Multi-band dysperiodicity analyses of disordered connected speech

A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen

Accepted Manuscript

Multi-band dysperiodicity analyses of disordered connected speech

A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen

PII: S0167-6393(10)00118-4
DOI: 10.1016/j.specom.2010.06.010
Reference: SPECOM 1906

To appear in: *Speech Communication*

Received Date: 3 April 2009
Revised Date: 2 April 2010
Accepted Date: 27 June 2010

Please cite this article as: Alpan, A., Maryn, Y., Kacha, A., Grenez, F., Schoentgen, J., Multi-band dysperiodicity analyses of disordered connected speech, *Speech Communication* (2010), doi: 10.1016/j.specom.2010.06.010

Multi-band dysperiodicity analyses of disordered connected speech

A. Alpan[1], Y. Maryn[2], A. Kacha[1], F. Grenez[1], J. Schoentgen[1,3]

[1] Laboratory of Images, Signals and Telecommunications Devices, 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, akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

Corresponding author:
Ali Alpan,
Laboratory of Images, Signals and Telecommunications Devices, Université Libre de Bruxelles, Brussels, Belgium

aalpan@ulb.ac.be
tel: +32.2.650.30.85 fax: +32.2.650.47.13
ULB - LIST CP165/51
Av. F. Roosevelt, 50
1050 Brussels
Belgium

Abstract

The objective is to analyze vocal dysperiodicities in connected speech produced by dysphonic speakers. The analysis involves a 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 only. In this article, analyses have been carried out on two corpora comprising over 250 and 700 speakers. This has enabled carrying out multi-frequency band and multi-cue analyses without risking over-fitting. The analysis results are compared to the cepstral peak prominence, which is a popular cue that indirectly summarizes vocal dysperiodicities frame-wise. A perceptual rating has been available for the first corpus whereas speakers in the second corpus have been categorized as normal or pathological only. For the first corpus, results show that the correlation with perceptual scores increases statistically significantly for multi-band analysis compared to conventional full-band analysis. Also, combining the cepstral peak prominence with the low frequency-band signal-to-dysperiodicity ratio statistically significantly increases their combined correlation with perceptual scores. The signal-to-dysperiodicity ratios of the two corpora have been separately submitted to principal component analysis. Results show that the first two principal components are interpretable in terms of the degree of dysphonia and the spectral slope respectively. The clinical relevance of the principal components has been confirmed by linear discriminant analysis.

Keywords: connected disordered speech, variogram, signal-to-dysperiodicity ratio, cepstral peak prominence, multi-band analysis, multi-variable analysis.

1. Introduction

Within the context of the assessment of laryngeal function, acoustic analysis has a central place because the speech signal may be recorded non-invasively and it forms 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 (Edgar, 2001), (Sapienza, 2002), (Schoentgen, 2003). Many acoustic features that have been used to assess vocal function reflect the deviation of the speech waveform from perfect periodicity. Jitter and shimmer, for instance, are frequently used to summarize perturbations of the speech cycle lengths and amplitudes, respectively.

Most often, acoustic markers of vocal dysperiodicities are obtained from stable fragments of sustained speech sounds. The reason is that most methods for estimating vocal dysperiodicities rely on the assumptions of stationarity and local periodicity of the signal. Sustained vowels can indeed be assumed, with good accuracy, to be produced by keeping time-invariant the characteristics of the voice source, vocal tract and articulators (Murry and Doherty, 1980) so that vocal perturbations and noise are easily measured. However, the widespread use of sustained vowels is due to the technical feasibility of the analysis rather than clinical relevance.

Disregarding the issue of technical feasibility, most clinicians consider connected speech to be more informative than sustained sounds. Arguments in favour of analysing connected speech are that the vibration of the vocal folds must be switched on and off continually, voicing must be maintained while the supra-laryngeal impedance is changing incessantly, especially during obstruents, and the larynx continually ascends and descends in the neck (Schoentgen, 2003). Lack of stationarity as well as the greater variability of the conditions under which vibration must take place are considered to be a greater challenge to a speaker’s larynx. Also, it has been speculated that speakers are less likely to compensate for their voice problems while producing connected speech than while sustaining sounds. It may therefore be the case that speakers are able to sustain quasi-normal sounds within a narrow interval of pitch and intensity, whereas their continuous speech waveform may be severely disturbed. In addition, the signal-to-dysperiodicity ratio obtained from sustained vowels depends somewhat on fundamental frequency and sound pressure. This dependence may be weaker on average when the analysis is carried out on continuous speech (Klingholtz, 1990).

In addition, sustained vowels may have other drawbacks. Some methods require that the analysis is carried out on recordings of long duration, which may be subject to pitch and intensity variations. As a consequence, the assumption of stationarity may be violated. The reliability of these methods may therefore depend on the ability of the speakers to sustain sounds at a constant pitch and intensity for a sufficient long time.

Vocal dysperiodicity estimation often relies on the measurement of speech cycle durations and amplitudes of voiced speech segments, or on pitch-synchronous spectral analysis. Vocal dysperiodicity designates any difference in length, amplitude or shape of neighboring speech cycles. This means that the validity of the acoustic measures relies on the accuracy of the measurement of the glottal cycle lengths. In the context of clinical applications, such methods are appropriate for documenting mildly to moderately hoarse voices only (Muta et al., 1988).

Up to now, comparatively few studies have investigated vocal dysperiodicities in continuous speech, even though clinical analysis of connected speech has been propounded early on by Fourcin and Abberton (1971, 1977), for instance. An overview of published studies is given in Table 1. Most of the studies have in common that they involve the detection of individual speech cycles, or individual pseudo-harmonics of the speech spectrum. But, the reliable detection of these is not always warranted in connected speech produced by severely hoarse speakers, resulting in omission and insertion errors biasing numerically the values of acoustic markers.

Qi et al. (1999) have proposed an analysis method of dysperiodicities in disordered speech based on a two-stage linear prediction. The cycle-to-cycle prediction error is assigned to the speech dysperiodicities. The cycle-to-cycle prediction is forward only, giving rise to large prediction errors across phonetic boundaries, boosting the signal-to-dysperiodicity ratio spuriously. The method is therefore not suited for tracking dysperiodicities in speech directly. However, a major advantage of this method is that it does not request isolating speech cycles individually.

To avoid predicting speech samples across phonetic boundaries and to enable speech to be analyzed directly, Bettens et al. (2005) has proposed a bi-directional multi-step predictive approach. Multi-step linear predictive modelling exploits the local periodicity of voiced speech sounds. Indeed, if the speech signal is cyclic and the cycle amplitudes change smoothly, it is possible to predict approximately the present cycle on the basis of some previous or subsequent cycle and the prediction error may be assigned to the cycle-to-cycle dysperiodicity. However, the weights that are involved in a multi-step prediction filter are not constrained to be positive. Multi-step linear predictive analysis may therefore invert the sign of lagged signal fragments, which is inconsistent with the definition of periodicity.

To avoid this inconsistency, Kacha et al. (2006a) has proposed a generalized variogram to estimate speech signal dysperiodicities. The generalized variogram 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 cycle length can be known a priori. A signal-to-dysperiodicity ratio (SDR) that summarizes the dysperiodicities has been shown to correlate strongly with the degree of perceived hoarseness, the correlations being stronger for segmental than global signal-to-dysperiodicity ratios (Jayant and Noll, 1984).

In (Kacha and al., 2006b), two variants of multi-band segmental analysis have been investigated and their performances compared to that of the conventional global signal-to-dysperiodicity ratio. Results have suggested that multi-band segmental signal-to-dysperiodicity ratios correlate better with scores of perceived hoarseness than global ones. However, variogram-based methods have been evaluated so far on a corpus comprising a total of 22 speakers only, normophonic and dysphonic subjects combined (Kacha et al., 2006a, 2006b; Alpan et al., 2007). The purpose of the present study is to evaluate such methods on much larger corpora (a total of over 900 speakers sustaining sounds and producing connected speech). The evaluation is carried out in terms of the ability of acoustic cues that report the size of the estimated vocal dysperiodicities to a) predict perceptual ratings of disordered voices; b) contend with or improve on existing spectral analysis methods; and c) discriminate between normophonic and dysphonic subjects. The substantial number of experimental subjects has enabled carrying out multi-cue as well as multi-frequency band analyses without risking overfitting. Indeed, vocal dysperiodicities have also been analyzed in different spectral bands. Reasons are the lack of knowledge with regard to their band-specific behaviour as well as the need to discover acoustic cues that correlate best with perceptual ratings that are likely to be influenced unequally by different frequency bands. In an earlier feasibility study, we had subdivided the acoustic frequency spectrum into three intervals (Kacha et al., 2006b). But, both the number of bands as well as the number of speakers have been too small to enable drawing firm conclusions. Here we therefore carry out multi-band analysis on much larger corpora by means of a subdivision of the frequency axis that is perceptually-inspired.

