Multi-band and Multi-cue Analyses of Disordered Connected Speech

A. Alpan1, Y. Maryn2, F. Grenez1, A. Kacha1, J. Schoentgen1,3

1 Laboratoire d’Images, Signaux et Dispositifs de Télécommunications, Université Libre de Bruxelles, Brussels, Belgium
2 Department of Otorhinolaryngology and Head & Neck Surgery, Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Bruges, Belgium
3 National Fund for Scientific Research, Belgium

aalpan@ulb.ac.be, Youri.Maryn@azbrugge.be, fgrenez@ulb.ac.be, akacha@ulb.ac.be, jschoent@ulb.ac.be

Abstract

The objective is to analyze vocal dysperiodicities in connected speech produced by dysphonic speakers. The analysis involves a speech variogram-based method that enables tracking instantaneous vocal dysperiodicities. The dysperiodicity trace is summarized by means of the signal-to-dysperiodicity ratio, which has been shown to correlate strongly with the perceived degree of hoarseness of the speaker. Previously, this method has been evaluated on small corpora. In the study that is reported here the corpus has comprised 28 normophonic and 223 dysphonic speakers. This has enabled carrying out the analysis in multiple frequency bands and submitting the signal-to-dysperiodicity ratios per band to multi-variable linear regression analysis with a view to predicting the perceptual ratings of the disordered speech fragments. The analysis results are compared to the cepstral peak prominence, which is a cue that indirectly summarizes vocal dysperiodicities frame-wise via the size of the first rhamonic of the speech cepstrum. Results show that the signal-to-dysperiodicity ratios obtained for low-frequency bands up to 1500 Hz contribute most to the prediction of the perceptual scores. Also, combining the cepstral peak prominence with the low frequency-band signal-to-dysperiodicity ratio increases their common correlation with perceptual scores to 0.8.

Index Terms: analysis of connected disordered speech, variogram analysis, signal-to-dysperiodicity ratio, cepstral peak prominence.

1. Introduction

Acoustic analysis of speech is non-invasive and enables clinicians to monitor and express numerically the degree of hoarseness of a speaker’s voice. Several categories of acoustic cues have been used to characterize speech of dysphonic speakers. The most popular category describes the deviations of the speech signal from strict periodicity [2]. The causes of observed disturbances are multiple. They include jitter of the speech cycle duration or amplitude (also called shimmer), additive noise owing to turbulence, diplophonia, biphonation, random vocal fold vibration, parasitic vibrations of the false vocal folds, as well as uncontrolled transients between different phonatory regimes.

Most acoustic clinically-relevant perturbative cues have been obtained for steady fragments of sustained vowels. The use of sustained vowels is motivated by technical feasibility rather than clinical relevance. Indeed, for sustained sounds from which onsets and offsets are omitted, the hypotheses of stationarity and pseudo-periodicity are valid for many speakers [1]. However, clinicians consider connected speech to be more informative than sustained vowels. Moreover, the perceptual evaluation of voice is mainly based on connected speech or sustained vowels including onsets and offsets, rather than on steady fragments of vowels, thus informing about the transient characteristics of vocal fold vibration.

Many existing acoustic signal analysis techniques attempt isolating individual cycles in the speech signal or harmonics in the speech spectrum on the base of an a priori estimate of the typical speech cycle length or vocal frequency [1]. The failure to obtain such estimates reliably for severely disturbed voices results in cycle, harmonic or rhamonic insertion or omission errors that bias the values of dysperiodicity cues. The generalized variogram method enables tracking cycle-to-cycle dysperiodicities (whatever their cause) in any speech sound produced by any speaker, because it is not based on the assumptions that the signal is locally periodic or that the average period length can be known a priori [3]. The signal-to-dysperiodicity ratio (SDR) that summarizes the dysperiodicities has been shown to correlate strongly with the degree of perceived hoarseness.

Cepstral peak prominence (CPP) is another acoustic cue that has been obtained for connected speech fragments. It summarizes indirectly the degree of disturbances via the size of the first rhamonic of the cepstrum of a speech frame. It has been shown to correlate strongly with perceived hoarseness, even though the detection of the first rhamonic may be error prone for severely hoarse speakers [4], [5]. One difference between the cepstral and variogram-based methods is that the latter enables obtaining the instantaneous dysperiodicities.

