Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals

Yingyong Qi
Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

Robert E. Hillman
Voice and Speech Laboratory, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts 02215 and Department of Otology and Laryngology, Harvard Medical School, Cambridge, Massachusetts 02139 and Communication Science and Disorders, MGH-Institute of Health Professions, Boston, Massachusetts 02215

(Received 4 November 1996; accepted for publication 18 March 1997)

The quantity, harmonic-to-noise ratio (HNR), has been used to estimate the level of noise in human voice signals. HNR estimation can be accomplished in two ways: (1) on a time-domain basis, in which HNR is computed directly from the acoustic waveform; and (2) on a frequency-domain basis, in which HNR is computed from a transformed representation of the waveform. An algorithm for computing HNR in the frequency domain was modified and tested in the work described here. The modifications were designed to reduce the influence of spectral leakage in the computation of harmonic energy, and to remove the necessity of spectral baseline shifting prescribed in one existing algorithm [G. de Krom, J. Speech Hear. Res. **36**, 254–266 (1993)]. Frequency-domain estimations of HNR based on this existing algorithm and our modified algorithm were compared to time-domain estimations on synthetic signals and human pathological voice samples. Results indicated a highly significant, linear correlation between frequency- and time-domain estimations of HNR for our modified approach. © *1997 Acoustical Society of America.* [S0001-4966(97)03307-9]

PACS numbers: 43.70.Aj, 43.70.Gr [AL]

INTRODUCTION

A valid and reliable method for estimating levels of noise in the human voice would be expected to provide useful information for the evaluation and management of voice disorders. Current techniques for measuring noise in the human voice treat the acoustic signal as a sum of two parts: a harmonic and a noise component. Based on this assumption, estimates of the harmonic-to-noise ratio (HNR) have been calculated. HNR estimation can be accomplished in two ways: (1) on a time-domain basis, in which HNR is computed directly from the acoustic waveform; and (2) on a frequency-domain basis, in which HNR is computed from a transformed representation of the waveform.

The time and frequency representations of a signal are usually related through Fourier transformations. Thus measurements in the time domain should have their equivalence in the frequency domain. The equivalency between time- and frequency-domain measurements, however, is not always apparent unless the signal is strictly periodic. For example, there are two major factors that complicate the estimation of HNR in human voice signals. First, the spectrum of a recorded acoustic signal has to be computed using a limited (windowed) segment of the signal. The application of a window broadens the harmonics of a signal, the extent to which depends on the shape and length of the window. Second, human voice signals are not truly periodic. There are period-to-period variations in fundamental frequency ($F0$ perturbation), which also tend to broaden the harmonics. This effect is more apparent in some disordered voices in which there is an increase in $F0$ perturbation. Combined together, these sources of spectral broadening or spectral leakage make it difficult to obtain accurate estimates of HNR for human voice signals in the frequency domain (Cox *et al.*, 1989). For similar reasons, it is equally difficult to measure HNR in human voice signals in the time domain. Thus it is necessary to employ approximation techniques to obtain frequency- and time-domain-based estimations of HNR in human voices. One approach to evaluating these approximation techniques would be to determine the extent to which the resulting HNR estimations preserve time and frequency equivalency.

A representative, time-domain approach for calculating HNR was proposed by Yumoto *et al.* (1982). To calculate HNR, an ''average'' wave for a single period is determined as the mean of a succession of periods. The energy of this average defines the harmonic component. Each individual period is then compared to this average. The variance of the period ensemble (all periods used to compute the mean) is used to define the noise component (see Yumoto *et al.* for details). Zero padding is used to time-normalize periods prior to computation of the mean and variance. This simple time-normalization procedure is difficult to apply to disordered voices that exhibit relatively large $F0$ perturbation, in part, because the computed variance will be significantly inflated by such perturbation. An example period ensemble and the mean and standard deviation of this ensemble are shown in Fig. 1. Note that there is a noticeable increase in the variance around sample 60 where zero padding has been used.

