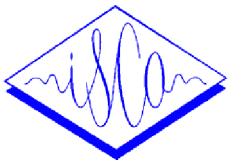

Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

PITCH AND NOISE ESTIMATION IN HOARSE VOICES

Claudia Manfredi, Piero Bruscaglioni*, Massimo D'Aniello, Luigi Pierazzi, Andrea Ismaelli*
Electronic Engineering Dept., Univ. of Firenze, Via S. Marta 3, 50139 Firenze, Italy.
E-mail: manfredi@die.unifi.it
*Physics Dept., Univ. of Firenze, Via S. Marta 3, 50139 Firenze, Italy. E-mail: bruscaglioni@dffs.unifi.it

Abstract
In pathologic voices, both slow and fast pitch variations within an utterance are indicative of the patient status. Moreover, the spectrogram of such voices usually shows high noise components, closely related to the degree of perceived hoarseness of the voice. In the present paper, both pitch and noise variations are tracked during an utterance. This is accomplished by means of a two-step procedure for finding f_0, based on robust estimation approaches, which allows selecting the varying optimal time window for analysis. The Normalised Noise Energy method [1] is revisited and an adaptive version is applied on optimised signal windows. Empty "dip" regions are avoided and the method results applicable both to sustained vowels and to words. Simulations show the good performance of the proposed approach. Its application to real data allows the physician objectively tracking important voice parameters.

INTRODUCTION

Everyone has a habitual pitch level that is used naturally on the average. However, pitch is shifted up and down in speaking in response to linguistic structure, emotional factors and other natural factors. Possible disease or malformation can also affect this pattern. Fast variations of pitch are referred to as jitter.
Any abnormality of the larynx that affects the vibration pattern of the vocal folds and the audible quality of the speech will be evident in the glottal waveform. Moreover, pathologic speech signals are corrupted by 'noise', which is directly related to the perceived roughness of voice. Commonly, subjective testing is performed by a physician, followed by direct or indirect laryngoscopy, which causes pain and discomfort to the patient. Hence, non-invasive robust methods are required capable of recovering the speech fundamental frequency from noisy speech signals and quantifying the degree of hoarseness objectively. In this paper, great attention is paid to properly defining the optimal time window length for analysis. Due to signal non-stationarity, a variable length is proposed, tailored to the varying speech characteristics. This problem is addressed by means of a two-step procedure. Each step is based on classical pitch estimation methods (Simple Inverse Filter Tracking (SIFT), Average Magnitude Difference Function (AMDF), Wavelets) revisited in order to enhance robustness to noise and jitter. Based on the previous step for pitch estimation, an adaptive version of the Normalised Noise Energy index [1] is proposed. The aim is that of giving the physician the possibility of tracking noise energy evolution within a complete word objectively. The importance in following this varying parameter is also in evaluating the effort made by the patient in speaking. The latter could be indicative of the effectiveness of the adopted operatory technique, as far as post-operatory functional recovery is concerned. The adaptive procedure is applied both to simulated signals, with different SNR and jitter values, and to real pathological voices, relative to vocal folds operated patients. The results show the good performance of the method, especially for highly degraded voices.

ANALYSIS OF THE NNE METHOD

The Normalised Noise Energy (NNE) acoustic measure [1] is a measure of the dysphonic component of the voice spectrum related to total signal energy. It was shown to be competitive with other approaches, mainly as far as robustness to jitter and shimmer are concerned, since it can be applied to short time intervals (seven pitch periods [1]). Given the speech signal $x(n)=s(n)+w(n)$, where $s(n)$ is the periodic component and $w(n)$ is the additive noise component, the NNE is defined

as a measure of the ratio between signal noise energy (dysphonic component of voice) and total signal energy. Let an M-points windowed speech signal in the m-th frame be represented by:

$$x_m(n)=s_m(n)+w_m(n),\ n=0,\ldots M-1 \quad (1)$$

and let $X_m(k)$ and $W_m(k)$ be the N-point Discrete Fourier Transform (DFT) of the signal $x_m(n)$ and noise $w_m(n)$ respectively. The NNE is evaluated according to the following equation:

