NON-LINEAR TECHNIQUES FOR DYSPHONIC VOICE ANALYSIS AND CORRECTION

Claudia Manfredi, Lorenzo Matassini*

Department of Electronics and Telecommunications, Università degli Studi di Firenze, Via S.Marta 3, 50139
Firenze, Italy. E-mail: manfredi@det.unifi.it

*Nonlinear Time Series Analysis Group, Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzstr.38,
D 01187 Dresden, Germany. E-mail: lorenzo@mpipks-dresden.mpg.de

ABSTRACT

This paper aims at finding suitable parameters for dysphonic voice analysis and classification. Moreover, a non-linear noise reduction scheme is proposed, for voice correction. Typical quantities from chaos theory and some conventional ones are evaluated, in order to provide entries for feature vectors in a feature space. Geometric signal separation is applied for voice classification, by means of a properly defined 'healthy index'. This allows moving the feature vector of pathological voices from the sick region to the healthy one, in order to enhance voice quality. The practical advantage is twofold: first, physicians are provided with a better understanding of voice dysfunctions for surgical and rehabilitation purposes, second non-invasive devices for voice denoising could be build up as an aid for dysphonic people.

1. INTRODUCTION

A variety of complex signals are produced by the voicing source in addition to periodic vibrations, as many sources of non-linearity are involved in the air-flow production and in the laryngeal vibration process [1]-[3]. For pathological voices, it was shown that bifurcations occur, as well as sudden jumps from one limit cycle to another one, with different period and amplitude [4]. Such situation should be reflected in quantities like entropy and fractal dimension of the attractor. Although normal phonation and voice disorders can be easily distinguished qualitatively, a quantification and data distribution of diseases is highly desirable. The geometric signal separation in a feature space is used here, due to its capability of identifying the dynamical state of complex systems, when the system itself is connected to a measurement device which records a scalar time series [5]. The choice of the entries of feature vectors takes into account the main speech signal characteristics, such as: jitter, shimmer, transition across phonemes, correlation dimension, autocorrelation, etc., according to previous results [6]-[8]. Neighbourhood relations in the feature space allows for classification and diagnosis tasks. Denoising is obtained as in [6], [15], taking into account the results in [7], [8], [13] as far as the choice of the window length is concerned. The proposed filter arises from chaotic deterministic systems field, the determinism yielding a criterion to distinguish the clean signal and the noise component. The effectiveness of the proposed procedure is successfully applied to real data, coming form healthy subjects and cordectomised patients.

2. GEOMETRIC SIGNAL SEPARATION

Given a recorded scalar time series, sufficiently long subsections of this series are transformed into feature vectors in a feature space \( V \). The entries of \( v \) contain the compressed information of the signal, relevant for the classification task to be performed, which can be estimated directly from the time series. In practice, healthy patients and sick ones should be associated to feature vectors that populate different regions of the feature space. It was shown in [2], [4] that the fractal dimension of the attractor is a good entry of the feature space, since healthy people should produce smaller values than pathological ones. The approach proposed in [5] is applied here, as computationally convenient when a pseudo-dimension \( D_M \) is considered, given by:

\[
D_M = 2\log_{10}(N - 2)
\]

with \( N \) = data length. Hence, about \( D_M = 8 \) is the maximum dimension for a sentence lasting 1s, with sampling frequency \( F_s = 22 \text{kHz} \), as in the present application. Another quantity useful for our purposes is the signal entropy, which essentially describes the amount of disorder in the system [5],[7]. Due to numerical problems involved with entropy computation (finite length of the time series and presence of noise), a pseudo-entropy \( h_2 \) is used here, given by:

\[
\boxed{h_2 = \sup \lim_{m \to \infty} \left( H_2(m+1, \varepsilon) - H_2(m, \varepsilon) \right)}
\]

In eq.(2), \( H_2(m, \varepsilon) = -\ln \sum_{i=1}^{m} p_{i_{1},i_{2},...,i_{m}}^{2} \) is the block entropy of block size \( m \) and partition radius \( \varepsilon \). \( p_{i_{1},i_{2},...,i_{m}}^{2} \) is the joint probability that at an arbitrary time \( n \) the observable falls into the interval \( I_{i_1} \), at time \( (n+1) \) falls into the interval \( I_{i_2} \) and so on. Both \( D_M \) and \( h_2 \) are averaged for \( \varepsilon \)-values of 5% to 10% of the variance of the data for embedding dimensions ranging between 2 and 8, the upper limit being suggested by eq.(1). Other quantities in the feature vector are:

- the spectral factor (SF) defined as the averaged ratio between the amplitude of frequencies under 1kHz and
frequencies between 4 and 6 kHz. In fact, sick patients have
to face a larger effort in speaking with respect to healthy
people, and hence a larger SF.

