

# Cepstral peak prominence: A comprehensive analysis

Rubén Fraile\*, Juan Ignacio Godino-Llorente

*Circuits & Systems Engineering Department, ETSIS Telecomunicación, Universidad Politécnica de Madrid, Campus Sur, Carretera de Valencia Km.7, 28031 Madrid, Spain*

## A B S T R A C T

An analytical study of cepstral peak prominence (CPP) is presented, intended to provide an insight into its meaning and relation with voice perturbation parameters. To carry out this analysis, a parametric approach is adopted in which voice production is modelled using the traditional source-filter model and the first cepstral peak is assumed to have Gaussian shape. It is concluded that the meaning of CPP is very similar to that of the first harmonic and some insights are provided on its dependence with fundamental frequency and vocal tract resonances. It is further shown that CPP integrates measures of voice waveform and periodicity perturbations, be them either amplitude, frequency or noise.

## 1. Introduction

Cepstral peak prominence (CPP) is an acoustic measure of voice quality that has been qualified as *the most promising and perhaps robust acoustic measure of dysphonia severity* [1]. Such a definite statement made by Maryn et al is based on a meta-analysis that considered previous results by Wolfe and Martin [2], Wolfe et al. [3], Heman-Ackah et al. [4], Halberstam [5], and Eadie and Baylor [6]; yet, later results are also consistent with that assertion, as those published by Awan et al. [7], Awan and Roy [8], Maryn et al. [9,10], Shue et al. [11], Alpan et al. [12], and Petterson et al. [13]. Conclusions from these have led some researchers to suggest the inclusion of CPP in the computation of some objective measures of dysphonia such as the acoustic indexes of dysphonia severity proposed by Awan and Roy [8,7], the Acoustic Voice Quality Index [9,10], and the Cepstral Spectral Index of Dysphonia [13]. Being a widely tested measure of dysphonia severity, CPP has also been proposed as a relevant measure to assess the effect of different treatments. For instance, Hartl et al. proposed to use CPP in combination with other parameters to assess the effects of surgical treatments [14], Awan and Roy considered CPP in the evaluation of the effects of a therapy based on manual circumlaryngeal techniques [8], and Solomon et al found that CPP was useful for following patients' voice evolution after thyroidectomy [15].

Apart from its correlation with overall dysphonia, the relation between CPP and specific voice disorders has also been studied:

Merk et al. proposed the variability in CPP to be used as a cue to detect neurogenic voice disorders [16], Rosa et al. reported that the combination of CPP with other acoustic cues is relevant for the detection of laryngeal disorders [17], Hartl et al. [18] [19] and Balasubramanian et al. [20] concluded that patients suffering from unilateral vocal fold paralysis exhibited significantly lower values of CPP than healthy individuals, Kumar et al. arrived to a similar conclusion for the case of vocal fold nodules [21], and Watts and Awan found CPP relevant for discriminating hypo-functional from normal voices [22]. The consistent performance of CPP in the clinical evaluation of voice quality has inspired some researchers to propose its application to other purposes such as the assessment of speech intelligibility [23], the detection of cognitive load [24], or even the evaluation of the sexual appeal of voice [25].

CPP was first introduced by Hillenbrand et al. [26] for the assessment of breathy voices. Later, Hillenbrand and Houde defined a variant of CPP called smoothed CPP (CPPs) that provided somewhat higher correlation with breathiness by adding smoothing operations both in temporal and cepstral domains [27]. Hartl et al. reported correlation between breathiness and CPP too [18], Shrivastav and Sapienza found that CPP has a more consistent behaviour in predicting breathiness than noise measures, jitter or shimmer [28], and Samlan et al. also found a significant relation between CPP and breathiness perceived in synthetic voices generated using a kinematic vocal fold model [29]. As for the relation between CPP and perceived breathiness, results reported by Alpan et al. indicate that such relation seems to be non-linear [30]. This observation is consistent with the conclusions of Samlan and Story stating that for small glottal gaps CPP and perceived breathiness do not have linearly related behaviours [31]. Consequently, the assumption that CPP is related to perceived breathiness

\* Corresponding author. Tel.: +34 913367830.

E-mail addresses: rfraile@ics.upm.es (R. Fraile), igodino@ics.upm.es (J.I. Godino-Llorente).

may presently be considered as well-founded, although the specific degree of correlation may depend on the language of the speaker [32] and on the linguistic experience of the listener [33]. Furthermore, CPP is also related to the physiological processes behind the production of breathiness, although variations in CPP occur due to several underlying anatomic and vibratory vocal-fold properties, so it is not feasible to identify the specific causes of a given change in CPP [29].

While CPP was firstly intended to measure breathiness, its use has been extended to the evaluation of overall voice quality, as mentioned before. Nevertheless, its usefulness for discriminating among voice qualities other than breathy seems to be limited, if any. Wolf and Martin [2] reported that CPP is a discriminant measure for distinguishing strain from other dysphonic voice types (hoarseness and breathiness) but that does not help in further distinctions among voice qualities; yet, the same authors later found that CPP calculated for band-pass filtered voices might be useful for the identification of voice qualities [3]. Similarly, Heman-Ackah et al concluded on the one hand that CPPs correlates with dysphonia severity more than NHR (noise-to-harmonics ratio), APQ (amplitude perturbation quotient), RAP (relative average perturbation) or smoothed PPQ (pitch perturbation quotient) [4,34] and also with breathiness and roughness but, on the other hand, that correlation with roughness is not relevant [4]. A similar conclusion was reached by Awan and Roy [35]. More specifically, Eadie and Baylor pointed out that CPPs only correlates with roughness for running speech but shimmer is a better predictor of roughness for vowels [6]. Coherently, Howard et al. found out that voice perturbation measures calculated in time domain are more correlated to specific perceptual features than CPP [36]. Results published by Moers et al. show that CPP and CPPs calculated in running speech provide higher correlations with perceived breathiness and hoarseness than perturbation and noise measures, but for roughness, noise measures provide somewhat higher correlations [37]. When analysing the prediction of voice qualities, Lowell et al. also concluded that dysphonic-rough voice quality is less accurately classified by cepstral-based measures than dysphonic-breathy and normal voice qualities [38]. From the point of view of phonetics, Espósito pointed out that CPP can help in distinguishing breathy from modal or creaky voices, but cannot help in discriminating between modal and creaky [39]. On the contrary, a relevant correlation between CPP and roughness (in addition to breathiness) has been reported by Cannito et al. [40] and Shue et al detected some correlation between CPP and pressed voice quality [11].

In spite of its demonstrated usefulness for the clinical evaluation of voice, to present there is not a definite explanation of what CPP actually measures. In fact, CPP shares with other cepstral measures the lack of an intuitive interpretation relative to the underlying physiology of vocal fold vibration [41]. Hillenbrand et al. assumed that the height of the cepstral peak used to compute CPP is affected by the periodicity of the signal (or harmonic organisation), the window size and the signal's total energy [26,27]. Their assumption was based on the high correlation measured between CPP and the autocorrelation peak for band-pass and high-pass filtered voices, though correlation was not so high for full spectrum signals. They also assumed that the CPP measure is similar in principle to a cepstrum-based signal-to-noise ratio calculation [26]; a similar assumption was also made by Awan et al. [42] when defining the CPP as the dominance of the first rahmonic with respect to the background noise level. The relation with the periodicity of the signal has also been assumed by Ferrer et al. [43]. In turn, the relationship between CPP and the noise level present in the voice signal has been rigorously reasoned by Murphy [44], who showed the linear relation between cepstral peak, i.e. first rahmonic, and the average of the harmonics to between harmonics ratio in the logarithmic spectrum. In order to have an additional insight into the

meaning of CPP, some authors have sought for correlations between CPP and other acoustic parameters. Heman-Ackah et al. reported greater correlations between CPP and pitch perturbation measures (RAP and sPPQ) than between CPP and measures of noise (NHR) and amplitude perturbation (APQ) [4]. On the contrary, Samlan et al. concluded that HNR is correlated to CPP [29]. Last, Cannito et al. measured relevant correlations between CPP and several measures of aperiodicity [40].

In this paper, we present an analysis of CPP that intends to provide an insight into its meaning and relation with perturbation parameters that on the one hand helps to interpret previous findings mostly reported from clinical studies and, on the other hand, complements previous studies, notably those published by Murphy [44], Alpan et al. [45] and Samlan and Story [31]. To carry out this analysis, we adopt a parametric approach in which we model voice production using the traditional source-filter model [46] so as to infer the meaning of the log-linear regression involved in the computation of CPP and we model the first cepstral peak (first rahmonic) as a Gaussian pulse in order to derive its meaning in spectral domain. Later, we use this combined approach to deduce the effect of signal windowing and sampling on the value of CPP and also to analytically seek for a relation between CPP and perturbation parameters such as shimmer, jitter and harmonics-to-noise ratio.

## 2. Analysis of CPP for infinitely long, continuous-time and noiseless voice signals

### 2.1. Definition of real cepstrum

Given a signal  $s(t)$ , its real cepstrum, or power cepstrum, is equal to the Fourier transform of the logarithm of its power spectrum, according to the first definition of cepstrum [47]:

$$C_r(q) = \mathcal{F}\{\log |S(f)|^2\} \quad (1)$$

where  $S^2(f)$  is the power spectrum of the signal:

$$S^2(f) = \mathcal{F}\{E[s(t) \cdot s^*(t - \tau)]\} \quad (2)$$

The cepstrum was primarily developed to detect echoes in seismic signals [47]. When a time signal is composed by echoes of an impulse (Fig. 1, top), the cross correlation between the original impulse and its echoes is a combination of impulses having its maximum located at the delay corresponding to the main echo. Being the cross correlation a sum of impulses, its Fourier transform is periodic. Its apparent frequency corresponds to the delay of the main echo, while secondary echoes surrounding the main one impose an amplitude modulation to the spectrum (Fig. 1, centre).

The logarithm previous to the Fourier transform in (1) allows converting the multiplicative effect of the modulating signal into an additive effect. Since the modulating signal in the spectrum is smoother than its periodic component, the cepstrum separates them, thus allowing a clearer identification of the delay of the main echo in cepstral domain (Fig. 1, bottom).

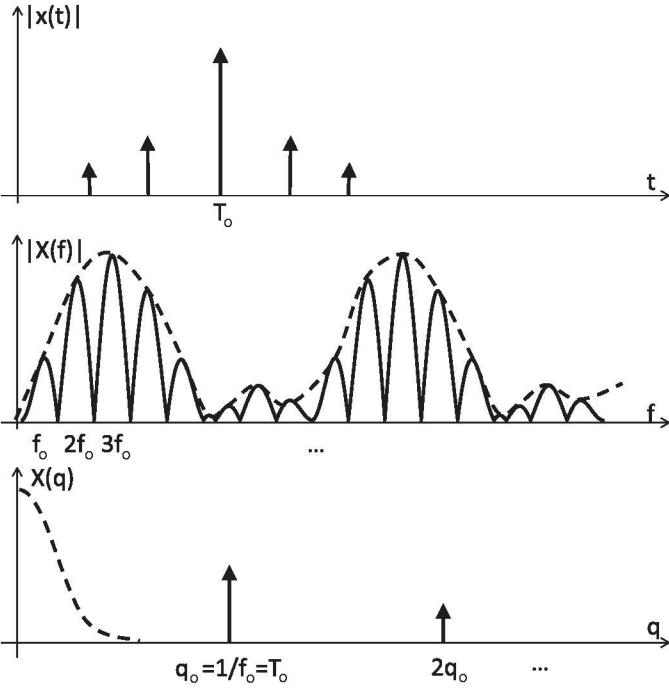
Later formalisation of cepstral analysis led to the definition of the complex cepstrum, which includes information of the phase spectrum [48]:

$$C_c(q) = \mathcal{F}\{\log S(f)\} \quad (3)$$

being:

$$\log S(f) = \log |S(f)| + j \cdot \chi(S(f)) \quad (4)$$

The relationship between the real and the complex cepstra is such that the real cepstrum is equal to four times the square of the even part of the complex cepstrum [48]. Consequently, the real cepstrum is an even function of quefrency.



