BLIND IDENTIFICATION OF NON-GAUSSIAN AUTOREGRESSIVE MODELS FOR EFFICIENT ANALYSIS OF SPEECH SIGNALS

Chunjian Li and Søren Vang Andersen

Department of Communication Technology, Aalborg University
DK-9220 Aalborg Ø, Denmark
e-mail: cl@kom.aau.dk, sva@kom.aau.dk

ABSTRACT

Speech signals, especially voiced speech, can be better modeled by non-Gaussian autoregressive (AR) models than by Gaussian ones. Non-Gaussian AR estimators are usually highly non-linear and computationally prohibitive. This paper presents an efficient algorithm that jointly estimates the AR parameters and the excitation statistics and dynamics of voiced speech signals. A model called the Hidden Markov-Autoregressive model (HMARM) is designed for this purpose. The HMARM models the excitation to the AR model using a Hidden Markov Model with two Gaussian states that have, respectively, a small and a large mean but identical variances. This formulation enables a computationally efficient exact EM algorithm to learn all parameters jointly, instead of resorting to pure numerical optimization or relaxed EM algorithms. The algorithm converges in typically 3 to 5 iterations. Experimental results show that the estimated AR parameters have much lower bias and variance than the conventional Least Squares solution. We also show that the new estimator has a very good shift-invariance property that is useful in many applications.

1. INTRODUCTION

Autoregressive (AR) modeling has been one of the most important techniques in speech signal processing. While the classical Least Squares (LS) solution, also known as LPC analysis, is computationally simple, it relies on a Gaussian AR model assumption. However, many important natural signals, including speech signals, are found to be far from Gaussian. The mismatch of a Gaussian model to a non-Gaussian signal causes an unnecessarily large variation in the estimates. This is supported by the fact that the Cramer-Rao bound for the variances of the AR estimators is lower in the non-Gaussian case than in the Gaussian case [1]. Smaller variances of AR estimators are desirable in many speech processing applications. As an example, in linear predictive coding, when a sustained vowel is segmented into overlapping frames that are subsequently encoded, small variance and shift-invariance property of the estimates of AR parameters are very beneficial in reducing the entropy and thus the needed bit rate for encoding the AR parameters. Non-Gaussian modeling of speech signals also reduces the bias of the AR estimator caused by the spectral sampling effect of the impulse train in voiced speech excitations. Applications in speech synthesis, speech recognition, and speech enhancement can benefit from these properties of non-Gaussian AR modeling.

This work was supported by The Danish National Centre for IT Research, Grant No. 329, and Microsound A/S.

We see the non-Gaussian AR model estimation problem as a blind system identification problem since the AR parameters and the non-Gaussian statistics of the excitation need to be estimated jointly. Reported works in this field include Higher Order Statistics (HOS) based methods (see [2] for a comprehensive review), Gaussian Mixture Model (GMM) based methods [1, 3, 4] and non-linear dynamical methods [5]. The HOS-based methods do not require explicit knowledge of the excitation probability density function (pdf), but tend to produce high-variance estimates when the length of the data record is small [3] and are associated with high computational complexity due to the bispectrum calculation. The GMM-based methods estimate their parameters using the Maximum Likelihood (ML) criterion. Since the exact ML solution for non-Gaussian signals typically involves solving a set of highly non-linear equations, it has to be solved by computationally complex numerical algorithms, or by solving for an approximation of the ML solution. In [1], the ML solution is solved by a conventional Newton-Raphson optimization algorithm. In [4], the AR parameters and the excitation probability density function (pdf) are separately estimated in a recursive manner to approximate the joint estimation in a tractable way. In [3], the AR parameters and the excitation pdf are estimated by a generalized EM (GEM) algorithm, which relaxes from the standard EM algorithm by breaking the multi-dimensional optimization into recursive one-dimensional optimizations. The price to pay for the GEM is a slower convergence rate than the EM. The non-linear dynamic method proposed in [5] estimates the coefficients of an inverse filter by minimizing a dynamic-based complexity measure called phase space volume (PSV). This method does not assume any structure of the excitation, but the computation of PSV is rather involved.

