Voice and Speech Assessment From Telephone Recordings Using Prosodic Analysis Based on μ-Law-Companded Features

Tino Haderlein[1], Anne Schützenberger[2], Michael Döllinger[2], Elmar Nöth[1]

[1]Lehrstuhl für Informatik 5 (Mustererkennung), Friedrich-Alexander-Universität Erlangen-Nürnberg, Martensstraße 3, 91058 Erlangen, Germany
E-mail: {Tino.Haderlein,Elmar.Noeth}@fau.de
Web: www5.cs.fau.de

[2]Phoniatrische und pädaudiologische Abteilung in der HNO-Klinik, Klinikum der Universität Erlangen-Nürnberg, Bohlenplatz 21, 91054 Erlangen, Germany
E-mail: {Anne.Schuetzenberger,Michael.Doellinger}@uk-erlangen.de
Web: www.hno-klinik.uk-erlangen.de/phoniatrie

Abstract

Objective assessment of voice and speech properties via telephone is desirable for rehabilitation purposes. 82 patients after partial laryngectomy read a standardized text on the phone. Five experienced raters assessed speech effort, match of breath and sense units, vocal tone, intelligibility, and overall voice quality perceptually based on these recordings. Objective evaluation was performed by the word accuracy and word correctness of a speech recognition system, and a set of prosodic features. The speech recognition system used μ-law features, i. e. modified Mel-Frequency Cepstrum Coefficients (MFCCs). The prosodic features were computed based on word hypotheses graphs produced by the speech recognizer. The human-machine correlation between these features and the perceptual evaluation show slightly better results for the system based on μ-law features than for the baseline MFCC system.

1 Introduction

Perceptual voice and speech evaluation for clinical and scientific purposes is biased and time-consuming. Automatically computed, objective measures help to reduce costs, and the problem of inter- and intra-rater variability is eliminated. In this way, it can be used as objective assessment method in voice and speech rehabilitation therapy. Available software usually evaluates isolated voice properties but not speech aspects [1]. However, the necessity for the analysis of more complex speech elements than vowels, especially for criteria like speech intelligibility or prosodic aspects, has been pointed out in the literature [2–4].

Prosodic analysis is widely used in automatic speech analysis on normal voices [5–8]. It can be used to assess voice and speech disorders as well [9, 10]. Prosodic measures were also applied to telephone speech of partially laryngectomized persons [11]. The telephone is a crucial part of social life. Voice and speech patients are often elderly persons who need a means of communication that does not require them to leave their home. Due to the band-limitation of the telephone channel, however, the voice is deteriorated even more, and no support for communication by facial or hand gestures is available. Hence, voice evaluation over a telephone reflects a situation of communication which is important for the patient. Objective rating of telephone speech as a part of clinical voice rehabilitation would be a step towards a global evaluation of deteriorated voice and speech. This would also be very comfortable for the affected persons, since they do not have to travel to the clinics just for an evaluation of their vocal abilities.

In this study, we evaluated voice and speech of partially laryngectomized persons via the telephone. We modified the prosodic analysis introduced for telephone speech in [12]: another type of features was computed in the underlying speech recognition system that has been proven successful for speech recognition in low signal quality [13]. This paper is organized as follows: Section 2 introduces the speech samples used for the experiments. The speech recognition system and the features computed in the speech recognizer will be described in Sect. 3. The prosody module and the prosodic features will follow in Sect. 4. The results will be discussed in Sect. 5.

2 Test Data and Subjective Evaluation

82 persons (68 men, 14 women) were recorded after partial laryngectomy due to laryngeal cancer. Their average age was 62.3 with a standard deviation of 8.8 years; the youngest speaker was 41, the oldest one was 86 years old. Informed consent had been obtained prior to the examination. The study respected the principles of the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects, and it has been approved by the ethics committee of our university. All persons read the German version of the tale 'The North Wind and the Sun' [14], which is widely used in medical speech evaluation in German-speaking and other countries. It consists of 71 distinct words and 108 words in total (172 syllables). The patients were recorded via a landline telephone, i. e. the frequency band was reduced to the interval between 300 and 3400 Hz. Other degradations, e. g. due to ambient noise, were avoided.