The objectives of the experiments that are reported are the following.

First, test whether separately tracking dysperiodicities in different frequency bands enables improving the correlation between measured signal-to-dysperiodicity ratios and perceptual scores of hoarseness, compared to a full-band analysis.

Second, compare the relevance of signal-to-dysperiodicity ratios to a popular cue of dysphonia, which is the cepstral peak prominence (*CPP*) (Table 1) and investigate their inter-correlation.

Third, test whether combining signal-to-dysperiodicity ratios with the cepstral peak prominence enables improving the overall correlation with perceptual scores of hoarseness.

Fourth, test whether the observed differences in correlation are statistically significant. Indeed, these differences are expected to be statistically significant only when the corpora are large because the correlations that are compared are already fairly high (0.7 – 0.8) and therefore comprised in a narrow interval.

Fifth, investigate multi-band signal-to-dysperiodicity ratios via a principal component analysis, because perceptual scoring is not available for all corpora. Principal component analysis is expected to combine linearly acoustic cues so that a small number of combinations "explain" the observed inter-stimuli differences. Given the lack of perceptual scores, the relevance of the combined cues have been expressed numerically via a linear discriminant analysis of the stimuli known to be "normal" and "pathological".

The results show for the first corpus that full-band, multi-band and multi-cue analyses differ statistically significantly with regard to the correlation with perceptual scores. Multi-cue analyses involve a combination of temporal and cepstral features that report signal dysperiodicity and their statistical analysis as a collective. Signal-to-dysperiodicity ratios have been submitted to principal component analysis for the two corpora. The results show that the first two principal components are interpretable in terms of the degree of dysphonia and the spectral slope, respectively. The relevance is confirmed numerically via a discriminant analysis in terms of normophonic and dysphonic speakers.

2. Methods

Sections 2.1 and 2.2 summarize the signal processing that enables estimating cycle-to-cycle dysperiodicities and their recapitulation via their segmental signal-to-dysperiodicity ratio (Kacha et al., 2006a; Alpan et al., 2007).

2.1. Generalized variogram analysis

For a periodic signal $x(n)$ of period T_0, one may write $x(n) = x(n+kT_0),\ k \in \mathbb{Z}$. For a locally-stationary speech signal $x(n)$, the deviation from strict periodicity over an analysis frame of length N can therefore be estimated via expression (1). Index n positions the samples within the frame.

$$\delta = \min_{T} \left\{ \sum_{n=0}^{N-1} \left[x(n) - \alpha x(n+T) \right]^2 \right\}, \quad -T_{\max} < T < -T_{\min} \ , T_{\min} < T < T_{\max} \ , \tag{1}$$

$$\alpha = \sqrt{\frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} x^2(n+T)}} \ . \tag{2}$$

Expression (1) involves the squared difference between a main analysis frame and an auxiliary frame shifted by a signed lag T comprised between ± 2.5 ms and ± 20 ms, (i.e 50 – 400Hz). For each main frame position, lag T is fixed so as to minimize the cumulated squared difference. Signed lags guarantee that, in connected speech, the shift of the auxiliary analysis frame across phonetic boundaries is avoided and only cycles that are internal to a phonetic segment are compared. Indeed, when a speech cycle is near the right boundary of a phonetic segment, the segment-internal cycles are expected to be to its left, that is, lag T is expected to be positive and vice versa for a speech cycle positioned near a left phonetic boundary. For voiced sounds, lag T is expected to be an integer multiple of the glottal cycle length. For unvoiced sounds, (1) can still be meaningfully computed but the interpretation of lag T in terms of glottal cycle lengths is not valid.

In connected 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 α is inserted into (1) to equalize the energies between main and auxiliary analysis frames. The expression between accolades in (1) is known as the variogram (Haslett, 1997) of the speech signal, when local gain α=1.

The analysis frame length is fixed to 2.5 ms, so that main and auxiliary frames do not overlap. The shift between successive main analysis frames is also fixed to 2.5 ms, thus enabling the sample-by-sample dysperiodicity (3) to be computed once and only once for each speech sample. Lag T_{opt} is the signed lag which minimizes energy-equalized variogram (1).

$$e(n) = x(n) - \alpha x(n + T_{opt}) . \qquad (3)$$

In addition, Alpan et al. (2007) have shown that equalizing main and auxiliary analysis frame averages enables discarding parasitic low-frequency transients, such as pop noise owing to the speaker's breath hitting the recording microphone. However, this option has not been implemented because it does not increase for the corpora that are analyzed here the correlation between computed acoustic cues and perceptual scores of hoarseness.

2.2 Segmental signal-to-dysperiodicity ratio

Speech signal *x(n)* as well as the corresponding dysperiodicity trace *e(n)* have been divided into intervals of length *Ls* equal to 5 ms (Kacha et al., 2006a). Then, a local signal-to-dysperiodicity ratio (4) has been computed for each interval.

$$SDR_{loc} = 10 \log \frac{\sum_{n=0}^{Ls-1} x^2(n)}{\sum_{n=0}^{Ls-1} e^2(n)} \tag{4}$$

The segmental signal-to-dysperiodicity ratio *SDRSEG* is obtained by averaging the SDR_{loc}s over all the intervals. The term “segmental” here refers to the subdivision of the signals into 5ms intervals prior to averaging. The segmental signal-to-dysperiodicity ratio has been favored in the framework of the evaluation of lossy speech coders, because it appears to correlate better with human-assigned scores of perceived quality. A possible explanation is that segmental ratios boost the contribution of short noisy segments, which seem to influence perceived timbre strongly (Jayant and Noll, 1984).

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 and segmental signal-to-dysperiodicity ratios (4) have been computed for each band. The ranges of the four mel bands (B1 – B4) have been (0 – 800 mel), (800 – 1600 mel), (1600 – 2400 mel) and beyond (Stevens et al., 1937). These mel-intervals correspond to the frequency bands (0 – 724 Hz), (724 – 2195 Hz), (2195 – 5188 Hz) and beyond. The filterbank has been designed by means of the Parks-McClellan method (Oppenheimer and Schafer, 1975). Figure 1 shows the frequency responses of the four-channel filterbank.

2.4 Cepstral analysis and cepstral peak prominence

The cepstral peak prominence (*CPP*) is a measure of the log-amplitude of the first rhamonic of the speech cepstrum (Hillenbrand and Houde, 1996). Usually, the speech cepstrum is defined as the inverse magnitude spectrum of the log-magnitude spectrum (Oppenheimer and Schafer, 1975).

The calculation of *CPP* involves the following steps.

1. Obtainment of the speech cepstrum for an analysis frame length of 2048 samples (for a sample frequency equal to 44.1 kHz). The shift between successive frames equals 10 ms.
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.

Cepstral peak prominences have been obtained by means of Hillenbrand's *CPPS* software (http://homepages.wmich.edu/~hillenbr/cpps.exe). Figure 2 shows the *CPP* computed for a clean and disordered voice, respectively.

2.5. Corpora

2.5.1. Dutch corpus

This first corpus has comprised sustained vowels [a] and two Dutch sentences ("Papa en Marloes staan op het station. Ze wachten op de trein.") produced by 28 normophonic and 223 speakers with different degrees of dysphonia. Diagnoses 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). The voiced segments of the two sentences have been extracted following Parsa and Jamieson (2001) and concatenated. From these, two additional artificial-stimuli corpora have been formed. The first sub-corpus comprises concatenations of the voiced segments followed by vowel [a]. The second comprises concatenations of the full sentences followed by vowel [a]. Except the sub-corpus involving voiced segments only that has been sampled at 22050 Hz, all the stimuli have been sampled at 44100 Hz. Five judges have evaluated the sub-corpus involving the concatenation of the full sentences and vowel [a] perceptually. The five judges have been professional voice therapists with at least five years of experience in clinical voice quality ratings. Each judge has rated, from 0 to 3, the item “grade” of the (G)RABS scale. "Grade" represents the degree of hoarseness or voice abnormality (Hirano, 1981). The five perceptual scores per stimulus have been averaged. Recordings, segmentation and perceptual evaluation have been carried out at the Sint-Jan General Hospital, Bruges, Belgium (Maryn et al., 2009).