Previously, the variogram-based method has been tested on small corpora. In this presentation, the signal-to-dysperiodicity ratios are obtained for a corpus of sustained vowels and connected speech fragments produced by 251 speakers. This enables performing multi-frequency band and multi-cue analyses, without risking over-fitting. The experiments that are reported therefore involve: a) obtaining the signal-to-dysperiodicity ratios for several frequency bands and investigating their ability to predict the perceptual scores of the speech stimuli; b) comparing the correlations with perceptual scores of temporal and spectral cues; c) pooling temporal and spectral cues and examining their combined ability to predict the perceptual scores of the speech fragments.
2. Methods

2.1. Generalized variogram analysis

For a periodic signal \( x(n) \) of period \( T_0 \), one may write \( x(n) = x(n - T_0) \). For a locally-stationary signal, the deviation from strict periodicity over an analysis frame of length \( N \) can therefore be estimated by the following expression. Index \( n \) positions the samples within the frame.

\[
\delta = \min \left\{ \sum_{n=0}^{N} [x(n) - x(n + T)]^2 : -T_{\min} < T < -T_{\max}, T_{\min} < T < T_{\max} \right\}
\]

(1)

The expression between accolades in (1) is known as the variogram of the speech signal. It involves the squared difference between a main analysis frame and a shifted auxiliary frame. Lag \( T \) may be positive or negative. In practice, it is permitted to vary between \( \pm 2.5 \) and \( \pm 20 \) ms. Signed lags guarantee that, in connected speech, the shift of the lagged analysis frame across phonetic boundaries is avoided and that only intra-segment cycles are compared.

Indeed, when a speech cycle is near the right-hand boundary of a phonetic segment the segment-internal cycles are expected to be to its left, that is, lag \( T \) is expected to be negative and vice versa for a speech cycle positioned near the left-hand phonetic boundary.

For each main analysis frame position, lag \( T \) is fixed so as to minimize the cumulated squared difference between the main and shifted frames. For voiced sounds, lag \( T \) is therefore an integer multiple of the glottal cycle length. For unvoiced sounds, (1) can be still meaningfully computed but the interpretation of lag \( T \) in terms of glottal cycle lengths is not valid anymore.

In running speech, the signal amplitude evolves deterministically owing to onsets and offsets, segment-specific loudness as well as accentuation. To remove these clinically non-relevant variations of the signal amplitude, a local gain \( \alpha \) that equalizes the energies between main and auxiliary analysis frames is inserted into (1).

\[
\delta = \min \left\{ \sum_{n=0}^{N} [x(n) - \alpha x(n + T)]^2 : -T_{\min} < T < -T_{\max}, T_{\min} < T < T_{\max} \right\}
\]

\[
\alpha = \frac{\sum_{n=0}^{N} x^2(n)}{\sum_{n=0}^{N} x^2(n + T)}
\]

(2)

The analysis frame length is fixed to 2.5 ms, so that main and lagged frames cannot overlap. The shift between successive analysis frames is also fixed to 2.5 ms, thus enabling the sample-by-sample dysperiodicity to be computed unambiguously (3).

\[
e(n) = x(n) - \alpha x(n + T_{opt})
\]

(3)

In (3), lag \( T_{opt} \) is the positive or negative shift that minimizes the cumulated squared difference in (2). Hereafter, the energy-equalized variogram (2) is called the generalized variogram to distinguish it from the conventional variogram in (1).

2.2. Global and segmental signal-to-dysperiodicity ratios

The global signal-to-dysperiodicity ratio in dB summarizes the relative amount of cycle-to-cycle noise in a speech recording of length \( L \).

\[
SDRGLOB = 10 \log \frac{\sum_{n=0}^{L} x^2(n)}{\sum_{n=0}^{L} x^2(n + T)}
\]

(4)

The segmental signal-to-dysperiodicity ratio \( SDRSEG \) consists in computing ratio (4) locally over intervals of 5 ms and then taking the average. The segmental signal-to-dysperiodicity ratio has been favored over the global one in the framework of the evaluation of lossy speech coders, because the segmental ratio appears to correlate better with human-assigned scores of perceived quality. A possible explanation is that segmental ratios boost the contribution of weak noisy segments, which seem to influence perceived timbre strongly [6].

2.3. Multi-band analyses

For each utterance, the speech signal as well as the corresponding dysperiodicity trace have been filtered by means of four-channel mel-spaced linear-phase filters. The ranges of the four mel bands (B1 – B4) have been (0 – 800 mel), (800 – 1600 mel), (1600 – 2400 mel) and 2400 mel and beyond. These mel-intervals correspond to the frequency bands (0 – 724 Hz), (724 – 2195 Hz), (2195 – 5188 Hz) and 5188 Hz and beyond. The filterbank was designed by means of the Parks-McClellan method.

The global and segmental signal-to-dysperiodicity ratios have been computed for each band. The global ratio has involved the log-ratio of the filtered signal and dysperiodicity trace energies, while the segmental ratio has involved the averages of the log-ratios over adjacent analysis frames of 5 ms.

2.4. Cepstral peak prominence

The cepstral peak prominence \( CPP \) is a measure of the amplitude of the first rhamonic of the speech cepstrum. It has been used to assess numerically vocal quality or dysphonia. Also, it has been shown to correlate with perceived degree of breathiness [4], [5]. The calculation of the cepstral peak prominence involves the following steps.