The automation of clinical evaluation methods requires a human evaluation reference. For this reason, four female speech therapists and one male ear-nose-throat physician listened to the samples in an evaluation session. An excerpt of an in-house evaluation sheet (Table 1) with clinically relevant voice and speech criteria was used for this purpose. The abbreviations for the voice criteria 'speech effort' *(effort)*, 'vocal tone' *(tone)*, and 'overall voice quality score' *(overall)* as well as for the speech criteria 'match of breath and sense units' *(brsense)* and 'overall intelligibility' *(intell)* will be used throughout this paper. The former four criteria were rated on a 5-point Likert scale, i. e. one out of 5 named alternatives had to be chosen. For automatic analysis and the computation of an average perceptual value among all raters, the scores had to be converted to integer numbers. These were not printed on the evaluation sheet. The overall voice quality score was not Likert-based: A gray bar with a width of 10 cm was printed

(1)	(2)	(3)	(4)	(5)
speech effort *(effort)*				
very high	high	moderate	low	none
match of breath and sense units *(brsense)*				
very good	good	moderate	low	none
vocal tone *(tone)*				
very pleasant	pleasant	moderate	un-pleasant	very un-pleasant
overall intelligibility *(intell)*				
very high	high	moderate	low	none
overall quality score *(overall)*				
very good				very bad

Table 1: Schematic diagram of the evaluation sheet; the Likert scales for the rating criteria were transformed to integer numbers (first line, not printed on the original sheet). The overall quality score was marked graphically in a box of width 10 cm and measured by hand. The abbreviations of the criteria (in italics) were also not printed on the sheet.

on the sheet. The raters were asked to mark their impression of the overall voice quality by a vertical line on this visual analog scale (VAS) without regarding their results for the criteria before. The distance of the drawn line from the left boundary was measured by hand with a precision of 0.1 cm and used as the value of the quality score, so possible values for this criterion were between 0.0 and 10.0.

3 The Speech Recognition Systems

3.1 Recognizers

The speech recognition system used for this study is based on semi-continuous Hidden Markov Models which define a statistical model for each different phoneme to be recognized. The basic acoustic features for the recognition are Mel-Frequency Cepstrum Coefficients (MFCCs) [15]. For each 16 ms frame, a 24-dimensional feature vector is computed. It contains short-time energy, 11 MFCCs, and the first-order derivatives of these 12 static features. The phoneme models are context-dependent. They take into account coarticulation effects and train different models for one core phone in different phone contexts. We use special polyphone models [16] where the context can be chosen arbitrarily large. The basic training set for the acoustic phone models for this study were downsampled broadband data recorded in the VERBMOBIL project [13, 17]. The 578 speakers (274 women, 304 men) were normal speakers from all over Germany. In this way a normal voice was defined as the reference for automatic evaluation. The average age of these persons was 27 years. About 80% of them were between 20 and 29 years old, less than 10% were over 40. This is significantly younger than the test speakers. However, an equivalent amount of elderly speech for training was not available. In order to be able to recognize telephone speech, we resampled the original 16 kHz data with 8 kHz and applied a band-pass filter (300 to 3400 Hz). This simulates telephone speech quality. The recognition vocabulary of the recognizer was changed to the 71 words of the standard text 'The North Wind and the Sun'.

The word accuracy (WA) between the recognized and the reference word sequence is usually used as the basic measure for the evaluation of a recognition system. If the number of words in the reference is denoted by n_{all} and the number of substituted (n_{sub}), inserted (n_{ins}), deleted (n_{del}), and correctly recognized words (n_{corr}) are also known, then the word accuracy in percent is computed as

$$\mathrm{WA} = 100 \cdot \left(1 - \frac{n_{\text{sub}} + n_{\text{del}} + n_{\text{ins}}}{n_{\text{all}}}\right) \quad . \qquad (1)$$

A related measure, the word correctness (WR), omits n_{ins}. Although both measures are usually given in percent, a high n_{ins} can cause the WA to become negative.

In order to reduce the computational complexity of recognition, a language model of possible speech input is usually added as another source of information. It contains probabilities about word sequences in natural language and can eliminate many errors from the pure acoustic recognition phase. However, for automatic assessment of intelligibility, this is a disadvantage. The more errors are corrected by using linguistic knowledge, the worse match human and automatic evaluation [18]. This makes WA and WR useless as measures for intelligibility, for instance. For this reason, our recognizer used only a unigram language model, i. e. the frequency of occurrence of single words in the text reference was known to the recognizer.

The baseline recognizer using MFCC features will be denoted as *base*. Another recognizer employed modified features. These will be introduced in the following section.