**Fig. 1.** When the spectrum of a signal is a periodic signal multiplied by an envelope (centre), the low quefrency part of the cepstrum (bottom) conveys information on the multiplicative envelope while the periodic component of the spectrum produces a corresponding sequence of impulses in cepstral domain (rahmonics).

## 2.2. Speech cepstrum

According to the source-filter model of speech production [46], a voiced speech signal  $s(t)$  can be modelled as:

$$s(t) = g(t) * v(t) * r(t) \quad (5)$$

where  $g(t)$  is the glottal signal,  $v(t)$  is the impulse response of the vocal tract and  $r(t)$  is the effect of the acoustic wave radiation at the lips, modelled as an impulse response. Both  $v(t)$  and  $r(t)$  vanish with time, while  $g(t)$  is usually modelled as an indefinitely long train of glottal pulses.

In spectral domain we have:

$$S(f) = G(f) \cdot V(f) \cdot R(f) = G(f) \cdot H(f) \quad (6)$$

where  $H(f)$  is the combined effect of vocal tract and lip radiation. The power spectrum can be estimated as:

$$S^2(f) = G^2(f) \cdot H^2(f) \quad (7)$$

According to the definition of cepstrum given in [47], the real cepstrum allows transforming the convolution in (5) and the product in (6) into an addition:

$$C_r(q) = \mathcal{F}\{\log |S(f)|^2\} = \mathcal{F}\{\log |G(f)|^2\} + \mathcal{F}\{\log |H(f)|^2\} \quad (8)$$

Yet, the computation of the inverse Fourier transform in (1) and (8) instead of the direct Fourier transform is usual, as indicated in [48,49]. According to the duality property of the Fourier transform [50], computing the inverse Fourier transform instead of the direct Fourier transform has the combined effect of reflection in the independent variable (quefrency) and multiplication by a constant. Since the speech signal  $s(t)$  is real,  $|S(f)|$  is positive and symmetric with respect to the vertical axis. Then,  $\log |S(f)|$  is real and symmetric. Therefore, its Fourier transform is real and symmetric too. As a consequence, reflection with respect to the quefrency axis has no effect and the only difference between both approaches, i.e. direct and inverse Fourier transforms, is a multiplicative constant. Additionally, due to the logarithm operation, taking out the square

exponent in the power spectrum only has the effect of a multiplicative constant. Thus, both approaches can be considered equivalent, except for some multiplicative factor.

## 2.3. Effect of vocal tract and lip radiation on speech cepstrum

The vocal tract filter  $v(t)$  in (5) is usually modelled as an all-pole system; in turn, lip radiation  $r(t)$  is commonly assumed to behave as a single-pole filter [46]. Therefore,  $H(f)$  in (8) can be modelled as an all-pole filter:

$$H(f) = \frac{H_0}{\prod_{p=1}^{n_p} (j\omega - s_p)}|_{\omega=2\pi f} = H(\omega)|_{\omega=2\pi f} \quad (9)$$

where  $n_p$  is the number of poles and  $s_p$  are the poles themselves. Taking the logarithm of the modulus we get:

$$\log |H(\omega)| = \log |H_0| - \sum_{p=1}^{n_p} \log |j\omega - s_p| \quad (10)$$

The real cepstrum can then be calculated as:

$$C_r(q) = \mathcal{F}\{\log |H_0|\} - \sum_{p=1}^{n_p} \mathcal{F}\{\log |j\omega - s_p|\} \quad (11)$$

Solving we get (see [Appendix A](#) for details):

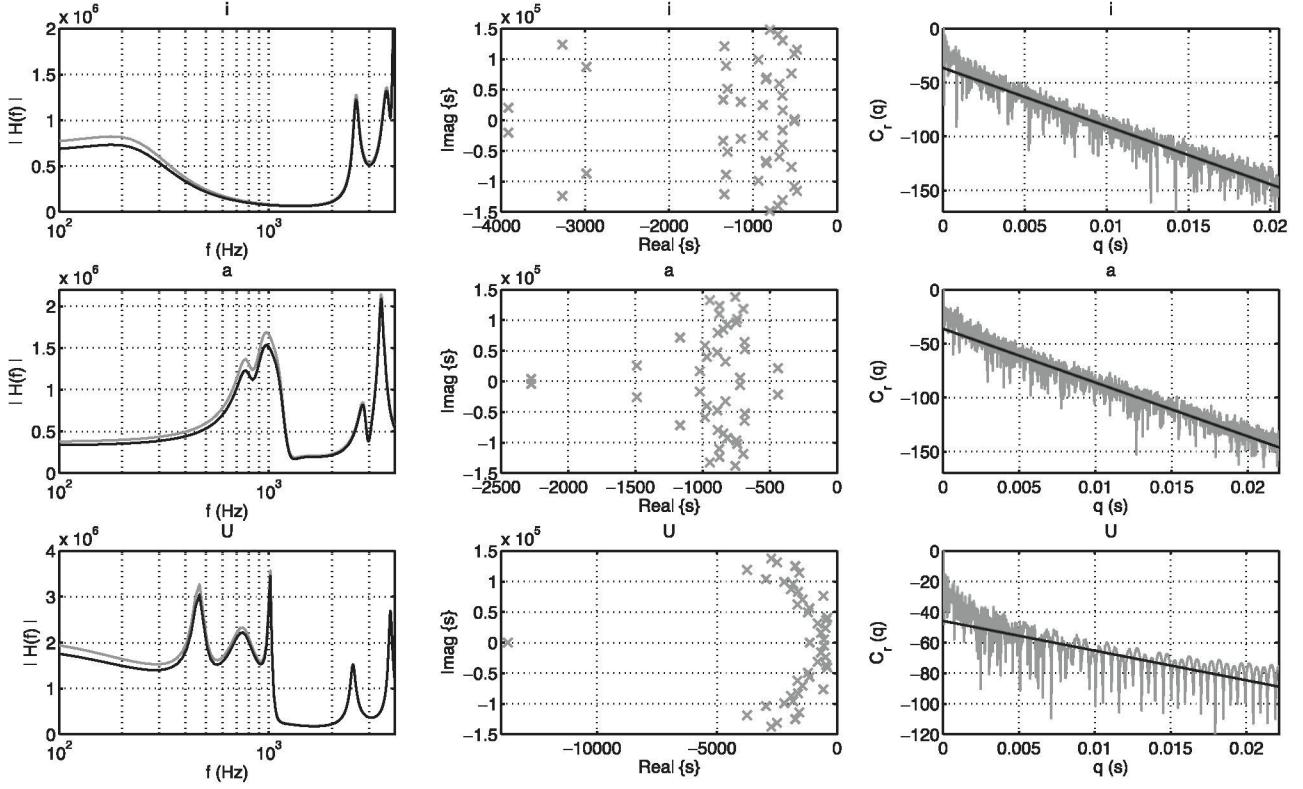
$$C_r(q) = 2\pi \cdot \log H_0 \cdot \delta(q) + \frac{1}{|q|} \cdot \left( \sum_{k=1}^{n_{rp}} \frac{e^{\sigma_k |q|}}{2} + \sum_{l=1}^{n_{cp}} \cos(\omega_l q) \cdot e^{\sigma_l |q|} \right) \quad (12)$$

where  $\delta(\cdot)$  is the Dirac delta function,  $n_{rp}$  is the number of real poles in (9),  $n_{cp}$  is the number of complex conjugate pole pairs in (9) and  $\sigma_l$  and  $\omega_l$  respectively are the real and complex parts of such poles. Note that the all-pole system that models the effect of the vocal tract plus lip radiation is represented in cepstral domain by the combination of an impulse centred at zero quefrency plus a sum of negative exponentials ( $\sigma_k < 0$  if we assume that the vocal tract behaves as a stable system) and damped sinusoids.

[Fig. 2](#) shows the responses of three simulated vocal tracts in cepstral domain. For quefrency values above a certain threshold the damped behaviour, linear decrease of the envelope in decibels, is self-evident.

## 2.4. Log-linear regression and vocal tract response

As outlined before, in cepstral analysis of speech it is commonly assumed that the glottal signal  $g(t)$  is periodic (or quasi-periodic) while vocal tract and lip radiation have a finite-length impulse response  $h(t)$ . The convolution of both signals results in a periodic speech signal  $s(t)$  having a fundamental frequency ( $f_0$ ) equal to that of the glottal signal. With these assumptions, the cepstrum of  $s(t)$  has two well differentiated parts ([Fig. 1](#)): for high quefrencies ( $q \geq q_0 = 1/f_0$ ) it consists of a series of peaks (rahmonics) placed at multiples of  $q_0$  corresponding to the transformation of the spectral harmonics; for low quefrencies ( $q < q_0$ ) it corresponds to the transformation of the envelope of harmonics' amplitudes. However, it should be noted that this is not strictly true: since the spectrum of a purely periodic signal is zero between harmonic locations, its logarithm cannot be computed and, consequently, its cepstrum does not exist. Yet, since signal windowing is needed for processing, the processed signal never is purely periodic and the cepstrum can always be computed. More rigorously, what is necessary for the cepstrum  $C_r(q)$  to have a peak at  $q = q_0$  is that the log-spectrum of the signal has a periodic component (having apparent period  $f_0 = 1/q_0$ ).



**Fig. 2.** Cepstral analysis of three different vocal tract shapes taken from [51]. Shapes correspond to vowels /i/ (i on the graph), /a/ (a) and /u/ (U). The grey lines corresponds to transforms of the impulse response of the vocal tract simulated using the Kelly-Lochbaum model and the simulator described in [52] (Fourier transform on the left and cepstrum on the right). The middle column corresponds to the pole plots in the s-plane. The black line corresponds to the linear regression of the cepstral values computed in dB (right) and to the effect of the subtraction of such regression in frequency domain (left). Only values for  $q > 1.5$  ms have been used for regression.

As the logarithm is a monotonic function, this is equivalent to stating that the power spectrum also has a periodic component with the same period. Furthermore, for this to happen, it is not necessary that the signal, more precisely its autocorrelation function, is periodic in time domain.

Regarding the specific case of CPP, one of the key factors in its definition seems to be the cepstral log-linear regression, according to Heman-Ackah [53]. This log-linear regression is obtained in order to subtract its value from the cepstral peak (or rahmonic). Awan and Roy proposed to consider only cepstral values corresponding to  $q > 2$  ms for regression, arguing that values for lower quefrequencies mainly correspond to the vocal tract [35]. Alpan et al put the limit in  $q = 1$  ms [12]. Recalling Eq. (12), if the vocal tract plus lip radiation response consisted in a one-pole system, subtracting the log-linear regression would imply removing the effect of the whole response from the cepstrum. Since the system is multiple-pole, its cepstral representation is more complex, as depicted in Fig. 2. However, for quefrequencies above a certain threshold a limited number of poles become dominant and the log-regression fits well the envelope of the cepstrum of the vocal tract plus lip radiation. Therefore, setting a minimum quefrency for the calculation of the log-regression implies not considering the part of the cepstrum which is affected by a greater number of poles, hence modelling only the most dominant ones.

