NON-LINEAR TECHNIQUES FOR DYNAMIC SPEECH RECOGNITION WITH DOUBLY RANDOM TIME SERIES MODELS

Lingnan GE, Katsuhiko Shira, Akira Kurematsu, Masaaki Honda
Computer and Network Department, School of Science and Engineering, Waseda University, Tokyo 169-8555, JAPAN

1. INTRODUCTION
At the beginning of words and phrases, the speech has an obvious ascending trend. Our experiments show that some consonants still possess the trends even taken twice differences, which indicates acoustic speech processes actually is not stationary. That’s why those signals cannot be distinguished in the 48-dimension parameters space of Mel-CEP or LPC-CEP, including twice differences. How to portray the non-stationary specialty and find more reasonable characteristic parameters?

We apply the rapid and simple non-parameter test methods, inverted sequence test and runs test, to get non-stationary measure and parameters of acoustic speech [1]. Non-stationary parameters are un-correlated with LPC-Cep(or Mel-Cep), which show that they are new parameters to describe speech signal characteristics.

Moreover we suggest non-linear models, doubly Random Time Series Model AR(p)-MA(q), to draw more reasonable features of speech. LPC-Cep and Mel-Cep are widely used in speech recognition. But LPC-Cep based on AR, some hypotheses of stability and linearity paradigm for the signals is necessary. In natural spoken language with variant noise, this situation is more serious. So we try to use those non-linear models to cut down both the dimension of parameter space and quantity of cut-branch to ensure running in real time.

2. MEASUREMENT OF NON-STATIONARITY
We apply the rapid and simple non-parameters test method, runs test and inverted sequence test, on acoustic speech to get their non-stationary measure and parameters.

\[ T = \frac{2N_1N_2}{N_1 + N_2} - \frac{2}{\sqrt{2K^3+3K^2-5K}} \]  \hspace{1cm} (2.1)

and

\[ U = A + 1/2 - K(K-1)/4 \]  \hspace{1cm} (2.2)

in the inverted sequence test [2], are defined as the measure of non-stationary, where \( n \) is the runs number, \( N_1 \) and \( N_2 \) numbers of zeros and 1’s in the quantified observations according to medium of the series in each flame respectively, and \( K \) is the group number to be tested. And all variables concerning with the above tests and the maximum lengths of the runs (defined as maxL) and the inverted sequences are regarded as non-stationary parameters.

For large sample case or \( K \geq 10 \),

\[ \max L = \frac{(N_1+N_2)^{3/2}}{N_1N_2} \left( \frac{1}{2} - \frac{N_1N_2}{N_1+N_2} \right) \]  \hspace{1cm} (2.3)

or

\[ U \]  \hspace{1cm} (2.4)

approximately obeys to the normal distribution where \( \xi_j \) is number of 0-runs.

Measure of non-stationarity is an important index when automatically taking up the type of AR(p)-MA(q).

We can prove that these 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.

3. NON-LINEAR MODELS
3.1 Doubly Random Time Series Model
A general type of AR(p)-MA(q) model considered here is that

\[
\begin{align*}
X_t &= \sum_{i=1}^{p} \Phi_i X_{t-i} + U_t, \\
\Phi_i &= \sum_{j=0}^{q} \theta_j U_{t-i-j}, \\
\theta_{0i} &= 1, i = 1, 2, ..., p \\
t &\in Z = \{0, \pm 1, \ldots\}
\end{align*}
\]  \hspace{1cm} (3.1)

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

1) \( U_t \) is a white noise, \( EU_t = 0 \), \( EU_t U_s = \delta_{tt} \sigma_u^2 \) \( \forall s, t \in Z \). 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

\[
\Pr(A_0 \cup A) = \Pr(A_0) \Pr(A), \quad \forall A_0 \in F^t \text{ and } \forall A \in N_t
\]

where \( N_t := \sigma(U_s, s \geq t) \) and

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


2) \( \{ \Phi_{it} \}, i = 1,2,\ldots,p, \) are independent processes each other and for every fixed \( i \) the \( \{ \Phi_{it} \} \) is stationary. All processes \( \{ \Phi_{it} \} \) are uncorrelated with \( \{ X_{it} \}; \)
3) The process \( \{ U_{it} \} \) is independent of \( \{ \Phi_{it} \} \) and \( \{ e_{it} \} \), and \( \{ U_{it} \} \) are white noise processes with \( Ee_{it} = 0 \) and \( Ee_{it}e_{jt} = \delta_{ij}\sigma^2_i \) where is Dirac’s function.
4) \( \theta_{ij} \) are constants to be determined.

