HNR EXTRACTION IN VOICED SPEECH, ORIENTED TOWARDS VOICE QUALITY ANALYSIS

François Severin, Baris Bozkurt, Thierry Dutoit

TCTS Lab, Faculté Polytechnique de Mons
Initialis 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-stationnary 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 developped 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 :

$$s(t) = e(t) * v(t) = (p(t) + a(t)) * v(t) \qquad (1)$$

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-liftered" 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 undesireable effects in the analysis due to truncation of higly-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 sinuisoid 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, substracting 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 :

$$r_x(\tau) \triangleq \int x(t)\, x(t+\tau)\, dt \,.$$

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

$$r_x(0) = r_p(0) + r_{ap}(0)\,,$$

$r_p(0)$ and $r_{ap}(0)$ being the respective energies of the periodic and aperiodic components.

The normalized autocorrelation is defined as:

$$r_x^{'}(\tau) = \frac{r_x(\tau)}{r_x(0)}\,.$$

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 :

$$r_p^{'}(0) = r_p^{'}(T_0) = r_x^{'}(T_0)\ ,$$

then the energy of the aperiodic component :

$$r_{ap}^{'}(0) = 1 - r_p^{'}(0) = 1 - r_x^{'}(T_0)\,.$$

The HNR is defined as :

$$HNR \triangleq \frac{r_p^{'}(0)}{r_{ap}^{'}(0)}\,.$$

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

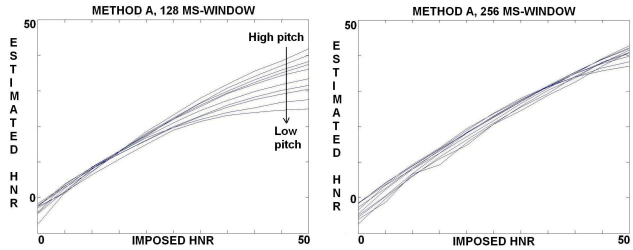

Figure 1. HNR estimation with method A, varying the pitch (80 to 300 Hz) and the imposed HNR (0 to 50 dB). Each curve corresponds to the variation of the HNR estimate for a constant pitch.

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 testings [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

3.2.1Database designed for global voice quality testings [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

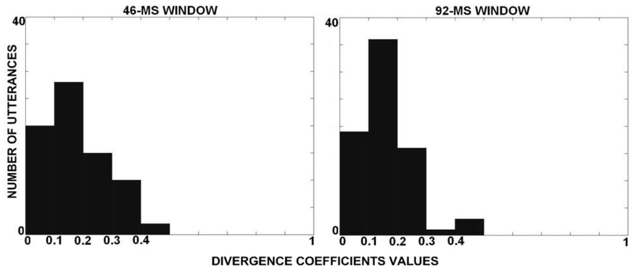

Figure 2. Distributions of the divergence coefficients : on the left, with a 46-ms window; on the right, with a 92-ms window.

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 (RHNR)* :

$$RHNR_{i,j} = \frac{HNR_{i,j} - HNR_{i,\min}}{HNR_{i,\max} - HNR_{i,\min}},$$

$HNR_{i,j}$ 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 RHNR are then compared between the three methods. For this, we calculate the maximum difference between the three RHNR of the same utterance :

$$\Delta_j = \max_{i,k}(RHNR_{i,j} - RHNR_{k,j}) \cdot$$

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 RHNR ("under/over estimation" means here that the RHNR 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 RHNR of voices that are both shouting and rough (their RHNR estimate is 0.34 for method A, and ranges from 0.51 to 0.72 for methods B & C).

Method C overestimates RHNR of voices with a very high pitch (i.e. the speaker adopts a 500-Hz pitch) : RHNR 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

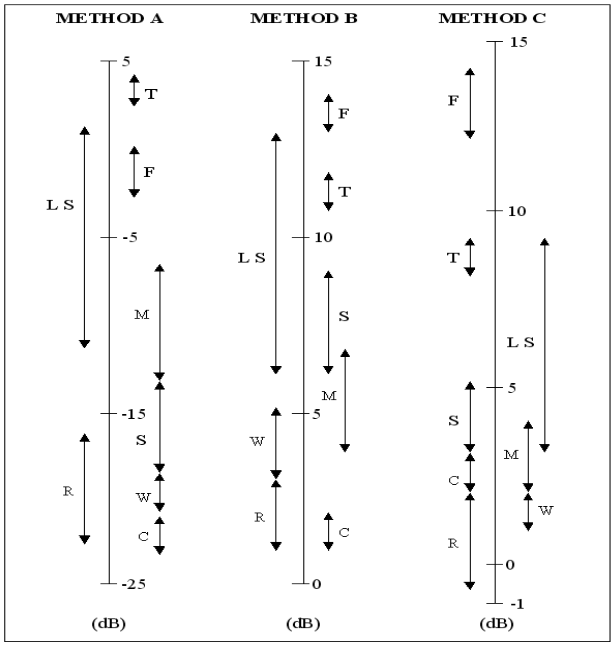

Figure 3. Distributions of the utterances according to their HNR estimates and to their voice quality - C : creaky, F : falsetto, LS : loud to shouting, M : modal, R : rough, S : soft, T : tense, W : almost whispering. The window length is 46 ms.

underestimations of RHNR, for which we could not define any rule.

About the overall RHNR 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 RHNR which confirms observations of [1], [2], [3].