Qi (1992) proposed a modification to the time-domain method of calculating HNR originally developed by Yumoto *et al.* In this modification, periods are time-normalized using dynamic time warping, a procedure that optimally aligns waveforms in time prior to computation. Qi and his col-

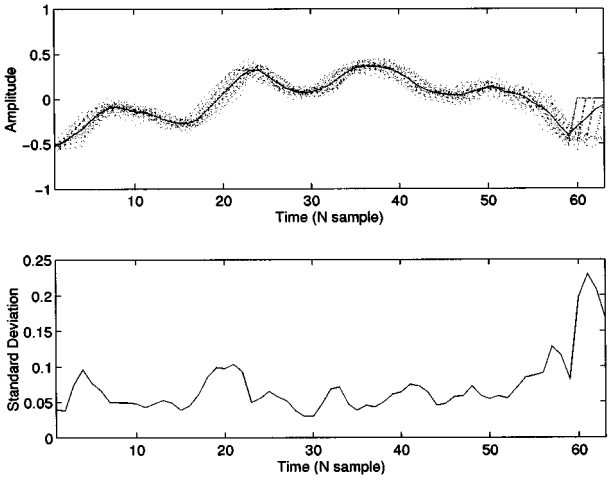

FIG. 1. An example period ensemble and its mean and standard deviation. Zero padding was used to time-normalize the period ensemble.

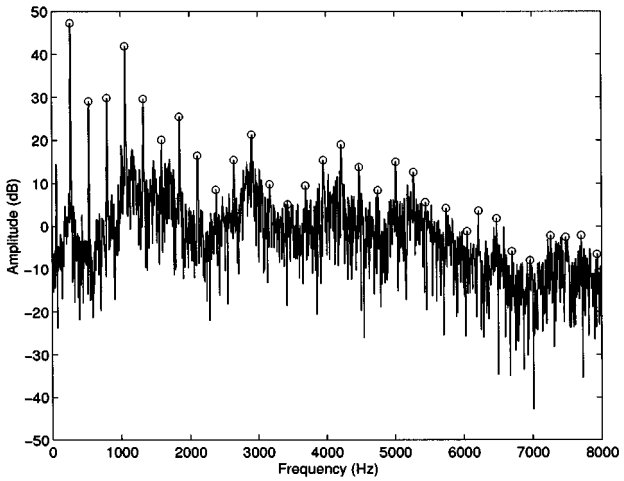

FIG. 3. The DFT spectrum of a signal with harmonic peaks marked.

leagues later demonstrated (Qi *et al.*, 1995b) that optimal time-normalization could also be accomplished using zero-phase transformations, and that an appropriate time-normalization minimizes the influence of $F0$ perturbation on the computation of HNR over a wide range of disturbances in voice signals. An example period ensemble after zero-phase transformation and its mean and standard deviation are shown in Fig. 2. Because the variances due to phase differences have been removed, the overall variance of the ensemble is reduced.

A number of techniques have been proposed for calculating HNR in the frequency domain (Kasuya *et al.*, 1986; Muta *et al.*, 1988; Cox *et al.*, 1989; Emanuel, 1991; de Krom, 1993). One recent, frequency-domain (or transformed-domain) approach to calculating HNR was proposed by de Krom (1993). In this method, the ''noise floor'' of a spectrum is estimated by cepstral liftering and spectral baseline shifting. Cepstral liftering is used to remove harmonic components in the spectrum, and baseline shifting is used to position the ''noise floor'' on a ''reasonable'' reference level. HNR is defined as the energy of the original spectrum relative to the energy of the cepstral-liftered and baseline-shifted spectrum. An example spectrum of a voice signal is shown in Fig. 3. The peaks of harmonics are marked. The cepstrum-liftered and baseline-shifted noise floor of the spectrum is shown in Fig. 4, together with the spectrum and its harmonic peaks.

HNR estimation in the frequency domain has a number of potential advantages over time-domain-based approaches. First, it is less difficult to estimate the mean $F0$ of a windowed signal segment required for the computation of HNR in the frequency domain than it is to identify individual period boundaries needed to compute HNR in the time domain. Second, HNR of different frequency ranges can be computed easily in the frequency domain. Finally, HNR in the frequency domain may be less sensitive to low-frequency am-

