INTERSPEECH 2009 BRIGHTON

Cepstral analysis of vocal dysperiodicities in disordered connected speech

A. Alpan[1], *J. Schoentgen* [1,2], *Y. Maryn*[3], *F. Grenez*[1], *P. Murphy*[4]

[1] Laboratory of Images, Signals & Telecommunication Devices, Université Libre de Bruxelles, Brussels, Belgium
[2] National Fund for Scientific Research, Belgium
[3]Department of Otorhinolaryngology and Head & Neck Surgery, Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Bruges, Belgium
[4]Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

aalpan@ulb.ac.be, jschoent@ulb.ac.be, Youri.Maryn@azbrugge.be, fgrenez@ulb.ac.be, peter.murphy@ul.ie

10.21437/Interspeech.2009-286

Abstract

Several studies have shown that the amplitude of the first rahmonic peak (R1) in the cepstrum is an indicator of hoarse voice quality. The cepstrum is obtained by taking the inverse Fourier Transform of the log-magnitude spectrum. In the present study, a number of spectral analysis processing steps are implemented, including period-synchronous and period-asynchronous analysis, as well as harmonic-synchronous and harmonic-asynchronous spectral band-limitation prior to computing the cepstrum. The analysis is applied to connected speech signals. The correlation between amplitude R1 and perceptual ratings is examined for a corpus comprising 28 normophonic and 223 dysphonic speakers. One observes that the correlation between R1 and perceptual ratings increases when the spectrum is band-limited prior to computing the cepstrum. In addition, comparisons are made with a popular cepstral cue which is the cepstral peak prominence (CPP).

Index Terms: Voice analysis, cepstrum, first rahmonic, correlation analysis, connected disordered speech.

1. Introduction

Within the context of the assessment of laryngeal function, acoustic analysis has a central place because the speech signal is recorded non-invasively and it is the base on which the perceptual assessment of voice is founded. Generally speaking, the goal of acoustic analysis is to document quantitatively the degree of hoarseness and monitor the evolution of the voice of dysphonic speakers.

Many voice disorders cause voiced speech to deviate from strict periodicity. Dysperiodicities may be caused by additive noise owing to turbulent airflow and modulation noise owing to extrinsic perturbations of the glottal excitation signal. Dysperiodicities may also be due to intrinsically irregular dynamics of the vocal folds and involuntary transients between dynamic regimes [1]. Many acoustic features that have been used to assess vocal function reflect the deviation of the speech waveform from strict periodicity. Jitter and shimmer, for instance, are frequently used to summarize perturbations of the speech cycle lengths and amplitudes, respectively.

A number of studies have shown that the amplitude of the first rahmonic peak in the cepstrum (R1) is a global descriptor of hoarse voice quality. In [2], Hillenbrand et al. have reported that the cepstral peak prominence (CPP) correlates well with perceptual ratings of breathiness. This cue consists in the log-amplitude of the first cepstral peak with regard to a linear regression line that is fitted to the log-cepstrum for normalization purposes. In [3] and [4], it has been independently reported that the cepstral peak prominence displays a better correlation with overall voice quality compared to cues such as jitter, shimmer, and several spectral tilt and noise measures.

The study that follows focuses on the amplitude of the first rahmonic (R1) obtained via the conventional cepstrum (i.e. the inverse Fourier Transform of the log-amplitude spectrum). It has been observed that increased levels of noise and perturbations in the voice signal decrease R1. In [5], Murphy has provided a theoretical description of cepstral analysis of voiced speech containing aspiration noise, which suggests that R1 is directly proportional to a geometric-mean harmonics-to-noise ratio, i.e. R1 provides an index of the overall richness of the harmonic spectrum in dB. Based on experiments with synthetic voice signals with various aperiodicities, a number of reasons for which R1 correlates highly with voice quality have been inferred, such as the approximate vocal frequency independence when extracted via period-asynchronous analysis.

In [6], analyses similar to those in this study have been carried out on human sustained vowels [a]. Working with sustained vowels enables computing the cepstrum of log-spectra averaged over the vowel duration. Indeed, when spectral averaging is carried out, the harmonics approach the true harmonic values better and the between-harmonics approach the true noise variance [5].

In this study, we focus on connected speech. In connected speech, the cepstrum is obtained frame-by-frame via the log-spectrum. A number of spectral analysis alternatives inspired by [5] have been implemented, including period-synchronous, period-asynchronous, harmonic-synchronous and harmonic-asynchronous band-limited analyses of real speech. The correlation between the amplitude of the first rahmonic (R1) and perceptual ratings has been calculated to compare the different options. Also, hypothesis tests have been carried out to check whether observed correlations are statistically significantly different. Comparisons are also made with the cepstral peak prominence (CPP).