A large RHNR 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 RHNR.

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

On the other hand, almost-whispering voices are found to have a low RHNR (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 RHNR 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 RHNR 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).

Classified \ Expected	Method	Calm voice	Modal voice	Loud voice
Calm voice	A	87.5%	12.5 %	0 %
	B	80 %	20 %	0 %
	C	22.5 %	25 %	51.2 %
Modal voice	A	10 %	72.5 %	17.1 %
	B	20 %	67.5 %	12.2 %
	C	45 %	40 %	14.6 %
Loud voice	A	2.5 %	15 %	82.9 %
	B	0 %	12.5 %	87.8 %
	C	32.5 %	35 %	34.2 %

Figure 4. Confusion matrix of the results of the three methods applied on a database designed for loudness analysis.

3.2.2 Further tests with added white noise

As an additional test, we measured the HNR of the same utterances in which we added white noise to the source signal (with equal HNR for all utterances) through an inverse-filtering/noising/LP-synthesis algorithm. The result is an increase of the estimated HNR, and the classification of the utterances according to their RHNR is well preserved. This shows the robustness of the methods to detect aspiration noise in naturally connected speech.

3.2.3 Database designed for loudness measurements

In the context of studying voice dynamics variations, we built a naturally connected speech database constituted of 10 French sentences pronounced three times (corresponding to three loudness levels : calm, modal and loud) by two male speakers and recorded with two different microphones. This makes a 120-utterances database (16-bit wav files sampled at 48 kHz). The average HNR of each utterance is measured with the three methods (observation window is 42 ms with methods A & B, 64 ms with method C ; time shift between two successive windows is 10 ms). Then the utterances are classified according to their HNR; the 1st third of them is classified as "calm voice", the 2nd third as "modal voice", and the 3rd third as "loud voice". A confusion matrix is presented on Fig. 4, showing that both methods A & B seem appropriate to classify the utterances according to their loudness, but that method C is not suitable for that purpose.

4. CONCLUSION

The main aim of this study was to compare three methods to estimate the HNR of voiced speech. As we do not have any reference method, the consistency of the results of the three coupled to the relevance of the estimations would be an indication of their reliability for voice quality classification tasks. For this comparison a first voice quality database is used. The three methods provide similar results, especially method B. Method A is the most divergent, but these divergences can be anticipated, in opposition to underestimations of method C. Then, a separate analysis of the results of the three methods leads us to the same conclusion : they are all good indicators of the presence of noise in voiced speech, though their accuracy is rather limited. This means that we can envisage applications like classification of the utterances according to their voice quality (sorting them according to only one criterion like loudness as it was done on the second database, or distinguishing creaky voices from modal ones, etc.), but we could not imagine a HNR measurement accurate to the tenth of dB for example. Moreover, establishing HNR classification scales like the ones of Fig. 3 should take into account the remarks of section 3.2.1 (high sensitiveness of method A for soft voices, high HNR estimate for falsetto voices, etc.).

5. ACKNOWLEDGEMENTS

This study is funded by Region Wallonne, Belgium, grant EUREKA # 03/1/5449.

The authors wish to thank the studio "5e Saison" (Paris, France) for their recording of the second database (cf. section 3.2.3).

REFERENCES

[1] D.H. Klatt and L.C. Klatt, L.C. (1990), "Analysis, synthesis and perception of voice quality variations among female and male talkers", in *Journal of the Acoustical Society of America*, 87 (2), pp. 820-857, February 1990.
[2] M. Fröhlich, D. Michaelis, H.W. Stube and E. Kruse, "Acoustic voice quality description : case studies for different regions of the hoarseness diagram", in *Advances in quantitative laryngoscopy, 2nd 'Round table'*, pp. 143-150, Erlangen (1997).
[3] T. Li, C. Jo, S.-G. Wang, B.-G. Yang, H.-S. Kim, "Classification of pathological voice including severely noisy cases", in *Proc. ICSLP 2004*, Jeju, Korea, pp. 77-80.
[4] G. Richard and C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component", in *Speech Communication*, 19 (1996), pp. 221-244.
[5] P.J.B. Jackson, "Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech", in *IEEE transactions on speech and audio processing*, vol. 9 (7), pp. 713-726, October 2001.
[6] C.d'Alessandro, B.Yegnanarayana and V. Darsinos, "Decomposition of speech signals into deterministic and stochastic components", in *Proc. ICASSP 1995*, Detroit, MI, USA, pp. 760-763.
[7] G. de Krom, *"Acoustic correlates of breathiness and roughness"*, PhD-thesis, published by LEd, Utrecht, 1994.
[8] J. Laroche, Y. Stylianou and E. Moulines, "HNS : speech modification based on a harmonic+noise model", in *Proc. ICASSP 93*, Minneapolis, USA.
[9] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound" in *IFA Proceedings* 17, 1993, pp. 97-110.
[10] E. Yumoto, W.J. Gould, T. Baer, "Harmonic-to-noise-ratio as index of the degree of hoarseness", in *Journal of the Acoustical Society of America*, 71 (1982), pp. 1544-1550.
[11] www.praat.org
[12] http://www.limsi.fr/VOQUAL
[13] G. Fant, J. Liljencrants, L. Qi-guang, "A four-parameter model of glottal flow", *STL QPSR*, 4, pp. 1-13, 1985.
[14] N. Henrich, "Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception", PhD-thesis, Paris VI university, 2001.