NONLINEAR RANDOM FEATURES OF NON-STATIONARY SIGNALS AND APPLICATIONS TO SPEECH RECOGNITION

Lingnan GE, Katsuhiko Shirai
School of Science and Engineering, Waseda University,
Tokyo 169-8555, JAPAN
Emails: Lingnan@shirai.cs.waseda.ac.jp, Shirai@shirai.cs.waseda.ac.jp

Yubo GE*
Department of Mathematical Science, Tsinghua University,
Beijing, 100084, CHINA
Emails: gyb-dam@tsinghua.edu.cn

Abstract
To describe and portray non-stationary signals characteristic more accurately, this paper suggests a dynamic technique to test the non-stability of real-time multimedia signal based on statistical theory and doubly random time series model. Applied this technique to acoustic speech processes stationary process repeatedly takes turn with non-stationary process, to select characteristic parameters and recognition models automatically. This technique can cut down the dimension of parameters, lighten the computation burden, release more store stacks, and reduce more candidates so that it can raise the real-time processing capability and recognition ratio.

1. Introduction
In current recognition system on digital signal processes alternating stationary pieces with non-stationary pieces, such as acoustic speech, image, senses, economic trend, and price of stocks, both the characteristic parameters drawn from such signals and the recognition models are always rigid. This treatment obviously is not possible to adapt dynamic signal processes. As you see, the speech has an obvious ascending trend. Our experiments show that some consonants still posses the trends even taken twice differences, which indicates that it is not reasonable to get Mel-CEP or LPC-CEP after making twice differences of these consonants.

So now we use new stochastic measurement to portray them and got satisfied results, based on the non-stationary specialty and outcome by using stochastic measurement and models. Consequently considerable attention has been paid to developing statistical theory and techniques to deal with such signal processes in the past two decades [1].

In the second section of this paper, we taste a dynamic technique testing non-stationarity of real-time acoustic speech processes and more efficiently features are suggested to deal with such signals. Instead of common linear time series a type of doubly random time series is recommended to find more reasonable characteristic parameters in Speech Recognition [3]. In Section 3, Moment estimation of parameters of a type of AR (p)-MA (q) (Auto Regression-Moving Average model) is given without the hypothesis of existence of higher order of moment than the second order. And the obtained estimation is easy to run. In the last section the paper shall give several available models to be automatically determined based on non-stationary measure developed in Section 2. In this section the paper also gives SS2-DSU-HMM and constructs our dynamic recognition experiments on Chinese words coming out of continuously speaking database, which show that our work greatly increases the correct recognition rate and ensures real time and simply performance.

2. Non-stationary Feature

2.1 Non-stationary Measure
We use two rapid and simple non-parameter test methods, inverted sequence test and runs test, to get non-stationary measure and parameters of acoustic speech signals. These two statistics are
\[ U = \frac{A+1/2 - K(K-1)/4}{\sqrt{2K^4+3K^2-5K}} \frac{1}{\sqrt{2}} \]

and

\[ T = (n - \frac{2N_1N_2}{N_1 + N_2})/(2N_1N_2(N_1 + N_2))^{3/2} \]

which came out of the inverted sequence test and runs test [4] and be defined as the measurement of non-stationarity. In the equation above, \( A \) is the total number of inverted sequence of rank statistics, \( K \) is the group number to be tested, \( N_1 \) and \( N_2 \) are numbers of zeros and 1s in the quantified observations according to medium of the series in each flame respectively, and \( n \) is the runs number. And all variables concerning with the above tests and the maximum lengths of the runs and the inverted sequences are regarded as non-stationary parameters.

For large sample case and \( K > 10 \), have that

\[ \sum_{i=0}^{d-1} \left( \frac{N_1 + N_2}{N_1 + N_2} \right) \left( \sum_{i=0}^{d-1} (\xi_i - \frac{N_1N_2}{N_1 + N_2}) \right) \quad \text{and} \quad U \sim \mathcal{N}(0,1) \]

where \( \xi_1 \) is the number of 0-runs.

Measure of non-stationarity consisting of \( U, T \) and differential order is an important index when automatically taking up certain type of AR (p)-MA (q). Non-stationary parameters are accurate portrayal of acoustic speech signals.

2.2 Analysis on Non-stationary

Based on runs test, we can prove that sonant indeed is not stationary. In fact, about the short and fluctuate frequently consonants, absolute values \( |T| \) locate between of 14 and 17. Made differences of the signal, value of \( |T| \) is still over 11.

