ON THE ESTIMATION OF SIGNAL-TO-NOISE RATIO IN CONTINUOUS SPEECH FOR ABNORMAL VOICES

Vijay Parsa, Donald G. Jamieson, Karen Stenning, and Herbert A. Leeper

National Centre for Audiology
University of Western Ontario, London, Ontario, Canada N6G 1H1.
parsa@nca.uwo.ca, jamieson@nca.uwo.ca, stenning@nca.uwo.ca

ABSTRACT
Acoustic measures of vocal function are attractive due to their noninvasive nature and due to the ease with which they can be obtained. In this paper, we developed an acoustic measure based on linear prediction modeling and filterbank analysis of continuous speech samples. The input speech sample was first modeled using a combination of short-term and long-term all pole linear prediction filters. The input speech sample and the residual signal were then divided into subbands using cosine-modulated or gammatone filterbanks. The Signal-to-Residue Ratio (SRR) was calculated as the weighted combination of the ratios of the input and residual signal energies in the subbands. The performance of the SRR parameter was evaluated with speech samples collected from patients suffering from vocal fold cancer before and after radiation therapy. Results showed that the SRR measure correlates better with the perceptual judgments of voice quality than the global SNR parameter.

1. INTRODUCTION
Acoustic measures are an integral part of the objective assessment, evaluation and tracking of pathological voice quality [1]. Clinically, acoustic measures are typically extracted from sustained vowel samples. Acoustic measures such as the fundamental frequency and amplitude perturbation coefficients (jitter and shimmer respectively), and Harmonics-to-Noise Ratio (HNR) are routinely used to characterize abnormal voice quality. However, sustained vowels lack important dynamic characteristics of running speech such as voice onset and terminations, voice breaks, and variations in fundamental frequency and amplitude. In addition, judgments of a talker's voice quality are usually based on the listener's perception of continuous speech. Acoustic measures extracted from continuous speech may therefore relate better with perceptual impressions of voice quality.

Hammarberg et al. [2] applied Long-term average speech spectrum (LTAS) and Fundamental Frequency Distribution (FFD) analysis on speech samples collected from 17 pathological talkers. Statistical analysis of the acoustic and perceptual data revealed that the LTAS spectral slope correlated highly with several voice scales. Klingholz [3] proposed a procedure for estimating the Signal-to-Noise Ratio (SNR) from a continuous speech sample. In this procedure, the harmonic power spectrum is synthesized and subtracted from the composite spectrum, and the ratio of the harmonic and noise spectra computed. Using a database of 50 normal and 74 pathological talkers, Klingholz [3] found that the SNR measure extracted from sustained vowels resulted in a 22.5% error rate in classifying pathological talkers, while the SNR computed from running speech samples provided a superior 5.6% error rate.

Several studies have shown that the acoustic measures extracted from linear prediction (LP) model are successful in discriminating between normal and pathological voices and in predicting perceptual judgments of pathological voices [4,5,6]. This paper concentrates on the estimation of the SNR parameter through LP modeling of the continuous speech sample.

2. SNR ESTIMATION

![Figure 1: Block diagram of the SNR estimation procedure using linear prediction analysis.](image)

Qi et al. [4] applied the block diagram shown in Figure 1 to estimate the SNR in a continuous speech sample. The input speech sample, $x(n)$, is segmented into blocks and a short-term LP analysis is applied. The input signal is then inverse-filtered using the short-term LP model to remove the short-term correlations. The short-term residual signal, $s(n)$, is then processed through an inverse filter derived from long-term LP modeling of the short-term residual signal. The SNR is estimated as the logarithmic ratio of the input and residual signal energies. The SNR was found to separate highly dysphonic and near-normal talkers very well, but disordered talkers in between these two extremes were not adequately characterized [4].

The SNR parameter is a global measure that puts equal emphasis on all frequencies of the signal spectrum. The performance of the SNR measure can be improved by emphasizing different frequency regions of the input spectrum in such a way that the correlation with perceptual ratings is maximized. Not only does this procedure result in a better procedure for predicting perceptual ratings, it also allows for the identification of frequency regions that are most important.
in characterizing pathological voice quality. One way to achieve this goal is to analyze the input and residual speech signals using filterbanks and linearly combine the SNRs in individual bands to form a weighted Signal-to-Residue Ratio (SRR). In this paper, we employed cosine-modulated and gammatone filterbanks to achieve the frequency separation.