3.2 μ-Law Features

One step in the computation of MFCCs is applying a logarithm to compress the Mel-filtered spectrum coefficients. This can be replaced by the μ-law (also 'mu-law' or 'u-law') coding that is usually used for data compression in telecommunications in order to achieve histogram equalization and a better signal-to-noise ratio:

$$f(x) = \text{sign}\, x \cdot \frac{\log(1 + \mu|x|/x_{\max})}{\log(1 + \mu)} \qquad (2)$$

When logarithmic compression is used, low values below 1 are set to a minimum threshold. The μ-law coding attenuates this problem. It 'compands' the input, i. e. it raises low values and compresses high values; the compression is even stronger than by a logarithmic function. A similar idea has also been used within the RASTA methodology [19, 20]. In our recognizer, $x_{\max}$ is set to 1 because an energy normalization precedes the companding step.

For the features, the factor $\mu = 10^5$ was chosen according to findings in [13, p. 89]. The respective recognizer will be denoted as *mu5*. Like the *base* recognizer, it was trained with downsampled close-talking speech. It was also polyphone-based and used a unigram language model.

4 Prosodic Features

In order to find automatically computable counterparts for the perceptual rating criteria, also a 'prosody module' was used to compute features based upon frequency, duration, and speech energy (intensity) measures. This is common in automatic speech analysis on normal voices [8, 21, 22]. The prosody module usually processes the speech signal itself and the output of the word recognition module (*base* and *mu5*). In this study, however, the boundaries between words were obtained by forced alignment with the original text as reference since the number of reading errors

features	context	
	WPW	W
Pause: before, Fill-before, after, Fill-after		•
En: RegCoeff, MseReg, Abs, Norm, Mean	•	•
En: Max, MaxPos		•
Dur: Abs, Norm	•	•
F0: RegCoeff, MseReg	•	•
F0: Mean, Max, MaxPos, Min, MinPos, Off, OffPos, On, OnPos		• •

Table 2: Local prosodic features; the context size denotes the interval of words on which the features are computed (W: one word, WPW: word-pause-word interval).

was negligible. 'Local' prosodic features are computed for each word position. Originally, there were 95 of them. After several studies on voice and speech assessment, however, a relevant core set of 33 features has been defined for further processing [23]. The components of their abbreviated names are given in parentheses:

- Length of pauses (Pause): length of silent pause before (–before) and after (–after), and filled pause before (Fill-before) and after (Fill-after) the respective word
- Energy features (En): regression coefficient (RegCoeff) and the mean square error (MseReg) of the energy curve with respect to the regression curve; mean (Mean) and maximum energy (Max) with its position on the time axis (MaxPos); absolute (Abs) and normalized (Norm) energy values
- Duration features (Dur): absolute (Abs) and normalized (Norm) duration
- F_0 features (F0): regression coefficient (RegCoeff) and mean square error (MseReg) of the F_0 curve with respect to 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.

The last part of the feature name denotes the context size, i. e. the interval of words on which the features are computed (see Table 2). They can be computed on the current word (W) or in the interval that contains the second and first word before the current word and the pause between them (WPW). A full description of the features used is beyond the scope of this paper; for details see [6, 24].

Besides the 33 local features per word, 15 'global' features were computed for intervals of 15 words length each. They were derived from jitter, shimmer, and the number of detected voiced and unvoiced sections in the speech signal [6]. They are summarized in Table 3.

Since all patients read the same text, the range of prosodic feature values among them was supposed to indicate the degree of voice or speech pathology. The human listeners gave ratings for the entire text. In order to receive also one single value for each feature that could be compared to the human ratings, the average of each prosodic feature over the entire recording served as final feature value.

5 Results and Discussion

Table 4 shows the average human evaluation values for the voice and speech criteria and the absolute WA and WR values for both recognizers *base* and *mu5*. For *mu5*, these values are slightly worse, except for the minimal WR. For the purpose of this study, however, not the recognition rates are crucial but their correlation to the perceptual results.

The inter-rater agreement, measured for a rater as the correlation of this person to the average of the other four raters, was $r = 0.86$ for *effort*, $r = 0.72$ for *brsense*, $r = 0.85$ for *tone*, $r = 0.84$ for *intell*, and $r = 0.89$ for *overall*. An automatically obtained feature or combination of features can be regarded as reliable as a human rater when its correlation to the human average rating reaches this value. Especially *brsense* is obviously not easy to judge for the human raters. This has to be kept in mind when looking at the human-machine correlations.

