

## Adaptive Inverse Filtering for High Accuracy Estimation of the Glottal Source

*Qiang Fu, Peter Murphy*

Department of Electronic and Computer Engineering  
University of Limerick, Limerick, Ireland

Email: [Qiang.Fu@ul.ie](mailto:Qiang.Fu@ul.ie) [Peter.Murphy@ul.ie](mailto:Peter.Murphy@ul.ie)

### Abstract

An adaptive, pitch-synchronous analysis method is proposed for the simultaneous estimation of vocal tract and voice source parameters from speech waveforms. A time varying autoregressive model with exogenous input (ARX) is chosen for vocal tract modeling because of the capability of such a model for characterising both the formants and antiformants of the vocal tract. The Liljencrants-Fant model for the voice source is integrated into an iterative adaptive estimation procedure. Furthermore, an adaptive inverse filtering technique is put forward to obtain high accuracy estimation of the glottal source waveform, which is necessary for the intended application of the method to pathological voice analysis. The technique is evaluated and compared with a number of other approaches using synthetic speech containing additive noise at the source. The results illustrate the superior performance of the new method.

### 1. Introduction

Voiced speech is typically modeled as the output of a linear, time-invariant, vocal tract filter excited by a quasi-periodic glottal volume velocity (GVV) signal [1]. Based on this source-filter model, a straightforward method to estimate the GVV is the *inverse filtering* technique. Wong et. al. [2] present a classic pitch synchronous closed phase covariance linear prediction (LP) algorithm. Alku [3] gives an iterative adaptive inverse filtering (IAIF) algorithm, in which the glottal spectrum and the vocal tract spectrum are estimated adaptively and iteratively. In [4], Alku uses the Discrete All-Poles (DAP) model instead of LPC in modeling of the vocal tract. Furthermore, the estimated glottal waveform resulting from the inverse filtering, can be parameterized using a glottal model, such as the most widely used *Liljencrants-Fant* (LF) [5] model, which is expressed in terms of four parameters.

Inverse filtering is basically a two-stage estimation method, which is subject to many restrictive assumptions, and is frequently impossible. Alternately, a joint estimation process can be developed to simultaneously estimate the parameters of the vocal tract and the glottal source. For example, in Krishnamurthy's Sum-of-Exponentials model [6], the glottal source is described using the LF model, and the vocal tract is modeled as a pole-zero system, with a different set of pole and zero locations in the closed and open phases to model the source-tract interaction effect. Ding and Kasuya [7] use a time-varying autoregressive with exogenous input (ARX) model to represent the vocal tract, and a parametric Rosenberg-Klatt model to generate a glottal waveform. Recently, Frohlich et. al. [8] also proposed a simultaneous inverse filtering and model matching (SIM) method, in which the LF model of the effective glottal source was integrated into the inverse filtering algorithm, thereby improving the estimation of the vocal tract resonance filter.

In this paper, we will propose a novel joint estimation algorithm. Given its capability for characterizing a time varying system, the autoregressive with exogenous input (ARX) model is chosen for vocal tract modeling. The LF model is integrated into an iterative adaptive estimation procedure. With the result of ARX model estimation, an adaptive inverse filtering method is incorporated to obtain a high accuracy automatic estimation of the glottal source waveform. The technique is evaluated and compared with a number of other approaches using simple synthetic pathological speech containing additive noise at the source. The results illustrate the superior performance of the new method.

## 2. Joint Estimation of Source and Vocal Tract based on ARX model

### 2.1. ARX Speech Modeling

It is clear that, in the closed phase, the input signal vanishes and the speech signal consists only of the free resonance of the vocal tract system. This behavior can be modeled by an all-pole filter, except in the case of nasals or nasalized vowels, where a pole-zero filter is more appropriate. A problem is that the closed phase can be short or in some case nonexistent.

In the open phase, the effective glottal signal forms a nonzero input to the vocal tract filter. Moreover, the system includes not only the vocal tract, but also the trachea and the coupling of the trachea and the vocal tract is time varying, due to the vocal-fold motion. Third, the source-tract interaction effect [9] increases the damping, shifts the resonance frequencies, and may introduce additional poles and zeros. Thus, the open phase model consists of an unknown source exciting a time varying pole-zero filter.

