DECONVOLUTION AND VOCAL-TRACT PARAMETER ESTIMATION OF SPEECH SIGNALS BY HIGHER-ORDER STATISTICS BASED INVERSE FILTERS

Wu-Ton Chen and Chong-Yung Chi

Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan, Republic of China

Abstract

In this paper, we propose a two-step method for deconvolution and vocal-tract parameter estimation of (non-Gaussian) voiced speech signals. In the first step, the driving input (a non-Gaussian pseudo-periodic positive pulse train) to the vocal-tract filter which can be nonminimum-phase is estimated from speech data by a higher-order statistics (HOS) based inverse filter. In the second step, autoregressive moving average (ARMA) parameters of the vocal-tract filter are estimated with the estimated input and speech data by a prediction error system identification method (an input-output system identification method). Finally, some experimental results with real speech data are provided to justify the good performance of the proposed method.

1. INTRODUCTION

It is well known that speech signals, denoted $x(k)$, can be modeled as the output of a linear system $h(k)$ as follows:

$$x(k) = u(k) * h(k) + w(k) \qquad (1)$$

where the driving input $u(k)$ is a white sequence for unvoiced speech and a non-Gaussian pseudo-periodic positive pulse train for voiced speech, $h(k)$ is the vocal-tract filter (including vocal-cord sound pulse) which can be nonminimum-phase, and $w(k)$ is measurement noise. Accurately estimating the pseudo-periodic positive pulse train $u(k)$ and parameters of $h(k)$, which is also a blind deconvolution problem, is crucial in such as speech recognition, speech coding, synthesis of speech, speaker identification and diagnosis of speech organs. A lot of methods for speech deconvolution have been reported such as pitch synchronous analysis [1], predictive deconvolution [2], homomorphic deconvolution [3], iterative inverse filtering [4], system identification based methods [5, 6] and Bernoulli-Gaussian model based deconvolution [7].

Recently, higher-order (≥ 3) statistics (HOS) (known as cumulants) based inverse filter criteria [8-12] have been used for the identification and deconvolution of nonminimum-phase systems with only non-Gaussian measurements because cumulants of non-Gaussian measurements can be used to extract not only the amplitude information but also the phase information of the unknown system $H(z)$; meanwhile, cumulants of additive Gaussian noise are totally zero. On the other hand, there are some spectral estimators [13, 14] which first estimate the unknown input by a whitening filter such as a (minimum-phase) linear prediction error (LPE) filter with a large order and then estimate the spectrally equivalent system $H_{MP}(z)$ with the estimated input, which actually corresponds to an estimate of the output of an allpass system $H(z)/H_{MP}(z)$ driven by the true input, and output measurements by an input-output system identification algorithm.

In this paper, we propose a two-step method for deconvolution and vocal-tract parameter estimation of voiced speech signals. Conceptually, the proposed method first estimates the non-Gaussian pseudo-periodic positive pulse train $u(k)$ input to the vocal-tract filter $h(k)$ by an HOS based inverse filter $1/H(z)$ and then estimates the autoregressive moving average (ARMA) parameters of $H(z)$ from the estimated $u(k)$ and measurements $x(k)$ by an input-output identification algorithm.

2. DECONVOLUTION AND VOCAL-TRACT PARAMETER ESTIMATION

First of all, let us briefly review a prediction error system identification method (an input-output system identification method) for succinct presentation of the new two-step deconvolution and vocal-tract parameter estimation method. Assume that $\widehat{u}(k)$ (corresponding to input measurements) is the given estimate of $u(k)$ and that the unknown vocal-tract filter $h(k)$ is an ARMA(p,q) model. The prediction error system identification method estimates the ARMA parameters of $h(k)$ by minimizing the following nonlinear objective function

$$\mathbf{J} = \sum_{k=k'}^{N-1} (\widehat{x}(k) - x(k))^2 \qquad (2)$$

0-7803-1238-4/93 $3.00 © 1993 IEEE

where N is the total number of measurements,

$$\widehat{x}(k) = \widehat{u}(k) * h(k)$$

$$= -\sum_{i=1}^{p} a_i\,\widehat{x}(k-i) + \sum_{i=0}^{q} b_i\,\widehat{u}(k-i) \qquad (3)$$

and $k' = \max\{p, q\}$. The initial conditions $\widehat{x}(k) = x(k)$, $k = 0, 1, ..., k'{-}1$, can be used in computing $\widehat{x}(k)$ for $k \geq k'$. Newton-Raphson type iterative numerical search algorithms can be used to find a local minimum of $\mathbf{J}$. Note that the reconstructed signal $\widehat{x}(k)$ can be viewed as an approximation to $x(k)$ and $\mathbf{J}$ is the sum of approximation error squares.

