

NONLINEAR, BIOPHYSICALLY-INFORMED SPEECH PATHOLOGY DETECTION

Max Little^{*ab}, Patrick McSharry^{ab}, Irene Moroz^a and Stephen Roberts^b

^(a)Mathematical Institute, ^(b)Engineering Science, Oxford University, UK

ABSTRACT

This paper reports a simple nonlinear approach to online acoustic speech pathology detection for automatic screening purposes. Straightforward linear preprocessing followed by two nonlinear measures, based parsimoniously upon the biophysics of speech production, combined with subsequent linear classification, achieves an overall normal/pathological detection performance of 91.4%, and over 99% with rejection of 15% ambiguous cases. This compares favourably with more complex, computationally intensive methods based on a large number of linear and other measures. This demonstrates that nonlinear approaches to speech pathology detection, informed by biophysics, can be both simple and robust, and are amenable to implementation as online algorithms.

1. INTRODUCTION

The linear source-filter theory of voice production states that the *vocal folds* oscillate during voiced speech, driving the *vocal tract* into resonance at specific frequencies. These acoustic pressure waves, radiating from the lips, are modelled as a further filtering of the flow rate signal. By spectral deconvolution and post-filtering of the sampled speech pressure signal, $p(n)$, it is possible to approximate the aerodynamic flow rate signal $u(n)$ at the top of the vocal folds [1].

However, studies of sustained vowel speech signals show that nonlinearity and turbulence are important features [2], even of normal speech [3]. Modelling and experimentation show that turbulent airflow and nonlinearity in the vocal fold dynamics are important, and that vocal pathologies often lead to apparently chaotic [4] and turbulent behaviour, manifesting as increased “hoarseness” [5]. For some types of vocal pathology, oscillation ceases altogether, or the oscillation is intermittent. The speech signal is then dominated by turbulent airflow noise.

Commonly used by speech pathologists are acoustic tools, recording acoustic pressure at the lips or inside the vocal tract. These tools [6], amongst others, can provide potentially objective measures of voice function, and could be used for automatic screening for vocal pathologies. However, there are many differing approaches to automatic screening. Most of these are based around a large set of measures, using many linear processing techniques such as the discrete Fourier transform, variance, autocorrelation, and other techniques such as peak-picking, and speech pitch detection [7].

These measures are then passed on to a neural network which classifies the speech signal as normal or pathological [5, 8]. However, such approaches are inherently complex, leading to two significant problems. The first is the risk of being *too flexible* and hence highly optimised to the training data [9], limiting generalisation performance on new data. The second problem is that such approaches require large computational resources, prohibiting their use in portable devices.

The origin of these problems is that speech signals exhibit very complex dynamics, which cannot, *prima facie*, be characterised by linear methods such as spectral analysis. Hence there is a need for a simpler treatment of the problem, grounded in more realistic biophysics of speech production, which does not transfer the complexity of the detection task into later post-processing.

2. METHODS

2.1. Method Design

While nonlinear signal processing methods can solve some of the problems highlighted above, they also have drawbacks. The most serious of these are:

- arbitrary algorithmic parameters whose value propagates to the results, making the validity and reproducibility of the method uncertain,
- sophistication that is unjustified for the problem, especially when simpler linear methods outperform [10],
- inappropriate application which without prior information on the dynamics often gives misleading results,
- much higher sensitivity to noise and other uncontrolled environmental factors than linear methods,
- failure to scale up computationally, quickly becoming infeasible, despite nonlinear effects often becoming apparent only on large data sets.

Nonlinear algorithms can be of value if they can overcome these pitfalls. Therefore, the current method makes use of non-parametric techniques or those that use as few algorithmic parameters as possible, using the simplest algorithm where possible. The choice of algorithms incorporates knowledge of the biophysics of speech production, and it is designed to work with signals collected in known, controlled circumstances.

2.2. Return Period Density Entropy (RPDE)

The return period density entropy is a simple method, which under certain assumptions, can be used to distinguish types of (non-transient) complex dynamical behaviour represented in a signal which might possibly be contaminated by highly correlated noise. It is based upon the theory of continuous dynamical systems [11], and makes use of *Poincaré sections* of the sampled trajectory $x(n)$ for time indices $n = 1, 2, 3 \dots$ of a differential flow which have been *time-delay embedded*, following Taken’s embedding theorem [12]. Here the time-delay embedding is implicit in finding the *numerical maxima return series* $y(i)$ for intersection sequence numbers $i = 1, 2, 3 \dots$ of the signal $x(n)$. This signal represents sampled experimental observations of an underlying differential flow [11]. The maxima operation was chosen as it is computationally simple yet

*Funded by the EPSRC, UK.

robust. Based upon analysis of the time index differences $d(i) = n_{i+1} - n_i$ between these maxima returns, it is possible to diagnose the behaviour of the underlying dynamics, as follows.