All parameters to be estimated in the model (3.1) are \( \beta_i, \sigma^2_i, \theta_{ij}, \sigma^2_j, i = 1,2,\ldots,p; j = 1,2,\ldots,q. \)

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 [3].

3.2 Moment Estimation for AR(p)-MA(q)

Suppose that the moments concerned exist and define that
\[
R(k) = EX_{t}X_{t+k} \text{ and } R_{\Phi_{it}}(k) = E\Phi_{it}\Phi_{it+k}.
\]

Firstly estimate parameters \( \sigma^2_h \) 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,
\[
R(k) = \begin{cases} 
\sum_{i=1}^{p} \beta_i R(k-i) & k > 0 \\
\sum_{i=1}^{p} \beta_i R(i) + \sigma^2_U & k = 0
\end{cases}
\]

from which we can obtain the estimators of \( \beta_i \)'s and \( \sigma^2_i \) when the estimators of \( R(i) \) given based on the samples.

Moreover estimate \( R_{\Phi_{it}}(j) \). Continuously using (3.1) and Assumption M, one can obtain
\[
R(k) = \sum_{i=1}^{p} R_{\Phi_{it}}(k)R(k) + 2\sum_{i<j} \beta_i \beta_j R(k-j+i) + \sum_{j=1}^{p} E(\Phi_{tj+k}X_{t+k-j}U_{jt}).
\]

In other wise, obtain that
\[
\begin{align*}
R(0) &= R_{\Phi_{it}}(0)R(0) + \beta_1 \beta_2 R(2) + \beta_1 \sigma^2 + \beta_2 R(2) + \sigma^2,
R(1) &= R_{\Phi_{it}}(1)R(1) + \beta_1 \beta_2 R(2) + \beta_1 \beta_3 + \beta_2 R(1), \\
R(k) &= R_{\Phi_{it}}(k)R(k-2) + \beta_1 \beta_2 R(k-3) + \beta_2 R(k-1)
\end{align*}
\]
and
\[
R(1) = \hat{\beta}_1 R(0) + \hat{\beta}_2 R(2) + R\Phi_{it}(2)R(3) + \beta_1 \beta_2 \sigma^2
\]

We have also \( \hat{R}_{\Phi_{it}}(k) \) from \( \hat{R}_{\Phi_{it}}(k) \) and \( \hat{\beta}_i \). Thus \( \hat{\theta}_{ij} \)
and \( \hat{\sigma}^2_i \) are the solutions of the system of equations,
\[
\begin{align*}
\hat{\sigma}^2_i(0) &= \beta^2_1 \sum_{j=1}^{p} \hat{\theta}^2_{ij} \\
\hat{\sigma}^2_i(k) &= \beta^2_1 \sum_{j=1}^{p} \hat{\theta}^2_{ij} + \sum_{j=1}^{p} \hat{\theta}_{ij} \hat{\theta}_{ij} \\
\hat{\sigma}^2_i(q) &= \beta^2_1 \sum_{j=1}^{p} \hat{\theta}^2_{ij}
\end{align*}
\]

3.3 Estimation of Order

Based on non-stationarity measure and differential technique, we took a pretreatment of the digital dada to be recognized. Though one can complete the estimation of the order of models based on the F-test for \( H_0: p = p_0, \quad H_1: p = p_0 - 1 \), we prefer take the AIC criterion
\[
AIC = N \log \hat{\sigma}^2 + 2(p_0 + 1)
\]

Having estimated \( \hat{\beta}_i \) and calculated
\[
\hat{\sigma}^2 = \frac{1}{N-p} \sum_{n=p+1}^{N} (X_n - \sum_{i=1}^{p} \hat{\beta}_i X_{n-ib})^2,
\]
we minimized the AIC to determine the \( \hat{p} \).

4. STRUCTURE OF EXPERIMENT SYSTEM

Based on our analyses of non-stability and trend, we 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 voice Units [4]. The parameters estimated in these models were regarded as the new characteristic parameters in our recognition experiments.

As a database Chinese syllables with the tones and size of 1312 were used. Further experiments will be done by use of database such as AURORA-2J, Japanese family/personal names (JF/PN), and out-of-vocabulary (OOV) database.

5. REFERENCES