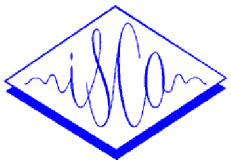

Models and Analysis of Vocal
Emissions for Biomedical Applications
(MAVEBA 1999)
Firenze, Italy, September 1-3, 1999

STOCHASTIC APPROACH TO VOCAL FOLD PATHOLOGY DIAGNOSTICS

Eugene I. Bovbel, Mikhail A. Toumilovich
Belarusian State University, Radiophysics Dept.
4, F. Skoryny av., Minsk, 220050, Belarus
Tel: +375 172 771 007, Fax: +375 172 770 890
E-mail: mike@rfe.bsu.unibel.by

Abstract. In this paper we consider a feature estimation approach for vocal fold pathology classification, based on digital signal processing theory. This problem is addressed by formulating a stochastic maximum likelihood (ML) estimation procedure, based on Estimation-Maximization (EM) algorithm. New spectral parameters of speech, noted as Spectral Pathology Component (SPC) is estimated. For classification purposes, the counterpropagation neural network (CNN) was proposed. A set of log Mel-frequency filter bank coefficients were used to parametrize the SPC spectral feature. An evaluation of CNN based classifier were performed using speech recording from healthy and pathology patients.

I. INTRODUCTION

It is well known that voice quality plays an important role for many peoples. The presence of laryngeal pathology can significantly affect on physical and emotion state of speaker. The main tool, used to detect laryngeal pathology is a subjective evaluation of patient voice, performed by a physician. It is necessary to note that the subjective evaluation of the human voice depends on the physician's experience and might not be capable to detect the presence of pathology in an earlier stage. Besides, this approach causes pain and discomfort to the patient. This reason stimulates efforts in the developing of voice analysis systems for speech pathology diagnostics with invasive procedure.

The method for vocal fold cancer detection from speech, based on digital signal processing methods was proposed in [1]. This problem is addressed using an iterative maximum likelihood procedure, based on EM algorithm [2]. The EM algorithm is a general method for ML estimation for so-called "incomplete" data problems. Because of its asymptotic optimal properties EM algorithm become one of the preferred methods of choice for maximum likelihood estimation. In this paper we derive approach, proposed in [1] and the modification of this method is used for the detection of vocal fold nodules, dysphonia, chorditis and vocal fold paralysis.

II. MODELS OF SPEECH SIGNALS

The voiced segment of speech $s(t)$ can be modeled as the response of a linear quasistationary system generated by an input train of pulses $p(t)$ convolved with the vocal tract response $h(t)$ and radiation at the lips $r(t)$:

$$s(t) = p(t) * h(t) * r(t) \qquad (1)$$

This is a general model, used for representing of healthy speech production.

For the model of speech production under pathology some assumptions are made. First, the factor causing pathology is stationary during speech production and reflected in glottal excitation. Second, the analysis has to be limited short frames, so that the signal parameters are constant. It is also assumed that the vocal tract frequency response and the radiation at the lips are derived from the healthy conditions. The model for the healthy vocal tract $h(t)$ and the radiation $r(t)$ is determined by applying inverse filtering techniques to the healthy speech signal. If this knowledge is based on another healthy speaker, vocal tract mismatch is occurring. In order to address this vocal tract mismatch a modeled error $n(t)$ is added:

$$s'(t) = \left\{ (p(t) * h'(t)) + n(t) \right\} * h(t) * r(t) \qquad (2)$$

Here, $h'(t)$ models the vocal fold pathology component. It is evident, that if the speaker under both pathology and healthy condition is the same, then modeled noise will be small.

III. ML PROBLEM FORMULATION FOR PATHOLOGY SPEECH

In the following we address ML estimation for stochastic signal model. The comprehensive studies of maximum likelihood related methods can be found in [3].

Generally, knowing the asymptotic distribution properties of some data vector $\bar{S}$, the likelihood function can be maximized over the model parameters collected into vector $\bar{\theta}$, taking its values in Θ:

$$f(X,\hat{\theta}) = \max_{\theta\in\Theta} \max f(X,\bar{\theta})$$

Under certain regularity conditions with respect probability function $f(X,\bar{\theta})$ and parameter space Θ, the resulting ML estimation will have desirable asymptotic properties.