Test on their period graphs \((C(f_j), f_j)\) gave the same conclusion, where

\[ C(f_j) = (n\sigma_a^2)^{-1} \left( \sum_{i=1}^{j} I(f_i) \right) \]

\[ I(f_i) = \frac{2}{N} \left( \sum_{i=1}^{N} a_i \cos(2\pi f_i t_i) \right)^2 + \left( \sum_{i=1}^{N} a_i \sin(2\pi f_i t_i) \right)^2 \]

\( f_i = i / N \) is the frequency and \( \{a_i\} \) is a time series with variance estimate \( \sigma_a^2 \).

2.3 Testing and Dealing with Non-stationarity

With the help of SAS software system, it can be proved that non-stationary parameters and LPC-Cep(or Mel-Cep) are un-correlated, which show the non-stationary parameters are new parameters to describe speech signal characteristics together with LPC-Cep or Mel-Cep.

Taken canonical correlation analysis, we confirm that to add non-stationary parameters to the former parameters will increase the ability of distinction each other among all DSU.

Calculated the first two canonical values of LPC, find that the accumulated distinguish ability reaches to 88.54%. While calculated the canonical values of LPC, Energy-Frequent-Quotient [12], and non-stationarity parameters, the accumulative distinguish ability increases up to 97.69% based on the first two canonical variables.

Besides the above statistical analyses, practical recognition experiment results also show the new characteristic parameters are vital and powerful.

3. Automatically Selecting Parameters

As we know, LPC-Cep and Mel-Cep are widely used in the signal processing now, for example in speech recognition by computer. However they are not best parameters. Since LPC-Cep is based on AR, some hypotheses of stability and linearity paradigm for the signals are necessary. Meanwhile, the order of AR model has to be selected high enough to simplify ARMA into AR. Although Mel-Cep based on short-time Fourier transform improved a little in deal with non-stationary signals, we cannot distinct some signals in the present parameter space with the dimension as high as 50 yet, including 16-order-Mel-Cep and their first two differentials. In natural spoken language with variant noise, this situation is more serious.

Thus, we suggest non-linear models to draw more reasonable feature of speech.

3. 1 Doubly Random Time Series Model

A general type of AR (p)-MA (q) model considered here
is that
\[
\begin{align*}
X_t = & \sum_{i=1}^{p} \Phi_i X_{t-i} + U_t, \\
\Phi_i = & \beta_i + \sum_{j=0}^{p} \theta_{ij} \epsilon_{t-j}, \quad \theta_{i0} = 1, i = 1, 2, \ldots, p \\
t \in Z := \{0, \pm 1, \ldots\}
\end{align*}
\]  
(3.1)

where \{\(X_t\)\} is a stationary series with mean zero and the following assumptions:

1) \(U_t\) is a white noise, \(EU_t = 0\), \(EU_t U_s = \delta_{st} \sigma_u^2\) \(\forall s, t \in Z\). And any event of the process \{\(U_t\)\} after time \(t\) is independent of any of the \{\(X_t\)\} before \(t\), exactly to say that
\[
Pr(A_0 A) = Pr(A_0) Pr(A), \quad \forall A_0 \in F^t \text{ and } \forall A \in N_t
\]

where \(N_t := \sigma[U_t, s \geq t]\) and
\[
F^t := \sigma\{X_s, s < t, \Phi_{it}, \forall t\text{ and } i\} = \sigma\{X_s, s < t, \epsilon_{it}, \forall t\text{ and } i\}.
\]

2) \{\(\Phi_i\)\}, \(i = 1, 2, \ldots, p\), are independent processes each other and for every fixed \(i\) the \{\(\Phi_i\)\} is stationary. All processes \{\(\Phi_i\)\} are uncorrelated with \{\(X_t\)\};

3) The process \{\(U_t\)\} is independent of \{\(\Phi_i\)\} and \{\(\epsilon_{it}\)\}, and \{\(\epsilon_{it}\)\} are white noise processes with \(E\epsilon_{it} = 0\) and \(E\epsilon_{it} \epsilon_{jt} = \delta_{ij} \sigma_i^2\) where is Dirac’s function.

4) \(\theta_{ij}\) are constants to be determined.

All parameters to be estimated in the model (2.1) are
\[
\beta_j, \sigma_u^2, \theta_{ij}, \sigma_i^2, i = 1, 2, \ldots, p, j = 1, 2, \ldots, q.
\]

About study of one dimensional AR (1)-MA (2), please refer [6]-[10]. This paper below gives a formula of parameter moment estimates for the considered AR(p)-MA(q) without the condition of existence of as high as 3 and over order of moments likely as [10]. As for multidimensional, refer [2]. In some special cases, the considered model can be explained as Bilinear Time Series and Multivariate Bilinear Model, which article [11] discussed in detail.