In the ARX model, speech production process can be modeled as a time-variant pole-zero system with an equation error described by the following equation:

$$s(n) + \sum_{i=1}^p a_i(n)s(n-i) = \sum_{j=1}^q b_j(n)u(n-j) + u(n) + e(n) \quad (1)$$

where  $s(n)$  and  $u(n)$  denote an observed speech signal and an unknown input glottal waveform at time  $n$ , respectively.  $a_i(n)$  and  $b_j(n)$  are time-varying coefficients.  $p$  and  $q$  are model orders, and  $e(n)$  is the equation error associated with the model.

As to  $u(n)$ , for the convenience of further discussion, we define

$$u(n) = u_s(n) + u_n(n) \quad (2)$$

where  $u_s(n)$  is the differentiated voicing source signal generated by the LF model in which the radiation characteristics of the lips are included.  $u_n(n)$  is the additive independent white noise. Thus, equation (1) can be further expressed as

$$s(n) + \sum_{i=1}^p a_i(n)s(n-i) = \sum_{j=1}^q b_j(n)u_s(n-j) + \sum_{j=1}^q b_j(n)u_n(n-j) + u_s(n) + u_n(n) + e(n) \quad (3)$$

Now let

$$\mathcal{E}(n) = \sum_{j=1}^q b_j(n)u_n(n-j) + u_n(n) + e(n) \quad (4)$$

then

$$s(n) + \sum_{i=1}^p a_i(n)s(n-i) = \sum_{j=1}^q b_j(n)u_s(n-j) + \mathcal{E}(n) \quad (5)$$

### 2.2. Voice Source Model

The LF model is a four-parameter model of the glottal flow derivative  $u(n)$ . The time-normalized glottal cycle is modeled in two sections: an exponentially weight sinusoid models the open phase until first collisional contact of the vocal folds, followed by an exponential return phase that prohibits an unrealistic abrupt termination of the flow. Suppose that the pitch period is  $M$  samples long, and that the open and closed glottal phases extend from sample 0 to  $N-1$  and sample  $N$  to  $M-1$ , respectively. Then the discrete-time version of LF model is defined as

$$u(n) = \begin{cases} A_{go} e^{\alpha_{go} n} \sin(\omega_{go} n + \phi_{go}), & n = 0, \dots, N-1 \\ -A_{gc} e^{-\alpha_{gc} (n-N)}, & n = N, \dots, M-1 \end{cases} \quad (6)$$

### 2.3. Joint Estimation

In the joint estimation algorithm, a Kalman filtering is applied for formant/antiformant estimation and an optimization procedure based on simulated annealing [11] is used for source parameter estimation. Both of the above procedures are based on the minimization of the predicted mean-square error (MSE) of the ARX model,

$$E = \frac{1}{N} \sum_{n=1}^N \{s(n) - \hat{s}(n)\}^2 \quad (7)$$

where the predictor is defined as:

$$\hat{s}(n) = -\sum_{i=1}^p a_i(n)s(n-i) + \sum_{j=0}^q b_j(n)u_s(n-j) \quad (8)$$

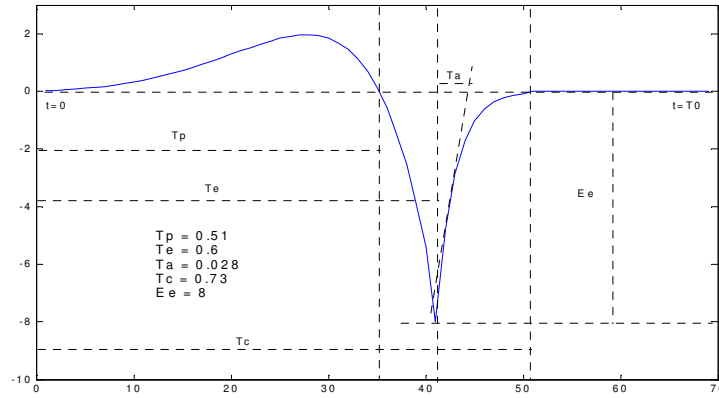


Figure 1 The LF model waveform

#### 2.3.1. Estimation of vocal tract parameters

By using the ARX model, we can construct a system identification architecture. In this structure, the input signal is a glottal waveform  $u_s(n)$  generated by the LF model, and the desired output signal is the reference speech signal  $s(n)$ .