Given an initial condition for the dynamical system, the resulting unique *trajectory* in the state-space can display many different kinds of behaviour. Of interest in this context are *periodic*, *period-doubled*, and *chaotic* behaviours [11]. It is assumed that the maxima operation produces an appropriate Poincaré section such that for periodic orbits, the trajectory intersects the section at the same point once per cycle, leading to a constant sequence of time index differences $d(i)$. Similarly, successive period-doubled orbits of length 2^k , $k = 2, 3, \dots$ will intersect the section at 2^k places, causing the differences $d(i)$ to form a sequence of length 2^k . For (hyperbolic) chaos, the trajectory intersects the section in a structured set of points that is often complicated and fractal [12]. In this case the time index differences will consist of subsequences of lengths 2^k for all k . Finally, for highly correlated noise, the series $d(i)$ will contain a random sequence of values (note that correlation is required to avoid finding spurious, noisy maxima).

Considering periodic to be the simplest possible type of behaviour, and chaotic the most complex, a measure of the degree of complexity is the entropy H of a *period density function* $f(d)$ constructed by normalising a histogram of the series $d(i)$:

$$H = - \sum_{d=1}^N f(d) \ln f(d) \quad (1)$$

where N is the maximum time index difference found in the series $d(i)$.

When $H = 0$, there is only one non-zero value of $f(d)$ indicating a periodic orbit. An increase in H indicates an increase in different periods d represented in the trajectory, or equivalently, noise-induced random variations in the values of $d(i)$. For an additive mixture of deterministic dynamics with correlated noise, the noise decreases the sharpness of the period density function $f(d)$ peaks with increasing noise variance. These changes in $f(d)$ propagate to the entropy H , thus noise of increasing variance increases the measured complexity of the signal $x(n)$.

Summing up, using H , it is possible to rank a set of signals on a scale of both deterministic and stochastic complexity. It may, however, not be possible to distinguish, from a finite duration signal $x(n)$ whether the dynamics is period-doubled with large k value, chaotic, or deterministic with high variance additive noise, and so an absolute scale of complexity is usually unobtainable in practice. Nevertheless, it is possible to rank a certain data set of signals of finite length, which is adequate for the current purposes. For a more detailed derivation of this technique, please see [13].

2.3. Detrended Fluctuation Analysis (DFA)

Detrended fluctuation analysis is a straightforward technique for identifying the extent of *fractal self-similarity* in a signal [14]. It is designed to calculate the scaling exponent α in nonstationary time series (where the statistics such as mean, variance and autocorrelation properties change with time).

First, the time series $x(n)$ is integrated:

$$y(n) = \sum_{j=1}^n x(j) \quad (2)$$

so that, for example, assuming $x(n)$ is independent and identically distributed, then $y(n)$ is a self-similar random walk. Then, $y(n)$ is

successively subdivided into windows of length L samples. For a times series of length M samples there will be the nearest integer to $\log_2 M$ scales. A least-squares straight line *local trend* is calculated by analytically minimising the squared error E^2 over the slope and intercept parameters a and b :

$$\arg \min_{a,b} E^2 = \sum_{n=1}^L (y(n) - an - b)^2 \quad (3)$$

Next, the root-mean-square deviation from the trend, or *fluctuation* is calculated over every window at every time scale:

$$F(L) = \left[\frac{1}{L} \sum_{n=1}^L (y(n) - an - b)^2 \right]^{1/2} \quad (4)$$

This process is repeated over all subdivisions of all lengths L . On a log-log graph of L against $F(L)$, a straight line indicates self-similarity expressed as $F(L) \propto L^\alpha$. The scaling exponent α is calculated as the slope of a straight line fit to the log-log graph of L against $F(L)$ using least-squares as above. An efficient algorithm is used here which shares the summation terms in equation (3) over all time scales. For a more in-depth presentation and discussion of self-similarity in time series in general, please see [12].