The results in Table 5 show the human-machine correlation between perceptual and automatic evaluation for WA, WR, and the single local and global prosodic features.

feature	description
StandDevF0	global standard deviation of F_0
MeanJitter	mean jitter in all voiced sections
StandDevJitter	standard deviation of jitter in all voiced sections
MeanShimmer	mean shimmer in all voiced sections
StandDevShimmer	standard deviation of shimmer in all voiced sections
#+Voiced	number of voiced sections
#–Voiced	number of unvoiced sections
Dur+Voiced	duration of voiced sections (in frames)
Dur–Voiced	duration of unvoiced sections (in frames)
DurMax+Voiced	maximum duration of voiced section
DurMax–Voiced	maximum duration of unvoiced section
RelNum+/–Voiced	ratio of number of voiced and unvoiced sections
RelDur+/–Voiced	ratio of duration of voiced and unvoiced sections
RelDur+Voiced/Sig	ratio of duration of voiced sections and duration of signal
RelDur–Voiced/Sig	ratio of duration of unvoiced sections and duration of signal

Table 3: The 15 global prosodic features

measure	unit	mean	st.dev.	min	max
effort	points	2.66	1.21	1.00	4.80
brsense	points	3.28	0.86	1.40	4.80
tone	points	4.30	0.61	2.20	5.00
intell	points	3.25	1.10	1.20	5.00
overall	VAS	6.39	2.30	1.90	9.54
WA (*base*)	%	47.0	19.6	–2.7	79.6
WR (*base*)	%	53.9	17.0	8.6	83.3
WA (*mu5*)	%	46.7	19.2	–4.5	79.1
WR (*mu5*)	%	53.1	16.6	12.1	81.8

Table 4: Subjective and objective evaluation results

criterion	*effort*		*brsense*		*tone*		*intell*		*overall*	
recognizer	*base*	*mu5*	*base*	*mu5*	*base*	*mu5*	*base*	*mu5*	*base*	*mu5*
WA	0.62	**0.63**	–0.66	**–0.67**	–0.58	–0.58	–0.69	–0.69	–0.63	–0.62
WR	0.65	0.65	–0.67	–0.66	–0.65	**–0.66**	–0.75	–0.75	–0.67	–0.67
Pause–beforeW	–0.56	**–0.57**	0.63	**0.64**	0.46	**0.47**	0.57	**0.59**	0.52	**0.54**
DurNormWPW	–0.60	**–0.61**	0.67	0.66	0.52	**0.53**	0.64	**0.65**	0.57	**0.59**
DurAbsWPW	–0.49	–0.48	0.53	0.53	0.37	0.36	0.46	0.45	0.43	0.42
EnNormWPW	–0.49	**–0.51**	0.57	**0.58**	0.40	**0.42**	0.50	**0.53**	0.44	**0.46**
EnMaxW	0.34	0.34	–0.21	–0.21	–0.37	–0.37	–0.42	–0.41	–0.40	–0.40
EnMeanWPW	0.33	0.33	–0.21	–0.20	–0.36	–0.36	–0.41	–0.40	–0.39	–0.39
EnMeanW	0.33	0.33	–0.20	–0.21	–0.35	–0.35	–0.40	–0.40	–0.39	–0.39
MeanShimmer	0.53	**0.57**	–0.46	**–0.47**	–0.47	**–0.51**	–0.60	**–0.61**	–0.51	**–0.53**
StandDevShimmer	0.51	**0.53**	–0.41	–0.41	–0.48	**–0.49**	–0.61	–0.61	–0.54	**–0.55**
#+Voiced	0.40	**0.43**	–0.33	**–0.35**	–0.37	**–0.39**	–0.50	**–0.52**	–0.41	**–0.44**
#–Voiced	–0.55	–0.52	0.46	0.40	0.59	0.56	0.58	0.55	0.63	0.59
Dur+Voiced	0.41	**0.43**	–0.28	**–0.30**	–0.43	**–0.45**	–0.51	**–0.53**	–0.48	**–0.50**
Dur–Voiced	–0.60	–0.59	0.49	0.47	0.55	0.55	0.66	0.65	0.61	0.60
DurMax+Voiced	0.42	**0.44**	–0.29	**–0.31**	–0.45	**–0.46**	–0.51	**–0.53**	–0.50	**–0.51**
DurMax–Voiced	–0.59	–0.59	0.48	0.47	0.54	0.54	0.65	0.65	0.60	0.60
RelNum+/–Voiced	0.30	0.30	–0.26	–0.26	–0.27	–0.26	–0.40	–0.39	–0.32	–0.31
RelDur+Voiced/Sig	0.58	0.57	–0.47	–0.45	–0.57	–0.56	–0.66	–0.66	–0.62	–0.61
RelDur–Voiced/Sig	–0.58	–0.57	0.47	0.45	0.57	0.56	0.66	0.66	0.62	0.61