The vocal tract estimation procedure is implemented by an adaptive algorithm based on the Kalman filter. The state transition equation and the measurement equation of the Kalman filter is expressed as equation (9) and (10), respectively

$$\mathbf{x}(n) = \mathbf{x}(n-1) \quad (9)$$

$$s(n) - u_s(n) = \mathbf{H}^T(n)\mathbf{x}(n) + \varepsilon(n) \quad (10)$$

where the state vector of Kalman filter is represented by the filter coefficients vector  $\mathbf{x}(n)$ , that is  $\mathbf{x}(n) = \{a_1(n), \dots, a_p(n), b_1(n), \dots, b_q(n)\}^T$ , the estimation of  $\mathbf{x}(n)$  is updated recursively; the measurement matrix is defined as  $\mathbf{H}(n) = \{-s(n-1), \dots, -s(n-p), u_s(n-1), \dots, u_s(n-q)\}^T$ .

#### 2.3.2. Estimation of source parameters

A signal flow graph of the optimization approach is shown in Figure 2. In the loop-1, the best set of the voice source and formant/antiformant parameters values are obtained. The loop-2 is to continue the analysis from period to period. The iteration to reach the best state is controlled by a temperature T and

an iterative counter  $I$ . When  $T$  becomes lower than a preset minimum value  $T_{\min}$ , the system is regarded as in its best state and the optimal voice source parameters are obtained.

### 3. Adaptive Inverse Filtering

One of the advantages of joint estimation is that we can simultaneously obtain the LF model parameters and the vocal tract filter coefficients. This may provide benefits in some applications, such as speech coding. For other applications such as pathological voice evaluation, people may hope to retain the high accuracy non-parametric glottal source waveform. In this case, the following adaptive inverse filtering equation will be used,

$$\hat{u}_s(n) = \sum_{i=0}^p \hat{a}_i(n)s(n-i) - \sum_{j=1}^q \hat{b}_j(n)\hat{u}_s(n-j) - \varepsilon(n) \quad (11)$$

where  $\hat{a}_i(n)$  and  $\hat{b}_j(n)$  are the estimated AR and MA parametric, respectively.