2.4. Gaussian Linear Discriminant Analysis (LDA)

For the purpose of discriminating between the two classes of normal and pathological cases, Gaussian linear discriminant analysis is a simple technique that allows *linear separation* by modelling the data conditional upon each class using joint Gaussian probability densities [15]. For a $J \times K$ *data matrix* $\mathbf{v} = v_{jk}$ of observation (measure) j and case k , these likelihood densities are parameterised by the means and *covariance matrix* of the data set:

$$\mu = E[\mathbf{v}], \mathbf{C} = E[(\mathbf{v} - \mu)(\mathbf{v} - \mu)^T] \quad (5)$$

where E is the expectation operator, and μ is the mean vector formed from the means of each row of \mathbf{v} . The class likelihoods are:

$$f_C(\mathbf{w} | C_i) = (2\pi)^{-J/2} |\mathbf{C}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{w} - \mu_i)^T \mathbf{C}^{-1} (\mathbf{w} - \mu_i) \right] \quad (6)$$

for classes $i = 1, 2$ and an arbitrary observation vector \mathbf{w} . It can be shown that, given this Gaussian class model, the maximum likelihood regions of the observation space \mathbb{R}^J are separated by a *decision boundary* which is a (hyper-)plane calculated from the difference of log-likelihoods for each class, which is the unique set of points where each class is equally likely [15]. The maximum likelihood classification problem is then solved using the decision rule that $l(\mathbf{w}) \geq 0$ assigns \mathbf{w} to class C_1 , and $l(\mathbf{w}) < 0$ assigns it to class C_2 , where:

$$l(\mathbf{w}) = \mathbf{a}^T \mathbf{w} - \theta$$

$$\mathbf{a} = \mathbf{C}^{-1} (\mu_1 - \mu_2), \theta = \frac{1}{2} \left(\mu_1^T \mathbf{C}^{-1} \mu_1 - \mu_2^T \mathbf{C}^{-1} \mu_2 \right)$$

In order to avoid overfitting, the *generalisation performance* of the classifier can be tested using *bootstrap resampling* [16]. The classifier is trained on K cases selected at random with replacement from the original data set of K cases. This trial resampling processes is repeated many times and the mean classification parameters $E[\mathbf{a}]$, $E[\theta]$ are selected as the parameters that would achieve the best performance on entirely novel data sets.

2.5. Automatically Detecting Vocal Pathologies

The formant spectrum $V(z)$ of the speech signal $p(n)$ is identified using linear prediction analysis (LPA) [17], and then removed using inverse filtering, i.e. by applying the filter $V(z)^{-1}$ to $p(n)$, to leave a residual signal $w(n)$. Subsequent application of a radiation impedance filter obtains an estimate for the vocal fold flow rate signal $u(n)$. This is a low-pass filter having unit magnitude response at zero frequency, i.e. $R(1) = 1$, with transfer function:

$$R(z) = (1 - r)^2 (1 - rz^{-1})^{-2} \quad (7)$$

Parameter r controls the resonance of the single pole. Using the zero-phase technique, i.e. applying the filter forwards over the signal, then reversing the output, and applying the filter once again cancels any phase delay [17]. In this paper, $r = 0.97$ was chosen to provide the best identification of $u(n)$. Next, the RPDE algorithm calculates the maxima series $y(i)$ of $u(n)$ and the period density function $f(d)$, from which the value of H is obtained. The DFA algorithm calculates α from the original, unprocessed speech signal $p(n)$. For classifier training, the data matrix \mathbf{v} receives K random selections of the H_k and α_k for all subjects k and the mean classification parameters are calculated over 1000 such selections. Subsequently, $l(\mathbf{w}_k) = [H_k, \alpha_k]^T$ gives the classification performance for cases correctly classified as normal (C_1 , *true negative*) and pathological (C_2 , *true positive*). The signals $p(n)$ were of length $M = 18000$, the LPA analysis order was $P = 15$ over 1000 samples, and 1000 samples were skipped at the beginning and the end after the low-pass filter to skip filter transients. The overall process is depicted in Fig. 1.

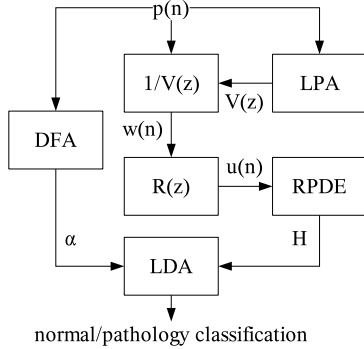


Fig. 1. Overall process of speech pathology detection.