The quefrency threshold below which the cepstrum of the vocal tract response departs from the log-linear behaviour depends on the specific configuration of the vocal tract and its associated all-pole model. Therefore, when a fixed threshold is selected for computing CPP (typically between 1 and 2 ms, as mentioned before) it is not surprising that the obtained value is affected by vowel type (i.e. vocal tract) [54]. This effect can be appreciated in Fig. 2. While for the first two vowels (/i/ and /a/) the log-linear descent of

the cepstral envelope happens for  $q \gtrsim 2$  ms, this threshold changes to  $q \gtrsim 4$  ms for the third one (/u/). Thus, setting a fixed threshold ( $q_{\text{thres}} = 1.5$  ms in this case) implies not being able to model the descent of the cepstral envelope equally well for all vocal tracts. It is then coherent that the averaging of CPP values for different articulatory configurations, that is, the computation of CPP from running speech, provides higher correlations with perceived dysphonia [55,5,4] and also measurements that are more robust against changes in utterances, both in length and phonemic content [56]. Yet, one should bear in mind that the calculation of CPP from sustained vowels and from running speech leads to different statistical distributions of the obtained measurements [34].

The regression in cepstral domain can be mathematically expressed as:

$$20 \cdot \log_{10} C_r^{\text{regr}}(q) = b_0 + b_1 \cdot |q| \rightarrow C_r^{\text{regr}}(q) = 10^{(b_0/20)} \cdot 10^{(b_1/20) \cdot |q|} = B_0 \cdot e^{B_1 \cdot |q|} \quad (13)$$

where  $b_0$  and  $b_1$  are the regression coefficients and the definitions of  $B_0$  and  $B_1$  can be easily deduced from the previous equation. By calculating the inverse Fourier transform, in spectral domain we get:

$$\log |S^{\text{regr}}(f)| = \frac{1}{2} \cdot \frac{1}{2\pi} \cdot \frac{-2B_0B_1}{B_1^2 + (2\pi f)^2} = -\frac{B_0}{2\pi B_1} \cdot \frac{1}{1 + (2\pi f/B_1)^2} \quad (14)$$

In spectral domain, the subtraction of the cepstral regression line implies dividing the spectrum by the exponential of the previous expression:

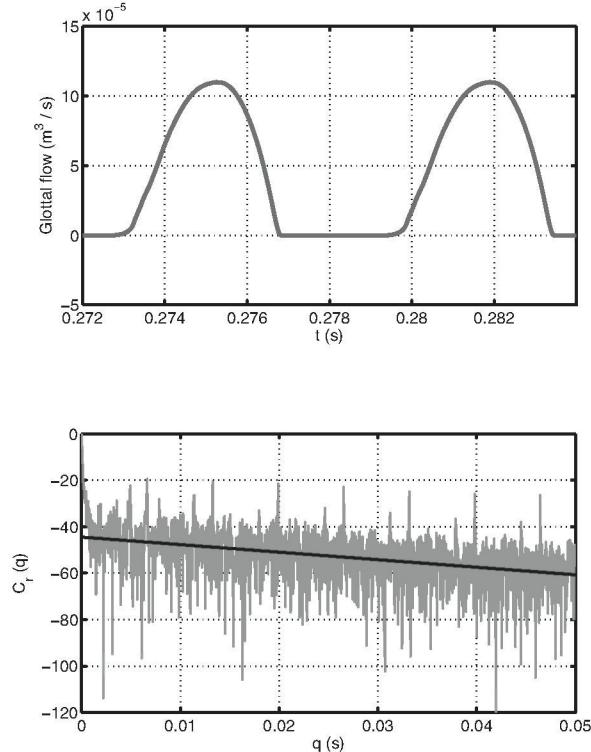
$$|S^{\text{regr}}(f)| = e^{-\frac{B_0}{2\pi B_1} \cdot \frac{1}{1 + (2\pi f/B_1)^2}} = \beta^{\frac{1}{1 + (2\pi f/B_1)^2}} \quad (15)$$

Thus, subtraction of the cepstral regression line is approximately the same as dividing the Fourier transform by  $\beta = e^{-(B_0/2\pi B_1)}$  for low frequencies ( $f \rightarrow 0$ ) and dividing it by 1 for high frequencies ( $f \rightarrow \infty$ ). The frequency threshold between both asymptotic behaviours approximately is  $f = (B_1/2\pi)$ . An estimate of the value for this threshold can be obtained from Fig. 2. In that graph, the slopes of the regression lines are  $b_1 \approx -5.4 \text{ dB/ms}$  for  $/i/$ ,  $b_1 \approx -5.0 \text{ dB/ms}$  for  $/a/$ , and  $b_1 \approx -1.9 \text{ dB/ms}$  for  $/u/$ , which correspond to frequency thresholds equal to 99, 91 and 36 Hz, respectively. Since the fundamental frequency of the voice signal frequently is above such thresholds and the formant resonances are above it, the subtraction of the regression line in the voice cepstrum has little effect on the overall shape of the spectrum. It only affects the low frequency components, usually below the fundamental frequency (see left plots in Fig. 2).

According to basic theory of linear systems [50], the poles associated to the longest responses (both in time and quefrency) are the nearest to the imaginary axis. In Fig. 2, the steepest regression corresponds to the vocal tract response having its poles furthest from the imaginary axis ( $/i/$ ) and vice-versa. Consequently, the longest responses are also associated with the highest resonances in spectral domain. Therefore, the subtraction of the cepstral log-regression in the calculation of CPP theoretically would imply the compensation of the greatest resonances in the signal spectrum; to some extent, it should be a spectral flattening operation. However, since only the tail of the vocal tract response is modelled by the regression, the flattening operation only affects the lowest frequencies of the spectrum.

## 2.5. Log-linear regression and glottal pulses

The glottal signal  $g(t)$  in (5) is usually modelled as the convolution of a fix pulse waveform  $p(t)$  and a series of impulses that



**Fig. 3.** Analysis of the effect of cepstral log-regression subtraction on a glottal signal obtained with the simulator described in [52]. The figure shows the pulse waveform in time domain (top left), its Fourier transform (right, grey line) and its cepstrum (bottom left, grey line). The regression line is plotted together with the cepstrum (bottom left, black line) and also the effect of its removal in spectral domain (right, black line). Only values for  $q > 1.5 \text{ ms}$  have been used for regression.

account for the instants  $t_k$  at which the glottal pulses happen (e.g. [46]):

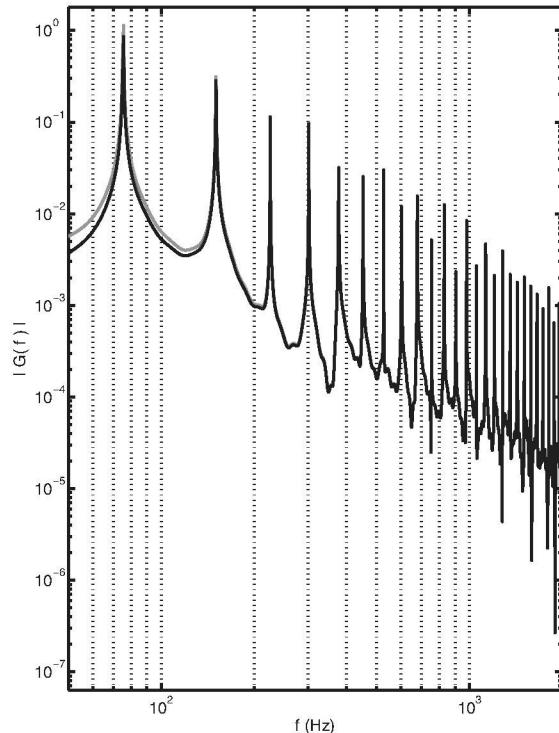
$$g(t) = p(t) * \sum_{k=-\infty}^{\infty} a_k \cdot \delta(t - t_k) \quad (16)$$

where  $a_k$  is the amplitude of the  $k^{\text{th}}$  glottal pulse. Similarly to the case of the vocal tract and the lip radiation, the glottal pulse waveform  $p(t)$  is usually modelled as an all-pole signal, typically having two or three poles [57]. As a consequence, the previous analysis is also valid for the case of the glottal pulse. The effect of the cepstral log-linear subtraction on the signal spectrum is also similar to the case of the vocal tract (see Fig. 3): only very low frequencies are affected.

Thus, the main effect of the subtraction of the cepstral regression is on very low frequencies; so its effect on the overall voice spectrum may be neglected. In other words, what CPP measures in spectral domain is basically the same as what is measured by the cepstral peak, or first rahmonic. This is consistent with the findings of Alpan et al. [45] regarding the similar correlations with perceptual rates that can be obtained CPP and the first rahmonic.

## 2.6. Relationship between cepstral peaks and spectrum

As illustrated in Fig. 1, the low-quefrency part of the cepstrum represents the smooth variations of spectral amplitude, which commonly are associated to  $H(f)$  for speech signals [46]. In contrast, the periodic part of the spectrum corresponds in cepstral domain to linearly spaced peaks called rahmonics. Having previously paid attention to the effect of  $H(f)$  on the cepstrum, we now analyse the case of the periodic component of the spectrum. To do that, let us assume that the high-quefrency part of the cepstrum can be



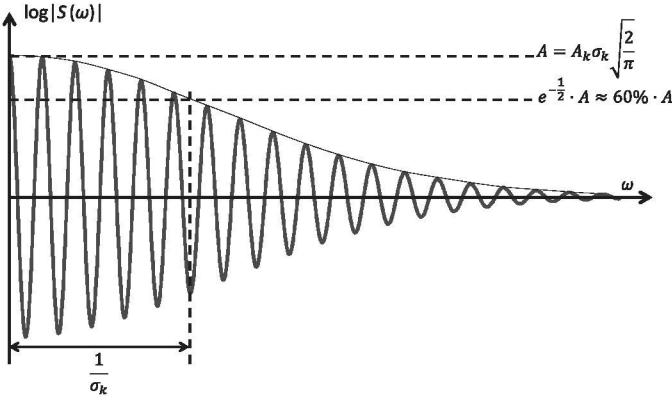


Fig. 4. Logarithmic spectrum corresponding to one Gaussian-shaped rahmonic having amplitude  $A_k$  and width  $\sigma_k$ .

modelled as the sum of a series of rahmonic peaks, the form of which correspond to Gaussian functions:

$$C_r(q) = \sum_{k=1}^{\infty} A_k \cdot (e^{-(q-k \cdot q_0)^2 / (2 \cdot \sigma_k^2)} + e^{-(q+k \cdot q_0)^2 / (2 \cdot \sigma_k^2)}) \quad (17)$$

where it has been assumed that the real cepstrum is symmetric, which is the case for speech signals, as justified before. The amplitude of each rahmonic peak is fixed by the coefficient  $A_k$  while its width is governed by  $\sigma_k$ . The spectral representation of such rahmonic series can be computed by taking the inverse Fourier transform:

$$\begin{aligned} \log |S(f)| &= \mathcal{F}^{-1}\{C_r(q)\} \\ &= \sum_{k=1}^{\infty} A_k \cdot \mathcal{F}^{-1}\{e^{-(q-k \cdot q_0)^2 / (2 \cdot \sigma_k^2)} + e^{-(q+k \cdot q_0)^2 / (2 \cdot \sigma_k^2)}\} \\ &= \sum_{k=1}^{\infty} A_k \cdot \frac{\sigma_k}{\sqrt{2\pi}} \cdot e^{-(\sigma_k^2 \omega^2 / 2)} \cdot 2 \cdot \cos(kq_0 \cdot \omega) \\ &= \sum_{k=1}^{\infty} A_k \sigma_k \cdot \sqrt{\frac{2}{\pi}} \cdot e^{-(\sigma_k^2 \omega^2 / 2)} \cos(kq_0 \cdot \omega) \end{aligned} \quad (18)$$

Therefore, each rahmonic corresponds to one damped sinusoid in the logarithmic spectrum (Fig. 4). The damping of the sinusoid is affected by the width of the rahmonic  $\sigma_k$  while the amplitude of the damped sinusoid is proportional to the product  $A_k \cdot \sigma_k$ , which is a measure of the rahmonic's energy.