Since

$$\varepsilon(n) = \sum_{j=1}^q b_j(n)u_n(n-j) + u_n(n) + e(n) \quad (12)$$

Here, we assume the prediction error  $e(n)$  is small enough, then

$$\hat{u}_n(n) = \varepsilon(n) - \sum_{j=1}^q b_j(n)u_n(n-j) \quad (13)$$

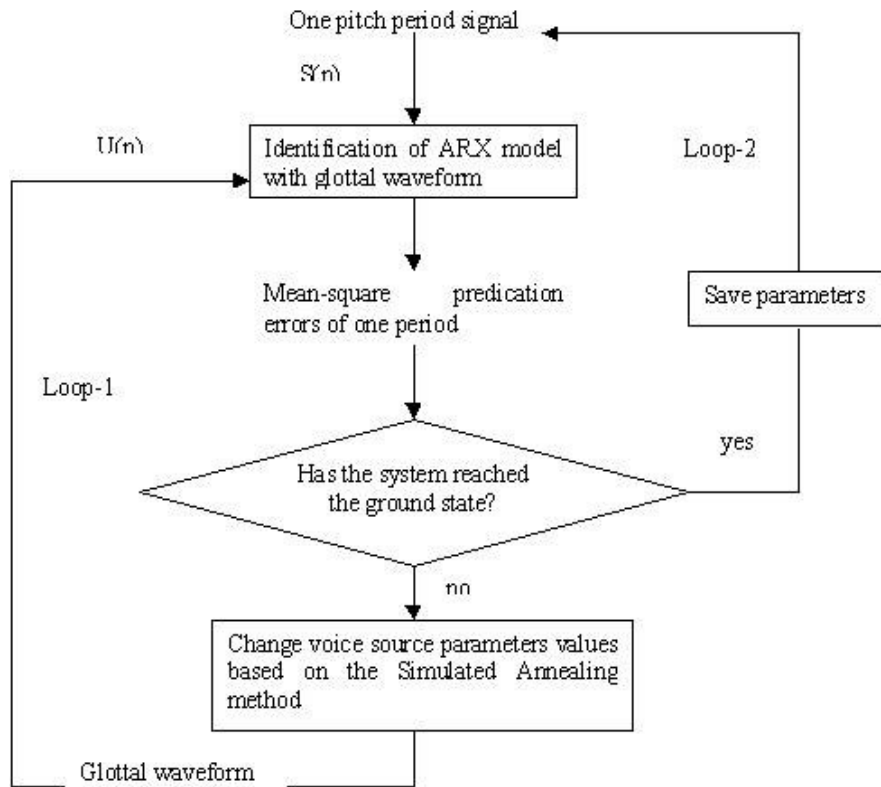


Figure 2 Flowchart of Joint Estimation

### 4. Experiments and discussion

A segment of synthetic speech is used for verifying the validity of this algorithm and performing a comparison with others existing inhere filtering algorithm, such as pitch synchronous closed phase covariance linear prediction (CPLP) and iterative adaptive inverse filtering (IAIF).

The results of the applying the proposed algorithm to synthetic vowels with different glottal noise are shown in Figure 3, 6 and 9, respectively. The synthetic glottal excitation is created using a LF

model and the vocal tract was represented by six formants. Synthetic vowels are analysed using the algorithm and the estimated values of the glottal flow and vocal tract function are compared to the original (real) values. From figure 3, in which the glottal noise is very small, the result shows that the estimated value of LF model was almost equal to its true values. Generally speaking, the proposed provides a similar performance as closed phase LP and generally better than that of IAIF, but it doesn't required the closed phase information, which is hard to get in many situations.

Furthermore, the algorithm can estimate not only the basic shape of the glottal source very accurately but also the glottal noise. This can be easily found out from figure 6 and 9. Compared with other inverse filtering techniques, this algorithm is able to separate the basic shape of glottal source and the glottal noise, thus the Signal to Noise (SNR) of the glottal source can be easily calculated. Table 1 gives the result of estimated SNR using the proposed algorithm.

Real SNR (dB)	40	20	10
Estimated SNR (dB)	37.3	20.1	10.1

Table 1 Estimated SNR of differentiate glottal source

## 5. Conclusions

An adaptive, pitch-synchronous analysis method is proposed for the joint estimation of vocal tract and voice source parameters from speech waveforms. Under a time varying ARX model and, this algorithm has capacity of estimating not only the formants of the vowels but also the antiformants, which is the important to the pathological speech evaluation [12]. Another advantage of this algorithm is that it doesn't require the closed phase information of speech, although it might be helpful for the algorithm. By using the adaptive inverse filtering, a non-parametric glottal source waveform and the glottal noise can also be separately obtained, which is unique among all existing inverse filtering algorithms. In a conclusion, compared with traditional inverse filtering techniques, the proposed algorithm provides an accurate and robust estimation for the glottal source and vocal tract.

## 6. References

- [1] O' Shaughnessy, D. *Speech Communication: Human and Machine*. Reading, Mass.: Addison-Wesley, 1987.
- [2] Wong, D. Y., Markel, J.D., and Gray, Jr. A. H., "Least squares glottal inverse filtering from the acoustic speech waveform", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 4. August 1979. pp. 350-355.
- [3] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", *Speech Communication*, 11, pp. 109-118, 1992.
- [4] Alku, P., Vilkman, E., "Estimation of the glottal pulseform based on discrete all-pole modeling", in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Yokohama, Japan, 1994, pp. 1619-1622.
- [5] Fant, G., Liljencrants, J., and Lin, Q., "A four parameter model of glottal flow", *STL-QPSR* 4, 1985, pp.1-13.
- [6] Krishnamurthy, A. K. "Glottal source estimation using a sum-of exponentials model", *IEEE Trans. on Signal Processing*, Vol. 40, No.3, March 1992, pp.682-686.
- [7] Ding, W., and Kasuya, H., "A novel approach to the estimation of voice source and vocal tract parameters from speech signals", in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.
- [8] Frohlich.M., Michaelis. D., and Strube. H.W., "SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals", *J. Acoust. Soc. Am.* 110(1), pp. 479-488.
- [9] Childers, D.G., and Wong, C.-F., "Measuring and modeling vocal source-tract interaction", *IEEE Trans. Biomed. Eng.*, Vol.41, pp.663-671, July 1994.
- [10] Childers. D.G., Principe, and Ting. Y.T. "Adaptive WRLS-VFF for speech analysis" *IEEE Trans. Speech and Audio Processing*, Vol.3. No.3, pp209-212. 1995.
- [11] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P., "Optimization by simulated annealing", *Science*, Vol. 220, pp. 671-680, 1983.
- [12] Hess W, *Pitch Determination of Speech Signals*. Springer Verlag Berlin, 1983.

