HNR EXTRACTION IN VOICED SPEECH, ORIENTED TOWARDS VOICE QUALITY ANALYSIS

François Severin, Baris Bozkurt, Thierry Dutoit

TCTS Lab, Faculté Polytechnique de Mons
Instituts Sci. Park, B-7000 Mons, Belgium
phone: (+32) 65.37.47.04, fax: (+32) 65.37.47.29, email: {francois.severin, bozkurt, thierry.dutoit}@tcts.fpms.ac.be
web: www.tcts.fpms.ac.be

ABSTRACT
This study tests three methods (algorithms of G. de Krom, C. d'Alessandro et al. and P. Boersma) to estimate the Harmonics-to-Noise Ratio (HNR) in speech. Tests are made on two databases of naturally connected speech designed for voice quality analysis. First, results of the three methods are compared, then the relevance of each method is analysed separately. The conclusion is that they are all good indicators of the amount of noise in speech, and though their accuracy is limited, they are efficient for voice quality analysis.

1. INTRODUCTION
The estimation of the aperiodic component in speech is very useful for voice quality analysis, as aperiodicity is known to characterize certain phonation types [1]. Noise measurement can also be used to classify voice pathologies ([2], [3]).
In this study we will focus on the two main sources of aperiodicity in the glottal flow component:
• Additive noise is caused by a constriction in the vocal system, leading to a turbulent flow. This study will focus only on aspiration noise, due to a constriction at the glottis. Additive noise is supposed to be quasi-stationary and gaussian [4] ; it characterizes especially breathy voices.
• Jitter and shimmer are noises which are not additive but structural. They correspond to random variations of the fundamental period and amplitude of the speech signal, and characterize rough voices.

Some methods have been proposed to quantify the presence of noise in speech. Jackson [5], d'Alessandro et al [6], de Krom [7], Stylianou [8] have designed algorithms to split the speech signal into its periodic and aperiodic components. The HNR (Harmonics-to-Noise Ratio) is defined as the log ratio of the energies of these two components. Other measures of the presence of noise were established without having to retrieve the two components ([9], [10], ...).

In our present work we have chosen to implement the two often referred algorithms developed by de Krom and d'Alessandro et al. Then we compare them to each other and to Boersma's method [9], used in the software Praat [11]. We consider none of the three methods to be a reference technique, however the consistency of the three methods in sorting utterances according to their HNR would allow us to use any of them for voice quality classification tasks.

2. DESCRIPTION OF THE ALGORITHMS
For convenience reasons we will call the de Krom’s method "method A", d'Alessandro’s "method B" and Boersma’s "method C".

2.1 Method A : de Krom's algorithm
It is a frequency-domain method, based on a harmonic analysis [7]. The following linear speech model is assumed:

\[ s(t) = e(t) * v(t) = (p(t) + a(t)) * v(t) \]

where \( s(t) \) is the speech signal, \( e(t) \) the excitation signal constituted of a quasiperiodic component \( p(t) \) and an aperiodic component \( a(t) \) ; this excitation is convolved with the vocal tract impulse response \( v(t) \).

The first step consists of windowing the signal. The window has to be large enough to allow accurate harmonics detection but not too much in order to respect the pseudo-stationarity hypothesis. The windowed speech segments are overlapping segments ; the Hanning window is chosen.

Then the real cepstrum (real part of the inverse Fourier transform of log spectrum) is computed. The harmonics of the spectrum give rise to cepstral peaks called rahmonics, the first one corresponding to the fundamental period. As a consequence of the log operation, rahmonics contain information about the periodic excitation only. The Fourier transform of the "rahmonic comb-lifted" cepstrum provides a representation of the aperiodic speech component's log spectrum. Then the periodic component spectrum and the HNR can be easily retrieved. The HNR is defined by de Krom as the difference between the log spectra of speech and aperiodic components ; he showed it to be sensitive to both additive noise and jitter. We prefer to compute the log ratio of the energies of the periodic and aperiodic components spectra.

2.2 Method B : d'Alessandro's algorithm
This method is also based on a harmonic analysis [6]. The assumed speech model is the model proposed in (1). A Linear Prediction (LP) analysis is first used to estimate the excitation signal, as its samples are much less correlated than the speech signal samples. The aim is to reduce undesirable effects in the analysis due to truncation of highly-correlated samples signals. The Hamming windowing function is then applied to the overlapping excitation segments.

