A NEW ACCURATE METHOD OF HARMONIC-TO-NOISE RATIO EXTRACTION

Ricardo J. T. de Sousa
School of Engineering , University of Porto, Rua Roberto Frias, Porto, Portugal

Keywords: Voice quality, Voice diagnosis, Harmonic-to-noise ratio, Hoarseness, Roughness.

Abstract: In this paper, an accurate method that estimates the HNR from sustained vowels based on harmonic structure modeling is proposed. Basically, the proposed algorithm creates an accurate harmonic structure where each harmonic is parameterized by frequency, magnitude and phase. The harmonic structure is then synthesized and assumed as the harmonic component of the speech signal. The noise component can be estimated by subtracting the harmonic component from the speech signal. The proposed algorithm was compared to others HNR extraction algorithms based on spectral, cepstral and time domain methods, and using different performance measures.

1 INTRODUCTION

1.1 Speech Assessment

In the audition tests which are include in speech assessment, perceptive parameters such as hoarseness, breathiness and roughness are evaluated in order to characterize the physiological changes of the vocal folds. In general, these physiological changes are indicative of the presence of structures such polyps, nodules and laryngeal cancer. In order to quantify the hoarseness and roughness phenomena in pathological voices, non-invasive noise measures such as Normalised Noise Energy (NNE) (Kasuya et al., 1986), Glottal to Noise Excitation (GNE) (Michaelis et al, 1997), Harmonic to Noise Ratio (HNR) (Yumoto and Gould, 1982) were developed and integrated in voice quality diagnosis.

The HNR measure contains information of both harmonic and noise components and is sensitive to several kinds of periodicities such jitter and shimmer (Murphy and Akande, 2006). This measure is defined as the ratio of harmonic component energy and noise component energy (Yumoto and Gould, 1982).

1.2 Existing HNR Extraction Methods

In this paper, three algorithms based on time, spectral and cepstral techniques are reviewed in order to compare the performance of the proposed algorithm to the existing HNR algorithms. Each algorithm is a representative example of different approaches to the estimation of important quality parameter of the voice signal. Several techniques such voice models based algorithms and adaptive techniques were found in this research.

As an example of time based method, the Boersma's algorithm (Boersma, 1993) is based on the second maximum of normalized autocorrelation function detection, which is used in the following equation (1).

$$\mathrm{HNR} = 10.\log\left(\frac{\mathrm{r}(\tau)}{1-\mathrm{r}(\tau)}\right), \qquad (1)$$

where $r(\tau)$ is the second local maximum of the normalized autocorrelation and τ is the computed by a pitch detector.

Spectral methods consider that the harmonic component information is concentrated in the spectral peaks and the noise is concentrated in the valleys. The method separates the spectrum of the voice signal into two regions (harmonic regions and noise regions). Yegnanararayana (Yegnanararayana et al, 1998) algorithm is an example of spectral based method. The cepstral approach considers that the harmonic information is concentrated in the rahrmonics peaks (mainly in the first one) and the noise information is concentrated in low quefrencies. Most of cepstral algorithms perform cepstral

J. T. de Sousa R. (2009).
A NEW ACCURATE METHOD OF HARMONIC-TO-NOISE RATIO EXTRACTION.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 351-356
DOI: 10.5220/0001552903510356

segmentation with short-pass or comb lifters carefully dimensioned. In general, these methods yield an estimation of the spectral baseline of the noise, from where the HNR value can be computed. Qi (Qi and Hillman, 1997) algorithm is an example of this method.

2 PROPOSED METHOD

2.1 HNR Extraction Method

The proposed HNR method consists of harmonic and noise component estimation. Initially, the signal is segmented into frames and a sine window with the following equation (2) is applied.

$$h(n) = \sin\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)\right] \;,\; 0 \le n \le N-1 \qquad (2)$$

The harmonic component is then estimated from the Odd Discrete Fourier Transform (ODFT) (Ferreira, 1998) of the signal by extracting the frequency, magnitude and phase of each harmonic. In the ODFT domain, the parameters of the harmonic structure are easily measured and are not much affected by noise. These parameters are used to synthesize the harmonic structure in ODFT domain. The harmonic component is subtracted from the complete signal which yields the noise component estimation, as is shown in the Figure 1.