## 2.7. Discussion on the relationship between cepstral peaks and spectrum

Murphy [44] interpreted the first rahmonic (i.e. the first cepstral peak) as a measure of the average of the harmonics to between harmonics ratio in the logarithmic spectrum. According to the reasoning above, if we consider that the first rahmonic is usually dominant over the second and following ones then the envelope of the harmonic peaks in the logarithmic spectrum is:

$$\begin{aligned} \log |S_{\text{harm}}(f)| &\approx \frac{1}{2} A_1 \sigma_1 \cdot \sqrt{\frac{2}{\pi}} \cdot e^{-(\sigma_1^2 (2\pi f)^2 / 2)} \\ &= A_1 \sigma_1 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-(\sigma_1^2 (2\pi f)^2 / 2)} \end{aligned} \quad (19)$$

On the opposite, the envelope of the inter-harmonics valleys is:

$$\log |S_{\text{inter-harm}}(f)| \approx -A_1 \sigma_1 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-(\sigma_1^2 (2\pi f)^2 / 2)} \quad (20)$$

The average of the harmonics to between harmonics ratio (*gmHNR*, using Murphy's notation) can the be approximated in logarithmic scale as:

$$\begin{aligned} \text{gmHNR} &\propto \int_0^{\infty} \log \left| \frac{S_{\text{harm}}(f)}{S_{\text{inter-harm}}(f)} \right| df \\ &= \int_0^{\infty} \log |S_{\text{harm}}(f)| df - \int_0^{\infty} \log |S_{\text{inter-harm}}(f)| df \\ &= 2A_1 \int_0^{\infty} \sigma_1 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-(\sigma_1^2 (2\pi f)^2 / 2)} df \end{aligned} \quad (21)$$

This integral corresponds to the value of a Gaussian distribution function at is median point [58]. Therefore:

$$\text{gmHNR} \propto 2A_1 \cdot \frac{1}{2} = A_1 \quad (22)$$

Since  $A_1$  is the magnitude of the first rahmonic, our analysis is completely coincident with Murphy's conclusions when the magnitudes of second and following rahmonics are negligible when compared to the first one.

Being the height of the cepstral peak a measure related to the harmonic structure of the voice signal, it easily follows its close relation with the glottal signal. In fact, when the glottal signal is absent, the cepstral peak loses its relevance for assessing voice quality. This is the case of tracheo-oesophageal voice, whose quality cannot be evaluated using CPP [59,60]. Similarly, while some researchers have calculated CPP from running speech without any attempt to previously remove unvoiced signal intervals (e.g. [27]), Lowell et all have recently shown that preserving unvoiced signal segments has the effect of compressing the average values of CPP mean and CPP standard deviation for all voice quality groups and, simultaneously, to increase their dispersion, hence providing more overlapped distributions [38]. Thus, CPP calculated from unvoiced signal segments seems not to be significant for the evaluation of voice quality. Similarly, it is known that increases in loudness are associated to increases in the relevance of the harmonic component of the voice signal with respect to non-periodic components [61]. Thus, voice intensity is related to the magnitude of the harmonics, as measured by  $A_k \cdot \sigma_k$  (see Fig. 4). This correlation between the cepstral peak, or CPP, and voice intensity has been measured by Awan et al. [54].

The relationship between CPP and breathiness has already been mentioned in the introductory section. Apart from CPP, breathy voice quality is know to be correlated with the noise energy, mostly at high frequencies, and with the relative height of the first harmonic with respect to the rest [26–28]. High levels of high frequency noise imply that the highest harmonics of the periodic component of the signal are less relevant in the spectrum. This is associated to a reduction in the value of  $(1/\sigma_k)$  in Fig. 4. Similarly, a relative increase in the relevance of the first harmonics with respect to the rest is also related to a narrowing of the spectral envelope in Fig. 4. For a rahmonic peak having a given energy  $A_k \cdot \sigma_k$ , any reduction in  $(1/\sigma_k)$  (increase in  $\sigma_k$ ) corresponds to a proportional reduction in  $A_k$ , i.e. a reduction on the height of the cepstral peak. So, the relation between CPP and breathiness can be explained by using this Gaussian model for cepstral peaks and assuming that the first one is much more relevant than the rest.

Complementarily, CPP has been shown to be correlated with vocal-fold closing speed through the analysis of acoustic and high-speed video-endoscopic measures [62]. This can also be explained

with the same model: any increase in the speed of a part of the glottal waveform is linked to an increase in the level of the high-frequency harmonics. In turn, this is related to an increase in  $(1/\sigma_k)$ , or decrease in  $\sigma_k$ . Since  $\sigma_k$  and  $A_k$  are inversely related, a more prominent cepstral peak is produced. This relationship between CPP and the glottal waveform was also deduced by Shue et al [11].

### 3. Effect of windowing and sampling

The processing of infinitely long, continuous time signals is not feasible. Instead, discrete-time windowed signals are processed by nowadays systems. Thus, the effect of these two operations (windowing and sampling) on the cepstral peak should not be disregarded.

#### 3.1. Windowing

A windowed signal can be expressed as:

$$s^W(t) = s(t) \cdot w(t) \quad (23)$$

where  $w(t)$  is null for  $t < 0$  and for  $t > \tau$ , being  $\tau$  the window length. If such a signal is sampled, the next discrete time signal is obtained:

$$s^W[n] = s[n] \cdot w[n] = s(nT_s) \cdot w(nT_s) \quad (24)$$

where  $T_s$  is the sampling period and  $w[n]$  is null for  $n < 0$  and for  $n \geq L$ , being  $L$  the length of the discrete window. Samples of the Fourier transform of  $s^W[n]$  can be obtained using the discrete Fourier transform (DFT) [63]:

$$S^W(f_k) = \sum_{n=0}^{L-1} s[n] \cdot w[n] \cdot e^{-j \cdot n \cdot 2\pi f_k \cdot T_s}; \quad f_k = \frac{k}{L \cdot T_s} \quad 0 \leq k < L \quad (25)$$

Let us suppose that  $s[n]$  is periodic, being  $N_0$  its fundamental period, and that the window is rectangular:  $w[n] = 1 \forall 0 \leq n < L$ . Let us also assume that the windowed signal comprises more than

one period but its length may not be a multiple of  $N_0$ , that is,  $L = mN_0 + \alpha N_0$ , where  $m \geq 1$  and  $0 \leq \alpha < 1$ . Then:

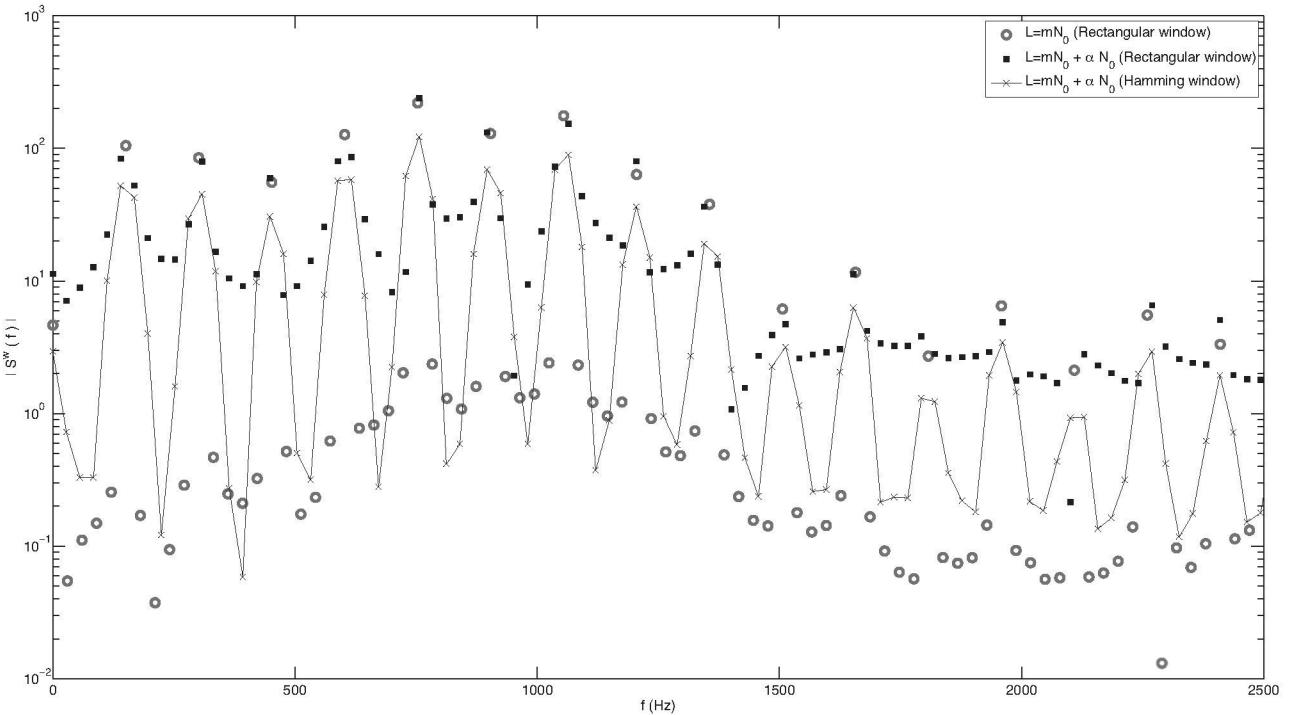
$$S^W(f_k) = \sum_{n=0}^{mN_0-1} s[n] \cdot e^{-j \cdot n \cdot 2\pi f_k \cdot T_s} + \sum_{n=mN_0}^{(m+\alpha)N_0-1} s[n] \cdot e^{-j \cdot n \cdot 2\pi f_k \cdot T_s} \quad (26)$$

Since we have assumed that  $s[n]$  is periodic, the first term in (26) is proportional to the coefficients of the Fourier series expansion corresponding to  $s[n]$  when  $f_k \cdot T_s = l/N_0$  for integer values of  $l$  and it is null for the rest of cases [50]. The second term is a result of sampling the Fourier transform of a fraction  $\alpha$  of the signal period:

$$S^W(f_k) = 2\pi m \cdot \sum_{l=0}^{N_0-1} S[l] \delta \left[ f_k T_s - \frac{l}{N_0} \right] + \sum_{n=mN_0}^{(m+\alpha)N_0-1} s[n] \cdot e^{-j \cdot n \cdot 2\pi f_k \cdot T_s} \quad (27)$$

being  $\delta[\cdot]$  the Kronecker delta.

Fig. 5 shows the DFT of the voice signal corresponding to the glottal pulses plotted in Fig. 3. The impulsive structure of the DFT can be easily appreciated, with impulses or harmonics appearing at one every five samples ( $N_0 = 5$ ), in the graph corresponding to a rectangular window having a length that is an exact multiple of the fundamental period ( $\alpha = 0$ ). When this condition does not happen, i.e.  $\alpha > 0$ , the second term in (27) increments the values of the DFT samples at inter-harmonic locations due to the nulls of the Fourier transform of the window not being coincident with the frequencies of the DFT samples [63]. According to the reasoning before, such an increment directly affects the amplitude of the first cepstral peak (see Fig. 6). However, the use of a window having lower side lobes in spectral domain can help in keeping a difference between harmonic and inter-harmonic values similar to the case of  $\alpha = 0$  (continuous line in Fig. 5 for a Hamming window). Note that with respect to the first case, this graph does not have the same impulsive waveform but one that resembles more a sinusoid. Yet, the



**Fig. 5.** DFT of a windowed voice signal corresponding to the glottal flow depicted in Fig. 3 (fundamental period  $N_0 = 106$ , sampling period  $T_s = 1/16,000$ ). Empty circles correspond to a rectangular window containing exactly 5 periods ( $m = 5, \alpha = 0$ ). Black squares correspond to a rectangular window that is not multiple of the signal period ( $m = 5, \alpha = 40/106 \approx 0.38$ ). Crosses linked with a continuous line corresponds to a Hamming window of the same length ( $m = 5, \alpha = 40/106$ ).