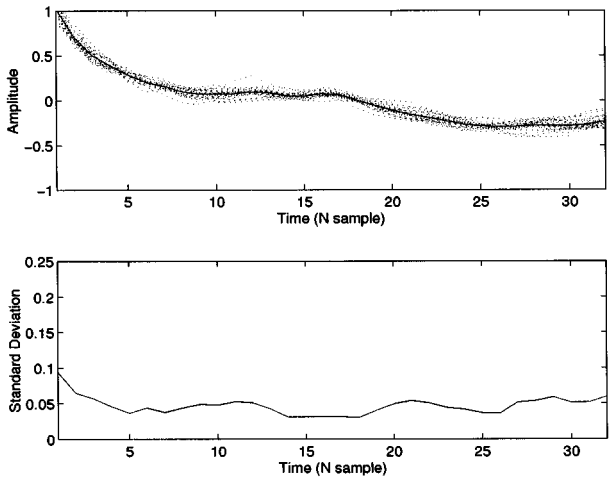

FIG. 2. The period ensemble of Fig. 1 after zero-phase transformation and its mean and standard deviation. Note that only half of the data points are shown. The other half is a mirror image of the first half.

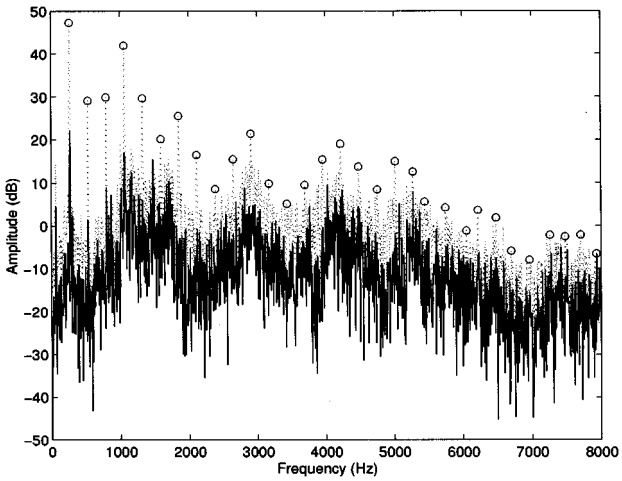

FIG. 4. The noise spectrum (solid line) obtained from comb cepstral liftering and baseline shifting. The original spectrum (dotted line) and its harmonic peaks are also shown.

plitude modulations. This advantage potentially could be useful in the analysis of voice signals from continuous discourse (Hillenbrand *et al.*, 1994).

Certain improvements/modifications, however, need to be made to the algorithm proposed by de Krom for computing HNR in the frequency domain. In the de Krom approach, it was implicitly assumed that all spectral components with amplitudes greater than a ''noise floor'' contribute to the energy of the harmonics. This assumption is not necessarily correct, in part, because of spectral leakages. In addition, the magnitude of baseline shift depends heavily on the estimation of spectral minima, which is strongly influenced by $F0$ and $F0$ perturbations. Because calculation of the magnitude of baseline shift is a crucial step in the computation of HNR, errors in baseline shift would introduce biases to the estimation of HNR. de Krom has, for example, pointed out that, ''Multiplication with a Hanning window increases the bandwidth of the harmonics. A consequence of this harmonic broadening is that we will obtain lower HNR values than could be expected... . In the spectra of signals with low fundamental frequencies, the harmonics are spaced close together, and the sidelobes of the smoothing window very much determine the minima to be reached in between harmonics.'' (page 261, de Krom, 1993)

The main purpose of the work to be described was to develop and test a modified algorithm for estimating HNR in the frequency domain. Specifically, procedures for estimating the levels of harmonics and ''noise floor'' were modified to reduce the influence of spectral leakage and to remove the necessity of calculating the magnitude of baseline shifting. Comparisons were made between this modified HNR estimation in the frequency domain and HNR estimation in the time domain.

I. METHOD

A. Estimation of HNR in the frequency domain

Estimation of HNR in the frequency domain requires computations of the magnitudes of harmonic and noise components. Numerically, both harmonic and noise components are estimated on the basis of discrete Fourier transform (DFT) of the signals. Further transformation of the signals, e.g., the cepstral transformation, may facilitate the separation of harmonic components from noise.

As explained previously, a major factor influencing estimation of the harmonic components in voice signals is the windowing effect. The DFT spectrum of a windowed signal segment is the convolution of the spectrum of the signal and the spectrum of the window. The resulting spectrum at any given frequency is a weighted sum of all spectral components of the signal, where the weighting is determined by the spectrum of the window function (Oppenhelm and Schafer, 1989). The spectrum of the window function is characterized by a main lobe and several side lobes. It is possible to balance the influence of the main lobe and the side lobes of the window function by manipulating the window shape. It is also possible to reduce the width of the main lobe by using a relatively long window. It is, however, practically impossible to eliminate the window effect (Oppenheim and Schafer, 1989).