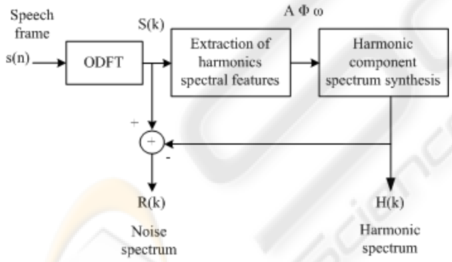
Speech frame
s(n)
S(k)
ODFT
A Φ ω
Extraction of harmonics spectral features
Harmonic component spectrum synthesis
+
-
R(k)
Noise spectrum
H(k)
Harmonic spectrum

Figure 1: Schematic diagram of harmonic and noise component estimation.

Finally, the HNR value of each frame is calculated from these two components using the equation (3).

$$HNR = 10\,.\log\frac{\sum_{k=1}^{N/2}\left|H(k)\right|^2}{\sum_{k=1}^{N/2}\left|R(k)\right|^2} \qquad (3)$$

2.2 Extraction of Harmonic Spectral Parameters

Each harmonic is modeled (Ferreira, 2001) by a sinusoid according to equation (4):

$$x(n) = A.\sin\left[\frac{2\pi}{N}(l+\Delta l)n+\varphi\right] \qquad (4)$$

where A is the sinusoid amplitude, N is the window length, l and Δl are respectively the integer part and the fractional part of the DFT bin which correspond the sine frequency. The bin fractional part represents the distance between the l bin and the true value of frequency. The algorithm computes the ODFT of the voice signal and next the local maxima are found by a peak picking algorithm. The maxima bins are the initials values which correspond to the integer part of the true frequency bin. The harmonic parameters are computed using the equations below (Ferreira, 2001), following the order of presentation.

$$\Delta l = \frac{3}{\pi}\arctan\left(\frac{\sqrt{3}}{1+2\left[\frac{\left|X_O(l-1)\right|}{\left|X_O(l+1)\right|}\right]^{1/G}}\right) \qquad (5)$$

$$\varphi = \angle X_O(l) + \pi.\left(1-\frac{1}{2\pi}\right) - \pi.\Delta l.(1+\frac{1}{N}) \qquad (6)$$

$$A = \frac{4.\left|X_O(l)\right|}{N}.\left[\frac{\sqrt{3}}{2.\cos\left[\frac{\pi}{6}(2.\Delta l-1)\right]}\right]^F \qquad (7)$$

where Xo is the ODFT, G and F are calibration parameters adjusted experimentally.

2.3 Harmonic Component Synthesis

The harmonic structure is estimated by performing a synthesis of each sinusoid parameters extracted from the ODFT representation of each frame of the signal (Ferreira, 2001) considering the windowing effect. In order to compute the sinusoid spectrum, the frequency response of a sine window without the frequency modulation implications (frequency shift) was calculated as shown in the equations which represents the spectrum phase (8) and magnitude (9) of h(n).

$$\angle H(\omega) = \frac{-\omega(1-N)}{2} \tag{8}$$

$$|H(\omega)| = A\frac{\left|\cos\frac{N\omega}{2}\right|\sin\frac{\pi}{2N}}{2} \times \left|\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}-\omega\right)} + \frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}+\omega\right)}\right| \tag{9}$$

The magnitude of the spectrum can be calculated for each k bin, carefully replacing ω by the following values.

$$\omega = \frac{2\pi}{N}\left(\Delta l - 0{,}5\right) \ , \ k = l \tag{10}$$

$$\omega = \frac{2\pi}{N}\left(\Delta l + 0{,}5\right) \ , k = l - 1 \tag{11}$$

$$\omega = \frac{2\pi}{N}\left(\Delta l - 1{,}5\right) \ , \ k = l + 1 \tag{12}$$

$$\omega = \frac{2\pi}{N}\left(0{,}5 - \Delta l - k\right) \ , \ 1 \le k \le l - 1 \tag{13}$$

$$\omega = \frac{2\pi}{N}\left(\Delta l - 0{,}5 + k\right) \ , \ l + 2 \le k \le N/2 \tag{14}$$

where A is the sinusoid spectral amplitude, N is the window length and Δ⁄ is the fractional component of the DFT bin. The phase of the spectrum is computed using the following equations:

$$\angle S(k) = \varphi \ , \ k = l \tag{15}$$

$$\angle S(k) = \varphi + \left(1 - \frac{1}{N}\right) \ , \ 1 \le k \le l - 1 \tag{16}$$

$$\angle S(k) = \varphi - \left(1 - \frac{1}{N}\right) \ , \ l + 2 \le k \le N/2, \tag{17}$$

where S(k) is the sinusoid spectrum and φ is the estimated phase. Finally, all sinusoid spectra are summed up yielding the synthetic harmonic structure.