In a cosine-modulated filterbank, a baseband prototype is first designed to match the necessary bandwidth and attenuation criteria. The prototype is then cosine modulated to obtain a filterbank with the specified number of channels [7]. On the other hand, the gammatone filterbank is a non-uniformly spaced filterbank with higher bandwidths at lower frequencies and smaller bandwidths at higher frequencies. The gammatone filterbank is based on psychoacoustical principles and is designed to approximate the auditory filter characteristics of the ear [8].

3. METHOD

3.1. Subjects

Eighteen adult males, aged 37 to 85 years, who were medically and histologically diagnosed with carcinoma of one vocal fold served as the subjects. For all these subjects, the cancer was restricted to one vocal fold, was on the glottic plane and had no anterior commissure involvement. All subjects were followed by and treated with radiation therapy at the London Regional Cancer Centre (LRCC) as prescribed by the attending radiation oncologist. All subjects received 6,000 cGy of radiation with 33 fractions through a pair of cobalt or 4 MeV fields at the LRCC [9].

3.2. Data collection

All subjects were assessed at the Vocal Function Laboratory at four different time periods: (1) a few days before the start of radiation therapy (P1), (2) three to four weeks after completion of radiation therapy (P2), (3) three months post-radiation (P3), and (4) six months post-radiation (P4). Ten of eighteen subjects were also evaluated one year after the radiation therapy (P5). A complete vocal function evaluation (morphological, acoustical, aerodynamic and perceptual) was administered as part of the assessment.

For acoustic data collection, each subject was seated in a quiet room in front of a microphone (Sony F-V150T) at a distance of 7.5 cm. Acoustic samples were recorded directly onto a computer hard drive at a 20 kHz sampling rate using a 12 bit A/D converter and Cspeech software. Samples of sustained /a/, /i/, /u/ vowels and the sentence the blue spot is on the key again were collected from each of the subjects. This study focused on the acoustic analysis of the sentence.

3.3. Perceptual evaluation

Categorical ratings of voice quality were obtained from eight listeners with extensive professional experience in listening to abnormal speech samples. The mean age of the listeners was 46.2 years with a range of 38 to 56 years. All listeners had hearing within normal limits for their age and gender.

The acoustic data were presented to the listeners using the EcosWin software [12]. Stimuli were output from the computer disk over the Tucker Davis Technologies (TDT) DD1 system, low pass filtered, amplified and presented binaurally over TDH-39 earphones. Each listener rated the stimulus by selecting one of the following seven descriptors: (1) normal, (2) mild, (3) mild-to-moderate, (4) moderate, (5) moderate-to-severe, (6) severe, and (7) aphonic. Six of the eight listeners were administered a retest with a gap of at least two weeks between sessions.

3.4. Acoustic analysis

A 20th order LP model was used to estimate the vocal tract parameters, which were computed using the Levinson-Durbin algorithm [11] on frames of 30 ms. A third order long-term LP model was used to remove the pitch periodicity from the residual of the short-term LP model. The final residual signal and the input to the modeling stage were put through sixteen channel cosine-modulated and gammatone filterbanks. For the cosine-modulated filter bank, the low pass prototype was a 233 tap linear-phase FIR filter which was designed using the window method with Kaiser window as the window of choice [8]. The gammatone filterbank was implemented using the Patterson-Holdsworth [9] formulae. The SRR was computed as a linear combination of the SNRs in each band.

4. RESULTS

Inter- and Intra-judge correlations for the perceptual ratings were relatively high. The test-retest correlation values ranged from 0.73 to 0.93, while inter-judge correlations ranged from 0.67 to 0.83. All the correlation coefficients were significant at $p < 0.01$ level.

Tests for within-subjects effects using the split-plot repeated measures analysis revealed that the effect of time period was significant ($F(3,378)=110.119, p<0.01$). Results from within-subjects contrasts showed that the differences in ratings between periods P3 and P1, and periods P4 and P1 were significant ($F(1,126)=151.027, p<0.01$, $F(1,126)=182.25, p<0.01$). Figure 1 depicts the mean perceptual ratings across the entire database at different time periods. There is a small but insignificant change between periods P1 and P2. The perceptual ratings at periods P3, P4, P5 were significantly lower than P1, indicating an improvement in the overall voice quality.