The pitch is defined as the frequency corresponding to
the first rahmonic of the residual cepstrum. Then the residual amplitude spectrum is equally separated into its periodic and aperiodic regions, delimited as the positive and negative regions of a sinusoid whose frequency is the pitch. The output of the algorithm is an estimate of the excitation's aperiodic component. For this, an extrapolation method is proposed, assumed that the aperiodic region is mainly constituted by noise spectrum, but the periodic region spectrum is due to both periodic and aperiodic components. The aperiodic component is then estimated, starting from the aperiodic region spectrum and going back and forth between frequency domain and time domain, while imposing finite duration constraint in the time domain, and known noise spectrum constraint in the frequency domain. Several iterations are needed to retrieve a correct estimation of the excitation's aperiodic component. Then, subtracting it from the excitation signal provides its periodic component. The speech periodic and aperiodic components can be recomposed by LP synthesis, then the HNR is computed. This HNR estimation has been shown to be sensitive to additive random noise as well as to jitter and shimmer.

As methods A and B require pitch calculation, we use a common pitch estimation method based on autocorrelation. This reduces the algorithms execution cost and simplifies the comparison. A voiced/unvoiced detector is also included.

2.3 Method C : Boersma’s algorithm

Boersma’s method does not include frequency domain processing; it uses the short-term autocorrelation function of speech to determine the pitch, then the HNR [9].

The autocorrelation of a signal is defined as:

\[ r_c(\tau) = \int x(t) x(t + \tau) dt . \]

The fundamental period \( T_0 \) is defined as the value of \( \tau \) corresponding to the highest maximum (index zero excluded) of the short-term autocorrelation function (called hereafter \( r_c(\tau) \)). The energy of the windowed speech signal is the value of the short-term autocorrelation function at its index zero:

\[ r_c(0) = r_c(0) + r_w(0) , \]

\[ r_c(0) \] and \( r_w(0) \) being the respective energies of the periodic and aperiodic components.

The normalized autocorrelation is defined as:

\[ r_c^*(\tau) = \frac{r_c(\tau)}{r_c(0)} . \]

Given the periodicity of the periodic component autocorrelation function and assuming an additive white noise (uncorrelated with itself), the energy of the periodic component is given by:

\[ r_c^*(0) = r_c^*(0) = r_c^*(0) . \]

then the energy of the aperiodic component:

\[ r_w(0) = 1 - r_c^*(0) = 1 - r_c^*(0) . \]

The HNR is defined as:

\[ \text{HNR} = \frac{r_c^*(0)}{r_w(0)} . \]

Though this algorithm is based on an additive white noise aperiodic component, it was shown to be correctly sensitive to jitter also.

![Figure 1. HNR estimation with method A, varying the pitch (80 to 300 Hz) and the imposed HNR (0 to 50 dB). Each curve corresponds to the variation of the HNR estimate for a constant pitch.](image)

3. TESTINGS

Methods A & B were first tested on synthetic speech, in order to check their efficiency to detect additive random noise and to set their parameters. Then they were applied together with method C to two naturally connected speech databases. The first one is designed for general voice quality tests [12], and the second one for loudness analysis.

3.1 Synthetic speech

Speech is generated by the convolution of an excitation signal and an all-pole vocal tract filter. We consider a four-formants vocal tract, and the excitation signal is a periodic LF-modelled source signal [13] plus additive white noise. Varying parameters are the window length, the noise duration (relative to the fundamental period) and the pitch. The average HNR is calculated on voiced utterances of 200 fundamental periods. Testing method A shows that the window duration has an influence on the HNR estimation, especially for high values of the HNR (Fig. 1), in the sense that both short fundamental periods and long observation windows lead to a better HNR estimation. The observation window has to be wide enough for an accurate HNR estimation, but narrow enough to respect the pseudo-stationarity hypothesis. Method B is observed to be less sensitive to the window length.