A commonly used and physically reasonable model for healthy speech signal is an autoregressive (AR) model, driven by white noise with variance σ^2. This model represents the contribution of the excitation component vocal tract and the radiation at the lips. In order to obtain the parameters of inverse AR filter, a linear prediction coding (LPC) analysis of order p is performed for the healthy speech signal [4]. After applying the inverse AR filter to both speech signals (healthy and pathology), we receive:

$$\begin{aligned} y(t) &= p(t) \\ y'(t) &= p'(t) + n(t) \end{aligned}, \qquad (3)$$

Where $p'(t)$ represents pure pathological component. Therefore, the goal is to separate the pathological component from the modeling error and estimate pathology features. To achieve this, a ML problem is formulated in order to estimate the parameters of speech pathology component.

To formulate a statistical ML problem we have to make the following assumptions. As already noted, the frame length of observation must be short enough so that parameters are constant, but it is also long enough so, that Fourier coefficients of signals $y(t), y'(t)$ at different frequencies are uncorrelated. From the statistical point of view, the signal $p'(t)$ is considered as a Gaussian random AR process. Under this assumption, the likelihood of the data vector $\bar{S}(y(t), y'(t), \mathrm{t} = 0,\ldots \mathrm{N}-1)$ with respect to the parameters $\bar{\theta}$, can be expressed in frequency domain:

$$\log f_{\bar{S}}(\bar{S},\bar{\theta}) = -\sum_{\omega} \begin{bmatrix} \log\det C(\omega,\bar{\theta}) + \\ \bar{S}^{+}(\omega) C^{-1}(\omega,\bar{\theta}) \bar{S}(\omega) \end{bmatrix} \qquad (4)$$

Where $C(\omega,\bar{\theta})$ - spectral density matrix, $\bar{S}(\omega)$ is Fourier transform of vector $\bar{S}$, "+" denotes the Hermitian operator. The unknown parameters are the spectral coefficients of signal $p'(t)$ and the parameters of modeling noise $n(t)$.

The general ML problem (4) is not only complicated, but it maybe ill posed, because every value of $C(\omega,\bar{\theta})$ may correspond to set of values for the parameters. Therefore, some constraints must be imposed to the parameters. As mentioned earlier, the signal $p'(t)$, that results from the pathological component, is modeled using AR Gaussian process of order p, with gain G Thus its power spectrum $P_p(\omega)$ is:

$$P_p(\omega) = \frac{G}{\left|1 - \sum_{k=1}^{p} c_k e^{-j\omega k}\right|^2} \qquad (5)$$

Let modeling noise $n(t)$ is an output of finite impulse response (FIR) filter of order q with transfer function

$$A(z) = \sum_{k=0}^{q} a_k z^{-k} \qquad (6)$$

driven the error $w(t)$. In other words, the noise component can be expressed in the time domain as:

$$n(t) = \sum_{k=0}^{q} a_k w(t-k) + e(t) \qquad (7)$$

where $e(t)$- white Gaussian noise with zero mean and variance σ_e^2. The signal $w(t)$ should be correlated with error residual, obtained after

inverse filtering both healthy and pathology signals.

We note, that even with these assumptions, the requested maximization of the likelihood function (4) with respect to the signal and filter parameters is still complicated. Therefore, the EM algorithm will be used.

IV. EM BACKGROUND

The EM algorithm establishes a general approach to iterative computation of ML estimation when the observations can be viewed as *incomplete* data. General solution of ML estimates is:

$$\hat{\underset{ML}{\theta}} = \arg\max_{\theta\in\Theta} \log f_{\bar{S}}(\bar{S},\bar{\theta}) \qquad (5)$$

Let us assume, that the observed vector $\bar{S}$ is *incomplete* and specify some *complete* data vector $\bar{X}$ related to $\bar{S}$ by $H(\bar{X})=\bar{S}$, where $H()$ is noninvertable transformation. The main idea behind the EM algorithm is that in some problems the estimation of $\bar{\theta}$ would be easy if the *complete* data $\bar{X}$ were available, while it is difficult based on *incomplete* data $\bar{S}$ only. Since only the *incomplete* data $\bar{S}$ is available, it is not possible to perform directly the optimization of *complete* data likelihood. Instead, reasonable to "estimate" $\log f_{\bar{X}}(\bar{S},\bar{\theta})$ from $\bar{S}$ and use the "estimated" likelihood function to obtain maximized $\bar{\theta}$. Since estimating the complete data likelihood $\log f_{\bar{X}}(\bar{S},\bar{\theta})$ requires $\bar{\theta}$, it is necessary to use an iterative approach.