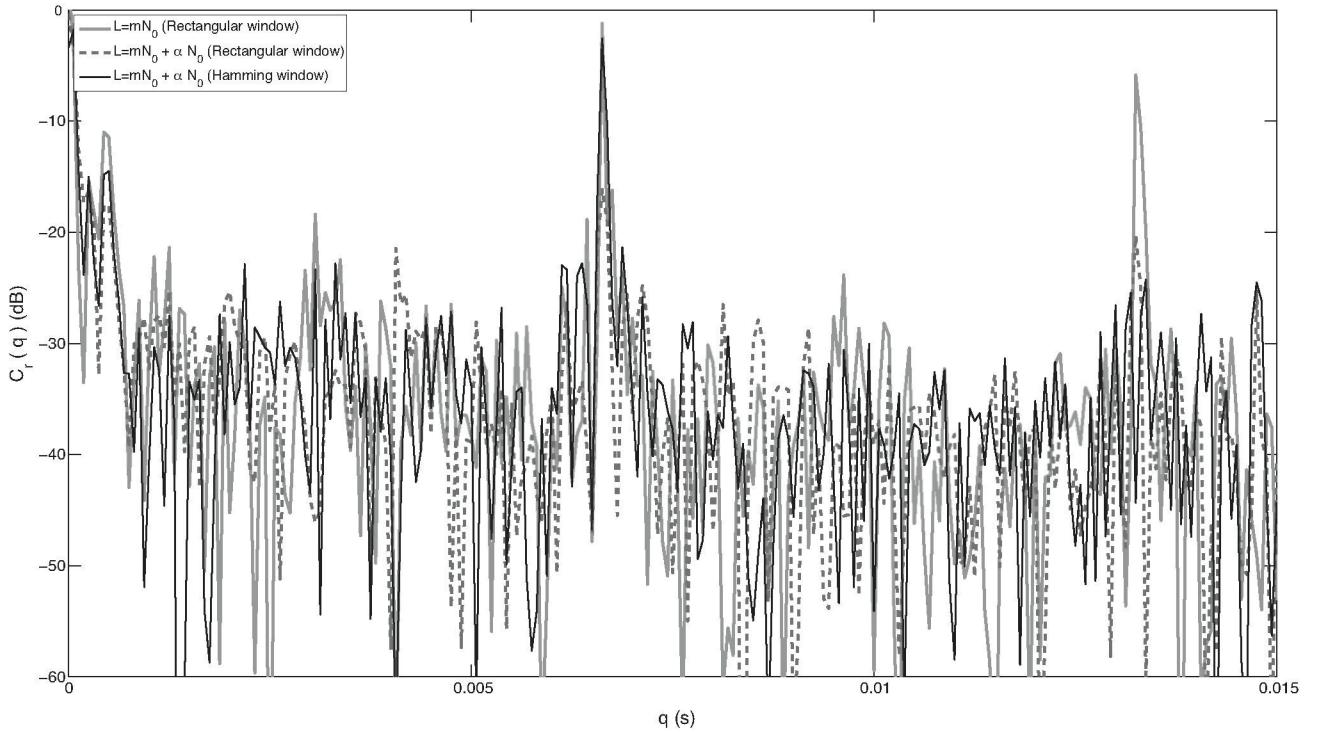


Fig. 6. Real cepstrum calculated from the spectral representation in Fig. 5.

amplitudes of both waveforms are similar. This results in the first rahmonic having a similar height but the next ones still differing (Fig. 6).

This analysis explains why the use of period-synchronous cepstral analysis leads to values of the cepstral peak that are higher than in the case of period-asynchronous analysis [45,44]. The factor  $m$  in the first term of (27) accounts for the increase of energy in time domain that results from taking several periods for calculating the Fourier transform. Normalisation would imply taking that factor out and multiplying the second term by a  $1/m$  factor. The effect of this would be an increase in the resolution of the harmonics representation in spectral domain. As pointed out by Murphy [44], such an increased resolution implies a flattening of the cepstrum that results in a reduction of the first rahmonic. However, the overall cepstral energy included in the whole set of rahmonics is increased as the number of periods included in the window grows. This is consistent with the behaviour of the sum of rahmonic amplitudes reported in [64].

The graphs in Figs. 5 and 6 also show that the choice of an appropriate window  $w[n]$  can help to reduce the impact of carrying out a pitch-asynchronous computation of CPP.

### 3.2. Sampling

Sampling in time domain corresponds to a windowing operation in spectral domain, that is, limiting the maximum frequency of the resulting signal [50]. Consequently, sampling has in cepstral domain effects similar to that of windowing in spectral domain:

- Limiting the spectrum (or log-spectrum) implies convolution in cepstral domain, which results in a widening of the rahmonic peaks and a reduction of their maximum values  $A_k$ . Thus, in principle sampling causes a reduction in the amplitude of the cepstral peak: the lower the sampling rate, the greater the reduction in the cepstral peak amplitude.

- Conversely, as the limit frequency (or sampling rate) is increased, the cepstral terms corresponding to the rahmonic structure grow (recall (27)) while the between-rahmonic values do not.
- A second effect of any increase on the limit frequency is an improved resolution in the rahmonics of the cepstrum.
- If frequency limitation is carried out harmonic-synchronously then the between-rahmonic values are reduced. Potentially, they would disappear for a perfectly periodic spectrum.

Therefore, limiting the spectrum, be it either by sampling or by spectral windowing, implies reducing the height of the cepstral peaks. However, according to the interpretation of the spectral meaning of  $A_1$  explained before, if it is removed only the part of the spectrum for which the harmonics are not significantly higher than the inter-harmonics level, the effect of spectrum limitation on the value of  $A_1$  can be diminished. Furthermore, the previously enumerated effects explain why the cepstral peak and the sum of cepstral peaks grow with fundamental frequency and also why a harmonic-synchronously limited spectrum also provides higher cepstral peaks [44,64].

## 4. Effect of perturbations

In this section we analyse the relations between CPP and typical perturbation parameters used for measuring aperiodicities on the acoustic voice signal: amplitude perturbation, frequency perturbation and noise.

### 4.1. Amplitude perturbation

By combining (5) and (16) we can write a periodic voice signal having fundamental period  $T_0$  as:

$$\begin{aligned}s_p(t) &= g^p(t) * v(t) * r(t) = \left( \sum_{k=-\infty}^{\infty} a \cdot \delta(t - kT_0) \right) * p(t) * v(t) * r(t) \\ &= \left( \sum_{k=-\infty}^{\infty} a \cdot \delta(t - kT_0) \right) * f(t)\end{aligned}\quad (28)$$

A signal with amplitude perturbations (i.e. shimmer) is a quasi-periodic signal in which the amplitudes of the periods vary according to a certain random distribution. If we call  $m_k$  the random values obtained from such a distribution, which has zero mean, then:

$$s_{\text{shim}}(t) = \left( \sum_{k=-\infty}^{\infty} a \cdot (1 + m_k) \cdot \delta(t - kT_0) \right) * f(t) \quad (29)$$

In spectral domain:

$$\begin{aligned}|S_{\text{shim}}(f)| &= \left| \sum_{k=-\infty}^{\infty} a \cdot (1 + m_k) \cdot e^{-j2\pi fkT_0} \right| \cdot |F(f)| = \left| \sum_{k=-\infty}^{\infty} a \cdot e^{-j2\pi fkT_0} \right. \\ &\quad \left. + \sum_{k=-\infty}^{\infty} a \cdot m_k \cdot e^{-j2\pi fkT_0} \right| \cdot |F(f)|\end{aligned}\quad (30)$$

Note that the second term in (30) is the Fourier transform of a random sequence. If such random sequence is white noise with zero mean and variance  $\sigma_{\text{SHIM}}^2$  then, on average [65]:

$$\begin{aligned}|S_{\text{shim}}(f)| &= \left| \sum_{k=-\infty}^{\infty} a \cdot e^{-j2\pi fkT_0} + a\sigma_{\text{SHIM}} \right| \cdot |F(f)| \\ &= \left| \frac{2\pi a}{T_0} \sum_{k=-\infty}^{\infty} F\left(2\pi \frac{k}{T_0}\right) \delta\left(2\pi f - 2\pi \frac{k}{T_0}\right) \right. \\ &\quad \left. + a \cdot \frac{\sigma_{\text{SHIM}}}{\sqrt{2B}} \cdot F(f) \right|\end{aligned}\quad (31)$$

where  $B$  is the signal bandwidth. For a windowed signal sampled at a rate  $f_s = 1/T_s$ , that is  $B = f_s/2$ :

$$\begin{aligned}|S_{\text{shim}}^w(f)| &= \left| \frac{a}{T_0} \sum_{k=-\infty}^{\infty} F\left(2\pi \frac{k}{T_0}\right) W\left(2\pi f - 2\pi \frac{k}{T_0}\right) \right. \\ &\quad \left. + a \cdot \frac{\sigma_{\text{SHIM}}}{2\pi\sqrt{f_s}} \cdot F(f) * W(f) \right|\end{aligned}\quad (32)$$

$W(f)$  is the Fourier transform of the time window  $w(t)$ . In what follows we will assume that it has a much narrower bandwidth than  $F(f)$ ; so  $F(f) * W(f)$  is a local averaging of  $F(f)$  and  $F(f) * W(f) \approx F(f)$ . Thus:

$$\begin{aligned}|S_{\text{shim}}^w(f)| &\approx \left| \frac{a}{T_0} \sum_{k=-\infty}^{\infty} F\left(2\pi \frac{k}{T_0}\right) W\left(2\pi f - 2\pi \frac{k}{T_0}\right) \right. \\ &\quad \left. + a \cdot \frac{\sigma_{\text{SHIM}}}{2\pi\sqrt{f_s}} \cdot F(f) \right|\end{aligned}\quad (33)$$

The first term in (33) corresponds to the spectral harmonic locations while the second one is associated to the between-harmonics

intervals. Therefore, for a voice signal affected by random shimmer, i.e. shimmer that can be modelled as white noise, to a great extent the harmonic amplitudes remain unaffected by shimmer variance. Such variance has its main impact on the between harmonic intervals. This result is consistent with the spectral interpretation of shimmer reported in [66].

The harmonic envelope is:

$$S_{\text{harm}}(f) = \frac{a}{T_0} F(f) W(f=0) = \frac{a}{T_0} F(f) W_0 \quad (34)$$

and its inter-harmonic counterpart:

$$S_{\text{inter-harm}}(f) \approx a \cdot \frac{\sigma_{\text{SHIM}}}{2\pi\sqrt{f_s}} \cdot F(f) \quad (35)$$

Recalling (21), the first cepstral peak is proportional to the geometric mean of the harmonics to inter-harmonics ratio. For the case of shimmer:

$$\begin{aligned}A_1 &\propto \int_0^{f_s/2} \log \left| \frac{S_{\text{harm}}(f)}{S_{\text{inter-harm}}(f)} \right| df \approx \int_0^{f_s/2} \log \left| \frac{(a/T_0)F(f)W_0}{a \cdot \frac{\sigma_{\text{SHIM}}}{2\pi\sqrt{f_s}} \cdot F(f)} \right| df \\ &= \int_0^{f_s/2} \log \left| \frac{2\pi\sqrt{f_s}W_0}{\sigma_{\text{SHIM}}T_0} \right| df\end{aligned}\quad (36)$$

Calculating the integral:

$$A_1 \propto \frac{f_s}{2} \log \left| \frac{2\pi\sqrt{f_s}W_0}{\sigma_{\text{SHIM}}T_0} \right| = \frac{f_s}{2} \log \left| \frac{2\pi\sqrt{f_s}W_0}{T_0} \right| - \frac{f_s}{2} \log \sigma_{\text{SHIM}} \quad (37)$$

Therefore, for certain values of  $C_0$  and  $C_1$  that depend on the fundamental period  $T_0$ , the sampling frequency  $f_s$  and the time window  $w(t)$ :

$$A_1 \propto C_0 - C_1 \cdot \log \sigma_{\text{SHIM}} \quad (38)$$

Consequently, the amplitude of the first rahmonic has an inverse relation to shimmer.