2. Methods

2.1. Corpus and perceptual ratings

The corpus has comprised the concatenation of two Dutch sentences with sustained vowel [a] (produced by the same

speaker) uttered by 28 normophonic and 223 dysphonic speakers, male and female. The stimuli have been sampled at 44.1 kHz. Diagnosed pathologies have been the following: functional dysphonia (81), nodules (42), polypoid mucosa (edema) (29), paralysis/paresis (18), polyp (11), cyst (8), acute laryngitis (5), others (34). Five judges have evaluated the stimuli perceptually. 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. The five perceptual scores per stimulus have been averaged. The recordings and evaluation have been made at the Sint-Jan General Hospital, Bruges, Belgium.

2.2. First rahmonic amplitude (R1)

Speech analysis has been carried out using period-asynchronous and period-synchronous frames. For the period-asynchronous analysis, frame lengths ranging from 1024 to 32768 samples have been used which correspond to frame lengths ranging from 23ms to 743ms. For the period-synchronous analysis, temporal frames with 2 to 16 cycles have been used. The frames have been hopped with an overlap of 50%. The average fundamental period has been extracted with Praat [7] on the vowel [a] part of the utterances.

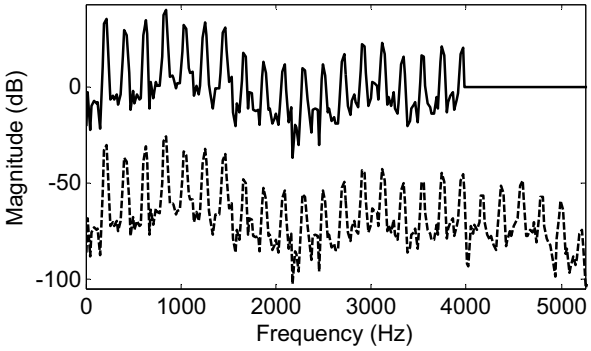

Figure 1: *Dashed line: Log-magnitude spectrum. Solid line: Band-limited and offset-removed log-magnitude spectrum for an analysis frame of a vowel [a].*

The computation of the amplitude of the first rahmonic (R1) has involved the following steps:

1. Computation of the log-magnitude spectrum for each Hamming-windowed frame.
2. Band-limitation of the log-magnitude spectrum by zeroing frequency samples above a cut-off frequency (4000Hz in Figure 1), followed by the removal of any offset. The spectral band-limiting has been harmonic-synchronous (depending on F0) or harmonic-asynchronous. In the harmonic-asynchronous case, the spectrum has been limited to frequencies ranging from 1 to 5kHz. For the harmonic-synchronous case, the spectrum has been band-limited to a fixed number of harmonics ranging from 2 to 16.
3. Computation of the cepstrum via the inverse Fourier transform of the band-limited log-magnitude spectrum.
4. Location of the first rahmonic using a peak-picking algorithm searching for the maximum in the quefrency range corresponding to a frequency range between 50Hz and 400Hz.
5. Obtainment of a global R1 amplitude by averaging the R1 amplitudes over all frames.

2.3. Cepstral peak prominence

The cepstral peak prominence (CPP) is a measure of the log-amplitude of the first rhamonic of the speech cepstrum [2].
The calculation of CPP involves the following steps.

1. Obtainment of the speech cepstrum for each analysis frame. Frame lengths of 1024, 2048, 4096, 8192, and 16384 samples have been used and they have been hopped with an overlap of 50%.
2. Fit of a linear regression line to the log-cepstrum between 1 ms and the maximum quefrency.
3. Obtainment, between the minimum and the maximum expected vocal quefrencies, of the height with regard to the regression line of the most prominent cepstral peak, which is the local (per-frame) cepstral peak prominence.
4. Obtainment of the global cepstral peak prominence (CPP) by averaging the local cepstral peak prominences over all analysis frames.

The cepstral peak prominences have been obtained by means of Hillenbrand's CPPS software [8].

3. Results

The correlations between perceptual ratings and the average amplitude of the first rahmonics (R1) and CPP have been calculated.

3.1. Full-band spectra

The cepstrum has been computed in the conventional way for the full-band spectrum. Period-asynchronous and period-synchronous frames have been used.