3 ALGORITHM TESTS

The proposed method was tested in the Matlab environment with synthesized voice sounds in order to characterize their accuracy and behaviour when submitted to an acoustic diversity. The algorithm also was submitted to real voices in order to characterize the behaviour under real conditions. For both experiments, the algorithm was calibrated so that they could measure with less error as possible. An analysis window length of 2048 points has been used.

3.1 Tests with Synthesized Voices

Test with synthesized voice sounds allow establishing a target (theoretical value) which can be used to compare to the measured value from the algorithm, yielding performance parameters for the algorithm evaluation. In this regard, a synthesized voice signal is generated with known harmonic and noise component and fundamental frequency. The synthesized voice signal is received by the algorithm which measures and returns a HNR value for each frame. Finally, the theoretical and measured values are compared resulting the algorithm performance measures scores.

3.1.1 Synthesis of Voice Sounds

The synthesis module creates the test voice signals with known features according to a source-filter model configuration to simulate some acoustic events which are aimed to be measured. This source-filter model simulates a real voice in stationary conditions. Looking at Figure 3, the harmonic component of the glottal impulse $g_h(n)$ is generated initially by creating a unitary pulse train i(n) signal with a specific fundamental frequency.

Next, this signal is applied to a filter, whose impulse response is an LF model (Fant et al, 1985) glottal impulse waveform. Basically, the noise component of the glottal impulse consists of white gaussian noise $g_r(n)$ with certain energy (Levison, 2005). These two signals are summed up, yielding as a result the complete glottal impulse signal g(n). The glottal impulse components are applied to the filter of the vocal tract and lip radiation so that the speech signal s(n) and its harmonic $s_h(n)$ and noise $s_r(n)$ component can be produced. The theoretical HNR value was adjusted through the energy of the noise component and the fundamental frequency (F0) trough the pulse generator. The vocal tract filter consisted of an all-pole IIR filter which yields formants at 664, 1027 and 2617 Hz simulating the /a/ phoneme. The lip radiation was modelled by a first order difference operator $R(z)=1-0.99z^{-1}$. Voice specialists use /a/ phoneme as stimulus in order to perform their perceptive assessment due to the fact that the associated vocal tract presents a configuration with less constrictions and obstructions. Synthesized voice sounds that simulate a female voice (F0=200 Hz) and male voice (F0=100 Hz) with several degree of hoarseness (5 dB, 10 dB, 15 dB, 20 dB, 25 dB) were used. These voice sounds were synthesized at sampling rate of 16 kHz.

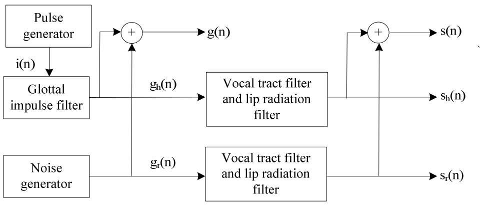

Figure 3: Production of the synthesized signals.

3.1.2 Comparison of Measured and Theoretical HNR Values

Both measured and theoretical HNR values are compared by the analysis module in order to compute the performance scores. Figure 4 shows an example of theoretical and estimated HNR values comparison, for each frame. The more these curves are close the more the algorithms are effective. Basically, the performance measures quantify the difference between these sequences of HNR values. In the experiments, the average of error (18) was considered as a suitable performance measure for the overall evaluation of the algorithms. This measure compares the mean value of theoretical HNR values and the mean value of the measured HNR values.