3. 2 Moment Estimation for AR (p)-MA (q)

Suppose that the moments concerned exist and define that
\[
R(k) = EX_k X_{k+1} \quad \text{and} \quad R_{\varphi}(k) = E\Phi_k \Phi_{k+1}.
\]

Firstly estimate parameters \(\hat{\sigma}_u^2\) and \(\hat{\beta}_i\), in model (3.1), which means the first equation of (3.1). In fact, since model (3.1) and assumption M, one can obtain the estimators of \(\beta_{1s}\) and \(\sigma_u^2\) when the estimators of \(R_{\varphi}(i)\) given based on the samples.

Moreover estimate \(R_{\varphi}(j)\). Continuously using (3.1) and Assumption M, one can obtain
\[
R(0) = \sum_{i=1}^{p} R_{\varphi}(0) R(0) + 2\sum_{i<j} \beta_i \beta_j R(j-i) + \sum_{i=1}^{p} \beta_i^2 R(0)
\]
\[
R(1) = \sum_{i=1}^{p} R_{\varphi}(1) R(1) + \beta_1^2 \beta_j R(0) + R(2) + \beta \beta_2 \sigma_u^2\]
\[
R(2) = \sum_{i=1}^{p} R_{\varphi}(2) R(2) + \beta \beta_2 \sigma_u^2
\]
\[
\quad \text{Specially, when } p = 2, \text{ have}
\]
\[
R(0) = \sum_{i=1}^{2} R_{\varphi}(0) R(0) + 2\beta_1 \beta_2 R(1)
\]
\[
R(1) = 2\sum_{i=1}^{2} R_{\varphi}(1) R(1) + \beta_1 \beta_2 R(0) + R(2) + \beta_2 \sigma_u^2
\]
\[
R(2) = \sum_{i=1}^{2} R_{\varphi}(2) R(2) + \beta_1 \beta_2 \sigma_u^2 + \beta_2 R(0)
\]
\[
\quad \text{In other wise, obtain that}
\]
\[
R(0) = R_{\varphi}(0) R(2) + \beta_1 \beta_2 R(3) + R_{\varphi}(0) \sigma_u^2 + \beta_2 R(2) + \sigma_u^2
\]
\[
R(1) = R_{\varphi}(1) R(1) + \beta_1 \beta_2 R(2) + \beta_2 \sigma_u^2 + \beta_2 R(1).
\]
\[
R(k) = \sum_{i=1}^{k+2} R_{\varphi}(k-i) R(i) + \beta_1 \beta_2 R(k-2) + \sigma_u^2
\]
(3.3)

and
\[
R(1) = \beta_1 R(0) + \beta_1 \beta_2 R(2) + \varphi_2(2) R(3) + \beta_1 \beta_2 \sigma_u^2
\]
(3.5)
Thus we have obtained the estimates of $R_i(k)$ by ones of $R(k)$ and $\beta_i$ from (3.3), (3.4) and (3.5). A performance could be the proposed procedure:

getting $\hat{R}_0$(1) and $\hat{R}_2$(2) respectively from (3.4) and (3.5); getting $\hat{R}_1$(0) from (3.4); and getting the rest from (3.3).

Finally, complete the estimates of parameters in (3.1).

Let $W_t = \Phi u - \beta_i$. Then we have a group of MA (q)'s, $Z_t = \sum_{j=0}^{q} \theta_j e_{t-j}$, $t = 1, 2, \ldots, p$. (3.6)

We have also $\hat{R}_{w_i}(k)$ from $\hat{R}_i(k)$ and $\hat{b}_i$. Thus $\hat{\theta}_i$ and $\hat{\sigma}^2$ are the solutions of the system of equations,

$$
\begin{align*}
\hat{R}_1(0) &= \hat{\sigma}^2 \left( \sum_{j=1}^{q} \theta_j^2 \right) \\
\hat{R}_i(k) &= \hat{\sigma}^2 \left( \theta_i + \sum_{j=1}^{q} \theta_j \theta_1 \theta_{i+j} \right) \\
\hat{R}_q(q) &= \hat{\sigma}^2 \left( \theta_{q+1} \right)
\end{align*}
$$

(3.7)

3.3 Estimation of Order

Based on non-stationarity measure and differential technique, we took a pretreatment of the digital data to be recognized. Though one can complete the estimation of the order of models based on the F-test for $H_0 : p = p_0$, $H_1 : p = p_0 - 1$, we prefer take the AIC criterion

$$
AIC = N \log \hat{\sigma}^2 + 2(p_0 + 1)
$$

Have estimated $\hat{\beta}_i$ and calculated

$$
\hat{\sigma}^2 = \frac{1}{N-p} \sum_{i=p+1}^{N}(X_i - \sum_{i=1}^{p} \hat{\beta}_i X_{i-1})^2
$$

We minimized the AIC to determine the $\hat{p}$.

