A New Iterative Algorithm for ARMA Modelling of Vowels and glottal Flow Estimation based on Blind System Identification

Milad LANKARANY
Department of Electrical and Computer Engineering, Shahid Beheshti University, Tehran, IRAN
(Now with the Department of Electrical and Computer Engineering of Concordia University
Montreal, Quebec, Canada)
and
Mohammad Hassan SAVOJI
Department of Electrical and Computer Engineering, Shahid Beheshti University
Tehran, IRAN

ABSTRACT

In this paper the voiced speech signal is modelled as an ARMA process with: 1- an AR filter whose coefficients are obtained using a new iterative model-based algorithm and 2- an MA filter whose input is the glottal excitation and its output is the linear prediction residual or the input of the AR filter. The AR filter is estimated using an iterative algorithm in which the Liljencrants–Fant (LF) model of the glottal flow is fitted, at each iteration, to the glottal flow derivative waveform extracted by closed phase inverse filtering. After calculating the AR filter, the MA filter coefficients are estimated using a new higher order statistics based blind system identification algorithm where an initial estimate of the original input, which is the obtained LF model, is whitened and used instead of the usual i.i.d input. We propose a new iterative algorithm using a constrained optimization that includes an objective function that is based on the diagonal slice of the third order cumulant of the MA filter's impulse response and, a constraint in which the mean square error between the estimated input and the initial model is kept lower than a limit. Finally, the efficiency of the algorithm is assessed on the real voiced speech sounds /a/ as a practical case example.

Keywords: Glottal Flow Estimation, ARMA Modelling, Blind System Identification, Constrained Optimization and LF Model.

1. INTRODUCTION

The voiced speech signal, in this paper, is modelled as an ARMA process whose input is the glottal flow with: 1- an AR filter whose coefficients are obtained using a new iterative model-based algorithm and 2- an MA filter whose input is the glottal excitation and its output is the residual of linear prediction (LP) analysis or the input of the AR filter. The proposed iterative model-based algorithm, for estimating the AR coefficients, deals with two main problems that all the other proposed approaches suffer from. The first problem relates to those methods using closed phase inverse filtering. The core problem in such algorithms is where the glottis is not sufficiently or completely closed, thus, not enough data exist for estimating the filter coefficients. The second problem that occurs in model-based algorithms is the inability of such algorithms to estimate the actual shape of the glottal flow derivative during the time interval of vocal-fold closure where a ripple may appear.

The method described in [2] is the first attempt to combine the model-based and closed phase inverse filtering approaches to alleviate these problems. We propose an iterative algorithm, for solving the mentioned problems, in which the LF model, at each iteration, is fitted to a glottal derivative waveform extracted by closed phase inverse filtering. Unlike the joint optimization methods (model-based algorithms) that use Kalman filter for updating the filter coefficients, the parameters of the inverse of the AR filter (LP analysis filter) is adaptively calculated using a normalized LMS adaptive filter to minimize the mean square error (MMSE) between the residual signal calculated as the analysis filter output excited by the speech signal and the LF model considered as the desired signal. The next iteration begins by obtaining a new LF model to fit the residual. The glottal flow derivative is obtained finally by inverse filtering the speech using the final estimation of the AR coefficients. This solution can tackle the problem of insufficient data during the closed phase by its iterative model-based property. On the other hand, using the closed phase inverse filtering, at its first estimation, relates this method to those solutions that use this type of inverse filtering.
After calculating the AR coefficients and obtaining its corresponding residual, which is considered as the initial estimation of the glottal flow derivative, the MA filter coefficients are estimated using higher order statistics (HOS) based blind system identification. It is clear that both the input signal and the MA filter coefficients are unknown. Therefore, we are dealing in fact with a blind system identification problem. As the MA filter is non-minimum phase, higher order statistics (HOS) based blind system identification algorithms related to identifying non-minimum phase blind systems, can be employed here to estimate the MA part of the speech production model.