A cepstral peak is detected between the minimum and the maximum expected vocal frequency (50 – 400 Hz). A linear regression line is fitted to the cepstrum. The line is computed between 1 ms and the maximum quefrency. The height of the cepstral peak, selected during the previous step, with regard to the regression line is taken as the local (per-frame) cepstral peak prominence.

The \emph{global} cepstral peak prominence is obtained by averaging the \emph{local} CPPs over several analysis frames. The analysis frame length is 46.4 ms (i.e. 2048 signal points for a sampling rate of 44.1 KHz) and the shift between successive frames equals 10 ms. Cepstral peak prominences have been obtained by means of Hillenbrand’s \textsc{Cpps} executable [7].
2.5. Corpora

The corpora comprise sustained vowel [a] (corpus 1) and two Dutch sentences (“Papa en Marloes staan op het station. Ze wachten op de trein.”) (corpus 2) produced by 28 normophonic and 223 dysphonic speakers. The voiced segments of the two sentences were extracted and concatenated. Based on these productions, two stimuli have been generated. The first has been a concatenation of the vowel [a] with the voiced segments obtained previously (corpus 3). The second has been a concatenation of the vowel [a] with the whole sentences (corpus 4). The concatenation of the voiced phonetic segments with vowel [a] has been sampled at 22050 Hz. The other stimuli have been sampled at 44100 Hz. Five judges have carried out the perceptual evaluation. Each judge has rated the item “grade”, (G) of the GRABS scale, from 0 (normal) to 3 (severe). The “grade” refers to the overall perceived abnormality of the speech stimuli, which have been the concatenation of the two (complete) sentences with vowel [a] (corpus 4). The five perceptual scores per stimulus have been averaged. The recordings and evaluation have been made at the Sint-Jan General Hospital, Bruges, Belgium.

2.6. Multi-band linear regression analysis

Linear regression analysis has been carried out to predict the degree of perceived hoarseness via a linear combination of the segmental or global signal-to-dysperiodicity ratios in different frequency bands. The analysis has been performed on the corpus combining sentences and sustained vowel [a] (corpus 4).

2.7. Multi-cue linear regression analysis

Linear regression analysis has been carried out to predict the degree of perceived hoarseness via a linear combination of the segmental signal-to-dysperiodicity ratio in the lowest frequency band and the cepstral peak prominence (CPP). The reasons for selecting the lowest frequency band are explained in Section 3.3. The analysis has been performed on the corpus combining sentences and sustained vowel [a] (corpus 4).

3. Results

3.1. Generalized variogram analysis and signal-to-dysperiodicity ratios

Table 1 shows the Pearson correlation coefficients between the perceived average grade and the global and segmental signal-to-dysperiodicity ratios for corpora 1 to 4. One sees that the segmental signal-to-dysperiodicity ratios correlate better than the global ones with the perceptual ratings. One observes the strongest correlation for the fragments combining sentences and vowel [a], for which the perceptual ratings have been obtained originally.

3.2. Cepstral analysis and cepstral peak prominence

Table 1 also shows the Pearson correlation coefficients between the average perceived grade and the cepstral peak prominence. Again one observes that the correlation is strongest for the fragments combining sentences and vowel [a].

Table 1: Pearson’s correlation coefficients between average grade scores and cues SDRGLOB, SDRSEG, and CPP for sustained vowel [a] (corpus 1), complete sentences (corpus 2), and the concatenations of voiced fragments (corpus 3) or complete sentences (corpus 4) with vowel [a].

<table>
<thead>
<tr>
<th>Corpus</th>
<th>SDRGLOB</th>
<th>SDRSEG</th>
<th>CPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus 1</td>
<td>-0.64</td>
<td>-0.63</td>
<td>-0.56</td>
</tr>
<tr>
<td>Corpus 2</td>
<td>-0.35</td>
<td>-0.64</td>
<td>-0.65</td>
</tr>
<tr>
<td>Corpus 3</td>
<td>-0.56</td>
<td>-0.69</td>
<td>-0.63</td>
</tr>
<tr>
<td>Corpus 4</td>
<td>-0.46</td>
<td>-0.70</td>
<td>-0.70</td>
</tr>
</tbody>
</table>

3.3. Multi-band generalized variogram analysis

Multi-band variogram analysis has been carried out for the corpus combining sentences and vowel [a] (Corpus 4), because this corpus has previously given the strongest correlation with perceptual ratings of abnormality. Only the first three bands have been entered into the linear regression analysis, which has been stepwise. The reason is that the correlation with the perceptual ratings of cue SDRSEG in band 4 has been close to zero. In that band, the speech signal is masked by noise and the SDRSEG values are low for most speakers.

