Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation

Peter J. Murphy and Olatunji O. Akande,

Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland.
{peter.murphy; olatunji.akande}@ul.ie

Abstract. Cepstral analysis is used to estimate the harmonics-to-noise ratio (HNR) in speech signals. The inverse Fourier transformed liftered cepstrum approximates a noise baseline from which the harmonics-to-noise ratio is estimated. The present study highlights the manner in which the cepstrum-based noise baseline estimate is obtained, essentially behaving like a moving average filter applied to the power spectrum for voiced speech. As such, the noise baseline, which is taken to approximate the noise excited vocal tract, is also shown to be influenced by the window length and the shape of the glottal source spectrum. Two existing estimation techniques are tested systematically for the first time using synthetically generated glottal flow and voiced speech signals, with *a priori* knowledge of the HNR. The source influence is removed using pre-emphasis to obtain an improved noise baseline fit. The results indicate accurate HNR estimation using the new approach.

1 Introduction

The cepstrum is used to estimate the harmonics-to-noise ratio (HNR) in speech signals [1], [2]. The basic procedure presented in [1] is as follows; the cepstrum is produced for a windowed segment of voiced speech. The rahmonics are zeroed and the resulting liftered cepstrum is inverse Fourier transformed to provide a noise spectrum. After performing a baseline correction procedure on this spectrum (the original noise estimate is high), the logarithm of the summed energy of the modified noise spectrum is subtracted from the logarithm of the summed energy of the original harmonic spectrum in order to provide the harmonics-to-noise ratio estimate (Fig.1), (the need for baseline shifting with this approach is clearly explained in the Method section).

A modification to this technique, [2], illustrates problems with the baseline fitting procedure and hence does not adjust the noise baseline but calculates the energy and noise estimates at harmonic locations only (Fig.2). In addition, rather than zeroing the rahmonics, the cepstrum is low passed filtered to provide a smoother baseline (the reason the baseline shifting is not required is due to the window length used as detailed under Method). However the noise baseline estimate is shown to deviate from the actual noise level at low frequencies (Fig. 2). Each of these approaches, [1], [2], provide useful analysis techniques and data for studies of voice quality assessment, however, to date, neither method has been tested on synthesis data

with *a priori* knowledge of the harmonics-to-noise ratio. The present study uses known amounts of random noise added to the glottal source to systematically test these techniques.

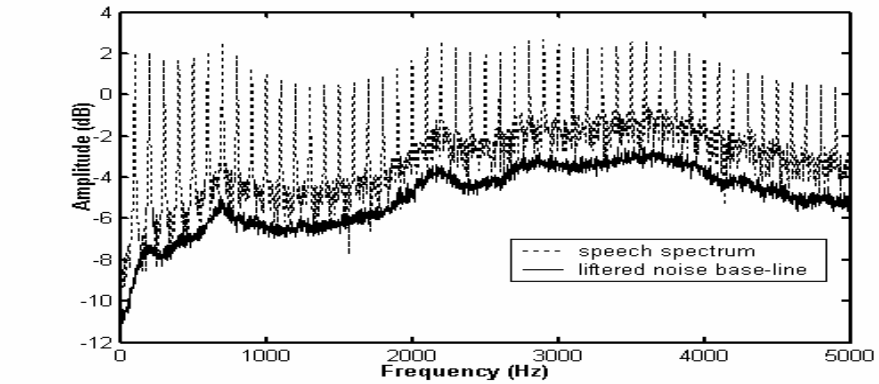

Fig.1. HNR estimation using de Krom [1] cepstral baseline technique using a window length of 1024 points. The noise level is underestimated due to the baseline shifting process, which detects minima at between-harmonic locations.

When such source signals are convolved with the vocal tract impulse response and radiation load the HNR is altered. However, an *a priori* HNR can still be estimated in the time domain by using a synthesized speech signal convolved with a noisy glottal source and one convolved with a noise-free source. The influence of the source spectrum on the noise baseline estimate is highlighted and is corrected for using a pre-emphasis technique.