HOS-based blind deconvolution and system identification, associated with nonlinear filtering, are classified as explicit and implicit solutions [10]. The implicit solutions include the well known Bussgang algorithm. In fact, implicit HOS algorithms, such as our proposed method, are relatively simple to implement and are generally capable of delivering a good performance, as evidenced by their use in digital communication systems. However, they suffer from two basic limitations in such application: a potential convergence to a local minimum and sensitivity to phase jitter. In contrast, explicit HOS-based solutions, being closed form, overcome the local minimum problem by avoiding the need for minimizing a cost function; unfortunately, they are computationally much more complex. Both explicit and implicit kinds of HOS-based solutions suffer from slow rate of convergence due to the fact that the time-average estimation of higher order statistics requires a large sample size.

Many algorithms have been proposed in the literature for the identification of FIR system using cumulants. Mendel in [11] categorized these methods in three groups: 1) closed-form solutions, 2) optimization based and 3) linear algebra solutions. Unlike the research works that concentrated mostly on identifying the MA filter coefficients with filter order lower than five [10], [12], [13], [14], [15], an optimization solution was proposed in [16], on the basis of third order statistics, to overcome the problem of estimating the MA filter coefficients for high order systems.

To meet our objective that is estimating the MA filter coefficients of the speech production model, we propose an iterative algorithm that uses the same objective function as in the method described in [16]. In the proposed method the LF model obtained in the AR process is used as initialization for our knowledge based blind system identification pivoted on a new viewpoint that overcomes the problem of needed long data sequences. In addition, unlike the conventional blind system identification algorithms where some assumptions on the statistical properties of the white source signal are needed to be made, here an initial estimate of the original input to be identified, based on some prior knowledge (LF model of the glottal flow), is whitened and used instead of the usual i.i.d input.

Organization of this paper is as follows: in section 2, estimating the AR coefficients of the speech signal is described. The problem of estimating the MA parameters of the speech signal is dealt with in section 3. And, finally in section 4, the experimental results are shown to demonstrate the accuracy and robustness of our proposed algorithm.

## 2. ESTIMATING THE AR FILTER COEFFICIENTS

The proposed iterative model-based algorithm for calculating the AR coefficients and an accurate estimation of GFD is depicted in the block diagram figure 2. The AR coefficients are first calculated using pitch synchronous closed phase linear prediction method. The glottal flow derivative is then obtained by inverse filtering.

As described in section 2.2, the parameters of the LF model are estimated using an algorithm that fits the model to the glottal flow derivative. Unlike the conventional model-based methods that update the AR coefficients using joint optimization algorithms based on Kalman filters, the parameters of the inverse of the AR filter is adaptively calculated using a normalized LMS adaptive filter to satisfy the minimum mean square error (MMSE) criterion between the linear prediction residual of the speech signal and the LF model that is considered as the desired signal in the adaptive filter. It is thus clear that the order of the inverse filter is the same as the AR filter. Also, the AR coefficients obtained by closed phase linear prediction method can be used as initialization of the adaptive filter. The new estimation of the inverse system coefficients results in a new estimation of the glottal flow derivative, at the next iteration. Then, a new LF model is calculated using this new signal. And, the inverse system parameters are updated. Therefore, a new estimation of the glottal flow derivative waveform is obtained at each iteration. Indeed, the algorithm stops when no considerable changes occur, in two consecutive iterations, in the glottal flow derivative. It is noted that the locations of the poles of the transfer function of the inverse system are monitored and the iteration is stopped as soon as any of the poles moves outside the unit circle. Obviously in this case the previous results are used.

### 2.1. Closed Phase Determination

The glottal closed and open phases, at first iteration, are identified using an initial estimate of the glottal flow derivative obtained by applying LP method over a whole period of the vowel sound. The closed and open phases are then modified once a new estimation of the glottal flow derivative is obtained.
2.2. LF Model

The LF model [17] is defined, over a single glottal cycle, by a set of parameters \( p = [T_o, T_c, T_s, \omega_o, \alpha, E_s, \beta] \) shown in figure 3. Here, the objective is to find the LF parameters subject to minimizing the MSE between the model and the residual signal. We use function “\textit{fminunc}” of MATLAB as the optimization tool for MMSE criterion between the linear prediction residual of the speech signal and the LF model.