2.5.2. *MEEI corpus*

The second corpus has been the Kay Elemetrics Voice Disorder Database developed by the Massachusetts Eye and Ear Infirmary (*MEEI*) Voice and Speech Labs (Kay Elemetrics Corp., 1994). This corpus comprises 53 normophonic and over 650 pathological utterances. The acoustic samples are sustained phonations of vowel [a] (3 - 4 s long) and the first 12 seconds of the Rainbow Passage (Fairbanks, 1960) (661 stimuli) spoken by normophonic subjects and patients with organic, neurological, traumatic, and psychogenic voice disorders at different stages (from early to fully developed). The speech samples have been recorded in a controlled environment at 25 kHz and 16 bits of resolution. Hereafter, the analyses are carried out on the 12-second continuous Rainbow Passage utterances. No perceptual assessment has been available.

2.6. Statistical Analysis

2.6.1. *Full-band and single-cue analyses*

Linear regression analyses have been carried out to predict the degree of perceived hoarseness via the segmental signal-to-dysperiodicity ratios in the whole frequency band (full-band analysis) first, and via the cepstral peak prominence (single-cue analysis) second. This analysis has been carried out on the four sub-corpora of the Dutch corpus.

2.6.2. 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 signal-to-dysperiodicity ratios in different frequency bands (1 to 4). This analysis has been carried out on the sub-corpus concatenating two Dutch sentences and sustained [a], for which a perceptual evaluation has been available.

2.6.3. Multi-cue linear regression analysis

Also, 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 later. This analysis has been carried out on the sub-corpus concatenating two Dutch sentences and sustained [a].

2.6.4. Principal component analysis and linear discrimination analysis

A principal component analysis has been carried out on 3 segmental signal-to-dysperiodicity ratios (the three lowest bands) of the MEEI corpus comprising normal and pathological utterances (12-second Rainbow Passage) (Jolliffe, 2002), for which no perceptual evaluation has been available. The *SDRSEG*s cues have been z-normalized prior to analysis. To assess the generality of the principal component representation, the analysis has also been carried out on the *SDRSEG*s obtained for the first three frequency bands for the sub-corpus concatenating two Dutch sentences and sustained [a]. A linear discriminant analysis has been carried out to assess numerically the discrimination performance of the first two principal components. Here, linear discrimination analysis is carried out to confirm the relevance of the principal component analysis.

3. Results and discussion

3.1 Correlation between acoustic cues and human-assigned perceptual scores

3.1.1 Full-band segmental signal-to-dysperiodicity ratio and cepstral peak prominence

Table 2 shows the values of the Pearson product-moment correlation ρ_P between the average perceived grade *G* (hoarseness) and the full-band segmental signal-to-dysperiodicity ratio (SDRSEG) for the four sub-corpora of the Dutch corpus. The null hypothesis ($\rho_P = 0$) has been rejected for each (two-tailed t-test, $\rho_{crit} = 0.21$, p < 0.001, t = 12.8 – 15.5). The strongest (absolute) correlation ($\rho_P = 0.7$) is observed for the sub-corpus concatenating Dutch sentences with vowel [a], for which the perceptual ratings had been carried out in the first place . The same table also shows the values of correlation ρ_P between the average perceived grade and the cepstral peak prominence (CPP). The null hypothesis ($\rho_P = 0$) has been rejected for each sub-corpus (two-tailed test, $\rho_{crit} = 0.21$, $p<0.001$, $t = 10.7 – 15.5$). Again, the correlation is strongest ($\rho_P = 0.7$) for the Dutch sub-corpus concatenating two sentences with vowel [a].

3.1.2 Multi-band analyses and contribution of different spectral bands to perceived hoarseness

Multiple cues obtained in different frequency bands or domains (temporal versus cepstral) have been combined to investigate whether linear combinations of cues enable improving the correlation of the combined cues with human-assigned perceptual scores. Given the size of the corpora, the fitting of multiple cues may be carried out safely without risking overfitting.

The multi-band analysis has been carried out for the sub-corpus concatenating Dutch sentences with vowel [a], because it has previously given rise to the strongest correlation with perceptual ratings of hoarseness. Only the *SDRSEG* cues for the first three bands have been entered into the linear regression analysis, which has been stepwise. Indeed, in the fourth band the correlation of the *SDRSEG* cue with the perceptual ratings has been close to zero. In that band, the harmonics are masked by noise and the *SDRSEG* values are low for most speakers.

Table 3 reports standardized regression coefficients β_i of the segmental signal-to-dysperiodicity ratios in frequency bands B1 to B3. The adjusted R^2 (0.58) is a conservative estimate of the percentage of the variance of the perceptual scores that can be attributed to the combined band-limited signal-to-dysperiodicity ratios. The multiple correlation R is statistically significant ($R_{crit} = 0.176$, $p < 0.05$, $F = 113.6$). The values of correlation R of perceptual scores with *SDRSEG*s differ significantly between full-band and multi-band analyses (two-tailed t-test, $t = 3.78$, $p < 0.001$) (Dunn and Clark, 1969; Hotelling, 1940). Figure 3 reports the same data graphically. It shows on the horizontal axis the average grade scores predicted by means of the *SDRSEG* cues for the first three frequency bands and on the vertical axis the average human-assigned scores.

One other objective has been to investigate whether different spectral bands contribute unequally to the perceived degree of hoarseness (grade G). In Table 3, the weights of the *SDRSEG* cues are shown to decrease for high-frequency bands and in Table 4 the first frequency band is shown to contribute most to the prediction of the perceptual scores via segmental signal-to-dysperiodicity ratios. Table 4 indeed displays (multiple) correlation coefficients R obtained for *SDRSEG*s for one (B1), two (B1, B3) and three bands (B1, B2, B3), which are in the order of the stepwise linear regression. Correlation R increases statistically significantly from 0.71 (for the first frequency band) to 0.76 (for the first three frequency bands) (two-tailed t-test, $t = 3.30$, $p < 0.001$) (Dunn and Clark, 1969; Hotelling, 1940).

This may suggest that most of the perceptually relevant information is comprised in the first frequency band (B1). A possible explanation is that because the spectral energy decreases with frequency owing to the spectral tilt, listener judgement is influenced most by the more intense frequency intervals that perceptually mask weaker ones. This is confirmed by Figure 4 that reports the variation of the correlation coefficient between perceptual scores and segmental signal-to-dysperiodicity ratios as a function of the upper cut-off frequency of the first band, which evolves from 200 Hz to 3200 Hz. The correlation reaches a plateau near 1000 Hz, which confirms that the first frequency band (B1) contributes most to the prediction of the perceptual scores.

3.1.3 Multi-domain (temporal versus cepstral) analyses

An additional multi-cue linear regression analysis has been carried out with a view to predicting perceptual scores by means of the cepstral peak prominence and the segmental signal-to-dysperiodicity ratio in the first frequency band. These two cues have been retained because their inter-correlation ($\rho_P = 0.61$) is moderate and the first frequency band contributes most to the prediction of the perceptual scores (Table 4).

Weights β_i reported in Table 5 suggest that cepstral peak prominence and segmental signal-to-dysperiodicity ratio contribute roughly equally to the prediction of the perceptual scores. The correlation between predicted and human-assigned perceptual scores increases from 0.7 and 0.71 (Tables 2 and 4) for the individual to 0.79 (Table 5) for the combined cues. The increase is statistically significant ($R_{crit} = 0.155$, $p < 0.05$, $F = 200.4$). In addition, Table 6 summarizes results of t-tests comparing the correlation between human-assigned scores and scores predicted via the combined *SDRSEG* and *CPP* to the correlation between human-assigned scores and scores predicted via single-band and multi-band *SDRSEGs* as well as *CPP* alone. The differences are statistically significant.

To sum up, regression analyses involving several cues increase observed correlations with perceptual scores. The increase is from 0.70 for full-band analysis to 0.76 for multi-band analysis (Table 4), and to 0.79 for multi-cue analysis involving the first-band *SDRSEG* and *CPP* (Table 5). The increase is expected because of the good correlation with perceptual scores ($\rho_P = 0.70$ or 0.71) of each cue individually (*SDRSEG* and *CPP*) and the moderate-only inter-correlation ($\rho_P = 0.61$) between the two.