Most of the reported non-Gaussian AR modeling techniques are for general purposes. While being applicable to any probability distribution, this also makes them less efficient in handling speech signals, whose production mechanism is well known and implies powerful structures in the signal. In this paper, we propose an algorithm that is designed to exploit the structure of voiced speech signals, aiming at better computational efficiency and data efficiency. The algorithm jointly estimates the AR parameters and the excitation statistics and dynamics based on a ML criterion. Here the voiced speech signal is modeled by a Hidden Markov-Autoregressive Model (HMARM), where the excitation sequence is modeled by a Hidden Markov Model (HMM) that has two states with Gaussian emission densities of different means but same variances and then convolved with an AR filter. The HMARM parameters can be learned efficiently by an exact EM algorithm consisting of a set of linear equations. This model is different from the Li-

near Predictive HMM (LP-HMM), or Autoregressive HMM (AR-HMM) used in [6] and [7]. The AR-HMM applies its dynamic modeling on tracking the AR model variation along frames, while the proposed HMARM applies dynamic modeling on tracking the impulse train structure of the excitation within a frame.

The remainder of this paper is organized as follows. Section 2 describes the problem formulation and derives the EM algorithm. The algorithm is evaluated with synthetic signals and speech signals in Section 3. Conclusion is made in Section 4.

2. THE METHOD

The speech production mechanism is well modeled by the excitation-filter model, where an AR(p) filter models the vocal tract resonance property and an impulse train models the excitation of voiced speech. To improve naturalness of the speech, a white noise component is added to the impulse train. This can be expressed in the following equations :

$$x(t) = \sum_{k=1}^{p} g(k)x(t-k) + r(t) \tag{1}$$

$$r(t) = v(t) + u(t), \tag{2}$$

where $x(t)$ is the signal, $g(k)$ is the kth AR coefficient, and $r(t)$ is the excitation. The excitation sequence is the sum of an impulse train $v(t)$ and a white Gaussian noise sequence $u(t)$ with zero mean and variance σ^2. This noisy impulse train structure is perfectly suitable for stochastic dynamic modeling. We design a two-state HMARM whose diagram is shown in Fig.1. The state q_t at time t selects according to the state transition probability $a_{q_{t-1}q_t}$ one of two states. The emission pdfs of the two states are Gaussian pdfs with identical variances σ^2, and a small mean $m_r(1)$ and a large mean $m_r(2)$ respectively. The small mean is close to zero, and the large mean is equal to the amplitude of the impulses. The emission outcome constitutes the excitation sequence $r(t)$, which is independent of $r(l)$ for $l \neq t$ and only dependent on the state q_t. The excitation $r(t)$ is then convolved with an AR(p) filter with coefficients $[g(1), \cdots, g(p)]$ to produce the observation signal $x(t)$. The objective of the algorithm is to learn the model parameters $\boldsymbol{\phi} = [\mathbf{A}, m_r(1), m_r(2), \sigma^2, g(1), \cdots, g(p)]$ given a frame of signal $\mathbf{x}$ with length T, where the state transition matrix $\mathbf{A} = (a_{ij})$, with $i, j \in (1, 2)$.

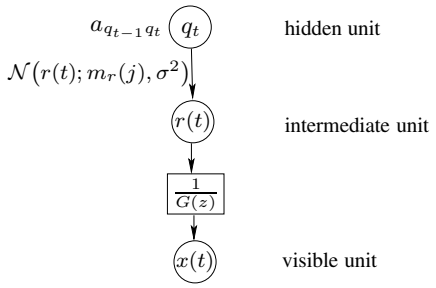


Fig. 1. A generative data structure of the HMARM.