An alternative cepstral-based approach for extracting a HNR from speech signals is estimated in [3]. However, this involves directly estimating the magnitude of the cepstral rahmonic peaks, leading to a geometric-mean harmonics-to-noise ratio (i.e. an average of the dB harmonics-to-noise ratios at a specific frequency locations), which is quite distinct from traditional harmonics-to-noise ratio estimators which reflect the average signal energy divided by the average noise energy, expressed in dB. This is shown in eqtn. 1 for an N-point DFT, giving harmonic amplitudes, $|S|_i$, and noise estimates, $|N|_i$.

$$HNR = 10\log_{10}\left\{\frac{\sum_{i}^{N/2}|Si|^2}{\sum_{i=1}^{N/2}|Ni|^2}\right\} \qquad (1)$$

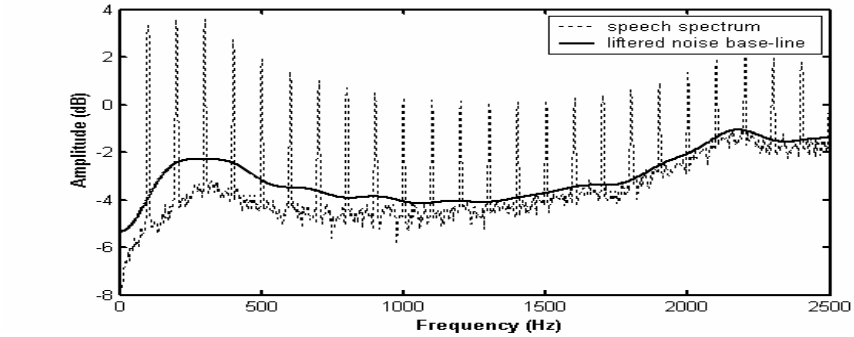

Fig.2. HNR estimation using Qi and Hillman [2] cepstral baseline technique (window length 3200 points).

2 Method

A periodic glottal source with additive noise, $g_{en}(t)$ can be written as

$$g_{en}(t) = e(t)*g(t)+n(t) \quad (2)$$

where $e(t)$ is a periodic impulse train, $g(t)$ is a single glottal pulse and $n(t)$ represents aspiration noise.

Applying a Hanning window, (w)

$$g_{en}^{w}(t)=(e(t)*g(t)+n(t))\times w(t) \quad (3)$$

The window function can be moved inside the convolution [4] to give

$$g_{en}^{w}(t)=e_w(t)*g(t)+n_w(t) \quad (4)$$

Taking the Fourier transform gives

$$G_{en}^{w}(f)=E_w(f)\times G(f)+N_w(f) \quad (5)$$

Taking the logarithm of the magnitude squared values and approximating the signal energy at harmonic locations, $log|G_{en}^{w}|_h^2$ and at between-harmonic locations, $log|G_{en}^{w}|_{bh}^2$, gives

$$log|G_{en}^{w}|_h^2=log|E_w(f)\times G(f)|^2 \quad (6)$$

$$log|G_{en}^{w}|_{bh}^2=log|N_w(f)|^2 \quad (7)$$

Although the noise spectrum is broadband, its estimation in the presence of a harmonic signal can be concentrated at between-harmonic locations i.e. in the spectrum of the glottal source signal energy dominates at harmonic locations and noise energy dominates at between-harmonic locations. This approximation becomes more exact if the spectra are averaged in which case the harmonics approach the true harmonic values and the between-harmonics approach the true noise variance [5].

The cepstral technique is described with reference to Fig.3. An estimate of the HNR is obtained by summing the energy at harmonic locations and dividing by the sum of the noise energy. Extracting the noise energy baseline via the cepstral technique can be viewed as an attempt to estimate the noise level for all frequencies, including harmonic locations. It is noted that the cepstrum can be applied to periodic glottal source signals, separating the slowly varying glottal spectral tilt from the fast variation due to harmonic structure. The baseline is estimated via either of the methods [1], [2] outlined in the Introduction; in either approach the rahmonics are removed in the cepstrum and the resulting cepstrum is inverse Fourier transformed to provide the noise baseline. However, the noise baseline estimate essentially behaves like a moving average (MA) filter and hence is dependent on both glottal and noise contributions (eqtns. (6) and (7), Fig. 3).