The algorithm iterates between estimating sufficient statistics of the *complete data* given the observation and a current estimate of the parameters (the E step), and maximizing the likelihood of the *complete data* using the estimated sufficient statistics (the M step). The E and M steps of the iterative EM algorithm can be formally expressed as:

E- step: compute

$$Q(\theta,\theta^{(k)}) = E(\log f_{\bar{x}}(\bar{X},\theta)/\bar{S},\theta^{(k)}) \quad (8)$$

M-step: choice

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} Q(\theta,\theta^{(k)}) \qquad (9)$$

The choice of *complete* data is motivated by the simple maximum likelihood solution available.

V. ALGORITHM IMPLEMENTATION

Using the results, obtained in section III, the observed signals can be written as:

$$\begin{aligned} y(t) &= v(t) \\ y'(t) &= p'(t) + \sum_{k=0}^{q} a_k v(t-k) + e(t) \end{aligned} \qquad (10)$$

The unknown parameters $\bar{\theta}$ are the coefficients of FIR filter $\{a_k\}$, spectral parameters of $p'(t)$ and the variance of white noise σ_e^2.

The likelihood function of the observed signal can be written in frequency domain:

$$L(\theta) = \sum_{\omega} \log f_{Y',Y}(Y'(\omega),Y(\omega);\theta) \quad (11)$$

where summation is made over all frequencies. For each frequency ω_k:

$$\begin{aligned} &\log f_{Y',Y}(Y'(\omega_k),Y(\omega_k);\theta) = \log f_Y(Y(\omega_k);\theta) + \\ &\log f_{Y'/Y}(Y'(\omega_k)/Y(\omega_k);\theta) \end{aligned} \qquad (12)$$

Note, that $\log f_Y(Y(\omega_k);\theta)$ is independent of θ, therefore it is assumed to be constant and did not affect on maximum of likelihood function (11). If we apply the general ML problem (4) to observed signals, that the maximizing the likelihood function (11) will be equivalent to minimizing

$$L'(\theta) = \left[\frac{\log(P_p(\omega_k)+\sigma_e^2) + \left|Y'(\omega_k) - A(\omega_k)Y(\omega_k)\right|^2}{P_p(\omega_k)+\sigma_e^2} \right] \qquad (13)$$

with respect to σ_e^2 and the coefficients $P_p(\omega)$ and $A(\omega)$.

As mentioned earlier, we will assume that desire signal $p'(t)$ is an AR process of

order p with coefficients $\{c_k\}$ and gain factor G, so that its power spectrum $P_p(\omega)$ is given by (5). Also we will assume that $A(\omega)$ is a frequency response of FIR filter of order q, i.e. it is of the form (6). Direct minimization of (13) is complicated and EM algorithm will be considered.

Let define the *complete* data as set of signals $\{p'(t), n(t), y(t)\}$. If these *complete* data are available, the maximum likelihood estimate of $\{a_k\}$ and σ_e^2 is achieved by least squares fitting $w(t)$ to $n(t)$. The spectral parameters of $p'(t)$ are estimated by solving the normal Yule-Waker equations using the sample correlation of $p'(t)$ [4].

We note, that the signals $p'(t)$ and $n(t)$ statistically independent The signals $p'(t)$ and $y(t)$ may be related in general, but this relation is arbitrary and unknown. Therefore, we will assume, that probability distribution of $p'(t)$ given $y(t)$ is a priory distribution of $p'(t)$, Thus the likelihood of complete data is:

$$\begin{aligned} L_c(\bar{\theta}) &= \log f_{p',n,y}(p'(t), y(t), n(t) : \bar{\theta}) = \\ &\log f_{p'/y}(p'(t), y(t), n(t) : \bar{\theta}) + \\ &\log f_{n/y}(n(t)/y(t), n(t) : \bar{\theta}) + \\ &\log f_y(y(t)) \end{aligned} \tag{14}$$

The $\log f_y(y(t))$ is independent of $\bar{\theta}$. The probability distribution of. $p'(t)$ is the distribution of random Gausssian AR process with power spectrum and it depends only on the spectral parameters of $p'(t)$. Thus we have:

$$\log f_{p'/y}(p'(t), y(t) : \theta) = \sum_{\omega} \left[\log P_p(\omega, \theta) + \frac{|S_p(\omega)|^2}{P_p(\omega, \theta)} \right] \tag{15}$$

and

$$\log f_{n/y}(n(t)/y(t), n(t) : \theta) = -\sum_{t=0}^{N-1} \left[\log \sigma_e^2 + \frac{1}{\sigma_e^2} (n(t) - \sum_{k=0}^{q} a_k w(t-k))^2 \right] \tag{16}$$

where $S_p(\omega)$ is the Fourier transform of $p'(t)$, i.e. $S_p(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} p'(t) \exp(-j\omega t)$

Finally, estimating θ by maximizing the likelihood of complete data (14) is equivalent to estimating the spectral parameters of $P_p(\omega)$ by minimizing (15) and estimating $\{a_k\}$ and σ_e^2 by minimizing (16).