In fact, according to what we found using a large number of simulations, the similarity between the higher order cumulants of the impulsive signals and the ideal impulse are more (smaller MSE) than that of the impulsive signals and the ideal impulse. In brief, the higher order cumulants of the impulsive signals are themselves impulsive but with very small MSE with respect to the ideal impulse. In other words, it can be said that, the differences between the impulsive signal and the ideal impulse are diminished in what regards the higher order statistics. Indeed, if the excitation of an unknown system is impulsive, the higher order cumulants of the observed output roughly contains only the influence of the higher order complexity of the system parameters. As a result, HOS based blind system identification can be applied to systems with impulsive excitations.

We use the mentioned prospective for identifying the parameters of the FIR blind systems in which an initial model that describes the general shape of the unknown original input exists. In our proposed method, a whitener is calculated to whiten the initial model. This whitener is then used for filtering the output and consequently providing a new signal called \( T_{in} \) in figure 5.

3. ESTIMATING THE MA FILTER COEFFICIENTS

Consider an unknown linear time invariant system, \( H \), with input, \( \{x(n)\} \), as depicted in figure 4. The input consists of an unobserved white data sequence with known probability density distribution.

\[
\begin{align*}
\text{unobserved data} & \rightarrow \text{observed output sequence} \ (u(n)) \\
\text{sequence} \ \{x(n)\} & \rightarrow \text{Original System} \ H
\end{align*}
\]

The problem is to estimate the system, \( H \) or equivalently, restore \( \{x(n)\} \) by calculating \( H^{-1} \) the inverse of the system \( H \), given the observed sequence \( \{u(n)\} \) at the system output. In fact, as far as the use of blind system identification or blind deconvolution algorithms are concerned, the usual i.i.d random white input signal with non-Gaussian distribution can be replaced with an impulsive excitation.

This is due to the fact that auto-correlation and higher order moments of an impulse are themselves impulses as with i.i.d white sources. Therefore, an ideal impulse satisfies all conditions considered for i.i.d signals and hence the blind deconvolution algorithms can be applied to a wide range of applications, such as seismic signal processing, to estimate the unknown impulsive input signal. It is understood that the impulsive signal is a signal with small difference, in the mean square error (MSE) sense, to the ideal impulse. The quantity of this error has not been discussed in previous researches but it is apparent that the signal with lower MSE with ideal impulse is more appropriate for the blind deconvolution algorithms to be used.

As we assume that the initial model is similar to input, the signal \( T_{in} \) can be interpreted as the output of a blind system whose impulse response is the same as the system, \( H \) and its input is an impulsive excitation. Figure 6 shows this system.

\[
\begin{align*}
\text{Impulsive Signal} & \rightarrow \text{Original System} \ H \\
\text{Original System} \ H & \rightarrow \text{Tin}
\end{align*}
\]

As described before, in this case, HOS based blind system identification can be applied to estimate the impulse response of the system \( H \).

3.1. Whitener

Despite most researches where linear prediction methods are used for whitening, FFT-based inverse filtering is used here as it results in better impulsive signals.

As far as our application of glottal flow waveform estimation is concerned, an iterative algorithm based on higher order statistics is proposed for calculating the coefficients of the MA filter of the speech production ARMA model. Here, both the vocal tract filter and the glottal excitation are unknown. On the other hand, physiological models, such as Rosenberg or LF model, are available for modelling the glottal excitation. In the context of accurate modelling of the speech signal, the AR model which gives a minimum phase (i.e., a minimum energy or anti-dispersive) system is not sufficient to illustrate the non-minimum phase property of the speech signal. This system must be completed with a dispersive MA system, as in an ARMA model, that is excited with the glottal flow, a natural signal that
is not energy compact (minimum phase). As the AR coefficients, initial GFD and the LF model are obtained in section II, the MA coefficients are estimated using our proposed iterative algorithm shown in the block diagram of figure 7. Also, it should be noted that the method described in [16], which is based on the minimization of the sum of squared differences between the observed cumulant (diagonal slice of output) and the cumulant calculated using the unknown parameters, is used as the core of our blind system identification algorithm.

The filter, \( h = [h(1) \ldots h(q)] \), shown in figure 7, must satisfy a nonlinear constraint optimization. The signal, \( T_{in} \), that is the result of filtering the output with the whitener, is carried through as the input of this filter whose role is to satisfy the criterion mentioned in [16]. Also, as far as the constraint optimization is concerned a nonlinear inequality should be satisfied simultaneously with that criterion depicted in figure 8.