3.2 Natural speech

3.2.1 Database designed for global voice quality tests [12]

A male speaker pronounces the same sentence (American English) with different voice qualities: modal, tensed, creaky, rough, laughing, etc. There are 75 utterances (16-bit wav files) sampled at 44100 Hz. Methods A & B are used with the same parameter settings as in tests on synthetic speech. We test two window lengths: 46 ms (2048 samples) and 92 ms (4096 samples); 92 ms is not theoretically suitable for connected speech analysis but our experiences on synthetic speech have shown the importance of using a large observation window. Time shift between two consecutive frames is 10 ms. For the remaining part of this paper, the variable we will refer to as HNR is the average HNR of all voiced frames for each utterance.

As the three methods do not provide the same range of HNR, we cannot compare their results directly. So we build a relative scale, based on the observation that for the three methods, the two utterances that give minimum and
maximum HNR are the same (for this we only consider utterances in which pitch analysis results in many voiced frames - about 80% of the utterances are concerned). For each method we score every utterance according to its relative position compared to these minimum and maximum, introducing the Relative HNR (RHNLR):

\[ RHNLR_{ij} = \frac{HNR_{ij} - HNR_{min}}{HNR_{max} - HNR_{min}}, \]

\( HNR_{ij} \) being the average HNR estimated for the utterance \( j \), method \( i \) (\( min \) and \( max \) are the indexes of the utterances giving the extreme values of HNR; they are the same for the three methods).

The RHNLR are then compared between the three methods. For this, we calculate the maximum difference between the three RHNLR of the same utterance:

\[ \Delta_j = \max_{i,j}(RHNLR_{ij} - RHNLR_{ji}). \]

We name these indexes the divergence coefficients: for the same utterance, they reflect the maximum difference between results of the three methods. Most of the divergence coefficients are contained between 0.1 and 0.2 (Fig.2) which illustrates a small (but not negligible) difference between the estimation methods. Moreover, methods are more consistent with each other when using the largest window. Most of time the divergence coefficients are high because one of the three methods provides an estimation that is very different from the two others. A detailed analysis leads to the following conclusions.

Method A "underestimates" the HNR ("under/over estimation" means here that the HNR is quite low/high compared to the two other methods) for soft or rough voices (see Fig. 3). This can be explained by comparing methods A and B, which is easier as they are similar in their basic principles. In method A, the aperiodic region width depends on the harmonics amplitude, which is smaller for soft or rough voices than for modal voices; with method B, the aperiodic region width is fixed. Though these two methods differ in the rest of their algorithm, this could partially explain the estimation differences. It is also observed that method A underestimates RHNLR of voices that are both shouting and rough (their RHNLR estimate is 0.34 for method A, and ranges from 0.51 to 0.72 for methods B & C).

Method C overestimates RHNLR of voices with a very high pitch (i.e. the speaker adopts a 500-Hz pitch) : RHNLR is about 0.96 for method C, and ranges from 0.63 to 0.72 for methods A & B. But more generally, method C presents underestimations of RHNLR, for which we could not define any rule.

About the overall RHNLR estimations, many observations are the same for the three methods and the two window lengths (though increasing this involves a decrease of the HNR estimate). Observations are summarized in Fig. 3.

The first observation is about rough, creaky, or whispering voices utterances : all of them have a low estimated RHNLR which confirms observations of [1], [2], [3].

A large RHNLR was estimated for utterances with a high pitch (about 500 Hz), especially for method C. The high pitch and the high open quotient associated to falsetto voices [14] make their waveform less influenced by non-harmonic fluctuations usually due to the first formant. This gives these signals a strong harmonic character, so a high RHNLR.

There are some loud voices in the database, labelled according to their degree of loudness. This classification can be retrieved by sorting the RHNLR. Moreover, most RHNLR of shouting voices are higher than RHNLR of load voices. (More observations about load voices are found in section 3.2.3.)

On the other hand, almost-whispering voices are found to have a low RHNLR (except for method B with the 46-ms window). But though all methods converge to the same result, this observation has to be taken with care because of the few voiced frames detected in this kind of voice. Method A estimates a very low RHNLR for all soft voices, for reasons we have already described. This is perceptually the most logical result, though it diverges from methods B and C.

The estimated RHNLR of the tense phonation is high. This was expected, but it has to be taken carefully as we selected only one tense voice in the database; this also explains the variations among the classifications of Fig. 3. (The same carefullness has to be taken about the creaky voice).
CONCLUSION

4. The main aim of this study was to compare these methods to...

3.2.2 Future work. We wish to address the noise in...