#### 4.2. Frequency perturbation

Using the same notation as in (29), a quasi-periodic voice signal affected by frequency perturbation (i.e. jitter) can be modelled as:

$$s_{\text{jit}}(t) = \left( \sum_{k=-\infty}^{\infty} a \cdot \delta(t - kT_0 - \tau_k) \right) * f(t) \quad (39)$$

where  $\tau_k$  is random variable. In what follows, it is assumed that  $\tau_k$  is uniformly distributed in the interval  $[-T_j/2, T_j/2]$ ; consequently, its average is zero and its variance  $\sigma_{\text{jit}}^2 = T_j^2/12$  [65].

According to the spectral jitter model proposed by Vasilakis and Stylianou [67], the spectrum of  $s_{\text{jit}}(t)$  when jitter is cyclic can be written as:

$$|G(f)|^2 = \frac{2\pi^2}{T_0^2} \sum_{k=-\infty}^{\infty} \left( 1 + \cos \left[ (T_0 - \tau_k)k \frac{\pi}{T_0} \right] \right) \delta \left( f - \frac{k}{2T_0} \right) \quad (40)$$

If such a signal is windowed:

$$\begin{aligned}
|G^W(f)|^2 &= \frac{\pi}{T_0^2} \sum_{k=-\infty}^{\infty} \left( 1 + \cos \left[ (T_0 - \tau_k) k \frac{\pi}{T_0} \right] \right) W \left( f - \frac{k}{2T_0} \right) \\
&= \frac{\pi}{T_0^2} \sum_{k=-\infty}^{\infty} \left( 1 + \cos \left[ k\pi - \tau_k \frac{k\pi}{T_0} \right] \right) W \left( f - \frac{k}{2T_0} \right) \\
&= \frac{\pi}{T_0^2} \sum_{k=-\infty}^{\infty} \left( 1 + \cos[k\pi] \cos \left[ \tau_k \frac{k\pi}{T_0} \right] \right) W \left( f - \frac{k}{2T_0} \right) \quad (41)
\end{aligned}$$

In  $|G^W(f)|^2$  the harmonics correspond to even values of  $k$ . Therefore, if  $W_0 = W(f=0)$  then the harmonic envelope is:

$$|G_{\text{harm}}^W(f)|^2 = \frac{\pi W_0}{T_0^2} (1 + \cos[2\pi f\tau_k]) \quad (42)$$

and in the inter-harmonic positions (odd values of  $k$ ):

$$|G_{\text{inter-harm}}^W(f)|^2 = \frac{\pi W_0}{T_0^2} (1 - \cos[2\pi f\tau_k]) \quad (43)$$

The average spectral envelopes are (see (B.3) in Appendix B):

$$|G_{\text{harm}}^W(f)|^2 = \frac{W_0}{T_0^2} \left( \pi + \frac{1}{fT_J} \sin(\pi fT_J) \right) \quad (44)$$

$$|G_{\text{inter-harm}}^W(f)|^2 = \frac{W_0}{T_0^2} \left( \pi - \frac{1}{fT_J} \sin(\pi fT_J) \right) \quad (45)$$

And the ratio between them:

$$\left| \frac{G_{\text{harm}}^W(f)}{G_{\text{inter-harm}}^W(f)} \right|^2 = \frac{\pi fT_J + \sin(\pi fT_J)}{\pi fT_J - \sin(\pi fT_J)} = \frac{(\pi fT_J / \sin(\pi fT_J)) + 1}{(\pi fT_J / \sin(\pi fT_J)) - 1} \quad (46)$$

The first cepstral peak is then proportional to:

$$\begin{aligned}
&\int_0^\infty \log \left| \frac{G_{\text{harm}}(f)}{G_{\text{inter-harm}}(f)} \right| df \\
&= \int_0^\infty \log \left| \frac{(\pi fT_J / \sin(\pi fT_J)) + 1}{(\pi fT_J / \sin(\pi fT_J)) - 1} \right| df \quad (47)
\end{aligned}$$

Since  $|(\pi fT_J / \sin(\pi fT_J))| > 1$  for  $f > 0$ , the logarithm can be written as a series [58]:

$$\begin{aligned}
&\int_0^\infty \log \left| \frac{(\pi fT_J / \sin(\pi fT_J)) + 1}{(\pi fT_J / \sin(\pi fT_J)) - 1} \right| df \\
&= 2 \sum_{n=0}^{\infty} \int_0^\infty \frac{1}{2n+1} \left( \frac{\sin(\pi fT_J)}{\pi fT_J} \right)^{2n+1} df \quad (48)
\end{aligned}$$

Using the solution of the integral reported in [68]:

$$\begin{aligned}
&\int_0^\infty \log \left| \frac{(\pi fT_J / \sin(\pi fT_J)) + 1}{(\pi fT_J / \sin(\pi fT_J)) - 1} \right| df \\
&= \frac{1}{T_J} \sum_{n=0}^{\infty} \sum_{r=0}^n \frac{(-1)^r (n-r+(1/2))^{2n}}{r!(2n-r+1)!} \quad (49)
\end{aligned}$$

Therefore, the first cepstral peak is proportional to the inverse of the standard deviation of period perturbations (i.e. jitter):

$$A_1 \propto \frac{1}{T_J} = \frac{1}{\sigma_{\text{jitter}} 2\sqrt{3}} \quad (50)$$

### 4.3. Noise

Glottal noise can be modelled as the combined effect of two components: pulsatile noise, that is, proportional in amplitude to the glottal pulse, and additive noise [69]. Therefore, a noisy voice signal can be expressed as:

$$\begin{aligned}
s_n(t) &= \left( \sum_{k=-\infty}^{\infty} a \cdot (1 + n_1(t)) \cdot p(t - kT_0) + n_2(t) \right) * h(t) \\
&= \left( \sum_{k=-\infty}^{\infty} a \cdot p(t - kT_0) \right) * h(t) + \left( \sum_{k=-\infty}^{\infty} a \cdot n_1(t) p(t - kT_0) \right) \\
&\quad * h(t) + n_2(t) * h(t) \quad (51)
\end{aligned}$$

The first term in (51) corresponds to the periodic component of the signal, the second term corresponds to a windowed glottal noise signal where the glottal pulse shape acts as a window and the third term is a white noise filtered by the vocal tract plus lip radiation response. Recalling that  $f(t) = p(t) * h(t)$ , the spectral representation of (51) is similar to that of a signal with shimmer plus an additive noise component:

$$\begin{aligned}
|S_n(f)| &= \left| \frac{2\pi a}{T_0} \sum_{k=-\infty}^{\infty} F \left( 2\pi \frac{k}{T_0} \right) \delta \left( 2\pi f - 2\pi \frac{k}{T_0} \right) + a \cdot \frac{\sigma_{n1}}{\sqrt{2B}} \cdot P \cdot F(f) \right. \\
&\quad \left. + \frac{\sigma_{n2}}{\sqrt{2B}} \cdot F(f) \right| \quad (52)
\end{aligned}$$

where  $P = \int_{-\infty}^{\infty} |P(f)| df$ .

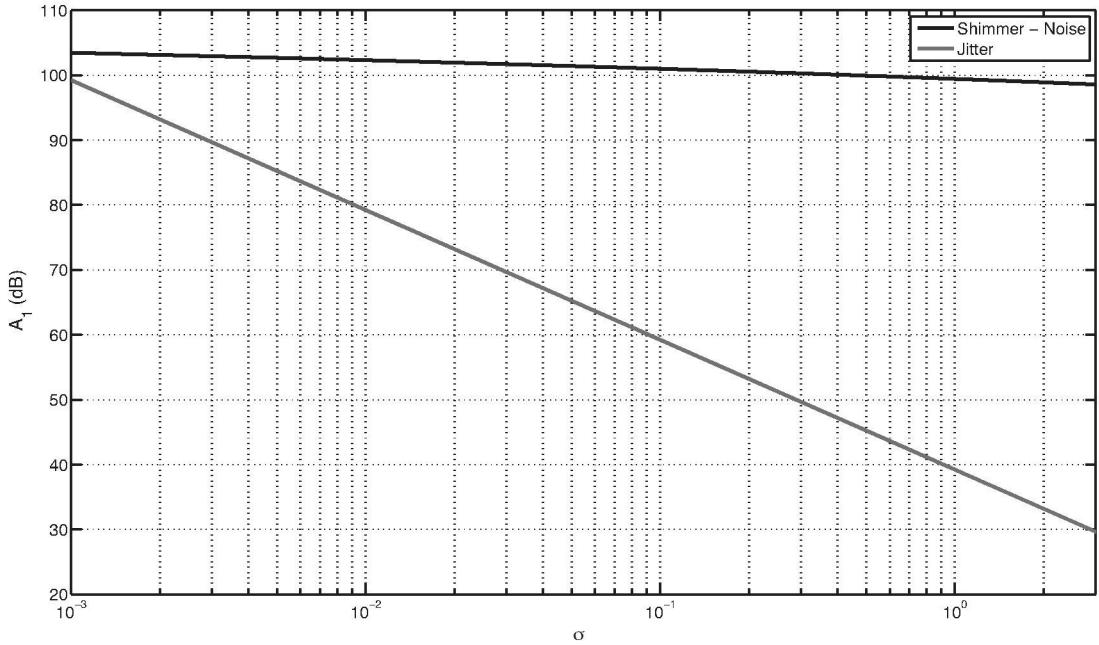
Using the result in (37):

$$\begin{aligned}
A_1 &\propto \frac{f_s}{2} \log \left| \frac{2\pi \sqrt{f_s} W_0}{(\sigma_{n1} \cdot P + \frac{\sigma_{n2}}{a}) T_0} \right| = \frac{f_s}{2} \log \left| \frac{2\pi \sqrt{f_s} W_0}{T_0} \right| \\
&\quad - \frac{f_s}{2} \log \left( \sigma_{n1} \cdot P + \frac{\sigma_{n2}}{a} \right) \quad (53)
\end{aligned}$$

Thus, the relationship between the first cepstral peak and the glottal noise power is similar to the relationship between the first cepstral peak and shimmer, except for the dependence on the glottal pulse shape and amplitude.

### 4.4. Discussion on the relation between cepstral peak and perturbation measures

Within this section we have analysed the relation between the amplitude of the cepstral peak and measures of amplitude, frequency and noise perturbations. We have shown that there is an inverse relation between these measures and the amplitude  $A_1$  of the cepstral peak. Fig. 7 shows the graphs corresponding to such dependences. The figure shows that the dependence between  $A_1$  and jitter is much more significant than the dependence between  $A_1$  and shimmer or noise. This is consistent with the findings of Murphy [44] and Heman-Ackah et al. [4]. Yet, the relation between  $A_1$  and shimmer or noise exists, although it is weaker. The relation with shimmer explains the correlation between CPP and roughness reported by Haderlein et al. [55]. The relation between CPP and harmonics to noise ratio (HNR) has also been reported [29] and the relation between  $A_1$  and the geometric average of the HNR has been reasoned by Murphy [44] and also shown in this paper. Interestingly, (53) shows a dependence between  $A_1$  and vocal intensity,



**Fig. 7.** Relation between cepstral peak  $A_1$  and standard deviation of shimmer, jitter and noise. The graph for shimmer and noise corresponds to (37) and (53), with  $\sigma = \sigma_{n1} \cdot P + (\sigma_{n2}/a)$  for the case of noise. The graph for jitter corresponds to (50) with a constant term in dB added for the sake of easing comparison between both graphs.

as measured by the amplitude parameter  $a$ , that has also been reported by Awan et al. [54]. Additionally, (53) highlights again the fact that CPP integrates information on noise level and pulse shape [62].