$$NNE = 10\log\left[\frac{\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}\left|\tilde{W}_m(k)\right|^2}{\frac{1}{L}\sum_{k=N_L}^{N_H}\sum_{m=1}^{L}\left|X_m(k)\right|^2}\right] \quad (2)$$

with $N_L=\lceil Nf_LT\rceil$, $N_H=\lceil Nf_HT\rceil$, where L is the total number of frames in the analysis interval, f_L and f_H are the lowest and highest frequencies of the frequency band of interest, $|\tilde{W}_m(k)|^2$ is an estimate of the unknown noise energy $|W_m(k)|^2$ and T is the sampling period. In [1], it was shown that $|\tilde{W}_m(k)|^2$ can be estimated in the harmonic dip regions D_i (i.e. where the harmonics have no or negligible component) by:

$$|\tilde{W}_m(k)|^2 = |X_m(k)|^2,\ k\in D_i \quad (3)$$

while, in the harmonics peak regions P_i, it can be obtained by interpolating between values in the deep regions on both sides of each peak region:

$$|\tilde{W}_m(k)|^2 = \frac{1}{2}\left\{\frac{1}{N_i}\sum_{r\in D_i}|X_m(r)|^2 + \frac{1}{N_{i+1}}\sum_{r\in D_{i+1}}|X_m(r)|^2\right\} \quad (4)$$

where $k\in P_i$ and N_i, N_{i+1} are the number of points in D_i and D_{i+1} respectively. The width of peak and dip regions is related to the choice of the window function. Here, as in [1], a Hamming window is considered. The approximate bandwidth of the Hamming window up to the first zero is 2N/M. Hence, with this choice:

$$P_i = \{k \mid k_i-2N/M \le k \le k_i+2N/M\}$$
$$D_i = \{k \mid k_{i-1}+2N/M < k < k_i-2N/M\} \quad (5)$$

where k_i and k_{i-1} are the i-th and (i-1)-th harmonic peak locations respectively. In order to avoid possible empty dip regions, in [1] the Hamming window length was set equal to $7T_p$, where T_p is the pitch period of the signal under study. However, especially for low-pitched voices, this choice may cause problems due to signal non-stationarity. The dip region D_i may disappear for a fixed value of window length M, as k_i and k_{i-1} get closer to each other, due to low fundamental frequency of phonation. From eq.(5), this situation occurs when the distance between k_i and k_{i-1} is less than 4N/M points, where N is the number of DFT points. Hence, in order to have non-empty dip regions, the following relation must be verified:

$$2\left(2\frac{N}{M}\right)+(d+1)\le f_0\frac{N}{F_s} \quad (6)$$

where: d is the number of points in the dip region, f_0 is the signal fundamental frequency and F_s is the sampling frequency. In order for eq. (6) to be verified also for low fundamental frequency values, the main lobe bandwidth of the window function must be narrower than the fundamental frequency value of the signal under consideration. It is thus of importance to analyse the relation between M and f_0. Eq. (6) can be rewritten as:

$$M \ge \frac{4N}{f_0\frac{N}{F_s}-(d+1)} \quad (7)$$

This relation gives, for varying f_0, the optimum time window length which guarantees at least d points in the dip regions, for fixed N and F_s. It can be verified that, for very low f_0 values, the required window size becomes extremely long. This may imply loss of stationarity in the signal under consideration and, consequently, incorrect evaluation of voice parameters. Moreover, it is evident that an increase of the number N of DFT points allows a shorter time window to be selected, with better results as far as NNE estimation is concerned. In the present application, N=16384. This choice for N is also effective as far as the number d of points in dip regions is concerned. In Fig.1, M is plotted as a function of f_0 and d, for fixed N=16384 and F_s=25kHz. Apparently, as d increases, the required window length shows little variation also for low f_0 values. The considerations made above suggest adaptively evaluating the NNE on subsequent signal frames, whose length takes into account possibly varying fundamental frequency values within an utterance. The algorithm proposed in the present work for noise energy evaluation consists of the two steps described below.