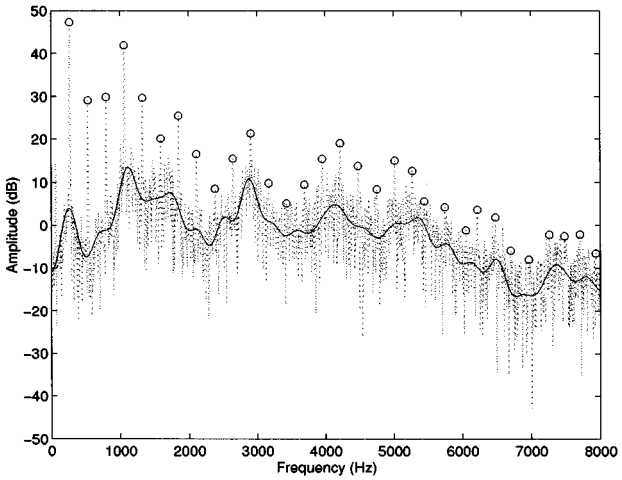

FIG. 5. The noise reference level (solid line) obtained from low-quefrency cepstral liftering. The original spectrum (dotted line) and its harmonic peaks are also shown.

Spectral leakage is inevitable and somewhat unpredictable. One way to estimate the energy of the harmonics is to sum all spectral components that exceed specified threshold levels as described by de Krom (1993). An alternative is to use the summation of only spectral peaks as an estimation of the harmonic energy. The latter approach has the following advantages:

(1) The window effect (spectral leakage) tends to introduce less error to the spectral peaks than to neighboring spectral components.
(2) The percentage of random magnitude fluctuations is usually smaller at the spectral peaks than at neighboring spectral components.
(3) The spectral peaks are relatively simple to measure.

Because of these advantages, the sum of spectral peaks was used as a measure of harmonics energy in our modified approach for estimating HNR in the frequency domain.

In de Krom's algorithm, the estimation of noise was obtained from the inverse transformation of the comb liftered cepstrum, followed by a heuristic baseline shift. The baseline shifting is complicated and is needed, in part, because all spectral energy above the noise floor is used to estimate HNR. In our modified algorithm, HNR is estimated as the energy of the spectral peaks that exceed a reference noise level. Thus rather than the whole noise spectrum, only a discrete set of noise reference levels at these peak frequencies need to be estimated. The average of this discrete set of noise reference levels can be estimated from the low-quefrency part of the cepstrum, thus the need for spectral baseline shifting is eliminated. An example of the low-quefrency, cepstrum-based estimation of the noise floor is shown in Fig. 5 together with the spectrum and its harmonic peaks.

B. Relationship between temporal and spectral estimations

Relationships between time-domain HNR and the level of added noise in synthetic signals has been extensively assessed in previous work (Qi, 1992; Qi *et al.*, 1995b). In the present work, the relationships between time- and frequency-domain estimations of HNR have been established using a limited set of synthetic signals and a relatively large set of human voice signals. We assumed that the establishment of highly significant, linear relationships between temporal and spectral estimations of HNR would help to confirm the equivalency (but not necessarily the validity) between time- and frequency-domain HNR measurements.

Obviously, a crucial factor in both time- and frequency-domain estimations of HNR is the size of the time window (signal length). The selection of window length is often made on the basis of a compromise between frequency versus time resolution. A short time window would introduce high spectral leakage, whereas a long time window would make it difficult to compute HNR for nonstationary voice signals (Qi and Shipp, 1992). Because our objective was to evaluate our modified approach for frequency-domain estimation of HNR relative to time-domain-based estimation, a fixed window size of 200 ms (3200 samples) was used.

