefore, the outcome of these experiments is that we found an objective evaluation method with inherent expert knowledge.

The data obtained in this study allow for the following conclusions:

1. It is possible to distinguish normal and pathologic voices by prosodic features automatically.
2. Automatically computed prosodic features can serve as measures for human evaluation criteria describing the quality of a substitute voice.
3. The correlation between human raters and the automatic prosody analysis is far beyond chance.

This is the first work where speech quality and suprasegmental properties of pathologic speech were evaluated automatically and where the results highly correlated with judgment from human experts for several different rating criteria.

Further work will include the evaluation of laryngectomees' telephone speech to test the feasibility of a fully remote evaluation procedure. Furthermore, as the current analysis programs were developed for laryngeal speech and proved to be successful on substitute voices, they might also be used for the evaluation of other speech and voice disorders. Current work examines the use of other techniques than averaging for dimension reduction.

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

References

1. Ananthakrishnan S, Narayanan S (2005) An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), Philadelphia, PA, vol I, pp 269–272
2. Baker JK (1975) The DRAGON system—an overview. IEEE Trans Acoust Speech Signal Process 23:24–29
3. Batliner A, Buckow J, Huber R, Warnke V, Nöth E, Niemann H (1999) Prosodic feature evaluation: brute force or well designed? In: Proceedings of the 14th international congress of phonetic sciences, San Francisco, vol 3, pp 2315–2318
4. Batliner A, Buckow J, Niemann H, Nöth E, Warnke V (2000) The prosody module. In: Wahlster W (ed) Verbmobil: foundations of speech-to-speech translation. Springer, Heidelberg, pp 106–121
5. Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2003) How to find trouble in communication. Speech Commun 40:117–143
6. Bellandese MH, Lerman JW, Gilbert HR (2001) An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers. J Speech Lang Hear Res 44:1315–1320
7. Brown DH, Hilgers FJM, Irish JC, Balm AJM (2003) Postlaryngectomy voice rehabilitation: state of the art at the millennium. World J Surg 27:824–831
8. Chen K, Hasegawa-Johnson M, Cohen A, Borys S, Kim S-S, Cole J, Choi J-Y (2006) Prosody dependent speech recognition on radio news corpus of American English. IEEE Trans Audio Speech Lang Process 14:232–245
9. Davis SB, Mermelstein P (1980) Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28:357–366
10. Devins GM, Stam HJ, Koopmans JP (1994) Psychosocial impact of laryngectomy mediated by perceived stigma and illness intrusiveness. Can J Psychiatry 39:608–616
11. Fröhlich M, Michaelis D, Strube HW, Kruse E (2000) Acoustic voice analysis by means of the hoarseness diagram. J Speech Lang Hear Res 43:706–720
12. Gallwitz F, Niemann H, Nöth E, Warnke V (2002) Integrated recognition of words and prosodic phrase boundaries. Speech Commun 36:81–95
13. Gandour J, Weinberg B (1983) Perception of intonational contrasts in alaryngeal speech. J Speech Hear Res 26:142–148
14. van Gogh CDL, Festen JM, Verdonck-de Leeuw IM, Parker AJ, Traissac L, Cheesman AD, Mahieu HF (2005) Acoustical analysis of tracheoesophageal voice. Speech Commun 47:160–168
15. Haderlein T, Steidl S, Nöth E, Rosanowski F, Schuster M (2004) Automatic recognition and evaluation of tracheoesophageal speech. In: Sojka P, Kopeček I, Pala K (eds) Proceedings of the 7th international conference on text, speech and dialogue (TSD 2004). Lecture notes in artificial intelligence, vol 3206. Springer, Heidelberg, pp 331–338
16. International Phonetic Association (1999) Handbook of the International Phonetic Association. Cambridge University Press, London
17. Jelinek F, Bahl LR, Mercer RL (1975) Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Trans Inf Theory IT-21:250–256
18. Junqua J-C (2000) Robust speech recognition in embedded systems and PC applications. Kluwer, Boston
19. Kießling A (1997) Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. PhD Thesis, Berichte aus der Informatik. Shaker, Aachen
20. Lohscheller J (2003) Dynamics of the laryngectomee substitute voice production. PhD Thesis, Kommunikationsstörungen—Berichte aus Phoniatrie und Pädaudiologie, vol 14. Shaker, Aachen
21. Moerman M, Pieters G, Martens JP, van der Borgt MJ, Dejonckere P (2004) Objective evaluation of the quality of substitution voices. Eur Arch Otorhinolaryngol 261:541–547
22. Nishio M, Niimi S (2006) Comparison of speaking rate, articulation rate and alternating motion rate in dysarthric speakers. Folia Phoniatr Logop 58:114–131
23. Nöth E, Batliner A, Kießling A, Kompe R, Niemann H (2000) Verbmobil: the use of prosody in the linguistic components of a speech understanding system. IEEE Trans Speech Audio Process 8:519–532
24. Reilly RB, Moran R, Lacy P (2004) Voice pathology assessment based on a dialogue system and speech analysis. In: Bickmore T (ed) Proceedings of the AAAI fall symposium on dialogue systems for health communication, Washington DC, pp 104–109
25. Robbins J, Fisher HB, Blom ED, Singer MI (1984) A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. J Speech Hear Disord 49:202–210
26. Schuster M, Kummer P, Eysholdt U, Rosanowski F (2003) Quality of life in laryngectomees after prosthetic voice restoration. Folia Phoniatr Logop 55:211–219
27. Schuster M, Haderlein T, Nöth E, Lohscheller J, Eysholdt U, Rosanowski F (2006) Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 263:188–193
28. Schutte HK, Nieboer GJ (2002) Aerodynamics of esophageal voice production with and without a Groningen voice prosthesis. Folia Phoniatr Logop, 54:8–18
29. Searl JP, Carpenter MA (2002) Acoustic cues to the voicing feature in tracheoesophageal speech. J Speech Lang Hear Res 45:282–294
30. Shriberg E, Stolcke A (2004) Direct modeling of prosody: an overview of applications in automatic speech processing. In: Proceedings of the international conference on speech prosody, Nara, Japan, pp 575–582
31. Stemmer G (2005) Modeling variability in speech recognition. PhD Thesis, Studien zur Mustererkennung, vol 19. Logos Verlag, Berlin
32. Wokurek W, Pützer M (2003) Automated corpus based spectral measurement of voice quality parameters. In: Proceedings of the international congress of phonetic sciences (ICPhS). International Phonetic Association (IPA), Barcelona, pp 2173–2176