We now define the notations for the HMARM model. Let $\alpha(j,t)$ and $\beta(i,t)$ denote the forward and backward likelihoods as defined in the standard HMM [8], a_{ij} denote the state transition (state i to state j) probability, $b_r(j,t)$ denote the observation pdf (emission pdf) of the excitation $r(t)$ given the state $q_t = j$, which is a Gaussian distribution

$$b_r(j,t) = \mathcal{N}\big(r(t); m_r(j), \sigma^2\big), \tag{3}$$

and $b_x(j,t)$ denote the observation pdf of the signal $x(t)$ given the state $q_t = j$. From (1) and (3), $b_x(j,t)$ can be shown to be a Gaussian process with a varying mean $m_x(j,t)$,

$$b_x(j,t) = \mathcal{N}\big(x(t); m_x(j,t), \sigma^2\big), \tag{4}$$

where

$$m_x(j,t) = \sum_{k=1}^{p} g(k)x(t-k) + m_r(j). \tag{5}$$

The forward and backward likelihood inductions are given by

$$\alpha(j,t) = \Big[\sum_{i=1}^{N} \alpha(i,t-1)a_{ij}\Big] b_x(j,t), \tag{6}$$

$$\beta(i,t) = \Big[\sum_{j=1}^{N} a_{ij} b_x(j,t+1)\beta(j,t+1)\Big], \tag{7}$$

respectively. Now define $\xi(i,j,t)$ to be the probability of being in state i at time t and in state j at time $t+1$, i.e. $\xi(i,j,t) = p(q_t = i, q_{t+1} = j|\mathbf{x}, \boldsymbol{\phi})$. One can evaluate $\xi(i,j,t)$ by

$$\xi(i,j,t) = \frac{\alpha(i,t)a_{ij}b_x(j,t+1)\beta(j,t+1)}{\sum_{t=0}^{T-1} a_{q_t q_{t+1}} b_x(q_{t+1}, t+1)}. \tag{8}$$

Define $\gamma(i,t) = \sum_{j=1}^{N} \xi(i,j,t)$. It can then be shown that the quantity $\sum_{t=1}^{T-1} \gamma(i,t)$ represents the expected number of transitions made from state i, and $\sum_{t=1}^{T-1} \xi(i,j,t)$ represents the expected number of transitions from state i to state j [8].

Now we derive the EM algorithm. Let bold face letters $\mathbf{x}$ and $\mathbf{q}$ denote a frame of signal and the state vector of the corresponding frame of excitation, respectively. We define the complete data to be $(\mathbf{x}, \mathbf{q})$. Instead of maximizing the log-likelihood $\log p(\mathbf{x}|\boldsymbol{\phi})$ directly, we maximize the expectation of the complete data likelihood $\log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi})$ over the states $\mathbf{q}$ given the data $\mathbf{x}$ and current estimate of $\boldsymbol{\phi}$, denoted by $\tilde{\boldsymbol{\phi}}$. So the function to be maximized in each iteration is written as :

$$Q(\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}) = \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) \tag{9}$$

$$= \sum_{\mathbf{q}} \frac{p(\mathbf{x}, \mathbf{q}|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \Big(\sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_x\big(q_t, x(t)\big) \Big) \tag{10}$$

$$= \sum_i \sum_j \sum_t \frac{p(\mathbf{x}, q_{t-1} = i, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log a_{q_{t-1}q_t} + \sum_j \sum_t \frac{p(\mathbf{x}, q_t = j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \log b_x\big(q_t, x(t)\big), \tag{11}$$

where (10) follows from the identity

$$p(\mathbf{x}, \mathbf{q}|\boldsymbol{\phi}) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_x\big(q_t, x(t)\big),$$

and (11) follows from the first order Markov assumption. The first term in (11) concerns only a_{ij} and the second term concerns the rest of the parameters. Thus the optimization can be done on the two terms separately. The re-estimation equation of a_{ij} is found by the Lagrange multiplier method, and is identical to the standard Baum-Welch re-estimation algorithm :