Table 5: Human-machine correlation between automatically computed features (WA and WR, local and global prosodic features) and the average perceptual rating; only features with $|r| \geq 0.4$ for one of the rating criteria are depicted. Cases where the μ-law features are better than MFCCs are indicated in bold face.

Only features reaching $|r| \geq 0.4$ for at least one rating criterion are mentioned in the table. A full discussion of all results is beyond the scope of this paper, so we will discuss only selected results that are related to former studies or show good results especially in this particular task.

While WA and WR are well-known good indicators for all of the perceptual criteria, the used μ-law features could only marginally improve the human-machine correlation when only the speech recognition results were considered. After all, they show consistently good results on the same level as MFCCs without any outliers.

For the prosodic features, it is apparent that the performance is better on duration-based measures when the underlying speech recognizer worked with μ-law features. Certain noise in the speech signal or in pauses between words might affect the speech recognizer but not a human listener. The companding function may attenuate this effect for the automatic analysis which agrees better with the perceptual results then. The improvements are small but consistent among the rating criteria, and they can be noticed both in the local and in the global features. Also for the normalized energy in a word-pause-word interval (EnNormWPW), it is an advantage to use the μ-law features in the recognizer. Again, the reason may be the different weighting of noise or signal parts. Other prosodic energy-based features do not benefit from this, but in general the correlation values are on the same level as for the MFCC-based system. When MeanShimmer is calculated using word hypotheses graphs based on μ-law features for speech recognition, *effort* (*base*: $r = 0.53$, *mu5*: $r = 0.57$) and *tone* (*base*: $r = –0.47$, *mu5*: $r = –0.51$) show the largest rise in correlation. However, this is also not significant.

The results might have been negatively influenced by the signal quality of the telephone transmission and the fact that the training data of the recognizers were just downsampled and not real telephone speech. The mismatch in the age of training and test speakers [25] is an aspect that must also be considered. However, this applies mainly to WA and WR. The prosodic analysis was less affected since it was not based on the recognition result but on forced alignment with the reference text. We have shown on similar data of partially laryngectomized persons that for the average patient a transcription of the recorded sample is not necessary because the reading errors have no significant negative effect on the prosodic analysis, at least not for the assessment of intelligibility [26].

Single features do not reach as high correlations to humans as humans among themselves, but the results clearly identified the most promising measures for voice and speech assessment. In the next step, all prosodic features, WA, and WR will be combined as input for Support Vector Regression (SVR), and the best feature set based on MFCC- and μ-law-based recognizers will be determined. For the intelligibility of close-talking and telephone recordings, this was proven successful for MFCC-based recognition [12]. The human-machine correlation for telephone recordings rose from $|r| = 0.75$ for WR alone (see also Table 5) to $r = 0.86$ for a set of four prosodic features and WR. This set comprised a modified global standard deviation of the F_0, the standard deviation of jitter in all voiced sections (StandDevJitter), the ratio of the duration of all voiced sections and the duration of the signal (RelDur+Voiced/Sig), and the silent pause before a word (Pause–beforeW). The latter two also appeared among the best single features in this study. We are optimistic that significant improvement can also be reached for μ-law features and the other rating criteria that have been examined.

Acknowledgments

Dr. Döllinger's contribution was supported by Deutsche Krebshilfe grant no. 111332.

References

[1] T. Dubuisson, T. Dutoit, B. Gosselin, and M. Remacle, “On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination,” *Journal on Advances in Signal Processing; Analysis and Signal Processing of Oesophageal and Pathological Voices*, vol. 2009, 2009. 19 pages.

[2] M. Vasilakis and Y. Stylianou, “Voice pathology detection based eon [sic] short-term jitter estimations in running speech,” *Folia Phoniatr Logop*, vol. 61, pp. 153–170, 2009.

[3] K. Umapathy, S. Krishnan, V. Parsa, and D. Jamieson, “Discrimination of pathological voices using a time-frequency approach,” *IEEE Trans Biomed Eng*, vol. 52, pp. 421–430, 2005.

