A CEPSTRUM-BASED HARMONICS-TO-NOISE RATIO IN VOICE SIGNALS

P.J. Murphy

Department of Electronic and Computer Engineering, University of Limerick, National Technological Park, Plassey, Limerick, Ireland.

ABSTRACT

A new cepstrum-based technique is developed in order to provide an alternative means of estimating the harmonics-to-noise ratio in voice signals. The geometric mean harmonics-to-noise ratio (GHNR) is defined as the mean of the individual spectral (i.e. at specific frequency locations) harmonics-to-noise ratios in dB. A heuristic development of the method treats the harmonic spectrum (in dB) of voiced speech taken over several cycles of the waveform as a more usual time domain signal, which is Fourier transformed. The sum of the resulting cepstral peaks (rahmonics) gives a direct estimation of the geometric mean harmonics-to-noise ratio (GHNR). The need for, inverse Fourier transform of the masked cepstrum back into the frequency domain, baseline correction and the usual harmonics-to-noise ratio (HNR) calculation is avoided by this approach. The technique is examined using synthetically generated voice signals.

1. INTRODUCTION

A cepstrum-based technique (where the cepstrum is defined as the inverse Fourier transform of the Fourier magnitude spectrum of the time-domain signal) for extracting the HNR in speech signals is given in [1]. In this calculation, the rahmonics (prominent peaks at integer multiples of 1/T for voiced speech of period T) in the cepstrum are removed, the resulting liftered (filtered) cepstrum is Fourier transformed to provide a noise spectrum, which is subtracted from the original log spectrum. This results in, what is termed here, a source related spectrum. After performing a baseline correction procedure on this spectrum, the modified noise spectrum is subtracted from the original log spectrum in order to provide the harmonics-to-noise ratio estimate. A modification to this technique has recently been introduced [2] in which the need for the baseline correction is removed, therefore providing a more efficient algorithm for HNR determination. These techniques do supply an estimate of the noise level at each frequency location in the spectrum, however three Fourier transforms are required in the estimation procedure. An alternative approach, based on direct measurement of the height of rahmonic peaks in the cepstrum, is taken in the present study. The method follows fewer steps, leading to a new index termed the geometric mean harmonics-to-noise ratio (GHNR).

2. METHOD

Idealised voiced speech (si(t), i.e. noiseless) is represented as the convolution of an impulse train excitation, e(t) (spaced at integer multiples of the fundamental period, T) with the impulse responses of a glottal shaping filter, the vocal tract transfer function and a radiation function (Eq. (1)). Taking the Fourier transform of Eq. (1) gives the speech spectrum in terms of the impulse excitation spectrum multiplied by the frequency response of the glottal, vocal tract and radiation components (Eq. (2)). These latter components are grouped together in Eq. (3) for convenience and absolute values are taken.

\[
\begin{align*}
\text{si}(t) &= e(t) \ast g(t) \ast v(t) \ast r(t) \\
\text{Si}(f) &= E(f) \times G(f) \times V(f) \times R(f) \\
|\text{Si}(f)| &= |E(f)||Vg,r(f)|
\end{align*}
\]

F represents Fourier transformation and vg,r(t) and Vg,r(f) represent the combined transfer characteristics of the vocal tract (v), glottal flow (g) and radiation (r), in the time and frequency domain respectively.

The combined transfer function is a slowly varying function while the Fourier transform of the periodic impulse excitation Fe(t) is a periodic impulse train E(f), spaced at 1/T.

Taking the logarithm of the spectrum gives (the logarithm is taken at the harmonic locations of |Si(f)| only, to avoid taking the log of zero)

\[
\log|\text{Si}(f)| = \log|E(f)||Vg,r(f)| = \log|E(f)| + \log|Vg,r(f)|
\]

In taking the logarithm, the multiplicative components consisting of the harmonic excitation (periodic) and filter function (DC or low “frequency”) are changed into additive.
components. The inverse (or forward) Fourier transform can now be applied, operating individually on the two additive components, conveniently separating the slow modulation (combined vocal tract function) from the constant amplitude harmonics. \( E(f) \) is usually considered to be of unit amplitude. This is avoided in the present development as \( \log(E(f)) \) at harmonic locations would then be zero. Any arbitrary constant not equal to one can be used.

\[
C(q) = \text{IDFT}[\log|S_i(f)|] \tag{5}
\]

where \( q \) represents quefrency, ‘C’ represents cepstrum and IDFT is the inverse discrete Fourier transform.

In the above development, the signal, \( \log|S_i(f)| \), is considered as a standard “time” signal with slowly varying modulation superimposed on an underlying periodic structure, the (inverse) Fourier transform of which, gives a high amplitude at locations in the “frequency domain” corresponding to these frequencies (Fig. 1). Since, the signal being transformed is in the frequency domain to begin with, new terminology (cepstrum (spectrum), rahmonics (harmonics), quefrency (frequency), etc.) is employed to reflect the distinction between the resultant and what would have occurred using actual time domain signals.

**Figure 1**: (a) Fourier spectrum of an idealized periodic speech waveform, and (b) its corresponding cepstrum with rahmonics spaced at integer multiples of \( T \), the fundamental period.

In the above development it has been shown that the cepstrum is in theory capable of separating the periodic impulse excitation (to within a multiplicative constant) from the combined transfer function of the vocal tract. Therefore, the cepstrum can separate out the periodicity characteristic of a perfectly periodic voiced speech waveform (and hence is useful for fundamental frequency extraction), but cannot separate out the glottal flow spectral characteristic, which remains combined within the low quefrency signal.