1. Synthetic signals

The vowel /ɑ/ was synthesized using a formant synthesizer (Fant, 1960). The synthesizer was a five-pole, autoregressive digital filter whose coefficients were determined by five given pairs of formant frequencies and bandwidth. The unperturbed excitation source to the synthesizer was generated from a parametric model of glottal output proposed by Fant *et al.* (Fant *et al.*, 1985; Klatt and Klatt, 1990; Childers and Lee, 1991). This model (hereafter referred to as the LF model) specifies differentiated glottal flow using four independent parameters. In the present work, all temporal parameters of the LF model, t_p, t_e, and t_c, were defined as a constant proportion of the fundamental period. The amplitude of the negative flow derivative, Ee, was set to a level so that the intensity of the synthetic signal is around 65 dB (Qi and Bi, 1994). Specifically, t_p, was at 45% of the fundamental period. t_e was at 60% of the fundamental period. t_c was at 100% of the fundamental period. Ee was set to 50. The synthesis was made by convolving this excitation input with the impulse response of the autoregressive digital filter (Qi *et al.*, 1995a). The sampling frequency for the synthesizer was 16 kHz.

Noise was introduced by adding a Gaussian random noise to the excitation source. The level of noise was controlled by the standard deviation of the Gaussian distribution. Perturbations in fundamental frequency were introduced by adding a uniformly distributed random number to the interval of each period of the excitation input. The level of $F0$ perturbation, i.e., the variance of the random number generator, was set at 1% of the fundamental period. Fundamental frequency levels of 120 Hz and 240 Hz were used, respectively. The synthetic signal was informally presented to three experienced listeners and was judged to be an acceptable /ɑ/. It was perceived as a noisy /ɑ/ when relatively large noise perturbations were introduced.

2. Human voice signals

Human voice samples of 48 subjects (24 men and 24 women) diagnosed with a wide range of pathological conditions were obtained from the Voice and Speech Laboratory of Massachusetts Eye and Ear Infirmary (Diaz *et al.*, 1993; Kay-Elemetrics, 1994). Each subject was asked to sustain (at least 3 s long) the vowel /ɑ/ three times at comfortable fundamental frequency and intensity levels. Each vowel sample was recorded using a condenser microphone (Sennheiser) and a digital tape recorder (Tascom, DA-30) in a sound treated booth.

All recordings were low-pass filtered ($f_c = 7.5$ kHz) and digitized into a computer at a sampling rate of 16 kHz and 16-bit A/D resolution. One of the three /ɑ/ vowels for each subject was selected as the representative production for that subject. A 200-ms, stable segment from each voice sample was visually identified and selected using a waveform editor. These selected voice samples were subjected to the acoustic analyses of interest.

3. HNR computation in the time domain

The modified approach described by Qi *et al.* (1995b) was used to obtain estimates of HNR in the time domain. Zero-phase transformation was used to achieve time-normalization of period ensembles. Zero-phase transformation is preferable to dynamic programming, in part, because it is simpler to implement, faster to compute, and, thus is potentially easier to apply in voice research and clinical evaluation of voice disorders. In addition, zero-phase transformation has an accuracy equivalent to dynamic programming based time-normalization, when errors in fundamental period determination are small (Qi *et al.*, 1995b).

Period boundaries of synthetic signals were obtained from the synthesis program. Period boundaries of human voice samples were identified using a time-delayed, peak-picking algorithm. Time delay was introduced to ensure correct determination of global, rather than local, maximum (or minimum) within a given range. Visual inspection indicated that boundaries between fundamental period were determined reasonably accurately for all human voice samples.

The computation of the zero-phase transformation is fully described in a previous report (Qi *et al.*, 1995b). Briefly, zero-phase transformation consists of the following:

Identifying period boundaries of a voice segment.
Computing the period-synchronized, zero-padded fast Fourier transformation (FFT) for each period.
Computing the magnitude spectrum and setting the phase of all frequency components to zero.
Inversely transforming the zero-phased magnitude spectrum.

Finally, HNR was computed as the ratio between the energy of the average waveform and the energy of the variance of the time-normalized, period ensemble.