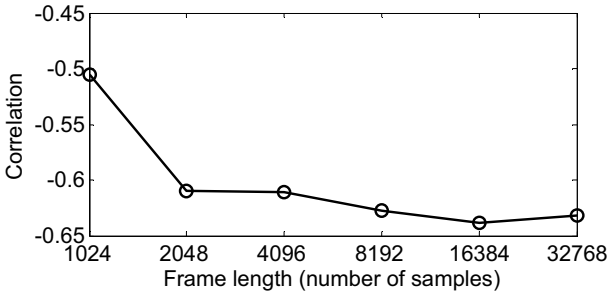

Figure 2: *Correlations between the perceptual ratings and R1 obtained for period-asynchronous temporal frames and the full spectrum.*

Figure 2 shows the correlation between perceptual ratings and R1 for several frame lengths. One sees that the highest absolute correlation of 0.64 has been obtained for a frame length of 16384 (372ms) whereas the lowest correlation has been obtained for a frame length of 1024 (23ms). Except for the latter, correlation values are similar.
Figure 3 shows the correlation between perceptual ratings and R1 for several frames comprising an increasing number of cycles.
One sees that the highest absolute correlation of 0.63 has been obtained for a frame length of 6 fundamental periods, whereas the lowest has been obtained for a frame length of 2 periods. A plateau is observed for a number of periods > 2.
For full-band cepstra, the correlations for period-synchronous and period-asynchronous analyses are similar. Hereafter, a temporal frame of 2048 samples (46ms) has therefore been chosen for the period-asynchronous option. Indeed, experiments have shown that correlations are best for that frame length (2048 samples), the full-band cepstrum (Figure

2) excepted. Also, a temporal frame of 6 cycles has been retained for the period-synchronous option.

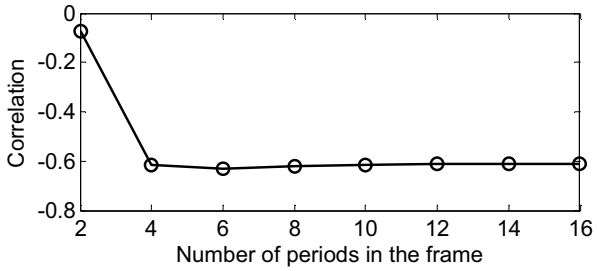

Figure 3: *Correlations between the perceptual ratings and R1 obtained for period-synchronous temporal frames and the full spectrum.*

3.2. Harmonic-asynchronously band-limited spectra

Prior to computing the cepstrum, the log-magnitude spectrum has been band-limited by setting to zero frequency samples above a cut-off frequency.
Figure 4 shows the correlation between the perceptual ratings and global R1 for cut-off frequencies ranging from 1000Hz to 5000Hz.

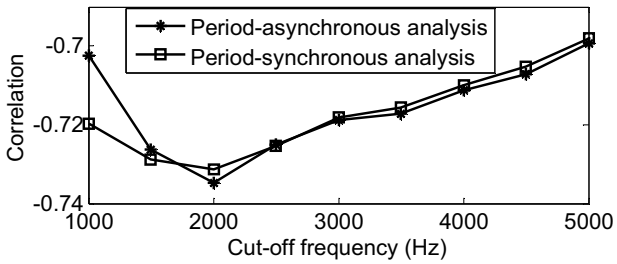

Figure 4: *Correlations between the perceptual ratings and R1 obtained for the harmonic-asynchronously band-limited spectra for period-synchronous and period-asynchronous analyses.*

For both the period-asynchronous and period-synchronous analyses, one sees that the highest absolute correlation of 0.73 is obtained for a cut-off frequency of 2000Hz.

3.3. Harmonic-synchronously band-limited spectra

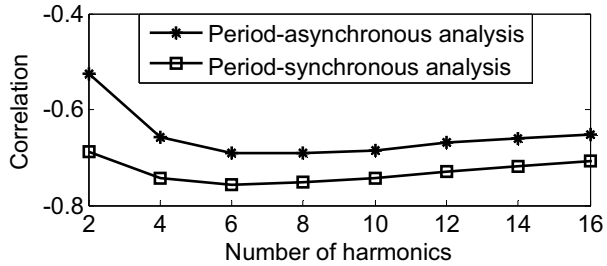

Figure 5: *Correlation between the perceptual ratings and R1 obtained for the harmonic-synchronously band-limited spectra for period-synchronous and period-asynchronous analyses.*

Prior to computing the cepstrum, the log-magnitude spectrum has been band-limited to a fixed number of harmonics.
Figure 5 shows the correlation between perceptual ratings and global R1 for the number of harmonics ranging from 2 to 16.
For both the period-asynchronous and period-synchronous analyses, one sees that the highest absolute correlation has been obtained for a number of harmonics equal to 6 ($\rho_P = 0.69$ and $\rho_P = 0.76$ respectively).