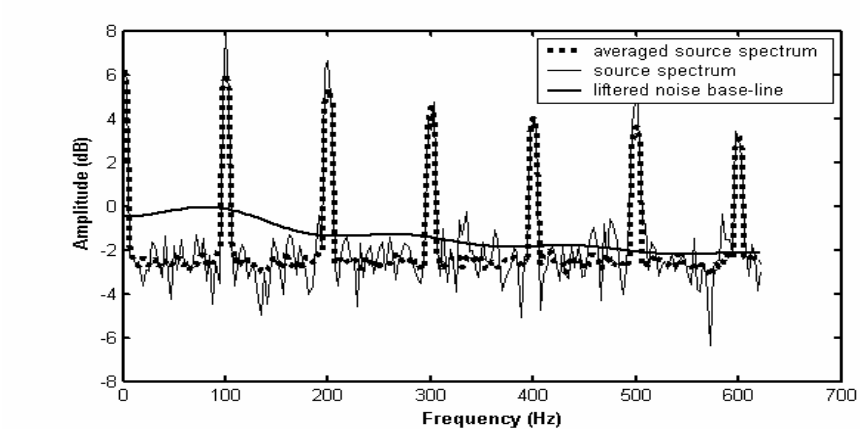

Fig.3. Spectrum of glottal source with 1% additive noise. The solid line represents a single spectral estimate. The dashed line represents an average of n spectral estimates. The liftered noise baseline is also shown.

The above analysis applied to a windowed segment of voiced speech,

$$s_{en}^{w}(t) = [(e_w(t)*g(t))+n_w(t)]*v(t)*r(t) \quad (8)$$

where $v(t)$ and $r(t)$ represent, respectively, the impulse response of the vocal tract and the radiation load, gives

$$log|S_{en}^{w}(f)|_h^2 = log|E_w(f) \times G(f)|^2 + log|V_R(f)|^2 \quad (9)$$

$$log|S_{en}^{w}(f)|_{bh}^2 = log|N_w(f)|^2 + log|V_R(f)|^2 \quad (10)$$

where $V_R(f)$ is the Fourier transform of $v(t)$ and $r(t)$ combined.

Again for the speech signal, HNR is estimated by summing the energy at harmonic locations and dividing by the summed noise energy estimated via the cepstral baseline technique. Now, the noise baseline (which is equivalent to a traditional vocal tract transfer function estimate via the cepstrum) is influenced by the glottal source excited vocal tract and by the noise excited vocal tract (Fig.2). It is the interpretation of the noise baseline as a MA filter that explains the need for baseline fitting in [1]; the liftered spectral baseline does not rest on the actual noise level but interpolates the harmonic and between harmonic estimates and hence resides somewhere between the noise and harmonic levels. As the window length increases (as per [2], for

example) the contribution of harmonic frequencies to the MA cepstral baseline estimate decreases. However, the glottal source still provides a bias in the estimate.
To remove the influence of the source, pre-emphasis is applied to the glottal source, $g_e^w(t)$ for the glottal signals and to $s_e^w(t)$ for the voiced speech signals (i.e. noiseless signals). $|G_{en}^w(f)|_h$ and $|S_{en}^w(f)|_h$ are estimated using periodogram averaging to provide estimates for $g_e^w(t)$ and $s_e^w(t)$ respectively (Fig.5).
A pre-emphasis filter,

$$h(z)=1-0.97z^{-1} \qquad (11)$$

is applied to these estimates in the frequency domain by multiplying each harmonic value by the appropriate pre-emphasis factor.

3 Analysis

3.1. Synthesis parameters

In order to evaluate the performance of the existing techniques along with the newly proposed method, synthesized glottal source and vowel /AH/ waveforms are generated at five fundamental frequencies (f0s) beginning at 80 Hz increasing in four steps of 60 Hz up to 320 Hz, covering modal register. The model described in [6] is adopted to synthesize the glottal flow waveform while the vocal tract impulse response is modeled with a set of poles. Lip radiation is modeled by a first order difference operator $R(z)=1-z^{-1}$. A sampling rate of 10 kHz is used for synthesis. Noise is introduced by adding pseudo-random noise to the glottal pulse via a random noise generator arranged to give additive noise of a user-specified variance (seven levels from std. dev. 0.125%, doubling in steps up to 8 %). The corresponding HNRs for the glottal flow waveform are 58 dB to 22 dB, decreasing in steps of 6 dB. However, the HNR for the corresponding speech signals vary with f0 due to the differential excitation of glottal harmonics, c.f. [7]. However, *a priori* knowledge of the HNR can be obtained by comparing clean synthesized speech to the glottal plus noise synthesis.