Overall then, any disorder of the vocal folds that causes more complex oscillations will be detected as a (relative) increase in H , and changes in turbulent airflow will be detected as a relative change in α . To ensure good control of environmental conditions, this method makes use of sustained vowels.

3. DATA

Used in this study are sustained vowel phonation samples from $K = 707$ subjects from the Kay Elemetrics Disordered Voice Database [18], 53 of which are from normal controls. This represents a wide variety of organic, neurological and traumatic voice disorders. Each sample was recorded under controlled acoustic conditions, and is on average around two seconds long, 16 bit uncompressed PCM. The normal control samples were recorded at 50kHz and then downsampled with anti-aliasing to 25kHz.

4. RESULTS AND CONCLUSIONS

Fig. 2 shows the results of the measures applied to a typical normal and pathological case. Fig. 3 plots the measures calculated over all the 707 cases, and shows the average decision boundary calculated over 1000 bootstrap trials. Finally, Fig. 4 shows the convergence of the classification performance for true positives, true negatives, and overall, as the number of bootstrap trials increases.

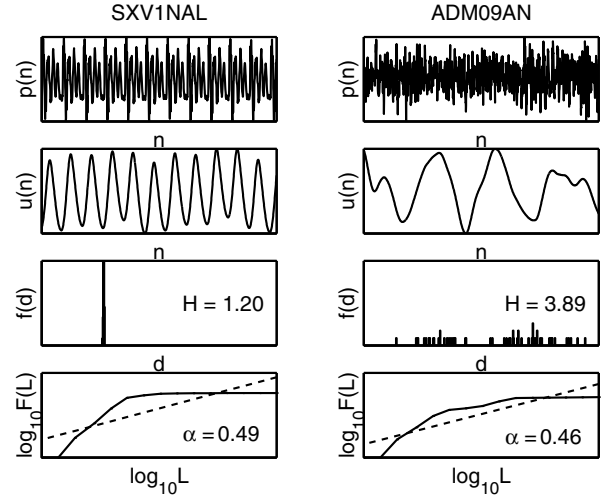


Fig. 2. Left column: normal case, right column: pathological case. Rows from the top: Speech pressure signals $p(n)$, vocal fold flow rate signals $u(n)$, return period densities $f(d)$ with entropy H , and log-log plot showing variation of fluctuations $F(L)$ with L with best-fit line and scaling exponent α .

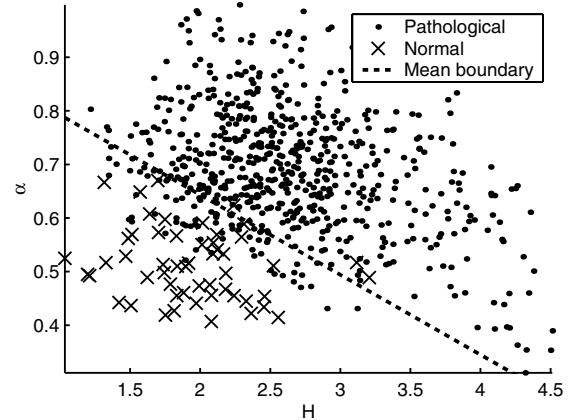


Fig. 3. Return period density entropy H against detrended fluctuation analysis α for all 707 subjects, showing mean linear discriminant analysis classification boundary calculated over 1000 bootstrap resampling trials.

Although the normal model for the probability density of these classification results is not exact, leading to some inconsistencies (such as some trials having greater than 100% performance), the normal density leads to 95% confidence intervals of $94.3 \pm 6.3\%$ true

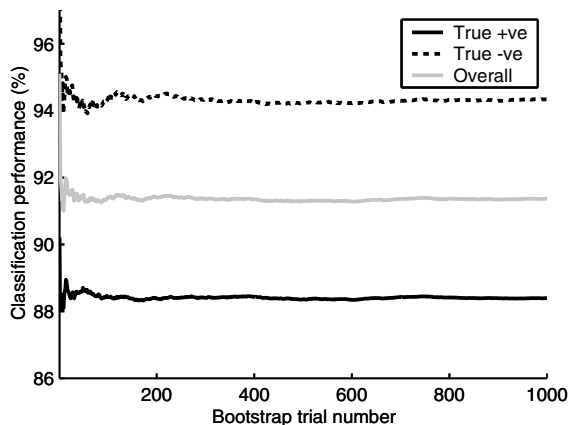


Fig. 4. Classification performance convergence.

negative, $88.4 \pm 3.1\%$ true positive, and overall $91.4 \pm 3.9\%$ performance. This result compares favourably with other approaches which achieve overall rates of 93.5% [5] and 92.8% [19], but less so with another study reporting an overall performance of 98.2% by combining both standard and nonlinear approaches [8].