The used objective function and the constraint (MMSE criterion) can be formulated as follows:

\[
\begin{align*}
J &= \sum_{m=0}^{q} \left[ c_{mm}(m,m) - \sum_{k=1}^{q} \rho(k)h(k+m) \right]^2 \\
\text{where } c_{mm}(m,m) &= E[L^2(m)L^2(0 + \Delta m + \Delta N)] \\
\text{Objective Function} &\rightarrow \min (J) \\
\text{Criterion based on Third order statistics} &\leq \alpha \\
\text{Initial Model} &\rightarrow \alpha \text{Error}
\end{align*}
\]

Fig7: The block diagram of proposed algorithm for estimating the MA coefficients

Considering our used objective function and optimization constraint (both convex functions), a global solution exists.

**Note1:** The maximum of the filter order is the output’s length. The simulations showed that the filter coefficients tend to negligible values as this order exceeds one-third of its maximum. Therefore, these coefficients can be ignored and the filter order can be assumed equal to one-third of the length of the output.

**Note2:** As we expect, the initial model is more similar to the input than the output (i.e., the final estimation of the LPC residual), therefore the constant \( \alpha \) is set as a percentage (e.g. 80\%) of the total square error between the output and the model.

**Note3:** The method described in [18], which is a novel algorithm based on adaptive filtering for deconvolution of the non-minimum phase FIR systems, is used for inversing the filter, \( h \), at each iteration of the constrained optimization.

**Note4:** The algorithm is stopped when no considerable changes are observed between two consecutive iterations.

4. EXPERIMENTAL RESULTS

The vowel /a/ pronounced by a male adult is used for evaluating our proposed algorithm for glottal flow estimation. The selected pitch marked speech waveform is shown in figure 9.a. The initial estimation of the LPC residue and its corresponding LF model is shown in figure 9.b. As stated before, closed phase inverse filtering makes more precise estimation of the glottal flow derivative which is considered as the desired input of the adaptive filter where the AR filter coefficients are updated. This signal is depicted in figure 9.c.

Using the adaptive filter in an iterative manner (section 2), a new estimation of glottal flow derivative is calculated. This signal, after 20 iterations, is shown in figure 9.e.

Also, this signal, after 10 iterations, is shown in figure 9.d in order to demonstrate the similarity between this signal and the same after 20 iterations (fig 9.e).

The iterative algorithm for the AR-filter calculation can be stopped using this similarity when no considerable change between two consecutive iterations is observed. As stated before, the signal that is obtained as input to the AR filter, once this is calculated, is used as the output of the blind MA system or equivalently the input of the knowledge based blind system identification algorithm.

The accurate estimation of glottal flow derivative is calculated and shown in figure 9.f. Finally, the glottal waveforms are calculated by integrating the estimated inputs as shown in figure 9.g.

5. CONCLUSION

An accurate estimation of the glottal flow derivative is obtained in this work where we model the voiced speech signal as an ARMA process driven by this signal. We proposed a new iterative model-based algorithm to adaptively calculate the AR part of the system in conjunction with an accurate estimate of the glottal flow derivative. This signal is, then considered as the output of an MA system where neither the input nor the system is known but where an accurate model for the input is available. This is referred to, in this work, as knowledge based blind
system identification. This concept is introduced as estimating the input of an unknown non-minimum phase FIR (MA) system using only the observed output and some knowledge of the original input. Here, unlike conventional blind system identification/deconvolution where some assumptions on the statistical properties of the white source signal are needed to be made, an initial estimation of the original input is whitened and used instead of the usual i.i.d input. Furthermore, a constrained optimization is used to estimate the MA system coefficients to satisfy more than just one criterion. We apply our proposed algorithms to estimate the glottal flow excitation of vowels in order to demonstrate its accuracy and efficiency. Although the results are encouraging, it must be emphasized that many issues such as the convergence of the algorithm or the theoretical validity of the results remain to be studied. In terms of ARMA modelling of speech, it is claimed that a more plausible input is arrived at using this algorithm.
REFERENCES