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1}=i, q_t=j|\tilde{\boldsymbol{\phi}})}{\sum_{t=1}^{T-1} p(\mathbf{x}, q_{t-1}=i|\tilde{\boldsymbol{\phi}})} = \frac{\sum_{t=1}^{T-1} \xi(i,j,t)}{\sum_{t=1}^{T-1} \gamma(i,t)}. \tag{12}$$

We denote the second term of (11) by $Q(\boldsymbol{\phi}, \hat{b})$. Following (1) and (4) we can write

$$Q(\boldsymbol{\phi}, \hat{b}) = \sum_j \sum_{t=1}^{T-1} \frac{p(\mathbf{x}, q_t=j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})} \Big(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \Big(x(t) - m_x(j,t) \Big)^2 \Big). \tag{13}$$

The re-estimation equations of the rest of the parameters are found by setting the partial derivatives of (13) to zero, and solving the equation system. For $g(k)$, we have p equations :

$$\sum_j \sum_{t=1}^{T-1} \gamma(j,t) \Big(x(t) - m_x(j,t) \Big) x(t-k) = 0, \quad k = 1, \cdots, p. \tag{14}$$

where $\gamma(j,t) = \frac{p(\mathbf{x}, q_t=j|\tilde{\boldsymbol{\phi}})}{p(\mathbf{x}|\tilde{\boldsymbol{\phi}})}$ is now interpreted as the posterior of state j at time t given the observation $\mathbf{x}$ and $\tilde{\boldsymbol{\phi}}$. For $m_r(j)$, we get two equations :

$$\sum_t^{T-1} \gamma(j,t) \Big(x(t) - m_x(j,t) \Big) = 0, \quad j = 1, 2. \tag{15}$$

For σ^2, we get

$$\widehat{\sigma^2} = \frac{\sum_j \sum_t^{T-1} \gamma(j,t) \Big(x(t) - m_x(j,t) \Big)^2}{\sum_j \sum_{t=1}^{T-1} \gamma(j,t)}. \tag{16}$$

Equation (14) and (15) form $p+2$ coupled linear equations which can be solved analytically. Then (16) can be solved by inserting the estimated $g(k)$ and $m_r(j)$.

In this model, $m_x(j,t)$ can be interpreted as the linear prediction of $x(t)$ taking into account the excitation dynamics, as shown in (5). The re-estimation equations also have intuitive interpretations. In (12), a_{ij} equals the expected number of transitions from state i to state j divided by the expected number of transitions made from state i ; Equation (14) is a multi-state version of the orthogonality principle ; Equation (15) tells that the prediction error weighted by state posterior is of zero mean ; and (16) calculates the mean of the prediction error power weighted by the state posterior as the variance of the stochastic element of the signal.

The existence of linear solutions to the maximization of the Q function makes fast convergence. This is a direct benefit from our proposed signal model. Compared to the GMM-based method in [3], which has no analytical solution to the maximization of Q function, the HMM in our model is constrained to have states with identical emission variance. It is this constraint that renders the set of non-linear equations linear, without compromising the validity of the model.

A GMM with similar constraint can be used in place of the HMM in our signal model, and the EM equations can be derived in the same way as shown above with proper changes in the definition of α and β (and $\xi(i,j,t)$ is not needed in the GMM). In our experience, this constrained GMM-AR model results in a slower convergence rate and slightly worse estimation accuracy than the HMARM. This is expected since the GMM lacks capability of dynamic modeling, while the impulse train does show a clear dynamic structure.