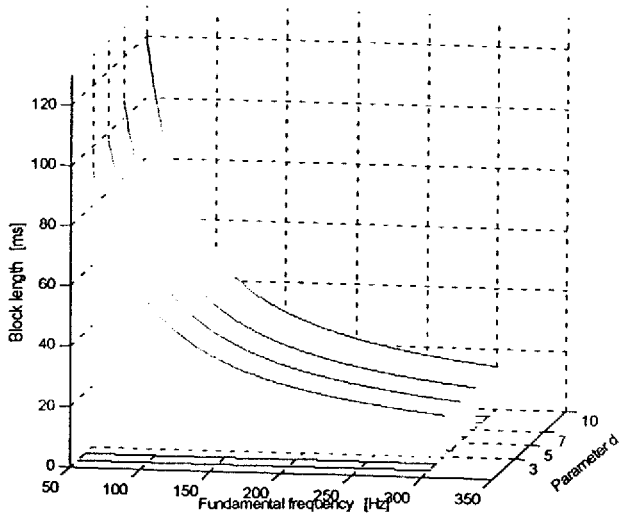

Figure 1 - Window length as a function of f_0 and d

Fist step: finding the range of variation for f_0.

As f_0 is unknown, the window length is chosen as $M = 3\,F_s / f_{min}$, where f_{min} is the minimum allowed f_0 value for the signal under consideration. In the present application (adult male voices) it was set equal to 50 Hz. This choice for M, though non optimal as far as signal stationarity and NNE evaluation are concerned, guarantees at least three pitch period in each data frame, thus allowing f_0 evaluation. This estimate is refined in the second step by means of a more accurate procedure. In a previous work [2] it was shown that a robust and fast procedure for estimating f_0 is the SIFT algorithm, when proper modifications are made to the method. Hence, this approach is applied here in order to obtain an initial rough f_0 estimation, which is the basis for the optimum window length definition. Basically, on each data window of length M, it consists of an identification step, for the correct model order p and parameters estimation, followed by an autocorrelation maximisation of the inverse filter residuals, which gives the estimated pitch value. In order to adaptively follow signal variations, the filter order is allowed here to vary on subsequent signal frames. Its estimation is addressed by means of the Singular Value Decomposition (SVD) approach, which was found to be a powerful technique in recovering signals embedded in noise [3]. The SVD approach consists in separating the signal subspace from the noise subspace in a properly organised data matrix A of rank k. The choice of the threshold for deciding which singular values belong to the signal subspace and which to the noise subspace is addressed here by means of a new approach, the Dynamic Mean Evaluation (DME) method, which defines a distance measure between the largest Σ singular values and the smallest ones. The DME was shown to be competitive with traditional approaches, mainly as far as short time windows are concerned, as in the present application [4]. The maximum of the autocorrelation sequence (AS) of the residuals is evaluated in the frequency range of interest (50-300Hz). The pitch value is then given by $f_0 = F_s/\tau$, where τ is the index corresponding to the maximum of the AS. This procedure gives the range of variation for f_0 during the pronounced utterance: $f_l \le f_0 \le f_h$, f_l and f_h being the minimum and the maximum estimated pitch value respectively.

Second step: tracking f_0 for adaptive M evaluation.

This step aims at defining the optimum and possibly varying window length M for NNE evaluation, which can be accurately defined if the actual f_0 value is known. The f_0 estimation is now performed on data frames whose length is adaptively defined by means of eq.7, in the frequency range $[f_l, f_h]$, where f_l, f_h are obtained in the previous step. In this work, N=16384, d=10 and F_s=25 kHz. This choice amounts to selecting a window length ranging between 15 ms and 120 ms. Data frames are now overlapped for ¾ of length, in order to track both f_0 and the NNE index more accurately. The signal is band-pass filtered (50Hz-300Hz) with the 'Mexican hat' Continuous Wavelet Transform (CWT) and its periodicity is extracted by means of the Average Magnitude Difference Function (AMDF) approach. This wavelet allows good time and frequency characterisation of the signal under study and realises a band-pass filter whose parameters depend on the scale parameter [5]. The AMDF analysis, which is directly carried out on signals in the time domain, can be used to detect fast and slow variations of the fundamental frequency f_0 [6].Given a signal frame of length M, {x(k)}, k=1,...,M, the AMDF is defined as:

$$AMDF(\eta) = \sum_{i=1}^{M} |x(i) - x(i+\eta)| \qquad \eta = 0,\ldots,M-1 \qquad (11)$$