[4] B. Halberstam, “Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels,” *ORL J Otorhinolaryngol Relat Spec*, vol. 66, pp. 70–73, 2004.

[5] S. Ananthakrishnan and S. Narayanan, “An Automatic Prosody Recognizer Using a Coupled Multi-Stream Acoustic Model and a Syntactic-Prosodic Language Model,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. I, (Philadelphia, PA), pp. 269–272, 2005.

[6] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, “The Prosody Module,” in Wahlster [17], pp. 106–121.

[7] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, “Prosody dependent speech recognition on radio news corpus of American English,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 232–245, 2006.

[8] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, “VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 519–532, 2000.

[9] A. Maier, *Speech of Children with Cleft Lip and Palate: Automatic Assessment*, vol. 29 of *Studien zur Mustererkennung*. Berlin: Logos Verlag, 2009.

[10] T. Haderlein, E. Nöth, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski, “Automatic Evaluation of Prosodic Features of Tracheoesophageal Substitute Voice,” *Eur Arch Otorhinolaryngol*, vol. 264, no. 11, pp. 1315–1321, 2007.

[11] T. Haderlein, K. Riedhammer, E. Nöth, H. Toy, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski, “Application of Automatic Speech Recognition to Quantitative Assessment of Tracheoesophageal Speech in Different Signal Quality,” *Folia Phoniatr Logop*, vol. 61, no. 1, pp. 12–17, 2009.

[12] T. Haderlein, E. Nöth, A. Batliner, U. Eysholdt, and F. Rosanowski, “Automatic Intelligibility Assessment of Pathologic Speech over the Telephone,” *Logoped Phoniatr Vocol*, vol. 36, no. 4, pp. 175–181, 2011.

[13] T. Haderlein, *Automatic Evaluation of Tracheoesophageal Substitute Voices*, vol. 25 of *Studien zur Mustererkennung*. Berlin: Logos Verlag, 2007.

[14] International Phonetic Association (IPA), “Handbook of the International Phonetic Association.” Cambridge University Press, Cambridge, 1999.

[15] S. Davis and P. Mermelstein, “Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. on Acoustics, Speech, and Signal Processing (ASSP)*, vol. 28, no. 4, pp. 357–366, 1980.

[16] E. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, “Automatic Speech Recognition without Phonemes,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, vol. 1, (Berlin), pp. 129–132, 1993.

[17] W. Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer, 2000.

[18] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, and M. Schuster, “Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010. 7 pages. Published Aug. 20, 2009.

[19] J. Köhler, N. Morgan, H. Hermansky, H. Hirsch, and G. Tong, “Integrating RASTA-PLP into Speech Recognition,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, (Adelaide, Australia), pp. 421–424, 1994.

[20] N. Morgan and H. Hermansky, “RASTA Extensions: Robustness to Additive and Convolutional Noise,” in *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, (Cannes, France), pp. 115–118, 1992.

[21] A. Rosenberg, *Automatic Detection and Classification of Prosodic Events*. PhD thesis, Columbia University, New York, 2009.

[22] A. Origlia and I. Alfano, “Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification,” in *Proc. 8th Int. Conf. on Language Resources and Evaluation (LREC’12)* (N. Calzolari *et al.*, eds.), pp. 997–1002, 2012.

[23] T. Haderlein, C. Schwemmle, M. Döllinger, V. Matoušek, M. Ptok, and E. Nöth, “Automatic Evaluation of Voice Quality Using Text-based Laryngograph Measurements and Prosodic Analysis,” *Comput Math Methods Med*, vol. 2015, 2015. 11 pages. Published June 2, 2015.

[24] T. Haderlein, C. Moers, B. Möbius, F. Rosanowski, and E. Nöth, “Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation,” in *Proc. TSD 2011* (I. Habernal and V. Matoušek, eds.), vol. 6836 of *LNAI*, (Berlin, Heidelberg), pp. 195–202, Springer, 2011.

[25] M. Eskenazi and A. Black, “A Study on Speech Over Telephone and Aging,” in *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, vol. 1, (Aalborg, Denmark), pp. 171–174, 2001.

[26] T. Haderlein, E. Nöth, A. Maier, M. Schuster, and F. Rosanowski, “Influence of Reading Errors on the Text-Based Automatic Evaluation of Pathologic Voices,” in *Proc. TSD 2008* (P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds.), vol. 5246 of *LNAI*, (Berlin, Heidelberg), pp. 325–332, Springer, 2008.