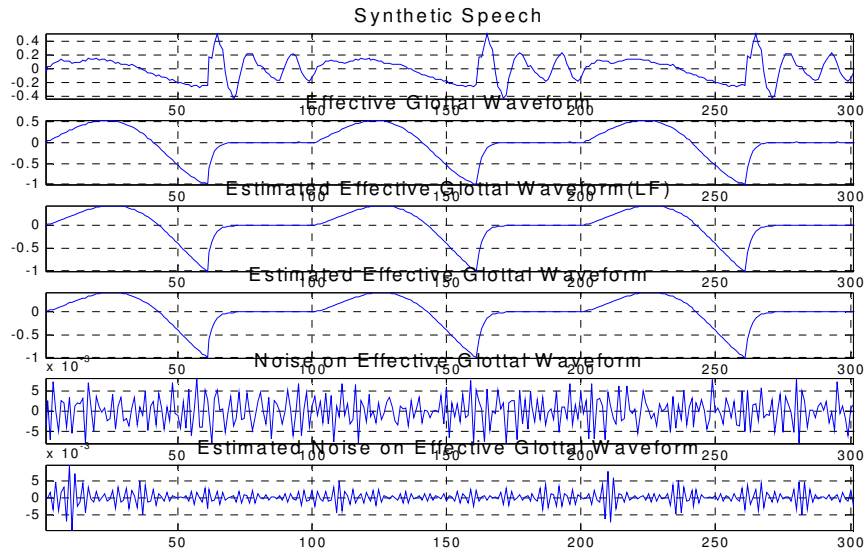


Figure 3 Results using proposed method with synthetic speech (40dB)

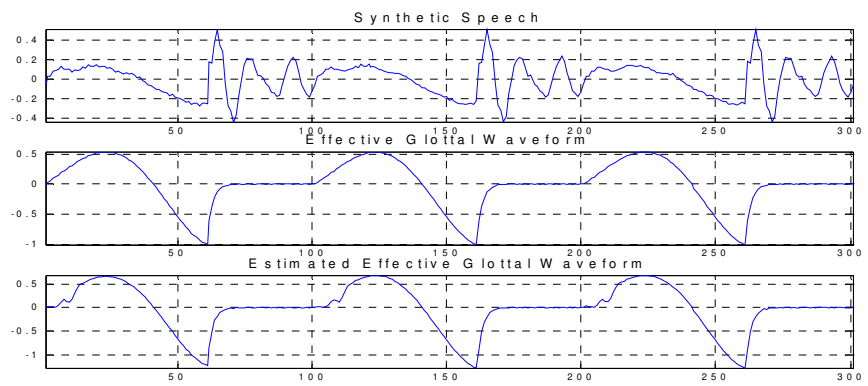


Figure 4 Results with synthetic speech using CPLP (40dB)

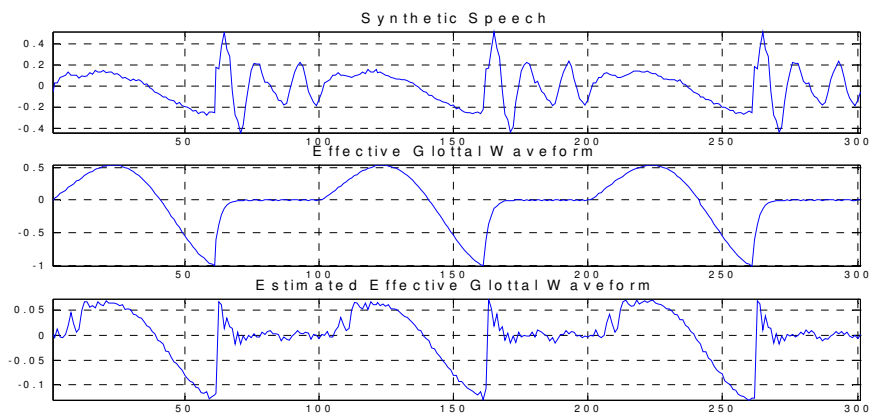


Figure 5 Results with synthetic speech using IAIF (40dB)