3.4. Cepstral peak prominence (CPP)

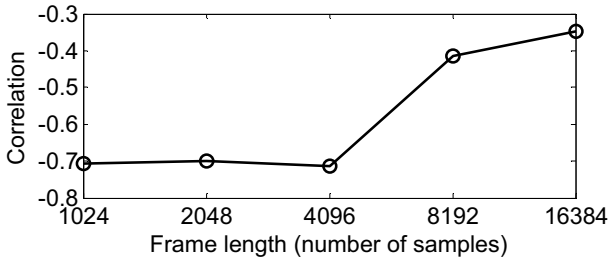

Figure 6: *Correlation between the perceptual ratings and CPP.*

Figure 6 shows the correlation between perceptual ratings and CPP for several frame lengths. One sees that the highest absolute correlation of 0.71 is obtained for a frame length of 4096 samples. The absolute correlation rapidly decreases for larger frame lengths.

3.5. Statistical tests

The null hypothesis ($\rho_P = 0$) has been rejected for all observed correlations (two-tailed t-test, $\rho_{crit} = 0.21$, $p<0.001$) except for the full-band cepstral peak (R1) obtained for a frame length of 2 fundamental cycles ($\rho_P = 0.07$) (Figure 3).

Table 1: *Significance of the differences between correlations obtained for the full-band spectra and the harmonic-asynchronously and harmonic-synchronously band-limited spectra with period-asynchronous analyses and cepstral peak prominence.*

	Period-asynchronous analysis		
	Harmonic-asynchronous ($\rho_P = 0.73$)	Harmonic-synchronous ($\rho_P = 0.69$)	CPP ($\rho_P = 0.71$)
Conventional R1 (full-band) ($\rho_P = 0.64$)	t = 4.39 significant p<0.0001	t = 1.50 not significant	t = 4.55 significant p<0.0001

Table 2: *Significance of the differences between correlations obtained for the full-band spectra and the harmonic-asynchronously and harmonic-synchronously band-limited spectra with period-synchronous analyses and cepstral peak prominence.*

	Period-synchronous analysis		
	Harmonic-asynchronous ($\rho_P = 0.73$)	Harmonic-synchronous ($\rho_P = 0.76$)	CPP ($\rho_P = 0.71$)
Conventional R1 (full-band) ($\rho_P = 0.63$)	t = 4.79 significant p<0.0001	t = 4.11 significant p<0.0001	t = 3.78 significant p<0.0001

Additional tests have been carried out to assess whether the correlations observed for different analysis options are statistically significantly different. When two correlation coefficients are calculated from the same sample, they are not statistically independent and a t-test with (n-3) degrees of freedom (n being the size of the sample) is used rather than the conventional test [9]. All the t-tests have been one-tailed.

Tables 1&2 summarize the outcome of statistical tests regarding differences in correlations between perceptual scores and rahmonic peak amplitudes (R1 or CPP). Full-band R1 is compared to band-limited R1 and CPP.
Additional tests show that differences between band-limited R1 and CPP are significant only when the R1 analyses are period and harmonic-synchronous (§ 3.3) (ρ_P = 0.76) (one-tailed t-test, t = 1.70, df = 248, p<0.05). The differences are not significant when the analyses are asynchronous.

4. Discussion and conclusion

a) One observes that when the log-magnitude spectrum is band-limited prior to obtaining the cepstrum, the correlation between R1 and the perceptual ratings significantly increases compared to the cepstrum obtained from the full-band spectrum (Tables 1&2). A possible explanation is that band-limiting enables focusing on spectral intervals that are perceptually relevant. The observation that band-limitation improves the correlation between perceptual ratings and acoustic cues of dysphonic speech has been confirmed for acoustic features that differ qualitatively from those who are investigated here [11].

b) The option that combines period-synchronous analysis and harmonic-synchronous spectral cutback gives the best correlation (ρ_P = 0.76). This can be explained by referring to [5], which suggests that for a period-synchronous and harmonic-synchronous analysis, R1 is approximately f0-independent and decreases linearly with regard to the noise level.

c) One notices that the correlation between global R1 and perceptual ratings increases from 0.73 for period-asynchronous and harmonic-asynchronous analysis to 0.76 for period-synchronous and harmonic-synchronous analysis. However, this increase is not statistically significant (one-tailed t-test, t = 1.26, df = 248).