Table 2 shows the results of the stepwise linear regression analysis carried out on the SDRSEGs for the first three frequency bands. The table reports on the left-hand side the standardized regression coefficients of the segmental signal-to-dysperiodicity ratios in frequency bands 1 to 3. The standardized regression coefficients can be compared directly. The multiple correlation coefficient R is statistically significant (R_{adj} = 0.176, p < 0.05, F =113.6).

Table 2: Results of the stepwise linear regression analysis carried out on segmental signal-to-dysperiodicity ratios for three frequency bands.

<table>
<thead>
<tr>
<th>β</th>
<th>β1</th>
<th>β2</th>
<th>R</th>
<th>R²</th>
<th>Adj.R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.574</td>
<td>-0.154</td>
<td>-0.171</td>
<td>0.761</td>
<td>0.580</td>
<td>0.575</td>
</tr>
</tbody>
</table>

To compare the contribution of the individual predictor variables, Table 3 displays multiple correlation coefficients R obtained for SDRSEGs for one (B1), two (B1,B3) and three bands (B1, B2, B3). One sees that the multiple correlation coefficient R increases from 0.71 (for the first frequency band) to 0.76 (for the first three frequency bands).

Table 3: Multiple correlation coefficients obtained from stepwise linear regression analysis carried out on SDRSEGs for one (B1), two (B1, B3) and three bands (B1, B2, B3).

<table>
<thead>
<tr>
<th>B1</th>
<th>B1, B3</th>
<th>B1, B2, B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.71</td>
<td>0.75</td>
<td>0.76</td>
</tr>
</tbody>
</table>

As an illustration, Figure 1 shows on the vertical axis the predicted average grade scores by means of the SDRSEG cues for the first three frequency bands and on the horizontal axis the original average grade scores. In addition, a stepwise linear regression analysis has been carried out combining global and segmental signal-to-dysperiodicity ratio cues in each frequency band. The correlation between predicted average grade scores and actual grade scores then increases to 0.78. The number of cues retained by the stepwise method, however, increases from 3 to
5. cues SDRSEG obtained for the first three frequency bands and cues SDRGLOB for the first two frequency bands.

Figure 1: Predicted average perceived grade scores (via the SDRSEG cues for the first three frequency bands) versus original average perceived grade scores.

Figure 2 shows the variation of the correlation coefficient between the perceptual scores and the segmental signal-to-dysperiodicity ratio for the first frequency band as a function of its upper cut-off frequency, which has been in the range 200 Hz to 3200 Hz. One sees that the absolute value of the correlation increases to a maximum for a cut-off frequency in the interval 1000-1500 Hz.

3.4. Multi-cue linear regression analysis

Multi-cue linear regression analysis has been carried out by means of the cepstral peak prominence and segmental signal-to-dysperiodicity ratio for the first frequency band with a view to predicting perceptual scores. These two cues have been retained because their correlation is 0.61 only and because, for the signal-to-dysperiodicity ratio, the first band contributes most to the prediction of the perceptual scores (Table 2). Table 4 shows the results of the linear regression analysis. Correlation $R$ between predicted and original perceptual scores increases from 0.71 (Table 3) to 0.79. It is statistically significant ($R_{crit} = 0.155$, $p < 0.05$, $F = 200.4$).

The weights reported in Table 4 suggest that cepstral peak prominence and segmental signal-to-dysperiodicity ratio contribute roughly equally to the prediction of the perceptual scores.

In addition, t-tests have shown that multiple correlation coefficient $R$ differs statistically significantly for multi-band (Table 3) and multi-cue (Table 4) analyses ($t=2.69$, $p<0.01$).

<table>
<thead>
<tr>
<th>$\beta_{SDRSEG}$</th>
<th>$\beta_{CPP}$</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adj. $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.46</td>
<td>-0.42</td>
<td>0.79</td>
<td>0.62</td>
<td>0.615</td>
</tr>
</tbody>
</table>

4. Discussion and conclusion

One notices in Table 2 that the weights of the SDRSEG cues decrease for high-frequency bands and Table 3 indicates that the first frequency band contributes most to the prediction of the perceptual scores by the signal-to-dysperiodicity ratios. This would suggest that most of the perceptually relevant information is comprised in the first frequency band. A possible interpretation is that the spectral energy decreases with frequency because of the spectral tilt and that, hence, listeners pay more attention to low than to high-frequency bands. Figure 2 reports the variation with the upper cut-off frequency of the correlation between perceptual scores and signal-to-dysperiodicity ratio. The correlation reaches its optimal plateau near 1000 Hz, which confirms that the first frequency band (B1) contributes most to the prediction of the perceptual scores via acoustic cues.

5. Acknowledgements

This research was supported by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme.

6. References