4. Recognition Experiment

4.1 Structure of Experiment system

Feature parameters and models:

Based on analyses of non-stability and trend, we respectively selected the $p = 3$ for the some classes and $p = 2$ for others. Then we considered three models, AR (2)-MA (2), AR (3)-MA (2) and AR (3)-MA (3) to adapt variant types of C-V Units in DSU system [5]. The estimated parameters of these models were added into feature parameters in recognition experiments.

Recognition models: SS2-DSU-HMM

We run the recognition experiments in the three criteria of recognition: Mahalanobis distance between characteristic parameters, the probability according to the construction of a traditional HMM and the probability in SS2-DSU-HMM. About Statistical Segment HMM, please refer [12] and as to DSU, refer [5]. The foundation of the 2nd-order HMM is the hypothesis

$$
\begin{align*}
P(X_{t+1} = j \mid X_t = i, O_t = o_t) &= P(X_{t+1} = j \mid X_t = i, X_{t-1} = i_t, Y_t = o_t) := g_{i_{t+1}, i} \\
f(O_{t+1} = y \mid X_t = i, O_t = o_t) &= f_{i, t, o_t}(y) := h_{i, o_t}(y)
\end{align*}
$$

where signs $X_t = \{X_1, X_2, \ldots, X_t\}$, $O_t = \{o_1, o_2, \ldots, o_t\}$ are similar to $O_t$ and $o_t$; $\forall O_t \in (-\infty, \infty)$. The $\{O_t\}$ is a visible or measurable data drawing from the acoustic speech process, e.g., a series of its characteristic parameters and $\{X_t\}$ an immeasurable (hidden) state process.

Database:
The compare recognition experiments on all C-V units of whole Chinese syllables with the tones and size of 1312, which were separated from bigger continuous acoustic data, have been finished. The group of parameters A is consist of the model parameters of the above 3 types of AR(p)-MA(q), non-stationary parameters and energy-frequency-quotient [8], and group B consist of the 12 order of Mel-Cep.

4.2 Experiment Results

The experiment results are listed in the following 3 tables.
Table 1 Recognition results for C-V units (%)

<table>
<thead>
<tr>
<th>Mahalanobis distance</th>
<th>Group A: inc.AR(2)-MA(2)</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained Set</td>
<td>70.2</td>
<td>56.5</td>
</tr>
<tr>
<td>Recognized</td>
<td>51.6</td>
<td>47.4</td>
</tr>
</tbody>
</table>

Table 2 Recognition results for C-V units (%)

<table>
<thead>
<tr>
<th>HMM</th>
<th>Group A: Mixed of 3 types</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained Set</td>
<td>93.4</td>
<td>81.4</td>
</tr>
<tr>
<td>Recognized</td>
<td>85.3</td>
<td>65.6</td>
</tr>
</tbody>
</table>

Table 3 Recognition results for C-V units (%)

<table>
<thead>
<tr>
<th>SS2-DSU-HM M</th>
<th>Group A: Mixed of 3 types</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained Set</td>
<td>99.5</td>
<td>90.4</td>
</tr>
<tr>
<td>Recognized</td>
<td>95.2</td>
<td>84.6</td>
</tr>
</tbody>
</table>

5. Conclusion

In natural spoken language with variant noise, we cannot distinguish some signals in the present parameter space with the dimension as high as 50 yet. The new measurements are shown to be useful and reasonable parameters to describe non-linear processes, especially the speech signal processes.

The doubly Random Time Series Model AR(p)-MA(q), Then non-linear model was used to adapt the three types of consonants, which cut down both the dimension of parameter space and quantity of cut-branch and develop more powerful recognition model to describe the state processes and the observable processes of the speech. Compared with current widely used feature space, our new feature space and auto select experiments greatly increased the correct recognition rate and ensures real time and simply performance.

6. References