With CWT and AMDF, the procedure adopted for f_0 estimation is the following:

- On each time window, whose length is adaptively evaluated in the previous step, the scale parameter s for the CWT is allowed to vary in the range 1÷80. This choice is achieved experimentally and is due both to the strong noise level present in the signal and to the need for finding f_0 in the frequency range of interest with the selected wavelet. The k parameter varies in the allowed range (linked to the variable window length). This gives a coefficient matrix for the CWT(k,s);
- From the coefficient matrix, the optimum scale value, $\hat{s}$, is selected as the one corresponding to the maximum entry: this in fact represents the best fitting of the wavelet to data.
- The AMDF technique is applied to CWT(k,$\hat{s}$), thus obtaining the estimate of f_0 as: $f_0=F_s/\eta_{min}$, where η_{min} is the AMDF minimum obtained as described above.

The choice of the AMDF instead of the AS is due to the non-stationarity and amplitude modulation of the signals under study. These aspects were shown to often cause misestimation of the true signal periodicity with the AS [2].The proposed approach has a low computational burden, thanks to the simple computations involved in the AMDF procedure. For long data frames it is in fact more efficient than the classical correlation technique.

ADAPTIVE ESTIMATION OF NNE

In this section, the following steps describe the implementation of the adaptive NNE algorithm:

1. The sampling frequency is set equal to F_s=25kHz, according to the data acquisition device. The number of DFT points is N=16384 and the number of points in the dip regions is chosen as d=10. This guarantees accurate noise energy estimation. With this choice, the adaptive time window length M is obtained (eq.7), tailored to varying f_0 values within the utterance. The f_0 value on each data frame is known, as it was obtained with the two-step procedure described in the previous section.
2. A Hamming window multiplies each data frame, and the power spectrum on that frame is evaluated.
3. On each data frame (partially overlapped for ¾ of length) the NNE is evaluated in the frequency band 50Hz-1500Hz by means of eq.5 and the following equation:

$$NNE(k)=10\log\left[\frac{\sum_{m=N_L}^{N_H}\left|\tilde{W}_m(k)\right|^2}{\sum_{m=N_L}^{N_H}\left|X_m(k)\right|^2}\right], \qquad k=N_L,\ldots,N_H \tag{13}$$

with the same notations as in eq.(2). Notice that the NNE index is now completely independent of the f_0 value for the analysed signal.

EXPERIMENTAL RESULTS

According to [6], the vocal tract is modelled as an interconnected series of coaxial lossless tubes of fixed length and varying diameter. Only non-nasal vocal sounds are taken into account in the present work. This corresponds to an all-pole model, when the effect due to the nasal cavity is disregarded. Each pair of complex-conjugate poles approximates the resonant frequency of a single tube. For simulations, data were generated according to an almost realistic model obtained as follows. A set of real data (healthy adult male, sustained /a/ vowel) underwent a Least Squares (LS) identification process, which gave the optimum AR model order p and parameters. Specifically, the Akaike Information Criterion [7], was applied for model orders ranging from 4 to 40. The 'best' AR model was found to be of order p=22, corresponding to 11 stable complex conjugate pole pairs. This result agrees with the operational rule: $p\cong F_s$ [6], as the sampling frequency was set equal to F_s=25 kHz, on analogy to real data. The input signal is a delta train whose periodicity is given by the fundamental period $T_0=1/f_0$. Additive noise (zero-mean white noise with uniform distribution) describes the effect due to pathology. Jitter is taken into account by increasing f_0 on subsequent time windows at 2% rate. Extensive simulation was carried out, with different SNR and jitter values, in order to test the method's robustness. Simulations show that the basic and the adaptive approaches give comparable results when no slow and/or fast f_0 variations are present in the analysed signal. For high noise level and especially for high jitter values, the basic NNE approach tends to overestimate the signal noise component, mainly due to the fixed window length, which doesn't allow following varying signal characteristics. In particular, fig.2 shows the

results obtained with an increasing jitter (linearly, 2% in successive periods) is applied to the signal. As the fundamental frequency increases, the optimal time window length decreases (+,o are the first and the second pitch estimates respectively, * is the varying window length). This allows a better tracking of signal characteristics. Notice that the basic NNE approach gives an NNE value of about -15dB. This is due to the fixed window length (71.5 ms), which causes large NNE fluctuations for low pitch values and an increasing trend for high pitch values. Moreover, a large number of empty dip regions (about 40%) was found with the basic approach, with consequent less accurate noise estimation. It is to be pointed out that the basic method often fails to find the correct f_0 value, which is an important diagnostic parameter. Critical situations occur also in this case when the pitch value is very low: in fact, a window length equal to $7T_p$ may imply loss of stationarity and, consequently, bad noise energy estimation.