To illustrate that equal values of correlation ρ_P may be interpreted differently, Figure 5 reports on the vertical axis the average perceptual scores of grade G assigned by human judges and on the horizontal axis the same scores predicted by linear regression of single-band *SDRSEG*, *CPP* and the combined *SDRSEG* and *CPP*. When comparing the scatter of samples in the two top plots in Figure 5, one notices that the trend is less non-linear for *CPP* and less dispersed around the bisector for *SDRSEG*. Figure 5.c suggests that the increased correlation is the outcome of a trend that is less non-linear blended with a decrease of the dispersion of the data points around the bisector. The combination of cepstral peak prominence and signal-to-dysperiodicity ratio appears to possess to a higher degree a property that linear regression analysis is used to detect, i.e. the preferential clustering of data points in the vicinity of a straight line. Therefore, we do not conclude at this stage that both cues are genuinely complementary, loose from any assumption of linearity.

3.2 Principal component and linear discrimination analyses of multi-band signal to dysperiodicity ratios

Principal component analysis has been carried out on the multi-band *SDRSEG* cues of the 12-second Rainbow Passage (MEEI corpus) for which no perceptual rating has been available. Table 7 shows the results of the analysis applied to the *SDRSEG*s of the first three frequency bands. Eigenvalues as well as individual and cumulative variances are shown to the left. Coefficients of the linear combinations of the z-normalized *SDRSEG*s are shown to the right. More than ninety percent of the total variance are explained by the first two principal components PC_1 and PC_2, which are respectively interpreted as the negative of the average of the z-normalized *SDRSEG*s and the difference between the z-normalized *SDRSEG*s in frequency bands B3 and B1. When the second versus the first principal component is reported in a 2D graph, normal and pathological utterances tend to cluster separately (Figure 6).

The crescent shape of the graph can be interpreted in terms of the spectral slopes of the speech spectra and the overall degree of vocal disturbance. To illustrate, Figure 7 shows four vowel [a] spectra that correspond to stimuli with $PC_1 < 0$ and $PC_2 < 0$, $PC_1 < 0$ and $PC_2 > 0$, $PC_1 > 0$ and $PC_2 < 0$, and $PC_1 > 0$ and $PC_2 > 0$. The spectra of speech and dysperiodicity traces are overlaid. When comparing spectra in the left (PC_1<0) (Figure 7.a and Figure 7.c) and right columns (PC_1>0) (Figure 7.b and Figure 7.d), one sees that spectra to the left correspond to normal voices. Indeed, their harmonic structure is well defined and the dB level of the speech spectrum is higher than the dB level of the dysperiodicity spectrum. This agrees with the interpretation of the first principal component as an average that reports the overall degree of dysperiodicity.

When comparing the spectra at the top ($PC_2 > 0$) (Figures 7.a and 7.b) and bottom ($PC_2 < 0$) (Figures 7.c and 7.d), one sees that in the latter the harmonics decrease more rapidly with frequency than in the former. This suggests that the second principal component depends on the spectral slope. Indeed, the spectral slopes for $PC_2 < 0$ are steeper (the slope is larger in absolute value) than for $PC_2 > 0$. This observation agrees with the interpretation of principal component 2 as a difference between the *SDRSEG*s in frequency bands B3 and B1. Signal-to-dysperiodicity ratios appear to report on the spectral slope indirectly via a more rapid decrease of the signal-to-dysperiodicity ratio in the higher bands when the slope is steep.

To assess the generality of the principal component representation, the same analysis has been carried out on the *SDRSEG*s obtained for the first three frequency bands for the Dutch sentences & vowel [a] sub-corpus. Figure 8 shows the principal component representation and, superimposed, the four grade intervals. The same crescent shape and clustering of clean stimuli to the left of the diagram are observed.

In addition, a linear discrimination analysis has been carried out to assess numerically the discrimination performance of the first two principal components obtained for the "Rainbow passage" corpus. Classification results are shown in Table 8. Out of 661 pathological stimuli, 596 have been correctly classified as pathological and 65 have been misclassified as normal. Similarly, out of 53 normal stimuli, 46 have been correctly classified as normal and 7 have been misclassified as pathological. Thus, the overall classification accuracy is 89.9%. This suggests that principal components combine dysperiodicity cues in a clinically meaningful way. One should however keep in mind that correct classification rates reported for the MEEI corpus are often higher than those reported for other corpora.

4. Conclusion

Generalized variogram and cepstral analyses has been used to estimate vocal dysperiodicities in disordered connected speech for two corpora with over 250 and 650 speakers respectively. The results are the following:

1. Multi-band segmental signal-to-dysperiodicity ratios correlate more strongly with the perceptual assessment of the degree of hoarseness than the full-band one.
2. The low frequency band contributes most to the prediction of perceptual scores.
3. The correlations between perceptual scores and cepstral peak prominence or full-band segmental signal-to-dysperiodicity ratio are similar.
4. Linear regression analysis of the low-frequency band segmental signal-to-dysperiodicity ratio combined with the cepstral peak prominence gives the best correlation with perceptual scores. The reason is the combined decrease of dispersion of the data points owing to the signal-to-dysperiodicity ratio and the increased linearity owing to the cepstral peak prominence.
5. The representation in terms of the first two principal components gives rise to a crescent-shape scatter of the speech samples. The first principal component reports on the overall degree of hoarseness and the second on the spectral slope.
6. Principal components combine dysperiodicity cues in a clinically meaningful way, which is confirmed by the overall classification accuracy of 89.9% of the speech stimuli in normal and disordered categories.

5. Acknowledgements

This research was supported by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme. We are grateful to Dr. Marc Remacle, Department of

Otorhinolaryngology and Head and Neck Surgery, University Hospital of Louvain at Mont-Godinne, Belgium, for providing the Kay Elemetrics Database.

6. References

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

Alpan, A., Schoentgen, J., Maryn, Y., Grenez, F., Murphy, P., 2009. Cepstral Analysis of Vocal Dysperiodicities in Disordered Connected Speech. in Proc. Interspeech, Brighton (U.K.), pp. 959-962.

Awan, S.N., Roy, N., 2005. Acoustic prediction of voice type in women with functional dysphonia. J. Voice 19, 268-282.

Bettens, F., Grenez, F., Schoentgen, J., 2005. Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis. J. Acoust. Soc. Am. 117, 328-337.

Dolansky, L., Tjernlund, P., 1968. On certain irregularities of voiced speech waveforms. IEEE Trans. Audio Electroacoust. 16, 51-56.

Dunn, O.J., Clark, V.A., 1969. Correlation coefficients measured on the same individuals. J. Am. Stat. Assoc. 64, 366–377.

Edgar, J.D., Sapienza, C.M., Bidus, K., Ludlow, C.L., 2001, Acoustic Measures of Symptoms in Abductor Spasmodic Dysphonia. J. Voice 15, 362-372.

Fairbanks, G., 1960. Voice and articulation drill book, New York: Harper Brothers.

Fourcin, A.J., Abberton, E., 1971. First applications of a new laryngograph. Medical and Biological Illustration 21, 172-182.

Fourcin A.J., Abberton E., 1977. Laryngograph studies of vocal fold vibration. Phonetica 34, 313-315.

Fredouille, C., Pouchoulin, G., Ghio, A., Revis, J., Bonastre, J.-F., Giovanni, A. Back-and-Forth Methodology for Objective Voice Quality Assessment: From/to Expert Knowledge to/from Automatic Classification of Dysphonia. Eurasip J. Adv. Sig. Proc. 2009.

Halberstam, B. 2004. Acoustic and Perceptual Parameters Relating to Connected Speech Are More Reliable Measures of Hoarseness than Parameters Relating to Sustained Vowels, ORL 66, 70-73.

Haslett, J., 1997. On the sample variogram and sample auto-covariance for non-stationary time series. The Statistician 46, 475–485.

Hecker, M.H.L., Kreul, E.J., 1971. Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency. J. Acoust. Soc. Am. 49, 1275-1282.

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

Heman-Ackah, Y.D., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., Hillenbrand, J., Sataloff, R.T., 2003. Cepstral peak prominence: a more reliable measure of dysphonia. Ann. Otol. Rhinol. Laryngol. 112, 324–333.

Heman-Ackah, Y.D., 2004. Reliability of calculating the cepstral peak without linear regression analysis. J. Voice 18, 203-208.

Hernandez-Espinosa, C., Gomez-Vilda, P., Godino-Llorente, J.I., Aguilera-Navarro, S., 2000. Diagnosis of vocal and voice disorders by the speech signal. in: Proc. IEEE Int. Joint Conf. on Neural Networks, Como (Italy), pp. 253-258.

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

Hinkle, D.E., Wiersma, W., Jurs, S.G., 1998. Applied statistics for the behavioral sciences, New York: Houghton Mifflin Company.

Hirano, M., 1981. Clinical Examination of Voice, Springer, Wien.