However, a hard decision boundary leads to unclassifiable, ambiguous cases, and Bayesian modelling obtains the *posterior probability* of correct classification. A parameter that affects the size of the ambiguity *reject region* may be integrated out to obtain an area under the *receiver operating characteristic curve* [15]. Experiments using an unsmoothed, Bayes optimal logistic classifier with moderation [15] revealed an area of 92.5%. LDA was favoured, and rejecting 15% of ambiguous cases lead to over 99% overall performance.

Comparing likely generalisation performance, this method has only five arbitrary parameters: the number of analysis samples N , LPA window length, LPA analysis order, low-pass filter transient skip length and parameter r . This is by far the best of the other studies [5, 8], which both mention explicitly at least 20 arbitrary parameters. These studies also rely on other methods, which themselves contain more arbitrary parameters. None of the other studies report confidence intervals, and all are significantly more computationally expensive and trained on far fewer patients (at most 400 by comparison to 707 for this study). For these reasons it is hard to have confidence in their generalisation performance on novel test data, and their utility in mass screening or portable applications.

Not all cases show self-similarity over the whole range of available scales (see the scaling curves in Fig. 2). Nonetheless, the mean and variance of α differs significantly between normal and pathological cases. Improvements to this method would resolve the choice of the arbitrary parameters, perhaps using model-based *surrogate data methods* [12].

5. REFERENCES

- [1] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least-squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans Acoust Speech Proc*, vol. 27, no. 4, pp. 350–355, 1979.
- [2] S. McLaughlin and P. Maragos, "Nonlinear methods for speech analysis and synthesis," in *Advances in nonlinear signal and image processing*, S. Marshall and G. Sicuranza, Eds., EURASIP Book Series on Signal Processing and Communications. Hindawi, 2006, to appear.
- [3] M. Little, P. McSharry, I. Moroz, and S. Roberts, "Testing the assumptions of linear prediction analysis in normal vowels," *J Acoust Soc Am*, 2005, accepted for publication.
- [4] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *J Acoust Soc Am*, vol. 97, no. 3, pp. 1874–1884, 1995.
- [5] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices," *IEEE Eng Med Biol Mag*, vol. 16, no. 4, pp. 74–82, 1997.
- [6] P. H. Dejonckere, "Perceptual and laboratory assessment of dysphonia," *Otol Clin North Am*, vol. 33, no. 4, pp. 731–, 2000.
- [7] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., pp. 495–518. Elsevier, Amsterdam; New York, 1995.
- [8] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer, "Automatic detection of pathologies in the voice by hos based parameters," *EURASIP J Appl Sig Proc*, vol. 4, pp. 275–284, 2001.
- [9] D. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, Cambridge, UK ; New York, 2003.
- [10] P. E. McSharry, L. A. Smith, and L. Tarassenko, "Prediction of epileptic seizures: are nonlinear methods relevant?," *Nat Med*, vol. 9, no. 3, pp. 241–2; author reply 242, 2003.
- [11] J. Guckenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, Springer-Verlag, New York, 1983.
- [12] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press, Cambridge; New York, 1997.
- [13] M. Little, P. McSharry, I. Moroz, and S. Roberts, "Stroboscopic method for detecting complex dynamics in disordered speech," *IEEE Trans Biomed Eng*, 2005, submitted.
- [14] C. K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time-series," *Chaos*, vol. 5, no. 1, pp. 82–87, 1995.
- [15] C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press ; Oxford University Press, Oxford New York, 1995.
- [16] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, Chapman and Hall, New York, 1993.
- [17] J. Proakis and D. Manolakis, *Digital signal processing: principles, algorithms, and applications*, Prentice Hall, Upper Saddle River, N.J., 3rd edition, 1996.
- [18] KayPENTAX, "Kay elemetrics disordered voice database, model 4337," 1996–2005.
- [19] J. B. Alonso, F. Díaz-de María, C. M. Travieso, and M. A. Ferrer, "Using nonlinear features for voice disorder detection," in *Proceedings of the 3rd International Conference on Non-Linear Speech Processing*, Barcelona, Spain, 2005, pp. 94–106.