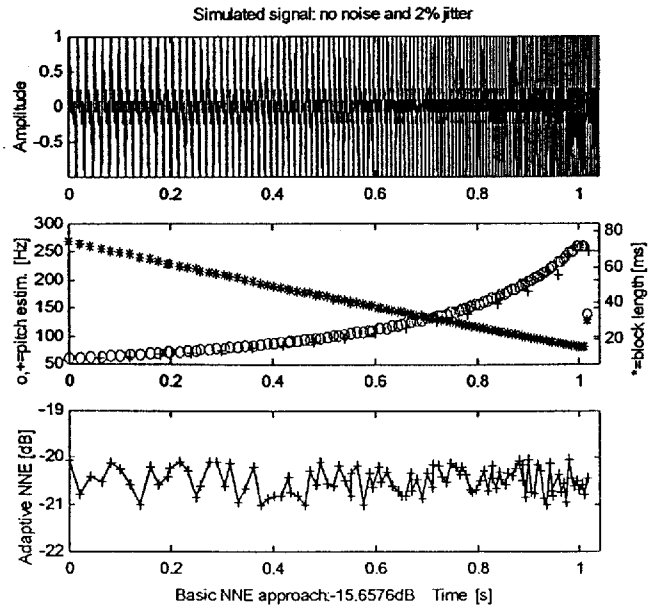

Figure 2 - Simulated signal with jitter

As for real data, voice samples were recorded in a quiet room at the Phoniatric Section of the Otholaryngoiatric Institute, Careggi Hospital, Firenze. About 30 voice samples were collected (10 healthy, 20 pathological). All patients are adult male subjects, affected by T1 glottis cancer and operated via endoscopic laser or traditional lancet technique. Subjects were required to pronounce the Italian word "aiuole" (flowerbeds), as it is made up with the five main Italian vowel sounds: "a", "e", "i", "o", "u". The use of a complete word instead of sustained vowels is due to the clinical interest in evaluating the effort made by the patient during the entire vocal emission, also as far as the glide between vowels ("ai", "iu", "uo") is concerned. However, the method is applicable to sustained vowels as well. As pointed out in [1], glottis T1 cancer (tumour confined to the glottis region with mobility of the vocal cords) spreads over a wide NNE range, with a distribution somewhat confused with that of normal subjects. The size and location of the tumour affects the NNE value thus making it difficult to correctly classify the pathology. This is still true for operated patients, whose functional recovery is influenced by multiple factors. We recall here that the aim of the present work is to give the physician an objective index quantifying both voice hoarseness and the effort made by operated patients (f_0 variations). Hence, no attempt is made to discriminate between healthy and pathological voices. Fig.3 refers to a laser-operated patient. The voice quality is rather good, as confirmed by the estimated noise values, which are comparable to those of healthy voices. The middle plot compares the two f_0 estimation procedures (+=first estimate, o=second estimate). In contrast to healthy voices, the plot shows a remarkable increase of f_0, due to the effort made by the patient in pronouncing the word. Notice that the second f_0 estimate, being a refinement of the first one, allows faster tracking of pitch variations, giving higher f_0 values corresponding to sudden f_0 increase. As a consequence, the adaptive window length (*) decreases, ranging from 35 ms to 20 ms. The lowest plot shows the adaptive NNE estimate. Notice that the transition between successive vowels can be objectively evaluated. The basic NNE method gives a value of about -20.5dB, thus overestimating the signal noise component. The analysis carried out on the available data set has confirmed the results reported here. All voice signals coming from laser cordectomised patients show an increase of f_0 during the pronounced word, possibly due to the smaller scar cord production with respect to lancet operated patients, with consequent incomplete glottal closure. This could be indicative of the effort made by operated patients in pronouncing a word. On the other hand, the adaptive noise estimate has shown an almost

constant noise value (middle part of the word) for healthy people, while large, random oscillations characterise cordectomised patients. In particular, lancet operated patients usually show almost constant but unstable pitch values and higher noise levels with respect to the other class.