Hotelling, H., 1940. The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters, Ann. Math. Stat. 11, 271–283.

http://homepages.wmich.edu/~hillenbr/cpps.exe

Jayant, N.S., Noll, P., 1984. Coding of waveforms: principles and applications to speech and video, Prentice-Hall, Englewood Cliffs.

Jolliffe, I., 2002. Principal Component Analysis (second ed), Springer, Berlin.

Kacha, A., Grenez, F., Schoentgen, J., 2006a. Estimation of dysperiodicities in disordered speech. Speech Com. 48, 1365-1378.

Kacha, A., Grenez, F., Schoentgen, J., 2006b. Multiband frame-based acoustic cues of vocal dysperiodicities in disordered speech. Biomed. Sig. Proc. 1, 137-146.

Kay Elemetrics Corp., Disordered voice database, Ver. 1.03, 1994.

Klingholtz, F., 1990. Acoustic recognition of voice disorders: a comparative study of running speech versus sustained vowels. J. Acoust. Soc. Am. 87, 2218-2224.

Klingholz, F., 1987. The measurement of the signal-to-noise ratio in continuous speech. Speech Com. 6, 1-12.

Laver, J., Hiller, S., Mackenzie, J., Rooney, E., 1986. An acoustic screening system for the detection of laryngeal pathology, J. Phonetics 14, 517-524.

Lieberman, P., 1963. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. J. Acoust. Soc. Am. 35, 344-353.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M., 2009. Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels, J. of Voice. 2009.

Murry, T., Doherty, E.T., 1980. Selected acoustic characteristics of pathologic and normal speakers. J. Speech Hear. Res. 23, 361-369.

Muta, H., Baer, T., Wagatsuma, K., Muraoka, T., Fukuda, H., 1988. A pitch-synchronous analysis of hoarseness in running speech. J. Acoust. Soc. Am. 84, 1292-1301.

Oppenheimer, A.V., Schafer, R.W., 1975. Digital Signal Processing, Prentice-Hall, Englewood Cliffs, New Jersey.

Parsa, V., Jamieson, D G., 2001. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech, J. Speech, Lang. Hear. Res. 44, 327-339.

Parsa, V., Jamieson, D.G., Stenning, K., Leeper, H.A., 2002. On the estimation of signal-to-noise ratio in continuous speech for abnormal voices. in: Proc. Int. Conf. Spoken Lang. Process, Denver (USA), pp. 2505–2508.

Qi, Y., Hillman, R.E., Milstein, C., 1999. The estimation of signal-to-noise ratio in continuous speech of disordered voices. J. Acoust. Soc. Am. 105, 2532-2535.

Sapienza, C. M., Cannito, M. P., Murry, T., Branski, R., 2002. Acoustic Variations in Reading Produced by Speakers With Spasmodic Dysphonia Pre-Botox Injection and Within Early Stages of Post-Botox Injection, J. Speech, Lang. Hear. Res. 45, 830-843.

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

Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8, 185-190.

Umapathy, K., Krishnan, S., Parsa, V., Jamieson, D.G., 2005. Discrimination of pathological voices using a time-frequency approach. IEEE Trans. Biomed. Eng. 52, 421-430.

Yiu, E., Worrall, L., Longland, J., Mitchell, C. 2000. Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. Clinical linguistics & phonetics 14, 295-305.

Figure 1: Frequency response of the mel-spaced filter bank used to decompose signal and dysperiodicity traces in four separate frequency bands.

Figure 2: Cepstral peak prominence for a normal (above) and pathological voice (below)

Figure 3: Original average perceived grade scores versus predicted average perceived grade scores (via the SDRSEG cues for the first three frequency bands).

Figure 4: Correlation between perceptual scores and SDRSEG for the first frequency band as a function of the upper cut-off frequency.

Figure 5: Original average perceived grade scores versus predicted average perceived grade scores via (a) SDRSEG in the first frequency band, (b) CPP, (c) linear combination of SDRSEG (in band 1) and CPP.

Figure 6: Principal components representation for the Rainbow Passage corpus.
Figure 7: Speech and dysperiodicity spectra obtained for [a] fragments. In black: speech spectrum, in grey: dysperiodicity spectrum. Vertical lines: boundaries between mel-spectral intervals.
Figure 8: Principal component representation for the full sentences with vowel [a] (Dutch corpus). The grey levels of the data points report on the perceived degree of abnormality (grade).

Table 1: Overview of published studies devoted to the acoustic characterization of connected speech produced by dysphonic speakers (man = manual, aut = automatic; te = temporal, sp = spectral, ce = cepstral, SNR = signal-to-noise ratio, CPP = cepstral peak prominence, NHR = noise-to-harmonic ratio, LTAS = long-term average spectrum, HNR = harmonic-to-noise ratio, SDR = signal-to-dysperiodicity ratio, R1 = amplitude first rahmonic).

Authors	Corpus size	Processing	Analysis domain	Dysperiodicity cues
Lieberman (1963)	31	man	te	cycle duration variation
Dolansky and Tjernlund (1968)	10	man	te	F0, intonation, ..
Hecker and Kreul (1971)	28	man	te	F0
Laver et al. (1986)	230	aut	te	jitter, shimmer
Klingholz (1987)	101	aut	sp	SNR
Muta et al. (1988)	6	man, aut	sp	SNR
Hillenbrand and Houde (1996)	25	man, aut	te, sp, ce	signal dysperiodicity (incl. CPP) spectral tilt size of first harmonic
Qi et al. (1999)	87	aut	te	SNR
Hernandez et al. (2000)	281	aut	te	subset of MDVP
Yiu et al. (2000)	30	aut	te, sp	subset of MDVP
Parsa and Jamieson (2001)	228	aut	te, sp	jitter, shimmer, LTAS-based cues, HNR, linear prediction-based cues
Parsa et al. (2002)	18	aut	te, sp	multi-band SNR
Heman-Ackah et al. (2002)	36	aut	te, sp, ce	NHR, jitter, shimmer, CPP
Heman-Ackah et al. (2003)	281	aut	te, sp, ce	NHR, jitter, shimmer, CPP
Halberstam (2004)	60	aut	te, sp, ce	CPP, subset of MDVP
Heman-Ackah (2004)	150	man, aut	ce	NHR, jitter, shimmer, CPP
Awan and Roy (2005)	134	aut	te, sp	CPP, jitter, shimmer, F0
Bettens et al. (2005)	22	aut	te	SDR
Umapathy et al. (2005)	212	aut	te	time-frequency analysis-based cues
Kacha et al. (2006a)	22	aut	te	SDR
Kacha et al. (2006b)	22	aut	te, sp	multi-band SDR
Alpan et al. (2007)	22	aut	te	SDR
Alpan et al. (2009)	251	aut	ce	R1

Fredouille et al. (2009)	80	aut	sp	linear- and mel-frequency spectral coefficients

Table 2: Pearson's correlation coefficients between average grade scores and cues *SDRSEG* and *CPP* for sustained vowel [a], sentences, and the concatenations of voiced fragments or full sentences with vowel [a] (Dutch corpus).

	[a]	Sentences	Voiced segments & [a]	Full sentences & [a]
SDRSEG	-0.63	-0.64	-0.65	-0.70
CPP	-0.56	-0.69	-0.63	-0.70

Table 3: Stepwise linear regression analysis carried out on segmental signal-to-dysperiodicity ratios for three frequency bands, β_i = standardized regression coefficients, R = multiple correlation coefficient (Dutch corpus).

β_1	β_2	β_3	R	R^2	Adj. R^2
-0,574	-0,171	-0,154	0,761	0,579	0,575

Table 4: Multiple correlation coefficients obtained from stepwise linear regression analysis carried out on segmental *SDR*s for one (B1), two (B1, B3) and three bands (B1, B2, B3) (Dutch corpus).

B1	B1, B3	B1, B2, B3
0,71	0,75	0,76

Table 5: Linear regression analysis involving cues *SDRSEG* (first band) and *CPP*, βi = standardized regression coefficients, R = multiple correlation coefficient (Dutch corpus).

β_{SDRSEG}	β_{CPP}	R	R^2	Adj. R^2
-0,46	-0,42	0,79	0,62	0,615

Table 6 : t-tests comparing the correlation with perceptual grade scores obtained for multi-cue regression analysis (*SDRSEG*$_1$ and *CPP*) to correlations obtained for full-band variogram analysis, cepstral peak prominence and multi-band variogram analysis respectively (Dutch corpus).