d) One also observes that differences between correlations obtained for the CPP (ρ_P = 0.71) and full-band R1 are statistically significant. This may be explained as follows. In [2], Hillenbrand defines the cepstrum unconventionally as the log-power spectrum of a log-power spectrum, whereas to compute R1, we define the cepstrum in the conventional way via the inverse Fourier transformed log-spectrum [10]. To examine the effect of this second log operation on the conventional global R1 (§ 3.1), we have taken the logarithm of the local R1s before computing the average over all frames. The correlation between log-R1 and perceptual ratings increases from 0.64 to 0.70. However, when the same operation has been carried out on harmonic-synchronously band-limited spectra with period-synchronous analysis frames (§ 3.3), no increase in correlation has been observed. A possible explanation is given in Figure 7 that briefly summarizes experiments that have not been reported in the results section. Those experiments suggest that the increase of the correlation of full-band log-R1 or CPP with perceptual scores compared to full-band R1 is due to a decrease of the non-linearity of the relationship between cue values and perceptual scores. The increase of the correlation of the band-limited R1 with perceptual scores compared to full-band R1 appears to be a consequence of the decrease of the scatter of the cue values because of the focus on the perceptually relevant spectral intervals as well as a less non-linear relationship between perceptual scores and acoustic cues.

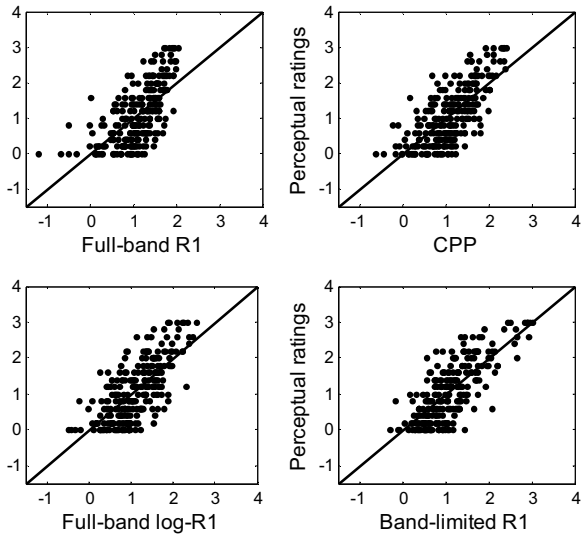

Figure 7: *Perceptual ratings versus predicted ratings by means of linear regression via the full-band R1, CPP, full-band log-R1 and band-limited (period-synchronous and harmonic-synchronous) R1.*

5. Acknowledgements

This research has been supported by COST ACTION 2103 "Advanced Voice Function Assessment" in the framework of a short-term scientific mission at the University of Limerick, and by the "Région Wallonne", Belgium, in the framework of the "WALEO II" programme.

6. References

[1] Schoentgen, J., "Spectral models of additive and modulation noise in speech and phonatory excitation signals", J. Acoust. Soc. Am., vol. 113, 553-562, 2003.

[2] Hillenbrand, J. and Houde, R. A., "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech", J. Speech Hear. Res. 39, 311-321, 1996.

[3] Heman-Ackah, Y.D., Michael, D.D. and Goding Jr, G.S., "The relationship between cepstral peak prominence and selected parameters of dysphonia", J. Voice, vol. 16, 20-27, 2002.

[4] Awan, S. N. and Roy, N., "Acoustic prediction of voice type in women with functional dysphonia", J. Voice, vol. 19, 268-282, 2005.

[5] Murphy, P. J., "On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals," J. Acoust. Soc. Am., vol. 120 (5), 2896–2907, 2006.

[6] Alpan, A., Schoentgen, J., Maryn, Y., Grenez, F. and Murphy, P., "Analysis of human voice signals via the first rahmonic", Advanced Voice Function Assessment International Workshop, AVFA 2009, Madrid (Spain), May 2009.

[7] Boersma, P., Weenink, D., "Praat: doing phonetics by computer (Version 4.6.09) [Computer program]", Retrieved June 26, 2007, from http://www.praat.org/

[8] http://homepages.wmich.edu/~hillenbr/cpps.exe

[9] Dunn, O. J. and Clark, V. A., "Correlation coefficients measured on the same individuals", J. Am. Stat. Assoc., vol. 64, 366–377, 1969.

[10] Oppenheimer, A.V. and Schafer, R.W., "Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[11] Alpan, A., Kacha, A., Grenez, F. and Schoentgen, J., "Assessment of vocal dysperiodicities in connected disordered speech", in Proc. Interspeech, pp. 1178-1181, Antwerp (Belgium), August 2007.