A cepstral analysis of non-idealised waveforms \( s(t) \) i.e. waveforms containing additive noise at the glottal source, as found in real speech signals (Eq. (6)), gives rise to cepstral rahmonics that contain information that is useful for HNR determination.

\[
s(t) = (e(t)*g(t)+n(t))*v(t)*r(t) \tag{6}
\]

\[
S(f) = (E(f)*G(f)+N(f))*V(f)*R(f) \tag{7}
\]
\[
\log|S(f)| = \log|E(f) \times G(f) + N(f)| + \log|V_r(f)|
\]  
(8)

where \(S(f) = F(s(t))\), \(E(f) = F(e(t))\), and \(V_r(f) = F(v_r(t))\)

where \(F\) represents Fourier transformation and \(v_r(t)\) and \(V_r(f)\) represent the combined transfer characteristics of the vocal tract \((v)\) and radiation \((r)\). ‘n’ and ‘N’ represent time and frequency domain additive noise respectively.

The concept behind the present technique follows directly from Eq. (8). Noise, although broadband, is considered negligible at harmonic locations (this approximation can be made more exact if the spectral estimates are averaged prior to taking the inverse Fourier transform), while the signal energy is zero at ‘between harmonic’ locations. In obtaining the cepstrum, the heights of the rahmonic peaks are dependent on the depth of the valleys between adjacent harmonic locations. The ‘between harmonic’ locations are no longer zero as they were for the perfectly periodic waveform. Therefore, the noise contributions can be considered to be contained within the height of the cepstral rahmonic peaks i.e. they limit their height (Consider taking the (inverse) Fourier transform of the spectrum shown in Fig. 2). Consequently, estimating the height of the rahmonics provides an alternative approach for extracting an HNR. The rahmonic amplitudes are not directly related to the HNR of the radiated speech waveform because they are independent of the actual envelope or ‘DC’ component of the original spectrum. As the averaging process of the Fourier transform is extracted from a dB signal, the sum of the rahmonics, where each individual rahmonic peak is obtained from the Fourier average of the frequency spectrum in dB, provides a geometrical mean HNR (GHNR).

In the present approach two Fourier transforms are required as opposed to three, and the usual HNR calculation is not required. The rahmonic peaks are simply summed. There is no problem with respect to adding the rahmonics as they are linear in amplitude even though their overall sum represents a dB derived ratio.

In any practical realization of the present method a window is applied to the signal, hence convolving the spectral estimates. In terms of the resulting cepstrum, this will tend to limit the height of the rahmonic peaks.

3. ANALYSIS

The vowel a/ is synthesized using an implementation of the discrete-time model for speech production. A sequence of glottal pulses is used as input into a delay line digital filter, where the filter coefficients are obtained based on area function data for the vowel a/. Radiation at the lips is modeled by the first order difference equation. The sampling frequency for the synthesis is 10 kHz.

Random additive noise is introduced by multiplying the glottal pulse by a random noise generator arranged to give signal dependent additive noise of a user specified variance (std. dev. 1 to 32 %). Signals are created for three levels of additive noise for frequencies beginning at 80 Hz and increasing in six, approximately equi-spaced steps of 60 Hz up to 350 Hz. Six noise levels are synthesized at 110 Hz. A 2048-point Hamming window is used for the spectrum analysis.

4. RESULTS

The set consists of three different noise levels for six different fundamental frequencies ranging from 80 Hz to 350 Hz and therefore covering the extremes of the expected fundamental frequency in modal register. Fig. 3 shows GHNR for speech waveforms with variation in additive noise of the glottal source of s.d. 4 %, 8 % and 16 %.

The observed trend of increasing HNR with increasing fundamental frequency for synthetic speech signals has been documented by other researchers (e.g. [1]). An explanation of the trend is given in [3]. It is due to the differential formant excitation by glottal source signals, which differ in fundamental frequency. Fig. 4 shows a linear relationship between GHNR and increasing noise levels for the 110 Hz signal.

![Figure 2: Typical harmonics-to-noise ratio (HNR) approximation via the spectrum.](image-url)
5. CONCLUSION

A new cepstrum-based technique for determining an HNR in speech signals has been investigated. The method is developed by considering noise at the glottal source, however, the technique is equally suited to determining the HNR for speech signals contaminated with extraneous noise.

![Graph showing geometric mean harmonics-to-noise ratio (GHNR) versus fundamental frequency for three level of additive glottal noise.]

Figure 3: Geometric mean harmonics-to-noise ratio (GHNR) versus fundamental frequency for three level of additive glottal noise

![Graph showing geometric mean harmonics-to-noise ratio (GHNR) versus additive glottal noise (six levels) for the 110 Hz signal.]

Figure 4: Geometric mean harmonics-to-noise ratio (GHNR) versus additive glottal noise (six levels) for the 110 Hz signal.

The proposed index (GHNR) is shown to give an indication of the HNR for the range of frequencies of interest in modal register. In defining GHNR, the meaning of the height of the rahmonic peaks in the cepstrum is explained for the first time. Future development of the technique may include auditory processing of the spectrum (although retaining harmonic structure), as it has been shown [4] that perceptually based HNRs correlate more closely with the perception of speech quality than traditional HNRs (average correlation coefficient changes from 0.77 to 0.93). The use of rahmonic peaks shows promise for efficient quantitative perceptual evaluation of connected speech.

6. REFERENCES