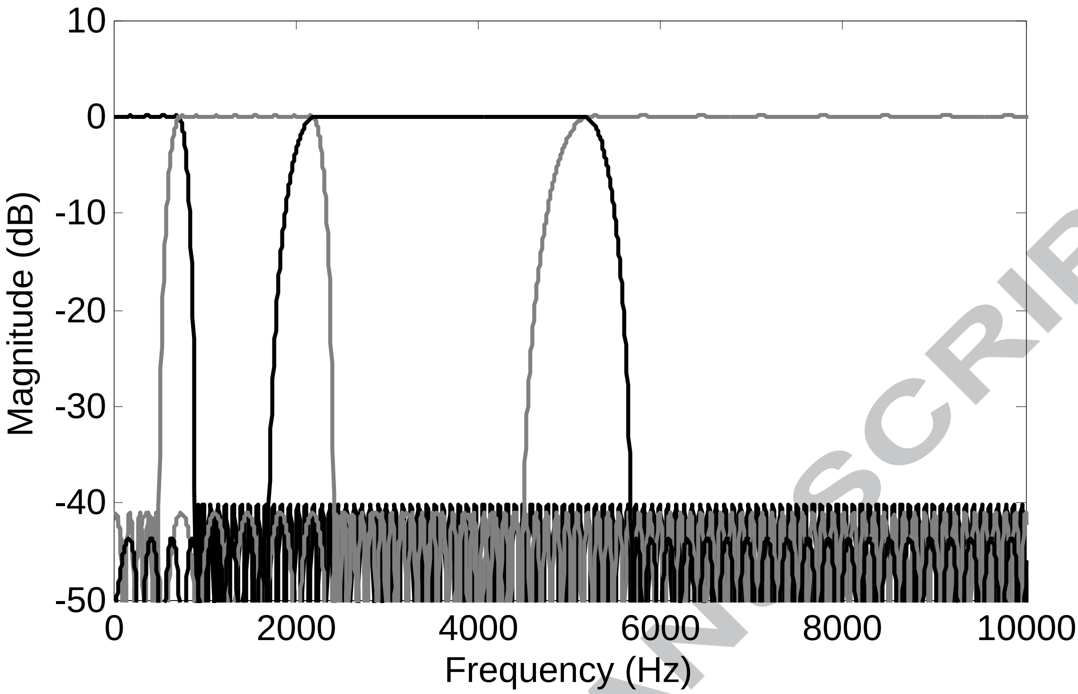
	Full-band *SDRSEG*	*CPP*	Multi-band *SDRSEG*
Multi-cue (*SDRSEG*$_1$ and *CPP*)	t=5.47, p<0.001	t=4.72, p<0.001	t=2.69, p<0.01

Table 7 : Results of the principal component analysis applied to the segmental signal-to-dysperiodicity ratios (*SDRSEG*) in the first three frequency bands for the MEEI corpus (12-second Rainbow Passage).

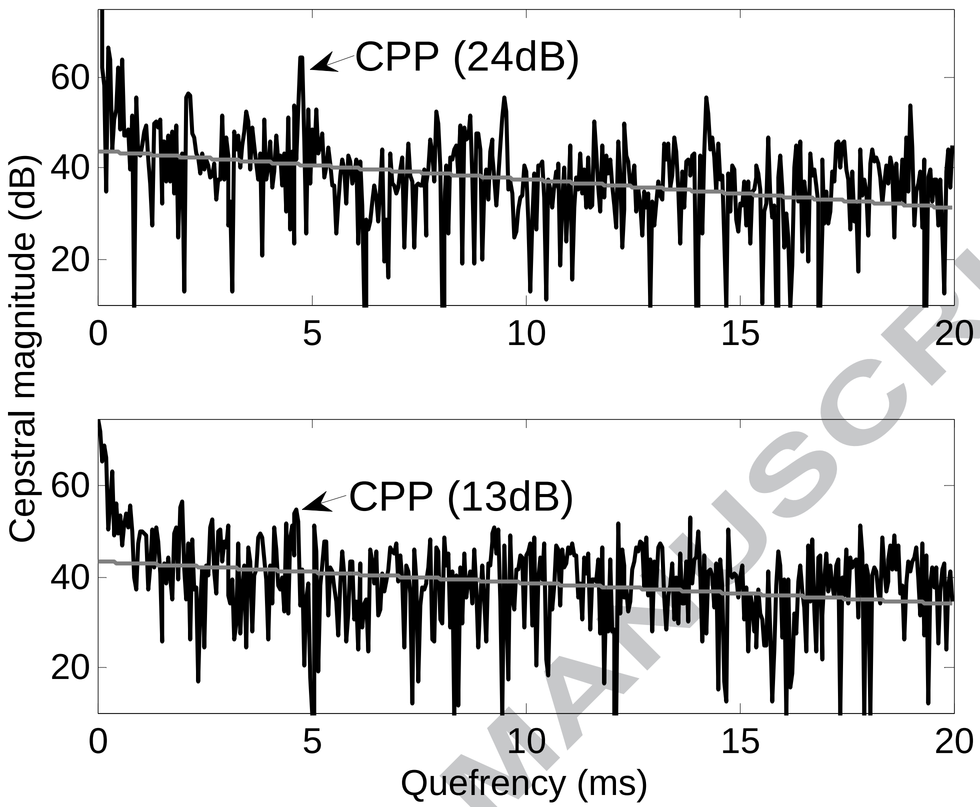
Component	Eigenvalues	Variance (%)	Cumulative variance (%)	Coefficients		
				Band 1	Band 2	Band 3
1	2.00	66.8	66.8	-0.56	-0.66	-0.50
2	0.79	26.1	92.9	-0.64	-0.05	0.77
3	0.21	7,1	100,0	-0.53	0.75	-0.40

Table 8 : Linear discrimination classification based on the first two principal components obtained for the MEEI corpus (Rainbow Passage). The overall classification accuracy is 89.9%.

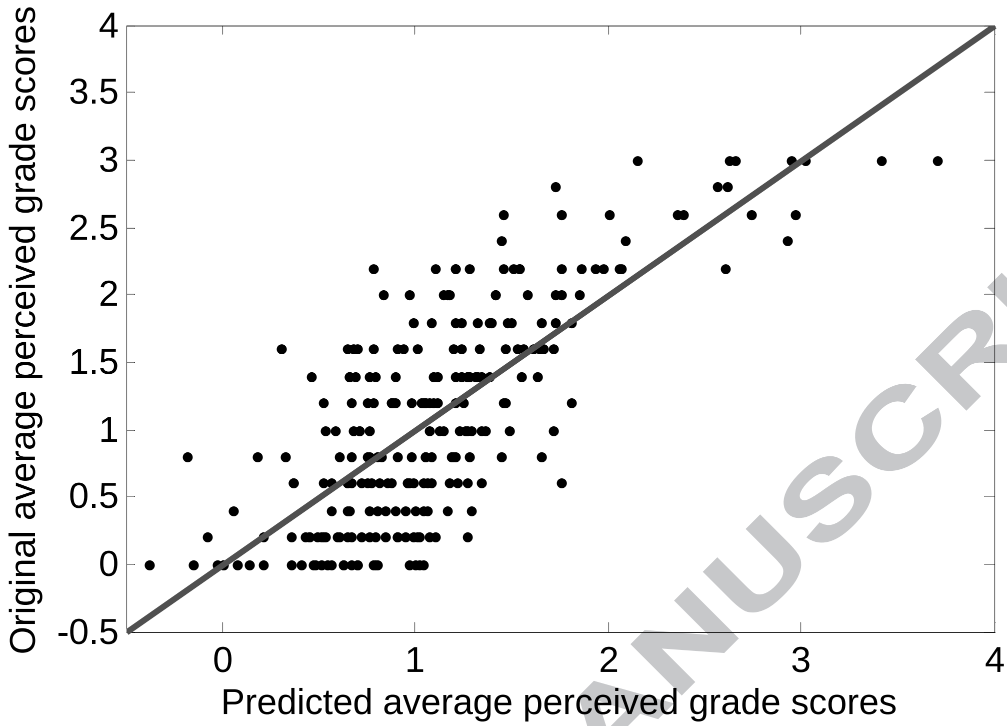
			Predicted		
			Pathological	Normal	Total
Original	Count	Pathological	596	65	661
		Normal	7	46	53
	%	Pathological	90,2	9,8	100,0
		Normal	13,2	86,8	100,0