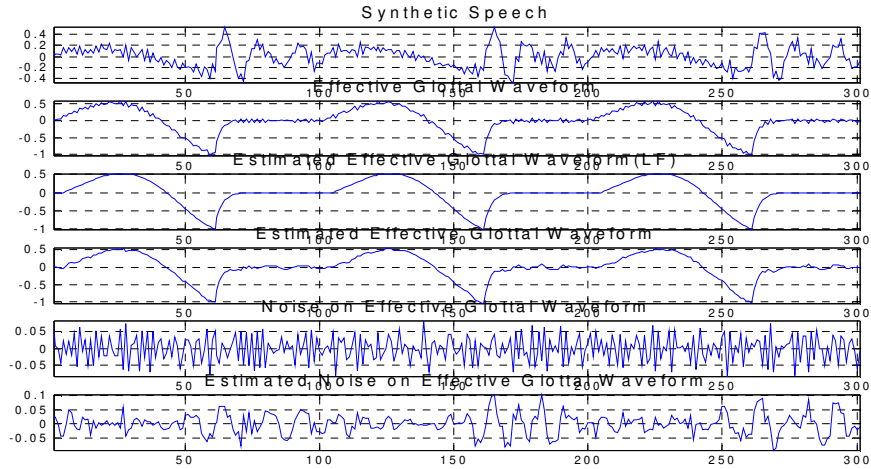


Figure 6 Results using proposed method with synthetic speech (20dB)

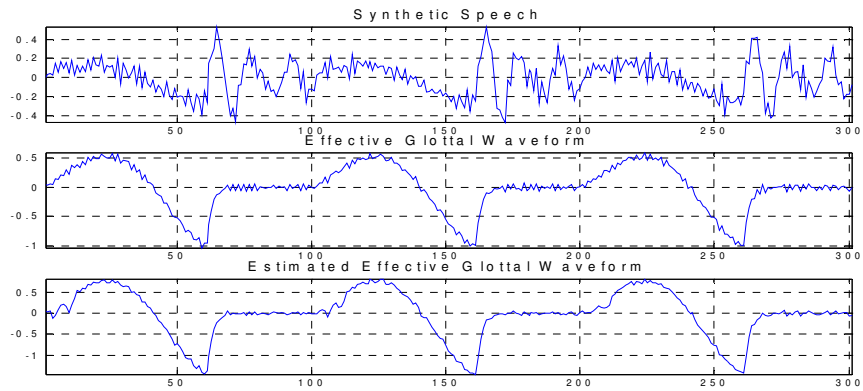


Figure 7 Results with synthetic speech using CPLP (20dB)

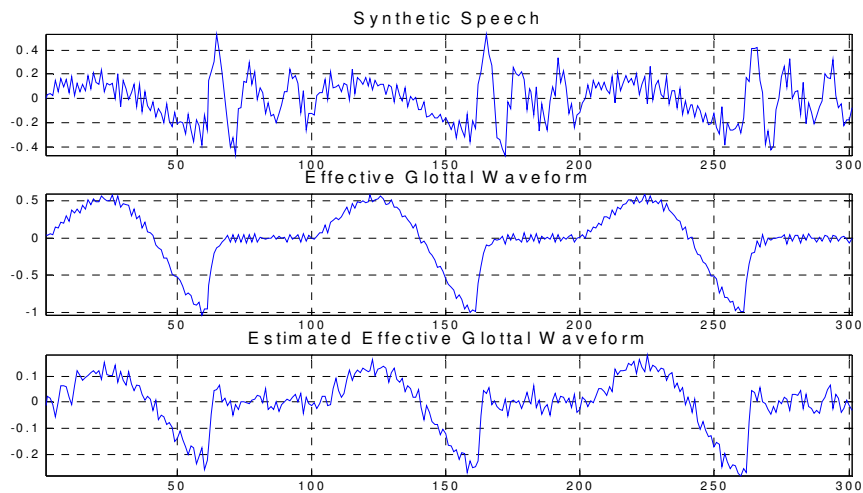


Figure 8 Results with synthetic speech using IAIF (40dB)