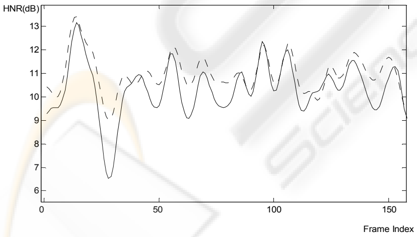

Figure 4: Theoretical (slashed line) and estimated (solid line) HNR.

$$E = \frac{1}{N_f}\sum_{i=1}^{N_f}\left[HNR_{theoretical}(i) - HNR_{measured}(i)\right] \qquad (18)$$

where N_f is the number of frames, i is the frame index, and $HNR_{theoretical}$ are $HNR_{measured}$ the theoretical and measured HNR value, expressed in dBs.

3.2 Tests with Real Voices

From a pathological voice sounds data base, six voice sounds of three male and three female were firstly selected among 17 voice patients (7 male, 10 female). These sounds were collected from speech therapy sessions regarding patients who have a certain level of hoarseness. For each gender, three levels of perceived hoarseness voice were defined (lowest, middle and highest levels of noise). This selection was made considering three voice sounds with very distinct perceptive amount of noise. Six non-voice specialists have evaluated the selected voices with the same loudness, and have agreed in regard to noise level order. With very distinct noise levels it is possible to guarantee that most people would establish the same noise level order. These voice sounds were recorded at sampling rate of 44100 Hz, with 16bit/sample accuracy and pre-processed so that the loudness could be the same.

4 RESULTS

4.1 Tests with Synthesized Voices Results

In this section, the results of HNR extraction algorithms with synthesized voices are presented on two tables showing the average error of the main existing methods (time, spectral and cepstral approaches) and the proposed method. The results of the algorithms performance were evaluated according to the error average level, the variation of average error in function of F0 and theoretical HNR value, and the general trend to underestimate or overestimate the HNR value. Mean of absolute value (MA), mean value (M) and standard deviation (SD) of the errors average were calculated to support this evaluation.

Analysing the Table 1 to 4 according to the absolute MA of the average error, it can be concluded that the proposed method presents low error level for sounds with F0=200 Hz (MA=0,21). However, the time based method presents low error level for sounds with F0=100 Hz (MA=0,11). In fact, the time based method shows major errors for high values of F0 due to detection faults of the second maximum of the normalized autocorrelation. The proposed algorithm is more effective to provide HNR estimative for higher values of F0, detecting the harmonic structure.

Table 1: Comparison of average error (dB) for the different HNR methods and F0=100 Hz.

Theoretical HNR (dB)	Time based	Spectral based	Cepstral based	Proposed method
5	0,19	-0,61	-1,15	0,37
10	0,12	-0,79	-0,63	-0,27
15	0,09	-0,67	-0,20	-0,40
20	0,07	-0,05	0,27	-0,27
25	0,06	0,50	0,82	0,27

Table 2: Mean of absolute value (MA), average (M) and standard deviation (SD) of the average of errors for F0=100Hz.

	Time based	Spectral based	Cepstral based	Proposed method
MA	0,11	0,52	0,61	0,31
M	0,11	-0,32	-0,18	-0,06
SD	0,05	0,54	0,77	0,35

Table 3: Comparison of average error (dB) for the different HNR methods and F0= 200 Hz.

Theoretical HNR (dB)	Time based	Spectral based	Cepstral based	Proposed method
5	0,80	-0,9	-2,45	0,12
10	0,65	-0,31	-1,87	-0,16
15	0,61	-0,34	-1,48	0,17
20	0,59	0,16	-1,12	-0,22
25	0,59	0,71	-0,71	0,40

Table 4: Mean of absolute value (MA), average (M) and standard deviation (SD) of the average of errors for F0=200Hz.

	Time based	Spectral based	Cepstral based	Proposed method
MA	0,65	0,49	1,53	0,21
M	0,65	0,60	-1,53	0,08
SD	0,09	0,49	0,67	0,25

Examining the average error variation in function of F0, the spectral method presents little difference between MA for 100 Hz and 200 Hz of F0. This means that the spectral has the same efficiency to detect the harmonics in female and male voices. However it yields higher error than the proposed method.

The analysis of variation of average error according to theoretical HNR value reveals better results for time-based method using both 100 Hz and 200Hz of F0. The autocorrelation is less vulnerable to the noise. The proposed algorithm yields some faults in the high frequency harmonics, which are particularly abundant in male voices. Although the proposed algorithm presents the second lower variation (SD=0,35 for F0=100Hz and SD=0,24 for F0=200Hz) under our test conditions.