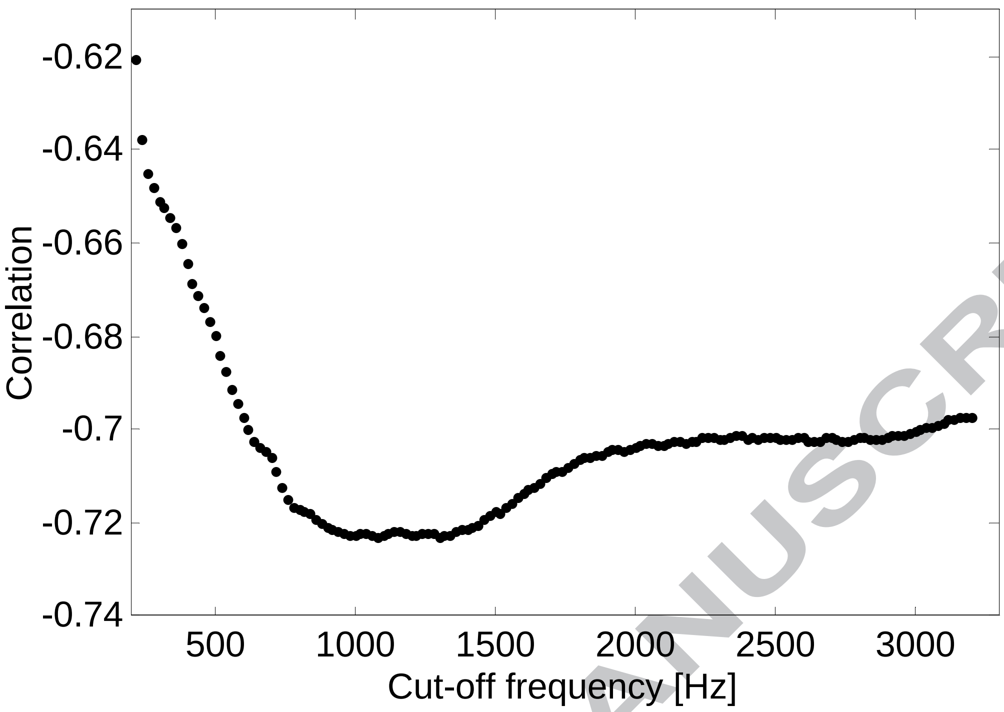


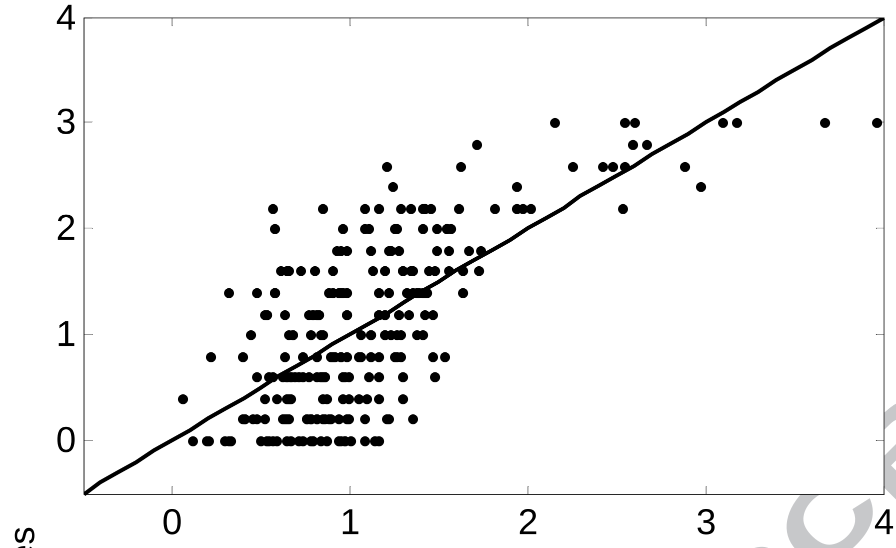
10
0
-10
-20
-30
-40
-50
Magnitude (dB)
0
2000
4000
6000
8000
10000
Frequency (Hz)



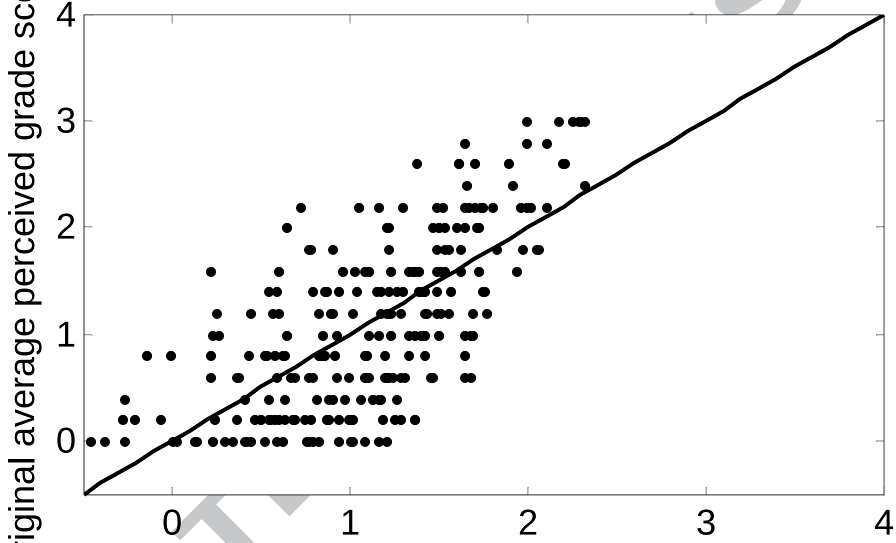
CPP (24dB)
CPP (13dB)
Cepstral magnitude (dB)
Quefrency (ms)
20
40
60
0
5
10
15
20


Original average perceived grade scores
Predicted average perceived grade scores
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
0
1
2
3
4

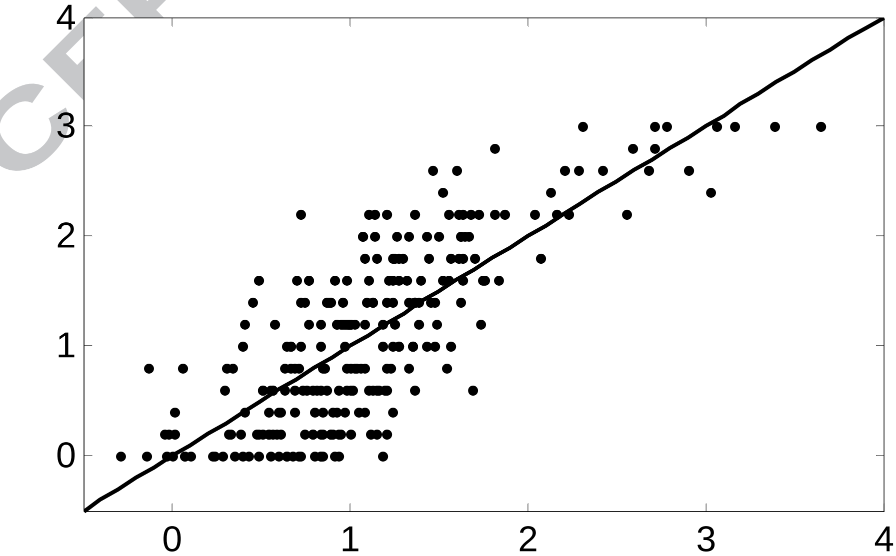

-0.62
-0.64
-0.66
-0.68
-0.7
-0.72
-0.74
Correlation
500
1000
1500
2000
2500
3000
Cut-off frequency [Hz]