3.2 Analysis procedure

The procedures in [1] and [2] are implemented as outlined in the Introduction. In [1] window lengths of 1024, 2048 and 4096 are chosen, while in [2] a window length of 3200 is chosen, as per the original algorithm descriptions. In the proposed approach, the spectrum (2048-point FFT) of the test signal is computed using an analysis window (Hanning) of 2048 points overlapped by 1024 points. The analysis is applied to a 1.6 second segment of synthesized speech, providing 14 spectral estimates. The resulting power spectra are averaged (in order to reduce the noise variance at harmonic locations) to give a single 2048-point FFT. Harmonic peaks and bandwidths in the averaged spectrum are identified, and are modified using a pre-emphasis filter. The between-harmonics, which are not pre-emphasized, approach the noise variance in the averaged spectrum. The cepstrum is applied to the log spectrum of the pre-emphasized harmonics with the non pre-emphasized between-harmonics. The noise floor is extracted using a rectangular low-pass liftering window to select the first 40 cepstral coefficients. In order to calculate the noise energy, the extracted baseline is transformed back to a linear power spectrum and summed at the harmonic points. A sum, representing the signal energy, is taken of the

harmonic peaks in the power spectrum of the signal (without pre-emphasis). The HNR is calculated as per eqtn.1.

4 Results

In order to illustrate the improvement offered by the new method over the existing cepstrum-based techniques, the liftered noise baseline is plotted together with the spectrum of the glottal source signal (with 1% additive noise) with (Fig.5) and without (Fig.4) pre-emphasis. It can be seen from Fig.4 that without pre-emphasis the estimated noise baseline deviates from the true noise floor as a result of the source influence on the liftered noise baseline. The result of removing the source influence before extracting the noise base line from the cepstrum is depicted in Fig.5 where the estimated noise baseline provides a much improved fit to the actual noise floor. The HNR plotted against f0 is shown for (a) deKrom [1] (Fig.6) (b) Qi and Hillman [2] (Fig.7) and (c) the new approach (Fig.8). The results of the HNR measurement for the synthesized vowel /AH/ with the new method is shown in Fig.9. In order to evaluate the performance of a method, the estimated HNR is compared to the original HNR (dotted curve) in the figures.

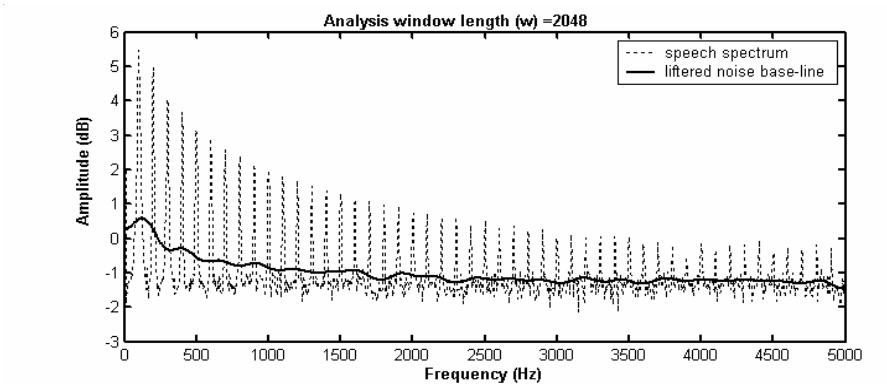

Fig.4. Spectrum of the glottal source and liftered noise baseline, where pre-emphasis is not applied in the baseline estimation procedure (not shown). The spectrum is calculated with an analysis window length of 2048 points.

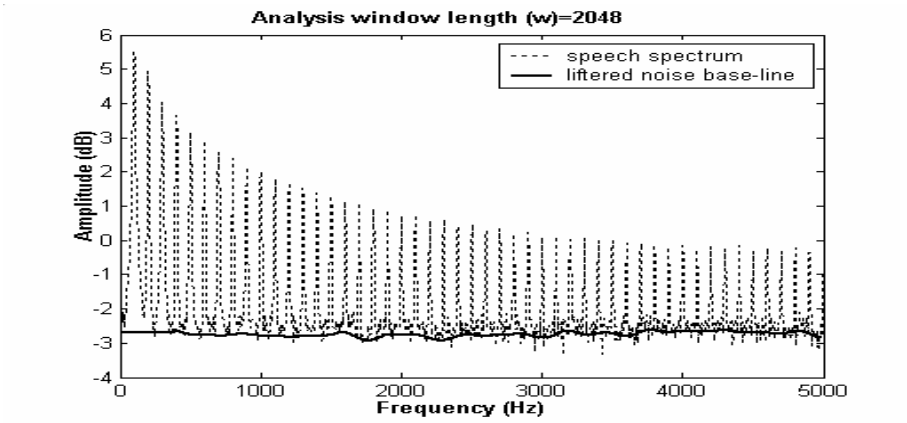

Fig.5. Spectrum of the glottal source and liftered noise baseline, where pre-emphasis is applied in the baseline estimation procedure (not shown). The spectrum is calculated with an analysis window length of 2048 points.