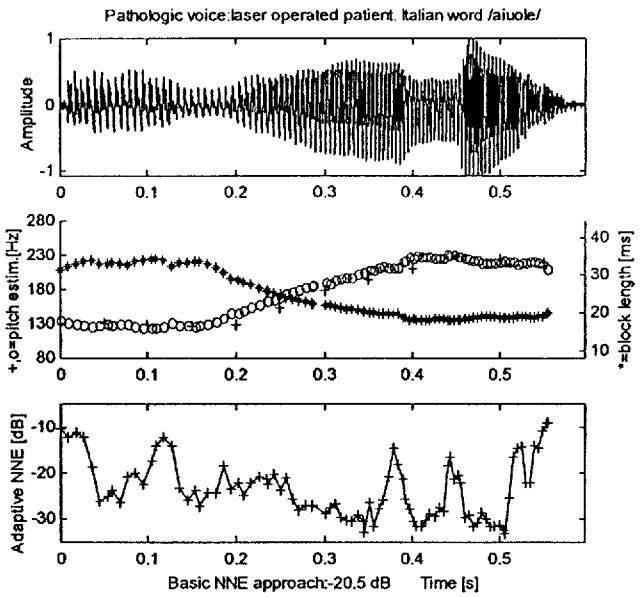

Figure 3 - Adaptive pitch and noise estimation

This approach can also be applied as far as the post-operatory evolution and possible rehabilitation are concerned, which are commonly performed on a subjective basis only. Finally, glottal functionality can be analysed by means of objective indexes other than visual inspection of the spectrogram.

FINAL REMARKS

In this paper, the problem of tracking both the pitch value and the noise level during an uttered word is considered. For pathological voices, these parameters vary during the word. In particular, pitch usually considerably increases or oscillates, due to the effort made by the patient in pronouncing the word. Hence, an adaptive procedure is proposed, capable of following fast pitch variations, which is robust against noise and jitter. The procedure allows definition of a variable time window length for evaluating the signal noise in an adaptive way. Modifications are proposed to the Normalised Noise Energy method [1], and an adaptive procedure is implemented, which allows estimation of the signal noise component on subsequent short time windows of varying length. Clinical interest lies in objectively evaluating the effort made by cordectomised patients during an utterance, as it could be indicative of patient status, also as far as post-operatory functional recovery is concerned. Extensive simulation has shown the good performance of the method, which was also applied to real data coming from T1A glottis cancer operated patients. Results obtained on real data have given to the physicians useful information and reliable indexes for these important voice parameters.

REFERENCES

[1]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 (5),** 1329-1334.

[2]Manfredi, C., D'Aniello, and M., Bruscaglioni, P. **(1998).** "Acoustic measure of noise energy in vocal folds operated patients", in *Proc. IX Europ. Signal Proc. Conf., EUSIPCO-98,* Island of Rhodes, Greece (S. Theodoridis, I. Pitas, A. Stouraitis N. Kalouptsidis) **2,**1141-1144.

[3]Rao, B.D., and Arun, K.S. **(1992).** "Model based processing of signals: a state-space approach", Proc. IEEE **80 (2),** 283-309.

[4]Fort, A., Ismaelli, A., Manfredi, C., and Bruscaglioni, P. **(1996).** "Parametric and non-parametric estimation of speech formants: application to infant cry", Med. Eng. Phys. **18 (8),** 677-691.

[5]Daubechies, I. **(1990).** "The wavelet transform, time-frequency localisation and signal analysis", *IEEE Trans on Inf. Theory* **36,** 961-1005.

[6]Deller, J.R, Proakis, J.G., and Hansen, J.H.L. **(1993).** *Discrete-time Processing of Speech Signals* (Maxwell McMillan, New York).

[7]Marple, S.L. **(1987).** *Digital Spectral Analysis with Applications* (Prentice-Hall, Englewood Cliffs NJ).