Figure 1: Mean and standard deviation of the perceptual ratings across different time periods.
Figure 2: Linear prediction analysis for SRR estimation.

Figure 2 displays the results from linear prediction analysis of a speech sample. The top panel in this figure depicts the input waveform. The middle panel shows the residual signal after the short-term prediction. The bottom panel shows the residual signal after long-term prediction was applied. The global SNR for this sample was 16.36 dB after short-term prediction and 19.64 dB after both short- and long-term predictions.

Figures 3 and 4 show the frequency response of the gammatone and linear filterbanks respectively. Note that the overlap between adjacent bands is more substantial for the gammatone filterbank than the linear filterbank. The input and residual signals were processed by these filterbanks and the SNRs in each band were computed. Stepwise linear regression was then performed to identify the frequency bands that best predict the perceptual ratings. Table 1 displays the results of the linear regression analysis using the gammatone filterbank. The SNRs obtained from frequency bands with centre frequencies of 246.7 Hz, 598.1 Hz, and 3387.0 Hz made significant contributions to the regression function. The results from linear regression analysis of the SNRs obtained from the cosine modulated filterbank are shown in Table 2. The linear combination of SNRs estimated in frequency bands with centre frequencies of 156.25 Hz, 468.75 Hz, 2656.25 Hz, 4218.75 Hz, and 4843.75 Hz resulted in a correlation of 0.80 ($R^2 = 0.65$). In comparison, the correlation coefficient between the perceptual ratings and the global SNR parameter was 0.47 ($R^2 = 0.22$).

5. CONCLUSIONS

The SNR of a continuous speech sample can be estimated using short-term and long-term linear prediction analysis. In this paper, we have extended this technique by filtering the signal and residual signals into different frequency bands, and weighting the SNRs in each band to result in a Signal-to-Residue Ratio (SRR) parameter. The weights were obtained through maximizing the correlation between the acoustic and perceptual ratings. We have investigated two different filterbank architectures: a uniformly spaced cosine modulated filterbank and a nonuniform gammatone filterbank. The performance of this technique was evaluated using perceptual ratings of voice samples obtained from patients with vocal fold cancer before and after radiation therapy. Our results showed that the filterbank based SRR measurement were superior to the global SNR, and that the cosine modulated filterbank performed better than the gammatone filterbank.

Table 1: Results from linear regression analysis using the gammatone filterbank.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band #6</td>
<td>0.708</td>
<td>0.502</td>
<td>0.496</td>
<td>0.9848</td>
</tr>
<tr>
<td>Bands #3,6</td>
<td>0.750</td>
<td>0.563</td>
<td>0.552</td>
<td>0.9285</td>
</tr>
<tr>
<td>Bands #3,6,14</td>
<td>0.774</td>
<td>0.598</td>
<td>0.583</td>
<td>0.8958</td>
</tr>
</tbody>
</table>

Table 2: Results from linear regression analysis using the cosine modulated filterbank.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band #2</td>
<td>0.685</td>
<td>0.469</td>
<td>0.462</td>
<td>1.0172</td>
</tr>
<tr>
<td>Bands #2,9</td>
<td>0.715</td>
<td>0.512</td>
<td>0.499</td>
<td>0.9813</td>
</tr>
<tr>
<td>Bands #2,9,14</td>
<td>0.753</td>
<td>0.568</td>
<td>0.551</td>
<td>0.9295</td>
</tr>
<tr>
<td>Bands #2,9,14,16</td>
<td>0.780</td>
<td>0.580</td>
<td>0.587</td>
<td>0.8907</td>
</tr>
<tr>
<td>Bands #1,2,9,14,16</td>
<td>0.806</td>
<td>0.649</td>
<td>0.626</td>
<td>0.8479</td>
</tr>
</tbody>
</table>

6. ACKNOWLEDGEMENTS

This work is dedicated to the memory of our colleague and good friend, Herbert “Andy” Leeper. This research was...
partially funded by the Canada Foundation for Innovation (CFI), and the Ontario Investment Trust (OIT).

7. REFERENCES