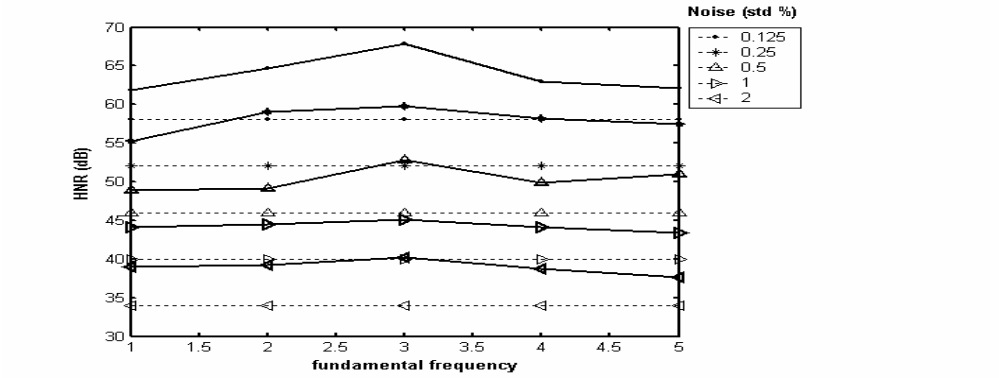

Fig.6. Estimated HNR (solid line, de Krom [1]) versus f0 for synthesized glottal source waveforms (dotted line – actual HNR).

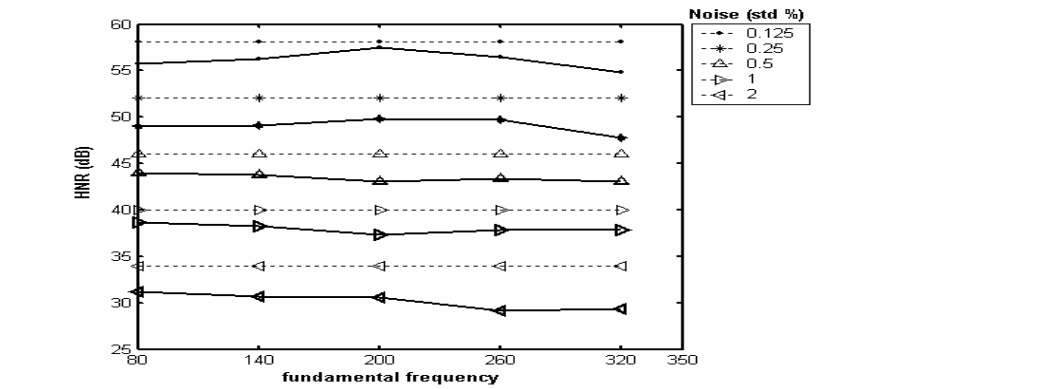

Fig.7. Estimated HNR (solid line, Qi and Hillman [2]) versus f0 for synthesized glottal source waveforms (dotted line – actual HNR).

5 Discussion

Increasing the window length moves the baseline closer to the actual noise level. The de Krom technique [1] tends to underestimate the baseline due to the fact that minima are estimated in the baseline fitting procedure. The Qi and Hillman approach [2] cannot match the noise level at low frequencies due to the influence of the source. Similar over- and under- estimates of the HNR for synthesized speech are also found (not shown). The effect of the source on the liftered noise floor is reduced by pre-emphasizing the harmonics of the test signal. HNR estimated with the new method tracks the corresponding input HNRs with marginal error as illustrated in Fig.8 and Fig.9.