- the first zero-crossing of the autocorrelation function (ZC), as sick people produce more correlated data due to their
difficulty in correctly pronouncing a word.
- the first Lyapounov exponent, giving the average
exponential rate of divergence of infinitesimally nearby
initial conditions. Jumps from one limit cycle to another one
are almost typical of sick voices.
- Jitter, i.e. the short-term (cycle-to-cycle) variation of the
fundamental frequency, which is lower for healthy people
(<1%).
- Shimmer, i.e. the short-term variation of the signal
amplitude, which assumes lower values for healthy people.
- Peaks in the phoneme transition: phoneme transition shows a
much longer transient for sick people than for healthy ones,
due to difficulties in switching from a dynamical regime to the
following. Hence, this parameter is larger for sick
subjects. This quantity was shown to be comparable to the
harmonics-to-noise-ratio (HNR), but more reliable [6], [7].
- Residual noise, evaluated with the filtering procedure to be
described below. The variance of the difference between the
original and the filtered signal is considered. Sick people
show larger values than healthy ones.

The feature vector is clearly redundant, its entries being more or
less correlated, but this is of help in distinguishing between
healthy and pathological voices. In order to quantify the degree of
illness, the centre of mass of the healthy and the pathological
clusters is computed (ch and cp respectively). Then, the distances
d(new, ch) and d(new, cp) of the voice sample under test (named
‘new’) are evaluated. Due to the different range covered by
different components of the feature vector, a weighted distance is
considered, the weights being the inverse of the standard
deviation of the distribution of the entries. The ‘healthy index’ H
is defined as follows:

\[ H(\text{new}) = 20 \log_{10} \frac{d(\text{new}, c_h)}{d(\text{new}, c_p)} \]  

(3)

Healthy voices are characterised by strong negative values, while
large positive values are typical of pathological voices. More
details can be found in [14].

3. NON-LINEAR NOISE REDUCTION

The first step consists in phase space reconstruction, i.e.
converting the observations into state vectors. This problem can
be solved by the method of delays, or related constructions [8]-
[10]. A delay reconstruction in m dimensions is made up by the
vectors \( b_n=(s_{n-m+1},\ldots,s_n) \), where \( s_n \) is the nth
signal sample. The time difference in number of samples \( \nu \) or in
time units VM between adjacent components of the delay vectors is
referred to as the lag or delay time. Human voice is quasi-
periodic and almost non-stationary. A delay reconstruction should
allow identifying the actual phoneme. This amounts to finding at
least the two parameters \( m \) and \( \nu \) in the equation for \( b_n \). To
this aim, recurrence plots are constructed on the basis of mutual
distances between points belonging to the same trajectory [11],
[12]: a dot in the plane of indices \( i \) and \( j \) means that \( |b_i-b_j|<\epsilon \). The
correct choice of \( \epsilon \) is of great importance: it must be larger than
the noise level, but such that the lines in the recurrence plot are as
long and as thin as possible and all the recurrences belong to lines.
The distance between two consecutive lines is the pitch
duration inside the phoneme. As a consequence, the product \( VM \)
must be larger than this value. Notice that the larger the \( \nu \) value
the worse the results, and a large value of \( m \) implies large
computational effort. Refer to [13] for more details.

In [1]-[7] it is shown how the complexities observed in
disordered voices are not caused by random external input to the
vocal apparatus, but by the intrinsic non-linear dynamics of vocal
fold movement. Normal phonation corresponds to an essentially
synchronised motion of all vibratory modes. A change of
parameters such as muscle tension or localised vocal fold lesions
may lead to de-synchronisation of certain modes, resulting in the
appearance of new features that look like noise. It is assumed that
the signal is corrupted by additive white noise: \( s_n=x_n+\eta_n \), where
\( x_n \) is the noisy signal, \( x_n \) is the clean signal, and \( \eta_n \) is the
superimposed noise. Denoising is performed by applying the
filter proposed in [6], [15], that arises from the chaotic
deterministic systems field. Delay vectors \( s_n=(s_{n-1},\ldots,s_0) \)
are formed and those close to \( s_n \) are found, for any \( n_0 \). The average
value of \( s_{n_0} \) is then used as the cleaned value \( \hat{x}_{n_0-m/2} \), which
is an estimate of \( x_{n_0-m/2} \). Hence:

\[ \hat{x}_{n_0-m/2} = \frac{1}{U(\epsilon(s_{n_0}))} \sum_{\epsilon \in U(\epsilon(s_{n_0}))} s_{n_0-m/2} \]  

(4)

where \( U(\epsilon(s_{n_0})) \) denotes the number of elements of the
neighbourhood \( U(\epsilon(s_{n_0})) \) of radius \( \epsilon \) around the point \( s_{n_0} \). The
filter is worth applying to correct voice disorders, by suitably
choosing the involved parameters. Each point belonging to a
structure is replaced by a local average of similar points coming
from different structures (this is done by searching for
neighbourhoods in the embedding space). As already pointed out,
the choice of \( \epsilon \) is critical: too large a value would result in a
drastic averaging and the resulting signal would sound artificial.
On the other hand, if \( \epsilon \) is too small, no correction can be
performed, as no recurrence is obtained. Tuning the parameters
according to the indications above, the filtered signal sounds
clearer than the original one.