The E and M step of the EM algorithm for minimizing (13) may now be stated explicitly. We defined the vector of parameters as $\bar{\theta} : \{c_k\}, \{a_k\}, \sigma_e^2$:

- $\{c_k\}$- the spectral parameters (which are the same as LPC parameters of AR model) that represents the pure pathological component;
- $\{a_k\}$- the coefficients of the filter, that models vocal tract mismatch between pathology and healthy speech;
- σ_e^2 - the variance of the error in the model of vocal tract mismatch.

Sufficient statistics are $n(t)$, $|S_p(\omega)|^2$, $e(t)$ and $|E(\omega)|^2$.

E step:

1) Initialize the vector of parameters $\bar{\theta}^{(0)}$ by arbitrary values.
2) Generate an intermediate signal

$$x(t) = y' - \sum_{k=0}^{q} a_k^{(n)} w(t-k)$$

where (n) refer to the estimate at the n-th iteration. If the true coefficients $\{a_k\}$ are known, the sequence $x(t)$ represents $x(t) = p'(t) + e(t)$. Therefore, the successive estimation can be obtained by applying an optimal Wiener filter.

2) Apply Wiener filter of $x(t)$ to obtain the estimation of $S_p(\omega)$, $|S_p(\omega)|^2$ and $E(\omega)$, $|E(\omega)|^2$

$$\hat{S}_p(\omega) = \frac{P_p^{(n)}(\omega)}{P_p^{(n)}(\omega) + (\sigma_e^2)^{(n)}} * X(\omega) \quad (17a)$$

$$\hat{E}(\omega) = X(\omega) - \hat{S}(\omega) \quad (17b)$$

$$E\left\{|S(\omega)|^2\right\} = \left|\hat{S}_p(\omega)\right|^2 + \frac{P_p^{(n)}(\omega) * (\sigma_e^2)^{(n)}}{P_p^{(n)}(\omega) + (\sigma_e^2)^{(n)}} \quad (17c)$$

$$E\left\{|E(\omega)|^2\right\} = \left|\hat{E}(\omega)\right|^2 + \frac{P_p^{(n)}(\omega) * (\sigma_e^2)^{(n)}}{P_p^{(n)}(\omega) + (\sigma_e^2)^{(n)}} \quad (17d)$$

where $E\{\ \}$ denotes the expectation, $X(\omega)$ is the Fourier transform of $x(t)$, $\hat{E}(\omega)$ is the Fourier transform of $\hat{e}(t)$

4) The expectation of $n(t)$ is:

$$\hat{n}(t) = \sum_{k=0}^{q} a_k^{(n)} w(t-k) + \hat{e}(t) \quad (18)$$

M-step:

1) Updating coefficients of FIR filter $\{a_k\}$.

It made by solving the least-squares problem of (16) with (18):

$$\left\{a_k^{(n+1)}\right\} = \arg\min_{\{a_k\}} \sum_{n=0}^{N-1} \left(\sum_{k=0}^{q} (a_k^{(n)} - a_k) * w(t-k) + \hat{e}(t)\right)^2 \quad (19)$$

1) Updating σ_e^2

$$(\sigma_e^2)^{(n+1)} = \frac{1}{N} \sum_{t=0}^{N-1} e^2(t) \quad (20)$$

2) Updating the spectral parameters of $p'(t)$ by solving normal solving the normal equations using the sample correlation of $p'(t)$.

The EM algorithm iterates, until some convergence criterion is met.

VI. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the pathology feature estimation procedure, database of speech records from healthy and pathology patients was tested. Speech records were digitized and sampled at 22 kHz. Speech signal analysis was carrying out by devising the input signal into sequence of overlapped frames.