The M of the proposed algorithm for both F0, lead to the conclusion that this algorithm has a low tendency to underestimate or overestimate. The other algorithms present a higher deviation in regard to the exact value, which in certain cases, are considerable, such the cepstral for 200 Hz of F0 (M=-1,53). This fact is caused by the non-fitting of the estimated noise base line. Regarding the proposed algorithm, an eventual deviation can be originated by the non-detection or a fake detection of some harmonics.

4.2 Tests with Real Voices Results

The measured HNR values are presented on two tables where they can be compared. The proximity of the HNR values returned by the tested algorithms for each voice sound was verified.

Table 5: Comparison of the HNR methods measures (dB) of female voice sounds.

Method	Real voice sounds		
	Voice sound 1	Voice Sound 2	Voice Sound 3
Time based	10,06	12,54	17,62
Spectral based	11,15	12,11	14,72
Cepstral based	10,48	13,71	18,40
Proposed method	9,53	12,16	14,39

Looking at the Tables 5 and 6, the values of measured HNR are very similar except in the cases of higher HNR in the female voices. The differences between HNR values which were produced by each algorithms for the same voice are expected considering the error that were estimated in the tests with synthesized voices.

Table 6: Comparison of the HNR methods measures (dB) of male voice sounds.

Method	Real voice sounds		
	Voice Sound 1	Voice Sound 2	Voice Sound 3
Time based	9,19	10,66	18,32
Spectral based	9,72	11,03	17,82
Cepstral based	9,52	10,76	19,75
Proposed method	9,61	10,43	18,88

The proposed algorithm yielded results that present the same order as the perceptual order (lowest, middle, highest) for female and male voices.

5 CONCLUSIONS

From the test with synthesized voices, it can be concluded that the proposed algorithm presents a fair level of accuracy, in particular for female voices which is very important feature in a diagnosis scene. The algorithm doesn't have significant tendency to underestimated or overestimate. This feature means that there is no need to calibrate or compensate the estimated HNR value. In some cases, calibration may not be effective due to the acoustic diversity of human voices. In spite of presenting the second best values regarding the variation of the errors according to F0 and theoretical HNR variation, it can be concluded that the measurement performed by the proposed algorithm is not substantially affected by the F0 and the noise level. This means that the proposed algorithm measures the female and male voices and several hoarseness levels with approximately the same accuracy.

From the tests with real voices, the proposed algorithm showed coherent HNR values, approximately similar to the HNR values of other algorithms. The harmonic plus noise model assumes good approximation of the harmonic and noise components which are present in pathological voices. Moreover, the results show that there is a correspondence between the artifacts that were associated to harmonic component and to noise component in each signal representation.

ACKNOWLEDGEMENTS

This work was developed in a doctoral program supported by the Portuguese Foundation for Science and Technology under the reference SFRH/BD/24811/2005.

REFERENCES

Levison, S. E., 2005. *Mathematical Models for Speech Technology*, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,West Sussex PO19 8SQ, England.

Yumoto, E., and Gould, W., 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness, *The Journal Acoustic Society of America*. 71(6):1544-1550.

Ferreira ,Aníbal J. S., 1998.An Odd-DFT based approach to time-scale expansion of audio signals. *IEEE Transactions on Speech and Audio Processing*, 7(4):441–453.

Ferreira ,Aníbal J. S., 2001.Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids. *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 47–50. IEEE.

Yegnanarayana, B., d'Alessandro, C., Darsinos, V. 1998. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *Speech and Audio Processing, IEEE Transactions* 6(1): 1-11.

Qi, Y., and Hillman, R. E., 1997. Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *The Journal of the Acoustical Society of America* 102(1): 537-543.

Boersma, P., 1993.Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. In Proceedings of the Institute of Phonetic Sciences* 17, 97-110.

Murphy, P. J. and. Akande, O., 2006. Noise estimation in voice signals using short-term cepstral analysis, *The Journal of the Acoustical Society of America*. 121 (3):1679-1690.

Michaelis ,D., Gramss, T., Strube H.W., 1997. Glottal-to-Noise Excitation Ratio a New Measure for Describing Pathological Voices. *Acustica-Acta Acustica* 83,700-706

Fant, G., Liljencrants, J. and Lin, Q., 1985. A four parameter model of glottal flow. *STL/QPSR 4/1985, French Swidish Symposium* 1,1-13