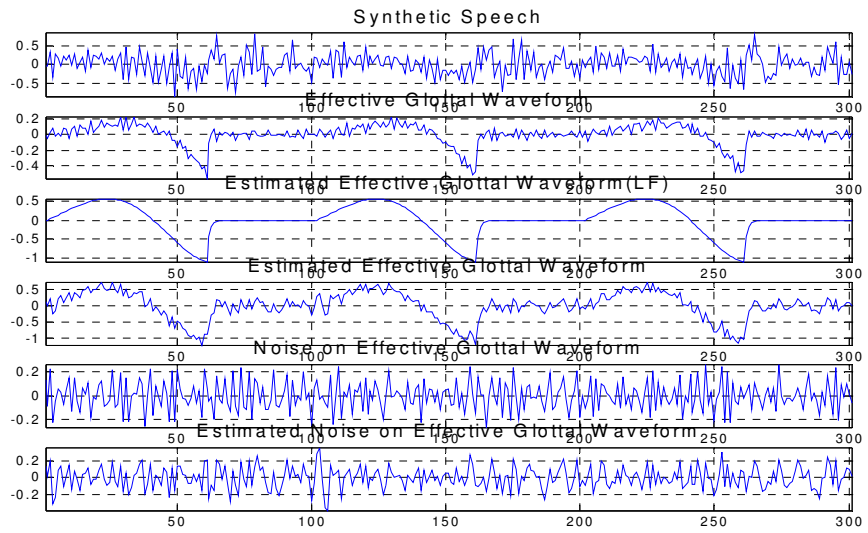


Figure 9 Results using proposed method with synthetic speech (10dB)

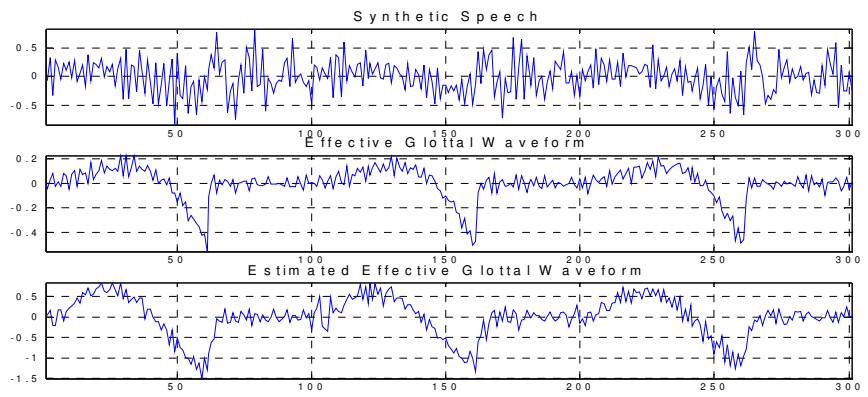


Figure 10 Results with synthetic speech using CPLP (10dB)

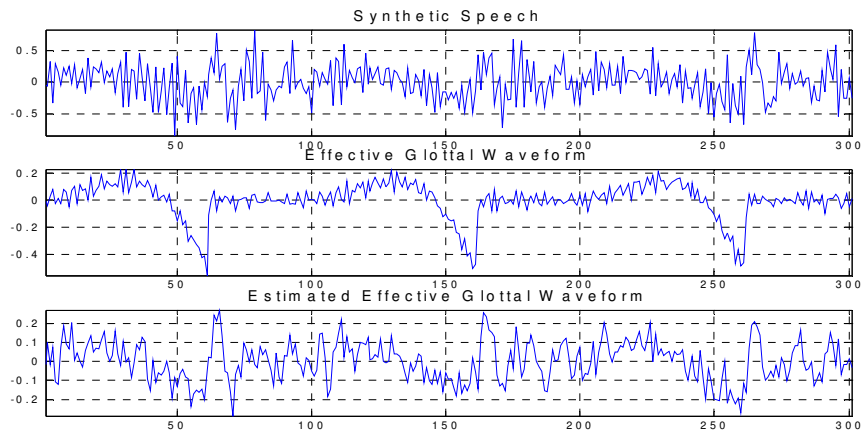


Figure 11 Results with synthetic speech using IAIF (10dB)