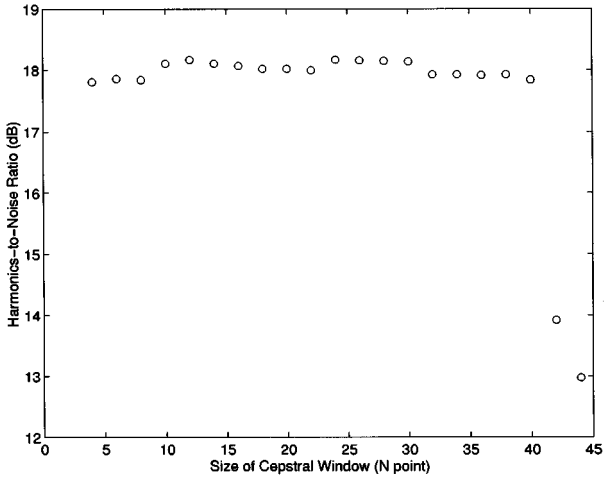

FIG. 6. The HNR as a function of cepstral window length.

4. HNR computation in the frequency domain

Computation of HNR in the frequency domain consists of the following:

Computing the DFT (3200 point) spectrum of a windowed signal segment.

Identifying the harmonic peaks using a frequency-delayed, peak-picking algorithm. The frequency delay was introduced to ensure that each peak located was global within a given frequency range of the spectrum.

Computing the cepstrum of the same signal segment.

Applying a cepstral window to lifter out the high-quefrency part of the cepstrum.

Inversely transforming the liftered cepstrum to obtain an estimation of the smoothed reference noise level.

Finally, HNR was computed as the mean difference between the harmonic peaks and the reference levels of noise at these peak frequencies.

An important parameter in this process is the size of the cepstral window, L. It determines the amount of high-quefrency components that are included in the estimation of reference noise levels. It should be large enough to adequately model the variations of the spectral envelope, but small enough to exclude cepstral peaks. Fortunately, it is relatively stable over a range of values. A constant $L=24$ (1.5 ms) was used in the present analysis. Examples of HNR estimates as a function of L are shown in Fig. 6 for a female voice sample ($F0=275$ Hz). This figure illustrates the relative stability of HNR estimates across a wide range of cepstral window sizes.

II. RESULTS

HNR estimations of the synthetic signals as a function of noise level are illustrated in Fig. 7(a) and (b). As shown, the time-domain estimates of HNR of synthetic signals appear to be more accurate than the frequency-domain HNR estimations of these synthetic signals. Frequency-domain HNR estimations are further away from the diagonal at low noise

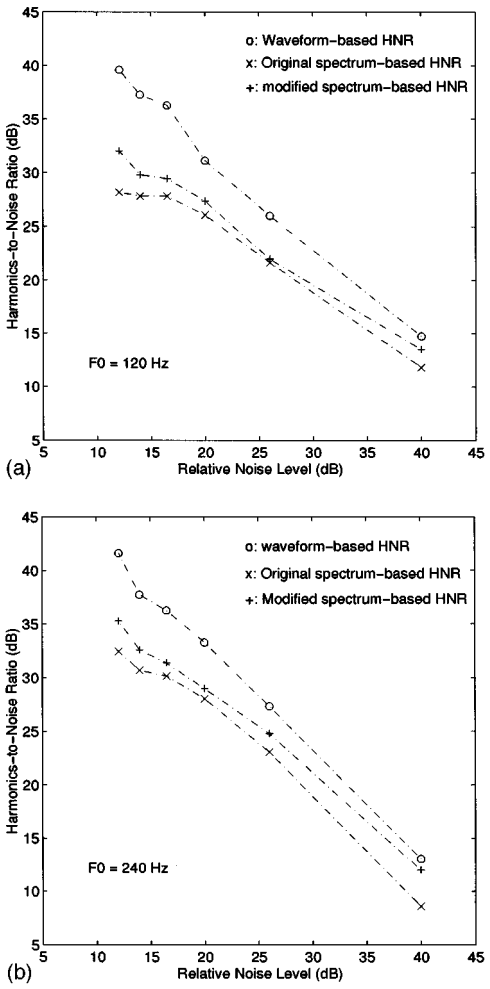

FIG. 7. The HNR as a function of noise level for synthetic signals.

levels than they are at high noise levels. Modified frequency-domain HNR estimation may be slightly better than the de Krom-based frequency-domain HNR estimation of these synthetic signals.

The HNRs of human voice signals that were computed using our modified algorithm in the frequency domain are shown as a function of HNRs computed using the zero-phase transformation based algorithm in the time domain (Qi *et al.*, 1995b) in Fig. 8. The dashed line in Fig. 8 is the least-square regression line. For comparison, the HNRs that were computed using the frequency-domain algorithm proposed by de Krom (a three-point bandwidth was used for liftering cepstral peaks) are plotted as a function of zero-phase transformation based time-domain estimates for the same set of human voice signals in Fig. 9.