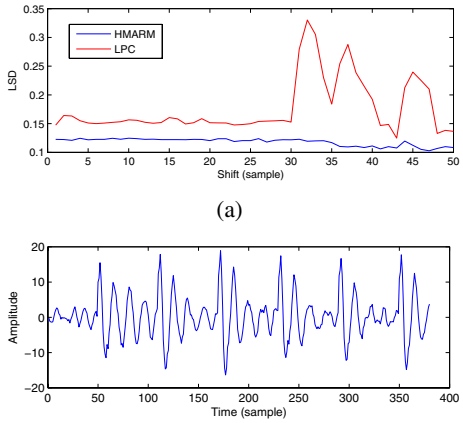
Finally, we point out an implementation issue of the HMARM estimation. Since the signal model is a causal dynamic model and the analysis is usually frame-based, the ringing from the last impulse of the previous frame has an undesired impact on the current frame estimates. This is because the estimator does not see the previous impulse but its effect is there. This could sometimes degrade the performance mildly. We therefore suggest to do a preprocessing that removes the ringing from the previous frame, or simply set the signal before the first impulse to zeros. The latter is used in our experiments.

3. EXPERIMENTAL RESULTS

We now experimentally compare the spectral distortion, the variance, and the bias of the AR parameters estimated by the proposed HMARM analysis and the LPC analysis. To get different realizations of an AR process, we shift a rectangular window along a long segment of the signal by one sample each time. Every shift produces a different realization frame of the AR process. A small variance of the estimates based on shifted realizations is also known as the shift-invariance property. The LPC analysis has a poor shift-invariance property when it is applied to voiced speech. This is because its underlying Gaussian model does not fit the non-Gaussian nature of the excitation of the voiced speech.

First, to have access to the true values of the AR parameters of a signal, we use a synthetic signal that mimics a voiced speech signal. The signal is analyzed by the HMARM and the LPC analysis respectively for 50 realizations with a frame length of 320 samples. The 50 realizations of estimated AR spectra are compared to the true AR parameters and the difference is measured by the Log-Spectral Distortion (LSD) measure. The LSD versus the shift is shown in Fig 2. It is clear that the proposed method has a flat distortion surface and this surface is lower than the LPC's. It is important to note that the LPC analysis encounters huge deviation from the true values in the second half of the plot. This is where a large "hump" in the signal comes into the analysis frame. The large humps in the signal are caused by the impulses in the excitation, which represent the non-Gaussian structure of the signal. The bias is 0.092 for the HMARM analysis, and compared to the 0.197 for the LPC analysis, accounts for an improvement of more than 6 dB. The variance is 0.128 for the HMARM and 9.69 for the LPC analysis, representing a variance reduction of 18 dB.

Second, we test the shift-invariance property with true speech signals. The AR spectra of four different sustained voiced phonemes are estimated 50 times with one sample shift each time. The frame length is set to 256 samples. The spectra are plotted in Fig 3. The estimates by the HMARM show good consistency, while the LPC analysis appears to be poor. In Fig. 4 we show the prediction residuals of the signal using the AR parameters estimated by the HMARM and the LPC respectively. It is clear that the residual of the HMARM has more prominent impulses, and less correlation in the valleys. From, as one example, a speech coding point of view,



(a)

(b)

Fig. 2. (a) : The Log-Spectral Distortion of the AR spectra. (b) : the synthetic signal waveform used in the test.

the lower variance of the AR estimates reduces the entropy of the AR parameters, and the more impulsive residual is also easier to code.

As it is well known that a properly chosen window can reduce the variance of the LPC estimates, we also conducted comparisons between the HMARM analysis and the Hamming-windowed LPC analysis. For the synthetic signal, the variance of the Hamming-windowed LPC is 1.197, which is still 9.7 dB higher than that of the HMARM. Although its variance is reduced, the Hamming-windowed LPC in general suffers from larger bias and lower spectral resolution. Due to space limit, more results will be presented in a following paper.

4. CONCLUSION

A non-Gaussian AR model is proposed to model the voiced speech signal. This model enables an efficient EM algorithm that consists of a set of linear equations. The algorithm jointly estimates the AR parameters of the signal and the dynamics of the excitation that is highly non-Gaussian in the voiced speech case. The experimental results using synthetic signals and real speech signals show that the algorithm has a good shift-invariance property, and the variance and bias are significantly smaller than the classical LPC analysis.