Last, it should be considered that the afore mentioned relations have been analysed independently from each other. However, in the analysis of real signals, perturbations do not happen independently; this may explain the differences in the conclusions obtained by diverse researchers when analysing different sets of voice recordings.

## 5. Conclusions

CPP has been reported to be one of the most reliable and robust acoustic cues of dysphonia [1]. To a significant extent, its robustness comes from the fact that it does not need previous pitch detection and tracking [26], an advantage that makes it outperform other acoustic measures when voice is recorded in non-controlled environments (e.g. office or clinic) [70]. To date, insights on what CPP actually measures had been provided via the analysis of the first cepstral peak, or first rahmonic [44], and the analysis of correlations with video-endoscopic measures [62], voice production model parameters [29,31], other acoustic measures [4,29,34] and perceived voice quality [1,26,40,45].

In this paper we have firstly analysed the meaning of the cepstral log-linear regression involved in the calculation of CPP following a parametric approach based on the classical source-filter model of voice production. From such an analysis we have concluded that the subtraction of the log-linear regression from the value of the first rahmonic in order to calculate the CPP has little impact on the spectrum of the signal. As a consequence, the conclusions of Heman-Ackah [53] regarding the relevance of the regression may well be more related to the relevance of calculating CPP following a systematic algorithm than to the effect of the regression itself. A second consequence of our analysis is that the interpretations on the meaning of the first rahmonic pointed out by Murphy [44] can be extrapolated to CPP.

In a second step, we have studied the relationship between the first cepstral peak and the spectrum by assuming that the cepstral peak has a Gaussian shape. Our results are fully coincident with the conclusions of Murphy [44] and they allow concluding that cepstral analysis based on CPP is only meaningful for voiced signals, hence discarding unvoiced segments or tracheo-oesophageal voice. This analysis also explains the relation between CPP and breathiness unveiled by Hillenbrand et al [26] and subsequent results.

Thirdly, the impact of signal windowing and sampling on the cepstral peak has been studied. The reasons why pitch-synchronous and harmonic-synchronous cepstral analyses provided better indicators of voice quality [44,45] have been illustrated. We have also shown that by an appropriate choice of the framing window, the effects of carrying out a pitch-asynchronous analysis can be reduced, thus keeping the advantage of CPP of not requiring pitch estimation. Additionally, the same analysis explains the direct relation between fundamental frequency and the amplitude of the cepstral peak. This relation together with the dependence between vocal tract response and the cepstral log-linear regression are plausible explanations for the variability of CPP with age [71] and sex [72].

Last, using an analytical framework we have shown that there is an inverse relation between the amplitude of the first cepstral peak and the variance of amplitude, frequency and noise perturbations of the voice signal. By considering both this last result and the above mentioned ones, one can confirm the previously published intuitions that CPP integrates measures of several features describing the aperiodicity and waveform of the acoustic voice signal [26,40,62]. In turn, this integration of several measures explains the relation between CPP and overall dysphonia [1] and, simultaneously, the fact that CPP is not particularly adequate for predicting specific aspects of voice quality, even when these are related to breathiness [73].

## Acknowledgement

This work has been carried out in the framework of project Grant TEC2012-38630-C04-01, financed by the Spanish Government.

## Appendix A. Real cepstrum of an all-pole function

The following relations are based on the properties of the Fourier transform (see e.g. [50]). As for the first term in (11), the Fourier transform of a constant is an impulse (Dirac delta):

$$\mathcal{F}\{\log |H_0|\} = 2\pi \cdot \log |H_0| \cdot \delta(q) \quad (\text{A.1})$$

Regarding the second term in (11), given a complex pole  $s_p$ , it can be expressed as  $s_p = \sigma_p + j\omega_p$ . Therefore:

$$\begin{aligned} \mathcal{F}\{\log |j\omega - s_p|\} &= \mathcal{F}\{\log |-\sigma_p + j(\omega - \omega_p)|\} \\ &= \mathcal{F}\{\log \sqrt{\sigma_p^2 + (\omega - \omega_p)^2}\} \\ &= \frac{1}{2} \cdot \mathcal{F}\{\log(\sigma_p^2 + (\omega - \omega_p)^2)\} \end{aligned} \quad (\text{A.2})$$

Additionally, if we apply the time differentiation property of the Fourier transform:

$$\mathcal{F}\{\log |j\omega - s_p|\} = \frac{1}{2} \cdot \frac{1}{jq} \cdot \mathcal{F}\left\{ \frac{2(\omega - \omega_p)}{\sigma_p^2 + (\omega - \omega_p)^2} \right\} \quad (\text{A.3})$$

Considering now the time shift property:

$$\mathcal{F}\{\log |j\omega - s_p|\} = \frac{1}{2} \cdot \frac{1}{jq} \cdot e^{-jq\omega_p} \mathcal{F}\left\{ \frac{2\omega}{\sigma_p^2 + \omega^2} \right\} \quad (\text{A.4})$$

Applying the frequency differentiation property to (A.4):

$$\begin{aligned} \mathcal{F}\{\log |j\omega - s_p|\} &= \frac{1}{2} \cdot \frac{1}{jq} \cdot e^{-jq\omega_p} \cdot \frac{-2}{j} \cdot \frac{\partial}{\partial q} \left( \mathcal{F}\left\{ \frac{1}{\sigma_p^2 + \omega^2} \right\} \right) \\ &= \frac{1}{q} \cdot e^{-jq\omega_p} \cdot \frac{\partial}{\partial q} \left( \mathcal{F}\left\{ \frac{1}{\sigma_p^2 + \omega^2} \right\} \right) \end{aligned} \quad (\text{A.5})$$

The last Fourier transform in (A.5) corresponds to one basic transform pair. Considering that  $\sigma_p < 0$  for the all-pole system to be stable:

$$\begin{aligned} \mathcal{F}\{\log |j\omega - s_p|\} &= \frac{1}{q} \cdot e^{-jq\omega_p} \cdot \frac{\partial}{\partial q} \left( \frac{1}{-2\sigma_p} \cdot e^{\sigma_p|q|} \right) \\ &= \begin{cases} \frac{1}{q} \cdot e^{-jq\omega_p} \cdot \frac{1}{2} \cdot e^{-\sigma_p q} & \text{if } q < 0 \\ \frac{1}{q} \cdot e^{-jq\omega_p} \cdot -\frac{1}{2} \cdot e^{\sigma_p q} & \text{if } q > 0 \end{cases} \\ &= -\frac{1}{2|q|} \cdot e^{-jq\omega_p} \cdot e^{\sigma_p|q|} \end{aligned} \quad (\text{A.6})$$

Since the speech signal is real valued, all poles either are real or they come in complex-conjugate pairs. If we assume that the system has  $n_{rp}$  real poles ( $\omega_p = 0$ ) and  $2 \cdot n_{cp}$  complex poles so that  $n_{rp} + 2 \cdot n_{cp} = n_p$  then:

$$\begin{aligned} \sum_{p=1}^{n_p} \mathcal{F}\{\log |j\omega - s_p|\} &= \sum_{k=1}^{n_{rp}} -\frac{e^{\sigma_k|q|}}{2|q|} + \sum_{l=1}^{n_{cp}} -\frac{e^{-jq\omega_l} + e^{jq\omega_l}}{2|q|} \cdot e^{\sigma_l|q|} \\ &= -\frac{1}{|q|} \cdot \left( \sum_{k=1}^{n_{rp}} \frac{e^{\sigma_k|q|}}{2} + \sum_{l=1}^{n_{cp}} \cos(\omega_l|q|) \cdot e^{\sigma_l|q|} \right) \end{aligned} \quad (\text{A.7})$$

## Appendix B. Average value of the spectral envelope of a jittered signal

According to (42) and (43), the envelopes of the spectrum of a windowed jittered signal can be expressed as

$$|G_{\text{env}}^w(f)|^2 = \frac{\pi W_0}{T_0^2} (1 - \cos \varphi \cos[2\pi f\tau_k]) \quad (\text{B.1})$$

with  $\varphi \in \{0, \pi\}$ . If  $\tau_k$  is uniformly distributed in  $[-T_j/2, T_j/2]$ , then the average spectral envelope can be calculated as:

$$|G_{\text{env}}(f)|^2 = \int_{-T_j/2}^{T_j/2} \frac{\pi W_0}{T_0^2} (1 + \cos \varphi \cos[2\pi f\tau_k]) \frac{1}{T_j} d\tau_k \quad (\text{B.2})$$

Solving:

$$\begin{aligned} |G_{\text{env}}(f)|^2 &= \frac{\pi W_0}{T_0^2} + \frac{\pi W_0}{T_0^2} \frac{1}{T_j} \cos \varphi \int_{-T_j/2}^{T_j/2} \cos[2\pi f\tau_k] d\tau_k \\ &= \frac{\pi W_0}{T_0^2} \left( 1 + \frac{1}{T_j} \cos \varphi \left[ \frac{1}{2\pi f} \sin(2\pi f\tau_k) \right]_{\tau_k=-T_j/2}^{T_j/2} \right) \\ &= \frac{\pi W_0}{T_0^2} \left( 1 + \frac{1}{2\pi f T_j} \cos \varphi \cdot 2 \sin \left( 2\pi f \frac{T_j}{2} \right) \right) \\ &= \frac{W_0}{T_0^2} \left( \pi + \frac{1}{f T_j} \cos \varphi \sin(\pi f T_j) \right) \end{aligned} \quad (\text{B.3})$$

## References

- [1] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, P. Corthals, Acoustic measurement of overall voice quality: a meta-analysis, *J. Acoust. Soc. Am.* 126 (2009) 2619–2634.
- [2] V. Wolfe, D. Martin, Acoustic correlates of dysphonia: type and severity, *J. Commun. Disord.* 30 (1997) 403–416.
- [3] V.I. Wolfe, D.P. Martin, C.I. Palmer, Perception of dysphonic voice quality by naive listeners, *J. Speech Lang. Hear. Res.* 43 (2000) 697–705.
- [4] Y.D. Heman-Ackah, D.D. Michael, G.S. Goding, The relationship between cepstral peak prominence and selected parameters of dysphonia, *J. Voice* 16 (2002) 20–27.
- [5] B. Halberstam, Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels, *ORL* 66 (2004) 70–73.
- [6] T.L. Eadie, C.R. Baylor, The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice, *J. Voice* 20 (2006) 527–544.
- [7] S.N. Awan, N. Roy, C. Dromey, Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model, *Clin. Linguist. Phonetics* 23 (2009) 825–841.
- [8] S.N. Awan, N. Roy, Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity, *J. Speech Lang. Hear. Res.* 52 (2009) 482–499.
- [9] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, M. De Bodt, Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels, *J. Voice* 24 (2010) 540–555.
- [10] Y. Maryn, M. De Bodt, N. Roy, The acoustic voice quality index: toward improved treatment outcomes assessment in voice disorders, *J. Commun. Disord.* 43 (2010) 161–174.
- [11] Y.L. Shue, G. Chen, A. Alwan, On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures, in: Proc. Interspeech, Makuhari, 2010, pp. 34–37.
- [12] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen, Multi-band dysperiodicity analyses of disordered connected speech, *Speech Commun.* 53 (2011) 131–141.
- [13] E.A. Peterson, N. Roy, S.N. Awan, R.M. Merrill, R. Banks, K. Tanner, Toward validation of the Cepstral Spectral Index of Dysphonia (CSID) as an objective treatment outcomes measure, *J. Voice* 27 (2013) 401–410.
- [14] D.M. Hartl, J. Vaisse, O. Laccourreye, D.F. Brasnu, Acoustic analysis of autologous fat injection versus thyroplasty in the same patient, *Ann. Otol. Rhinol. Laryngol.* 112 (2003) 987–992.
- [15] N.P. Solomon, S.N. Awan, L.B. Helou, A. Stojadinovic, Acoustic analyses of thyroidectomy-related changes in vowel phonation, *J. Voice* 26 (2012) 711–720.
- [16] M. Merk, W. Ziegler, B. Brendel, Acoustic assessment of neurogenic voice disorders in a clinical setting, in: International Workshop on Models and Analysis