Spearman correlational analyses of measurements of the human voice signals indicate that our modified frequency-domain-based estimates were significantly correlated with the time-domain-based estimates of HNR ($r=0.88$, $p<0.0001$). There was also a significant correlation between the frequency-domain estimate of HNR proposed by de Krom (1993) and the modified time-domain-based estimation ($r=0.57$, $p<0.0001$). It is obvious, however, that our modified frequency-domain approach yielded a much stronger correlation with the time-domain estimate of HNR than de Krom's (1993) method. Our method accounts for roughly

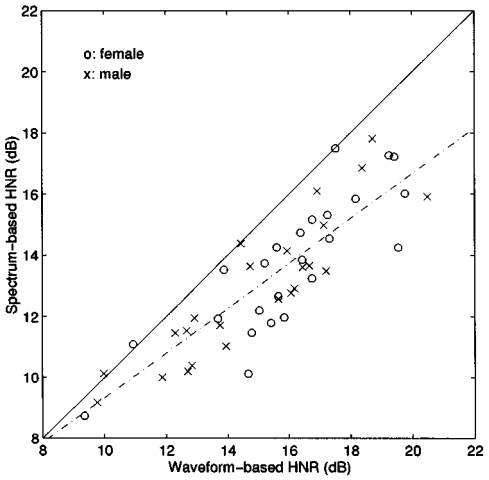

FIG. 8. The modified frequency-domain HNR as a function of the time-domain HNR for human voice signals.

77% of the variance ($r^2=0.77$) in the time-domain-based estimates, compared to 33% ($r^2=0.33$) accounted for by de Krom's approach. Finally, it should be noted that the slope of the least-square regression lines was smaller than unity in both sets of data (see Figs. 8 and 9), although it is closer to unity for our modified method than for de Krom's approach. Overall, estimates of HNR for the human voice signals in the frequency domain tend to be smaller than those in the time domain.

III. DISCUSSIONS AND CONCLUSIONS

A modified algorithm for computing HNR in the frequency domain was developed and tested. Modifications were aimed at reducing the influence of spectral leakage in the computation of harmonic energy and removing the necessity of spectral-baseline shifting prescribed in the existing algorithm (de Krom, 1993). The modified frequency-domain estimation of HNR was compared to time-domain estimation. A significant linear relationship between time and frequency estimations of HNR was established for human voice signals, suggesting equivalency between our modified time- and frequency-domain estimations of HNR. Because frequency-domain HNR has a number of advantages discussed previously, it may be preferable for many practical applications.

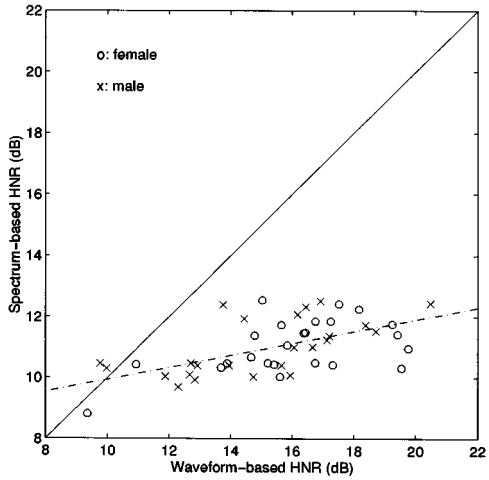

FIG. 9. The original frequency-domain HNR as a function of the time-domain HNR for human voice signals.

The correspondence between time- and frequency-domain estimation of HNR in human voice signals was not perfect, i.e., the slope of the regression line relating the two was not unity for human voice signals. Two factors may contribute to reducing the slope of the regression line between time- and frequency-domain estimates:

(1) Time-normalization of the period ensemble may have exaggerated the similarity between individual periods of the waveform, resulting in a slight overestimation of HNR in the time domain.
(2) Spectral leakage may have reduced the magnitude of harmonics in the process of estimating HNR in the frequency domain, resulting in a slight underestimation of HNR.

It is important to be aware of the differences between frequency- and time-based estimation of HNR and the possible origins of such differences.

