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

Tino Haderlein1, Anne Schützenberger2, Michael Döllinger2, Elmar Nöth1

1Lehrstuhl 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
2Phoniatri sche 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/phoniatrics

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 used in this study. The speech recognition system used $\mu$-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 $\mu$-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 bandwidth 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, 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...
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 measured as the value of the quality score, so possible values for this criterion were between 0.0 and 10.0.

### 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.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>speech effort (effort)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very high</td>
<td>high</td>
<td>moderate</td>
<td>low</td>
<td>none</td>
</tr>
<tr>
<td><strong>match of breath and sense units (brsense)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very good</td>
<td>good</td>
<td>moderate</td>
<td>low</td>
<td>none</td>
</tr>
<tr>
<td><strong>vocal tone (tone)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very pleasant</td>
<td>pleasant</td>
<td>moderate</td>
<td>unpleasant</td>
<td>very unpleasant</td>
</tr>
<tr>
<td><strong>overall intelligibility (intell)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very high</td>
<td>high</td>
<td>moderate</td>
<td>low</td>
<td>none</td>
</tr>
<tr>
<td><strong>overall quality score (overall)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>very good</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of words in the reference is denoted by \( n_{\text{all}} \) and the number of substituted (\( n_{\text{sub}} \)), inserted (\( n_{\text{ins}} \)), deleted (\( n_{\text{del}} \)), and correctly recognized words (\( n_{\text{corr}} \)) are also known, then the word accuracy in percent is computed as

\[
WA = 100 \cdot \frac{1 - \frac{n_{\text{sub}} + n_{\text{del}} + n_{\text{ins}}}{n_{\text{all}}}}{n_{\text{corr}}}
\]  

A related measure, the word correctness (WR), omits \( n_{\text{ins}} \). Although both measures are usually given in percent, a high \( n_{\text{ins}} \) 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 \( \mu \)-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 \( \mu \)-law (also \( \text{‘} \mu \text{’-law} \) or \( \text{‘} \mu \text{-law} \)’) coding that is usually used for data compression in telecommunications in order to achieve histogram equalization and a better signal-to-noise ratio:

\[
f(x) = \text{sign} x \cdot \frac{\log(1 + \mu|x|/x_{\text{max}})}{\log(1 + \mu)}
\]  

When logarithmic compression is used, low values below 1 are set to a minimum threshold. The \( \mu \)-law coding attenuates this problem. It ‘compresses’ 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, \( x_{\text{max}} \) 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 \( \mu \text{5} \). 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 \( \mu \text{5} \)). 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
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
- **F0 features (F0):** regression coefficient (RegCoeff) and mean square error (MseReg) of the F0 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 F0 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 b sense, $r = 0.85$ 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 b sense 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.

### Table 2: Local prosodic features

<table>
<thead>
<tr>
<th>features</th>
<th>context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pause: before, Fill-before, after, Fill-after</td>
<td>□</td>
</tr>
<tr>
<td>En: RegCoeff, MseReg, Abs, Norm, Mean</td>
<td>□</td>
</tr>
<tr>
<td>Dur: Abs, Norm</td>
<td>□</td>
</tr>
<tr>
<td>F0: RegCoeff, MseReg</td>
<td>□</td>
</tr>
<tr>
<td>F0: Mean, Max, MaxPos, Min, MinPos, Off, OffPos, On, OnPos</td>
<td>□</td>
</tr>
</tbody>
</table>

### Table 3: The 15 global prosodic features

<table>
<thead>
<tr>
<th>feature</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>StandDevF0</td>
<td>global standard deviation of $F_0$</td>
</tr>
<tr>
<td>MeanJitter</td>
<td>mean jitter in all voiced sections</td>
</tr>
<tr>
<td>StandDevJitter</td>
<td>standard deviation of jitter in all voiced sections</td>
</tr>
<tr>
<td>MeanShimmer</td>
<td>mean shimmer in all voiced sections</td>
</tr>
<tr>
<td>StandDevShimmer</td>
<td>standard deviation of shimmer in all voiced sections</td>
</tr>
<tr>
<td>#+Voiced</td>
<td>number of voiced sections</td>
</tr>
<tr>
<td>#-Voiced</td>
<td>number of unvoiced sections</td>
</tr>
<tr>
<td>Dur+Voiced</td>
<td>duration of voiced sections (in frames)</td>
</tr>
<tr>
<td>Dur–Voiced</td>
<td>duration of unvoiced sections (in frames)</td>
</tr>
<tr>
<td>DurMax+Voiced</td>
<td>maximum duration of voiced section</td>
</tr>
<tr>
<td>DurMax–Voiced</td>
<td>maximum duration of unvoiced section</td>
</tr>
<tr>
<td>HeatNum+/-Voiced</td>
<td>ratio of number of voiced and unvoiced sections</td>
</tr>
<tr>
<td>HeatDur+/-Voiced</td>
<td>ratio of duration of voiced and unvoiced sections</td>
</tr>
<tr>
<td>HeatRel+/-Voiced/Sig</td>
<td>ratio of duration of voiced and unvoiced sections and duration of signal</td>
</tr>
</tbody>
</table>

### Table 4: Subjective and objective evaluation results

<table>
<thead>
<tr>
<th>measure</th>
<th>unit</th>
<th>mean</th>
<th>st. dev.</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>effort</td>
<td>points</td>
<td>2.66</td>
<td>1.21</td>
<td>1.00</td>
<td>4.80</td>
</tr>
<tr>
<td>b sense</td>
<td>points</td>
<td>3.28</td>
<td>0.86</td>
<td>1.40</td>
<td>4.80</td>
</tr>
<tr>
<td>tone</td>
<td>points</td>
<td>4.30</td>
<td>0.61</td>
<td>2.20</td>
<td>5.00</td>
</tr>
<tr>
<td>intell</td>
<td>points</td>
<td>3.25</td>
<td>1.10</td>
<td>2.00</td>
<td>5.00</td>
</tr>
<tr>
<td>overall</td>
<td>VAS</td>
<td>6.39</td>
<td>2.30</td>
<td>1.90</td>
<td>9.54</td>
</tr>
<tr>
<td>WA (base)</td>
<td>%</td>
<td>47.0</td>
<td>19.6</td>
<td>-2.7</td>
<td>79.6</td>
</tr>
<tr>
<td>WR (base)</td>
<td>%</td>
<td>53.9</td>
<td>17.0</td>
<td>8.6</td>
<td>83.3</td>
</tr>
<tr>
<td>WA (mu5)</td>
<td>%</td>
<td>46.7</td>
<td>19.2</td>
<td>-4.5</td>
<td>79.1</td>
</tr>
<tr>
<td>WR (mu5)</td>
<td>%</td>
<td>53.1</td>
<td>16.6</td>
<td>12.1</td>
<td>81.8</td>
</tr>
</tbody>
</table>
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 $\mu$-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 $\mu$-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 $\mu$-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 $\mu$-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 $\mu$-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 $\mu$-law features and the other rating criteria that have been examined.

### Acknowledgments

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