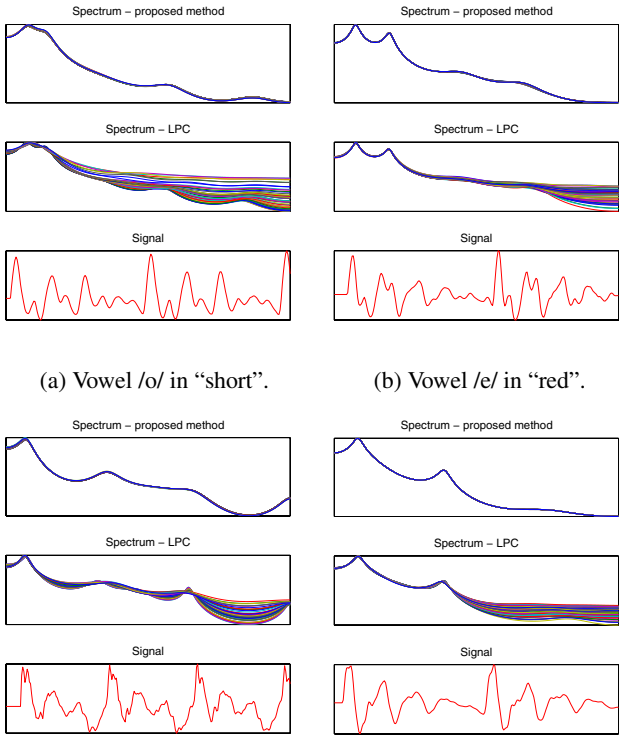
5. REFERENCES

[1] D. Sengupta and S. Kay, "Efficient estimation of parameters for non-Gaussian autoregressive processes," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37. No.6, pp. 785–794, 1989.

[2] C. L. Nikias and M. R. Raghuveer, "Bispectrum estimation : a digital signal processing framework," *Proc. IEEE*, vol. 75, pp. 869–891, 1987.

[3] S. M. Verbout, J. M. Ooi, J. T. Ludwig, and A. V. Oppenheim, "Parameter estimation for autoregressive Gaussian-Mixture processes : the EMAX algorithm," *IEEE Trans. on Signal Processing*, vol. 46. No.10, pp. 2744–2756, 1998.

[4] Y. Zhao, X. Zhuang, and S.-J. Ting, "Gaussian mixture density modeling of non-Gaussian source for autoregressive process," *IEEE Trans. on Signal Processing*, vol. 43. No.4, pp. 894–903, 1995.



(a) Vowel /o/ in "short". (b) Vowel /e/ in "red".

(c) Vowel /a/ in "began". (d) Vowel /ou/ in "shoulder"

Fig. 3. The AR spectra estimated by HMARM and LPC analysis.

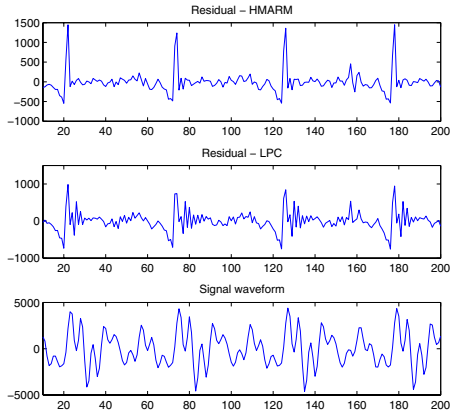


Fig. 4. Prediction residuals by the HMARM and the LPC analysis.

[5] H. Leung, S. Wang, and A. M. Chan, "Blind identification of an autoregressive system using a non-linear dynamical approach," *IEEE Trans. on Signal Processing*, vol. 48. No.11, pp. 3017–3027, 2000.

[6] A. Poritz, "Linear predictive hidden Markov models and the speech signal," *ICASSP'82*, vol. 7, pp. 1291–1294, 1982.

[7] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive Hidden Markov Models for speech signals," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-33. No.6, pp. 1404–1413, 1985.

[8] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Model," *IEEE ASSP Magazine*, pp. 4–16, Jan. 1986.