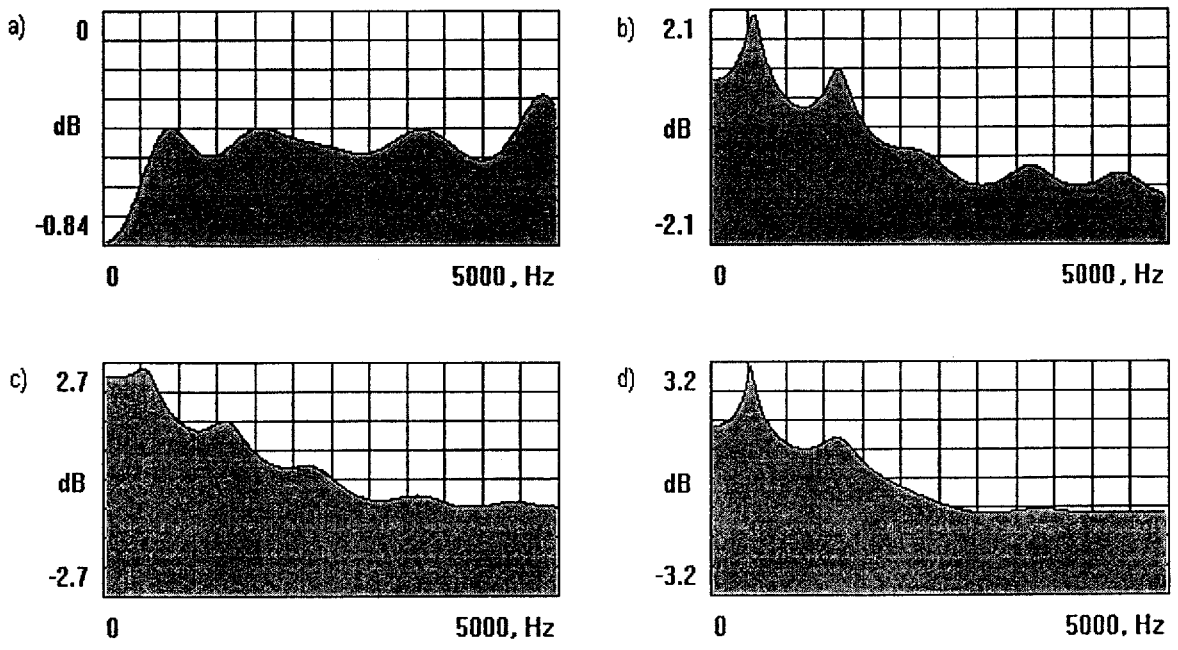

Figure 1. Typical examples of the SPC feature obtained for healthy and pathology conditions: a) for healthy; b) for paralysis; c) for dysphonia; d)for chorditis.

The frame window length was set to about 20 mseconds (512 samples) with frame overlap 30%. The analysis was focused on the sustained Russian vowels /e/, /u/, produced by 10 healthy and 40 pathology patients. The SPC feature parameters were estimated by applying EM algorithm. The order of AR model was set to 12, and order of FIR filter was set to 10.

Figure 1 shows typical examples of SPC parameters obtained for healthy and pathology conditions. As figure indicates, that healthy speech has consistent differences from speech under pathology.

For the classification purposes a CNN based on Kohonen self-organizing map was developed. To reduce the number of parameters for CNN, spectral parameters of SPC were parametrized to log Mel-frequency filter bank coefficients. The Mel scale is defined as

$$Mel(f) = 2595 \log_{10}(1 + f/700)$$

The analysis order was set to 20 coefficients.

VII. SUMMARY

In conclusion an iterative EM algorithm for maximum likelihood estimation of vocal fold pathology detection was developed. Despite some drawbacks of a general EM algorithm, such as slow linear convergence and influence of initial estimating on convergence, developed algorithm is versatile procedure. The number of iterations used before convergence was achieved for present study was 8.

The obtained feature SPC has been used to classification task. The combination of the EM algorithm and a CNN proposed for vocal fold pathology diagnostics provided high accuracy.

REFERENCES

[1] L. Gavidia-Ceballos, J.H.L. Hansen. IEEE Transactions on Biomedical Engineering, vol. 43, No. 1, pp. 35-45, January 1996.

[2] A. P. Dempster *et al.* Journal of Royal Statistical Society, B-39, pp. 1-38, 1977.

[3] Le Cam L. Int. Statist. Rev., "Maximum Likelihood: An Introduction, 1990, pp.153-171.

[4] S. L. Marple, Jr., Digital Spectral Analysis with Application, Prentice-Hall, Englewood Cliffs, 1987.