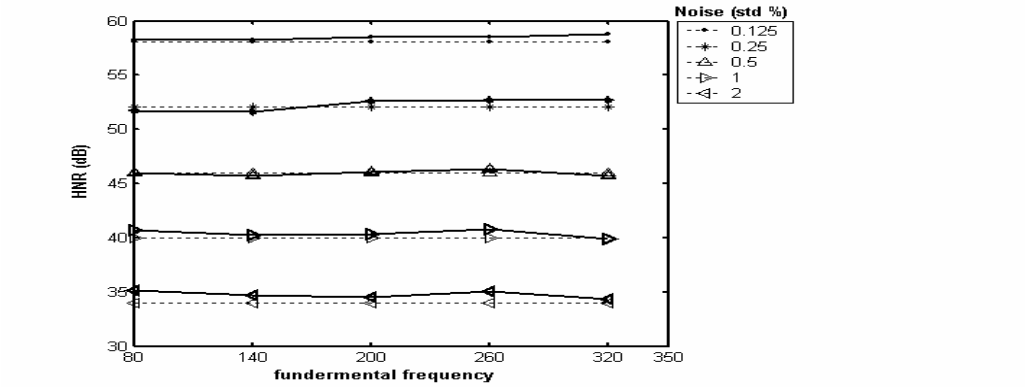

Fig.8. Estimated HNR (solid line, with the new method) versus f0 for synthesized glottal source waveform (dotted line – actual HNR).

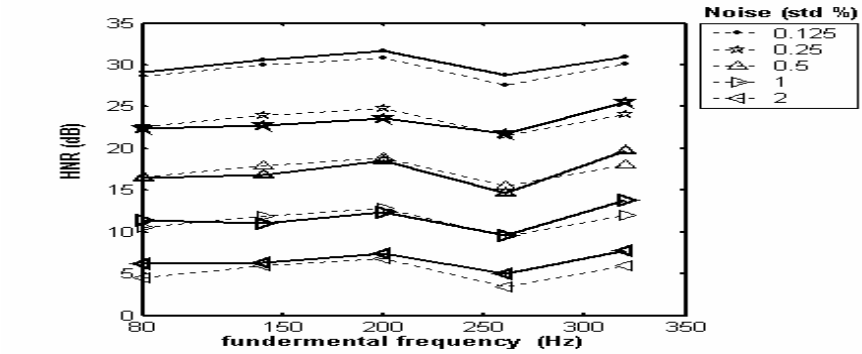

Fig.9. Estimated HNR (with the new method) versus f0 for synthesized vowel /AH/ (dotted line – actual HNR).

6 Conclusion

Two existing cepstral–based HNR estimation techniques are evaluated using synthesized glottal waveforms and speech signals with *a priori* knowledge of the HNR for these signals. The methods provide reasonably consistent estimates of the HNR, however, HNR tends to be overestimated in [1] due to the baseline fitting procedure underestimating the noise levels and [2] tends to over-estimate the HNR due to the underestimate of noise levels due to the influence of the glottal source on the noise baseline. A combination of appropriate window length and pre-emphasis is shown to remove the bias due to the glottal source, providing an accurate noise baseline from which to estimate the HNR. Further work will apply the technique to human voice signals and will investigate adapting the technique for use with shorter analysis windows with a view to analyzing continuous speech.

7 Acknowledgements

This work is supported by Enterprise Ireland Research Innovation Fund, RIF/037.

References

[1] de Krom, G. “A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals”. *J. Speech Hear. Res.* 36(2):254-266, 1993.

[2] Qi, Y. and Hillman, R.E. “Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals”. *J. Acoust. Soc. Amer.* 102(1):537-543, 1997.

[3] Murphy, P.J. “A cepstrum-based harmonics-to-noise ratio in voice signals”, Proceedings International Conference on Spoken Language Processing, Beijing, China, 672-675, 2000.

[4] Schafer, R.W. and Rabiner, L.R. “System for automatic formant analysis of voiced speech,” *J. Acoust. Soc. Amer.* 47:634-648, 1970.

[5] Murphy, P.J. “Averaged modified periodogram analysis of aperiodic voice signals”, Proceedings Irish Signals and Systems Conference, Dublin, 266-271, June, 2000.

[6] Fant, G., Liljencrants, J. and Lin, Q. G. “A four parameter model of glottal flow”, STL-QPSR 4, 1-12, 1985.

[7] Murphy, P.J. “Perturbation-free measurement of the harmonics-to-noise ratio in speech signals using pitch-synchronous harmonic analysis”, *J. Acoust. Soc. Amer.* 105(5):2866:2881, 1999.