The establishment of a highly significant, linear relationship between time- and frequency-domain estimations of HNR suggests to us that it should be productive to further explore the use of these measurements to define physical properties of human voice signals. Claims about the validity of these measurements are not warranted. Validation of these measurements depends on our ability to synthesize signals that are true representations of normal/pathological voices. This is difficult to accomplish because of our lack of complete understanding of normal/pathological voice production. For example, voice signals usually are synthesized using a linear, source-filter system that ignores the interactions between source and filter. Such synthetic signals are relatively simple to analyze, but are not complete representations of human voice signals. It is even more challenging to synthesize various pathological voices. Our aim was to demonstrate that the frequency-domain HNR measurement developed and tested here is in better compliance with the time and frequency equivalency principle of signal analysis than the existing frequency-domain approach.

ACKNOWLEDGMENTS

This work was supported, in part, by grants from the National Institute of Deafness and Other Communication Disorders: (1) DC01440, Analysis and Improvement of Alaryngeal Speech, and (2) DC00266, Objective Assessment of Vocal Hyperfunction.

Childers, D., and Lee, C. (**1991**). ''Vocal quality factors: Analysis, synthesis, and perception,'' J. Acoust. Soc. Am. **90,** 2394–2410.

Cox, N. B., Ito, M., and Morrison, M. D. (**1989**). ''Technical considerations in computation of spectral harmonics-to-noise ratios for sustained vowels,'' J. Speech Hear. Res. **32,** 203–218.

de Krom, G. (**1993**). ''A cepstrum-based techniques for determining a harmonics-to-noise ratio in speech signals,'' J. Speech Hear. Res. **36,** 254–266.

Diaz, J., Hillman, R., Gress, C., Bunting, G., and Doyle, P. (**1993**). ''Voicebase: A clinical and research database for voice disorders,'' ASHA **35**, 183.
Emanuel, F. (**1991**). ''Spectral noise,'' Sem. Speech Language **12**, 115–130.
Fant, G. (**1960**). *Acoustic Theory of Speech Production* (Mouton, S-Gravenhage).
Fant, G., Liljencrants, J., and Lin, Q. G. (**1985**). ''A four-parameter model of glottal flow,'' STL-QPSR **4**, 1–12.
Hillenbrand, J., Cleveland, R., and Erickson, R. (**1994**). ''Acoustic correlates of breathy vocal quality,'' J. Speech Hear. Res. **37,** 769–778.
Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S. (**1986**). ''Normalized noise energy as an acoustic measure to evaluate pathologic voice,'' J. Acoust. Soc. Am. **80,** 1329.
Kay-Elemetrics (**1994**). *Voice and Speech Laboratory—Massachusetts Eye and Ear Infirmary—Voice Disorders Database* (Kay Elemetrics, Lincoln Park, NJ).
Klatt, D., and Klatt, L. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87,** 820–857.
Muta, H., Baer, T., Kikuju, W., Tervo, M., and Fukuda, H. (**1988**). ''A pitch-synchronous analysis of hoarseness in running speech,'' J. Acoust. Soc. Am. **84,** 1292–1301.
Oppenheim, A., and Schafer, R. (**1989**). *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
Qi, Y. (**1992**). ''Time-normalization in voice analysis,'' J. Acoust. Soc. Am. **92,** 2569–2576.
Qi, Y., and Bi, N. (**1994**). ''A simplified approximation of the four-parameter LF model of voice source,'' J. Acoust. Soc. Am. **96,** 1182–1185.
Qi, Y., and Shipp, T. (**1992**). ''An adaptive method for tracking voicing irregularities,'' J. Acoust. Soc. Am. **91,** 3471–3477.
Qi, Y., Weinberg, B., and Bi, N. (**1995a**). ''Enhancement of female esophageal and tracheoesophageal speech,'' J. Acoust. Soc. Am. **97,** 2461–2465.
Qi, Y., Weinberg, B., Bi, N., and Hess, W. J. (**1995b**). ''Minimizing the effect of period determination on the computation of amplitude perturbation in voice,'' J. Acoust. Soc. Am. **97,** 2525–2532.
Yumoto, E., Gould, W., and Baer, T. (**1982**). ''Harmonics-to-noise ratio as an index of the degree of hoarseness,'' J. Acoust. Soc. Am. **71,** 1544–1550.