- of Vocal Emissions for Biomedical Applications, MAVEBA, Florence, 1999, pp. 83–85.
- [17] M.O. Rosa, J.C. Pereira, M. Greller, A.C.P.L.F. Carvalho, Signal processing and statistical procedures to identify laryngeal pathologies, in: Proc. IEEE Internat. Conf. Electronics, Circuits, and Systems, ICECS'99, vol. 1, Pafos, 1999, pp. 423–426.
- [18] D.M. Hartl, S. Hans, J. Vaissière, D.F. Brasnu, Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia, *Eur. Arch. Oto Rhino Laryngol.* 260 (2003) 175–182.
- [19] D.M. Hartl, S. Hans, J. Vaissière, M. Riquet, D.F. Brasnu, Objective voice quality analysis before and after onset of unilateral vocal fold paralysis, *J. Voice* 15 (2001) 351–361.
- [20] R.K. Balasubramanian, J.S. Bhat, S. Fahim III, R. Raju III, Cepstral analysis of voice in unilateral adductor vocal fold palsy, *J. Voice* 25 (2011) 326–329.
- [21] B.R. Kumar, J.S. Bhat, N. Prasad, Cepstral analysis of voice in persons with vocal nodules, *J. Voice* 24 (2010) 651–653.
- [22] C.R. Watts, S.N. Awan, Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts, *J. Speech Lang. Hear. Res.* 54 (2011) 1525–1537.
- [23] T. Haderlein, C. Moers, B. Möbius, F. Rosanowski, E. Nöth, Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation, in: I. Habernal, V. Matoušek (Eds.), *Text, Speech, and Dialogue*, Volume 6836 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2011, pp. 195–202.
- [24] T.F. Yap, J. Epps, E. Ambikairajah, E.H.C. Choi, Voice source features for cognitive load classification, in: Proc. IEEE Internat. Conf. Acoustics, Speech, and Signal Processing – ICASSP 2011, Prague, 2011, pp. 5700–5703.
- [25] R.K. Balasubramanian, J.S. Bhat, M. Srivastava, A. Eldose, Cepstral analysis of sexually appealing voice, *J. Voice* 26 (2012) 412–415.
- [26] J. Hillenbrand, R.A. Cleveland, R.L. Erickson, Acoustic correlates of breathy vocal quality, *J. Speech Lang. Hear. Res.* 37 (1994) 769–778.
- [27] J. Hillenbrand, R.A. Houde, Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech, *J. Speech Lang. Hear. Res.* 39 (1996) 311–321.
- [28] R. Shrivastav, C.M. Sapienza, Objective measures of breathy voice quality obtained using an auditory model, *J. Acoust. Soc. Am.* 114 (2003) 2217–2224.
- [29] R.A. Samlan, B.H. Story, K. Bunton, Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling, *J. Speech Lang. Hear. Res.* 56 (2013) 1209–1223.
- [30] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, P. Murphy, Cepstral analysis of vocal dysperiodicities in disordered connected speech, in: Proc. Interspeech, Brighton, 2009, pp. 959–962.
- [31] R.A. Samlan, B.H. Story, Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling, *J. Speech Lang. Hear. Res.* 54 (2011) 1267–1283.
- [32] B. Blankenship, The timing of nonmodal phonation in vowels, *J. Phonetics* 30 (2002) 163–191.
- [33] C.M. Esposito, The effects of linguistic experience on the perception of phonation, *J. Phonetics* 38 (2010) 306–316.
- [34] Y.D. Heman-Ackah, R.J. Heuer, D.D. Michael, R. Ostrowski, M. Horman, M.M. Baroody, J. Hillenbrand, R.T. Sataloff, Cepstral peak prominence: a more reliable measure of dysphonia, *Ann. Otol. Rhinol. Laryngol.* 112 (2003) 324–333.
- [35] S.N. Awan, N. Roy, Acoustic prediction of voice type in women with functional dysphonia, *J. Voice* 19 (2005) 268–282.
- [36] D.M. Howard, E. Abberton, A. Fourcin, Disordered voice measurement and auditory analysis, *Speech Commun.* 54 (2012) 611–621.
- [37] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, T. Haderlein, Vowel- and text-based cepstral analysis of chronic hoarseness, *J. Voice* 26 (2012) 416–424.
- [38] S.Y. Lowell, R.H. Colton, R.T. Kelley, S.A. Mizia, Predictive value and discriminant capacity of cepstral- and spectral-based measures during continuous speech, *J. Voice* 27 (2013) 393–400.
- [39] C.M. Esposito, An acoustic and electroglossographic study of White Hmong tone and phonation, *J. Phonetics* 40 (2012) 466–476.
- [40] M.P. Cannito, M. Doiuchi, T. Murry, G.E. Woodson, Perceptual structure of adductor spasmodic dysphonia and its acoustic correlates, *J. Voice* 26 (2012), 818.e5–818.e13.
- [41] D.D. Mehta, R.E. Hillman, Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods, *Curr. Opin. Otolaryngol. Head Neck Surg.* 16 (2008) 211–215.
- [42] S.N. Awan, N.P. Solomon, L.B. Helou, A. Stojadinovic, Spectral-cepstral estimation of dysphonia severity: external validation, *Ann. Otol. Rhinol. Laryngol.* 122 (2013) 40–48.
- [43] C.A. Ferrer, M.S. De Bodt, Y. Maryn, P. Van de Heyning, M.E. Hernández-Díaz, Properties of the cepstral peak prominence and its usefulness in vocal quality measurements, in: Internat. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVEBA 2007, Florence, 2007, pp. 93–96.
- [44] P.J. Murphy, On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals, *J. Acoust. Soc. Am.* 120 (2006) 2896–2907.
- [45] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, P. Murphy, Assessment of disordered voice via the first rahmonic, *Speech commun.* 54 (2012) 655–663.
- [46] L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
- [47] A.M. Noll, Cepstrum pitch determination, *J. Acoust. Soc. Am.* 41 (1967) 293–309.
- [48] D.G. Childers, D.P. Skinner, R.C. Kemerait, The cepstrum: a guide to processing, *Proc. IEEE* 65 (1977) 1428–1443.
- [49] A. Oppenheim, R. Schafer, *Homomorphic analysis of speech*, *IEEE Trans. Audio Electroacoust.* 16 (1968) 221–226.
- [50] S. Haykin, *B.V. Veen, Signals and Systems*, John Wiley & Sons, 2001.
- [51] B.H. Story, I.R. Titze, Parameterization of vocal tract area functions by empirical orthogonal modes, *J. Phonetics* 26 (1998) 223–260.
- [52] R. Fraile, M. Kob, J.I. Godino-Llorente, N. Sáenz-Lechón, V.J. Osma-Ruiz, J.M. Gutiérrez-Ariola, Physical simulation of laryngeal disorders using a multiple-mass vocal fold model, *Biomed. Signal Process. Control* 7 (2012) 65–78.
- [53] Y.D. Heman-Ackah, Reliability of calculating the cepstral peak without linear regression analysis, *J. Voice* 18 (2004) 203–208.
- [54] S.N. Awan, A. Giovinco, J. Owens, Effects of vocal intensity and vowel type on cepstral analysis of voice, *J. Voice* 26 (5) (2012), 670.e15–e20.
- [55] T. Haderlein, C. Moers, B. Möbius, E. Nöth, Automatic rating of hoarseness by text-based cepstral and prosodic evaluation, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), *Text, Speech, and Dialogue*, Volume 7499 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2012, pp. 573–580.
- [56] S.Y. Lowell, R.H. Colton, R.T. Kelley, Y.C. Hahn, Spectral- and cepstral-based measures during continuous speech: Capacity to distinguish dysphonia and consistency within a speaker, *J. Voice* 25 (2011) e223–e232.
- [57] D.G. Childers, C.K. Lee, Vocal quality factors: analysis, synthesis, and perception, *J. Acoust. Soc. Am.* 90 (1991) 2394–2410.
- [58] M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, 1972.
- [59] R.P. Clapham, C.J. Van As-Brooks, M.W.M. Van den Brekel, F.J.M. Hilgers, R.J.H. Van Son, Automatic tracheoesophageal voice typing using acoustic parameters, in: Proc. Interspeech, Lyon, 2013, pp. 2162–2166.
- [60] K. Nagle, T. Eadie, Determining time- and frequency-based acoustic correlates of listener effort in tracheoesophageal speech, in: Proc. Internat. Conf. Advances in Quantitative Laryngology – AQL 2013, Cincinnati, 2013, pp. 99–100.
- [61] P. Dejonckere, Effect of louder voicing on acoustical measurements in dysphonic patients, *Logoped. Phoniatrics Vocal.* 23 (2) (1998) 79–84.
- [62] D.D. Mehta, S.M. Zeitels, J.A. Burns, A.D. Friedman, D.D. Deliyiski, R.E. Hillman, High-speed videoendoscopic analysis of relationships between cepstral-based acoustic measures and voice production mechanisms in patients undergoing phonomicrosurgery, *Ann. Otol. Rhinol. Laryngol.* 121 (2012) 341–347.
- [63] A.V. Oppenheim, R.W. Schafer, J.R. Buck, *Discrete-Time Signal Processing*, Prentice Hall, 1999.
- [64] P.J. Murphy, Periodicity estimation in synthesized phonation signals using cepstral rahmonic peaks, *Speech Commun.* 48 (2006) 1704–1713.
- [65] K.S. Shanmugam, A.M. Breipohl, *Random Signals: Detection, Estimation, and Data Analysis*, Wiley, 1988.
- [66] P.J. Murphy, Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals, *J. Acoust. Soc. Am.* 107 (2000) 978–988.
- [67] M. Vasilakis, Y. Stylianou, Spectral jitter modeling and estimation, *Biomed. Signal Process. Control* 4 (2009) 183–193.
- [68] R.G. Medhurst, J.H. Roberts, Evaluation of the integral  $I_n(b) = \frac{2}{\pi} \int_0^{\infty} \left( \frac{\sin x}{x} \right)^n \cos(bx) dx$ , *Math. Comput.* 19 (1965) 113–117.
- [69] S. Fraj, F. Grenez, J. Schoentgen, Synthesis of breathy and rough voices with a view to validating perceptual and automatic glottal cycle pattern recognition, in: Internat. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications – MAVEBA 2011, Florence, 2011, pp. 135–138.
- [70] K. Leong, M.J. Hawkshaw, D. Dentchev, R. Gupta, D. Lurie, R.T. Sataloff, Reliability of objective voice measures of normal speaking voices, *J. Voice* 27 (2013) 170–176.
- [71] R. Vipperla, S. Renals, J. Frankel, Ageing voices: the effect of changes in voice parameters on ASR performance, *EURASIP J. Audio Speech Music Process.* 2010 (2010) 525783.
- [72] G. Chen, X. Feng, Y.L. Shue, A. Alwan, On using voice source measures in automatic gender classification of children's speech, in: Proc. Interspeech 2010, Makuhari, 2010, pp. 673–676.
- [73] R. Shrivastav, A. Camacho, A computational model to predict changes in breathiness resulting from variations in aspiration noise level, *J. Voice* 24 (2010) 395–405.