4. EXPERIMENTAL RESULTS

Voice samples come from healthy people (17 samples),
dysphonic subjects (affected by T1A glottis cancer, 12 samples),
and subjects that underwent surgical treatment (endoscopic laser
or traditional lancing, 4 samples). Voice samples were recorded in
a quiet room, Phoniatic Section, Otolaryngology Dept., Careggi
Hospital, Univ. of Firenze, Italy. Patients were asked to
pronounce the Italian word 'aiuole' (flower-beds), made up by
the five main Italian vowels. The clinical interest was that of
evaluating the effort made by patients during the entire utterance,
and hence a larger SF. Drastic averaging and the resulting signal would sound artificial.
On the other hand, if \( \epsilon \) is too small, no correction can be
performed, as no recurrence is obtained. Tuning the parameters
according to the indications above, the filtered signal sounds
clearer than the original one.
the three data sets under investigation. Fig. 1 shows the projection of the feature space onto these two entries. As expected, for healthy voices $D_M \leq 3$, while $D_M > 5$ for dysphonic patients. Similarly, $ZC \geq 1.2\text{ms}$ for pathological voices, as dysphonic subjects are not capable of isolating each vowel well, thus producing more correlated time series with respect to healthy subjects ($0.2\text{ms} \leq ZC \leq 0.4\text{ms}$). Cordectomised patients lie between the healthy and the dysphonic ones, with small $D_M$ values, but large $ZC$ values. This result, although preliminary, shows that cordectomised subjects still require larger effort in speaking with respect to healthy ones, due to the limited extension of the produced scar fold. Similar results were obtained with entropy and other entries of the feature vector, as pathological voices produce more disordered series, and larger prediction error.

As for classification, the healthy index $H$ in eq. 3 was evaluated for the full data set. Fig. 2 shows the results concerning healthy and dysphonic voices only. The peak on the left hand side of the plot is relative to healthy voices, while the one on the right side concerns dysphonic voices. The two sets are thus perfectly separated. For cordectomised patients, it was found that $-7 \leq H \leq 4$. Notice that $H$ is based on all the entries in the feature vector. Reducing the number of parameters causes a smoothed distribution of the results, and hence worse classification.

In fig. 5, the projection of the feature space onto $h_2$ and SF is reported. Healthy and pathological voices populate different regions, and voices coming from patients that underwent surgical treatment correctly lie almost in the same region as healthy ones. These results were obtained with $m=30$, $\nu=4$ and $\varepsilon=0.3\sigma$, $\sigma$ being the data variance. Dysphonic voices spread around the average value ($5, 1.8$). The centre of mass of the healthy cluster is ($35, 0.5$). Surgical treatment produces an average correction located in

Figure 1: 2-D projection of the feature space onto $D_M$ and $ZC$.

Figure 2: Distribution of healthy (left) and pathological (right) voices according to the healthy index $H$.

Figure 3: Recurrence plot of a healthy voice. The distance between adjacent lines corresponds to the pitch period (number of points).

Figure 4: Recurrence plot with the non-linear filtering procedure for denoising described above, applied to a dysphonic voice. The main shape of sub-structures is preserved, while the noise-like features are smoothed. The four structures (each lasting about 80 points) are neighbourhoods, as the corresponding sequences of points are similar. The filtered signal sounds less hoarse than the original one.
The proposed non-linear noise reduction procedure moves the centre of mass of hoarse voices to (15, 0.7), i.e. produces an effect similar to that obtained with the surgical treatment.

**Figure 4:** Correction performed by the non-linear noise reduction procedure applied to a dysphonic voice. Dotted line: original signal. Solid line: filtered signal.

**Figure 5:** 2-D projection of the feature space onto $h_2$ and SF. Dotted lines link points before and after the correction obtained with non-linear filtering.

Notice that the algorithm is sensitive to the choice of $\epsilon$. The length of the dotted lines in fig. 5 is somewhat proportional to $\epsilon$, but does not evolve along a straight line. Hence, for too large values of $\epsilon$, the averaging procedure described in eq.4 would destroy the original signal almost completely, producing a pseudo-entropy and a spectral factor close to zero.

5. **CONCLUSIONS**

This paper presents a method for voice classification, based on the construction of an appropriate feature space. The entries of the feature space are chosen on the basis of physicians' suggestion, and according to linear and non-linear dynamical system theory. A proper choice of entries shows that a net separation between healthy and pathological voices can be obtained, also by means of a properly defined healthy index. Moreover, a procedure is proposed for noise reduction, based on recurrence plots and the previously defined feature space.

The approach could be of clinical interest, also as far as post-surgical rehabilitation is concerned. Glottal functionality can in fact be analysed by means of objective indexes other than visual inspection. Finally, the denoising procedure could be implemented in a portable device, in order to help disphonic people in reducing effort required in speaking.

6. **REFERENCES**