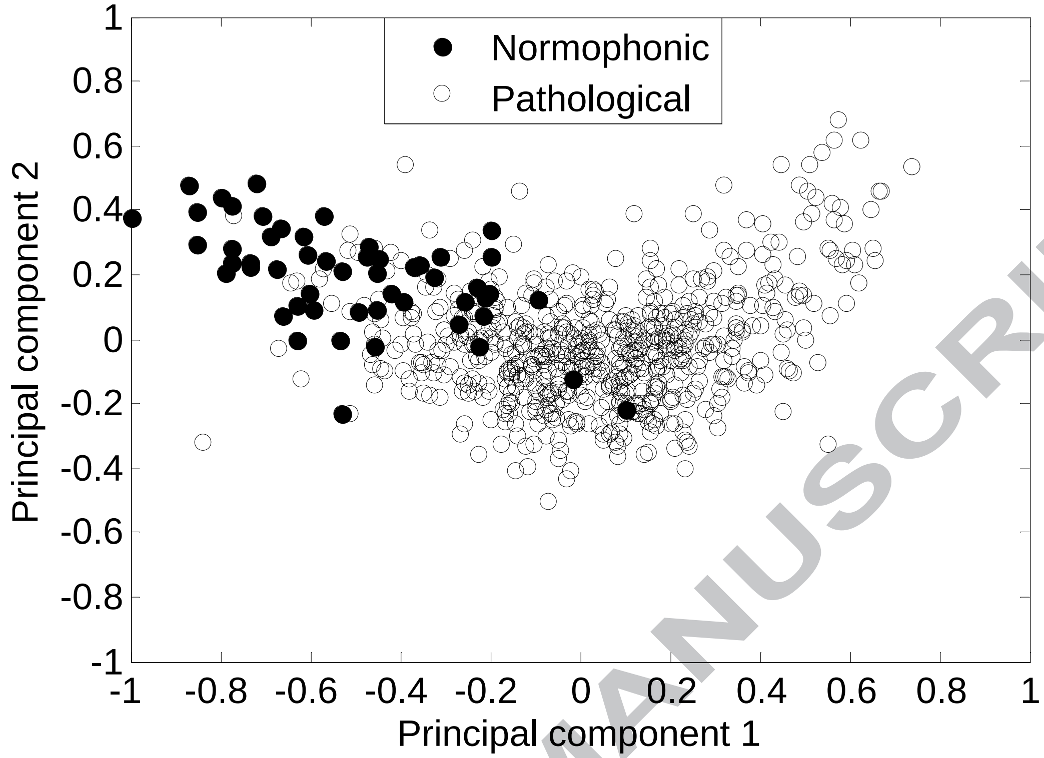
(a) Predicted scores (SDRSEG)

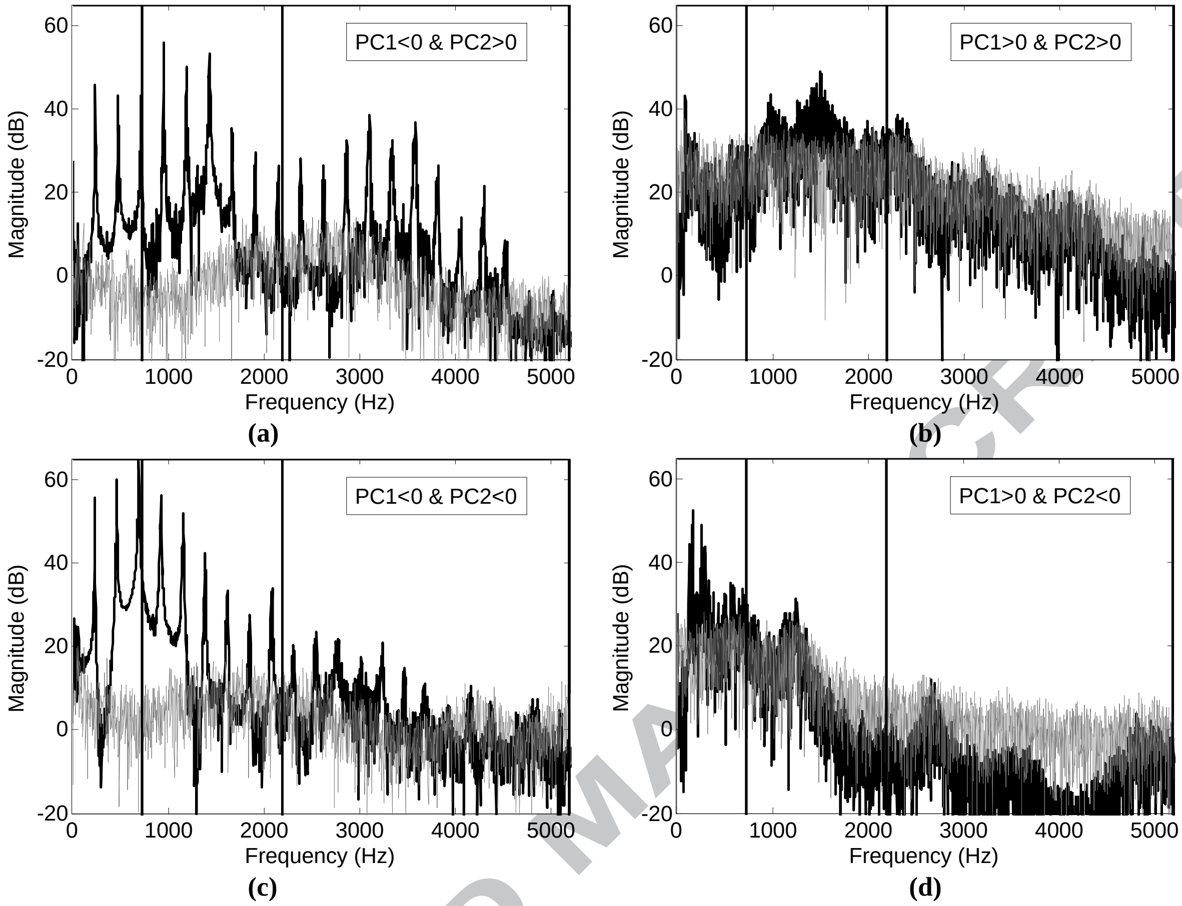


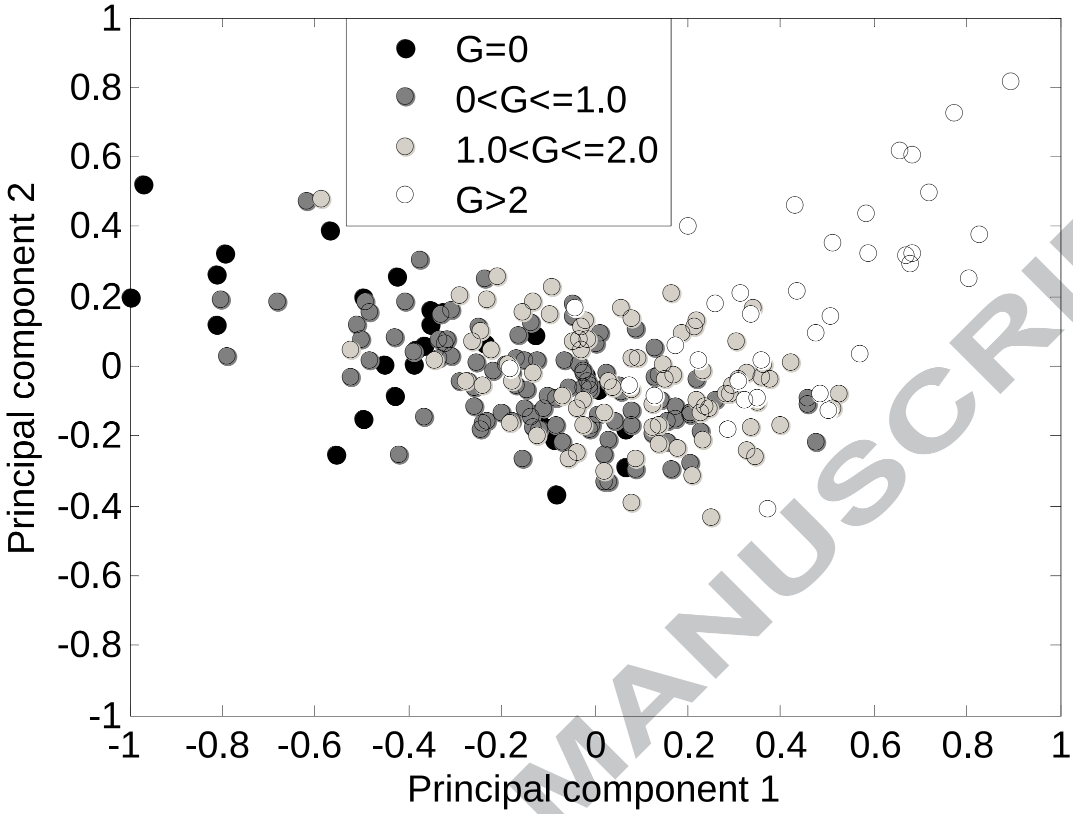
(b) Predicted scores (CPP)



(c) Predicted scores (SDRSEG and CPP)


Normophonic
Pathological
Principal component 2
Principal component 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


PC1<0 & PC2>0
PC1>0 & PC2>0
PC1<0 & PC2<0
PC1>0 & PC2<0
Magnitude (dB)
Frequency (Hz)
60
40
20
0
-20
0
1000
2000
3000
4000
5000
(a)
(b)
(c)
(d)


G=0
0<G<=1.0
1.0<G<=2.0
G>2
Principal component 1
Principal component 2