The new two-step deconvolution and parameter estimation method is described in the following.

Step 1. Estimation of $u(k)$ by HOS Based Inverse Filters (Deconvolution):
Assume that $g(k)$ is the inverse filter of $h(k)$, i.e., $h(k) * g(k) = \delta(k)$. One can use existing HOS based inverse filters such as those reported in [8-12], which will provide an estimate $\widehat{f}(k) = \alpha\, g(k-\tau)$ where $\alpha \neq 0$ is a scale factor and τ is an unknown integer (time delay), to obtain the input estimate $\widehat{u}(k{-}\tau) = x(k) * \widehat{f}(k)$ except for a scale factor.

Step 2. Estimation of ARMA Parameters of $h(k)$:
Assume that $\widehat{f}(k)$ is a causal FIR filter of order L. Let $T = \{k_1, k_1{+}1, ..., k_2\}$ where $0 \leq k_1 \leq k_2 \leq L$. For $n = k_1, k_1{+}1, ..., k_2$, find the minimum $\mathbf{J}(n)$ given by Eq. (2) with the input estimate $\widehat{u}(k)$ equal to $\widehat{u}(k{-}\tau{+}n)$. The desired ARMA parameters of $h(k)$ are those associated with $\mathbf{J}(\widehat{\tau}) = \min\{\mathbf{J}(n), n \in T\}$.

Regarding the determination for the set T in Step 2, an intuitive choice is $T = \{1, ..., L\}$ without considering computational load. Two more practical methods for determining T are suggested as follows.

The first method is based on the assumption that either the positive peak or the negative peak of the true inverse filter $g(k)$ is located near the origin. Let τ_1 and τ_2 be the integers associated with $\widehat{f}(\tau_1) = \max\{\widehat{f}(k) > 0,\ k = 0, 1, ..., L\}$ and $\widehat{f}(\tau_2) = \min\{\widehat{f}(k) < 0,\ k = 0, 1, ..., L\}$, respectively. The integers k_1 and k_2 can be chosen as $k_1 = \min(\tau_1, \tau_2) - \Delta_1 \geq 0$ and $k_2 = \max(\tau_1, \tau_2) + \Delta_2 \leq L$ where $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ are integers.

Let $\widehat{u}'(k)$ denote the estimated $u(k)$ obtained through minimum-phase - allpass (MP-AP) deconvolution [15] which includes a whitening process with an LPE filter followed by a process with an HOS based allpass system deconvolution filter. The second method is based on the fact that $\widehat{u}'(k)$ is a scaled estimate of the true $u(k)$ without any time delay, although $\widehat{u}(k{-}\tau)$ approximates the quasi-periodic positive pulse train much better than $\widehat{u}'(k)$ when L is large because the performance of whitening filters used in the MP-AP deconvolution is sensitive to noise. Assume that τ' is the time delay between $\widehat{u}'(k)$ and $\widehat{u}(k-\tau)$. The integers k_1 and k_2 can be chosen as $k_1 = \tau' - \Delta_1 \geq 0$ and $k_2 = \tau' + \Delta_2 \leq L$ where $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$ are integers. Therefore, $\widehat{u}'(k)$ too must be obtained in addition to $\widehat{u}(k{-}\tau)$ in Step 1 if this determination method for T is used.

3. EXPERIMENTAL RESULTS WITH REAL SPEECH DATA

Let us present some tests to the proposed two-step deconvolution and parameter estimation method with the speech data shown in the top part of Figure 1(a) which were obtained by lowpass filtering the continuous speech signal with the cutoff frequency equal to 3 kHz followed by the A/D conversion with a sampling rate equal to 10 kHz.

In Step 1, the inverse filter $f(k)$ was assumed to be a causal FIR filter of order $L = 80$. Wiggins' algorithm [8] was used to find the optimum $\mathbf{f} = [f(0), f(1), ..., f(L)]'$ such that the varimax norm defined as

$$\mathrm{V}(\mathbf{f}) = \frac{\sum_{k=L}^{N-1} y^4(k)}{\{\sum_{k=L}^{N-1} y^2(k)\}^2} \qquad (4)$$

is maximum where $y(k)$ is the output of the inverse filter with input $x(k)$ (speech data). The estimate $\widehat{u}(k-\tau)$ (the optimum $y(k)$) and the associated optimum $\widehat{f}(k)$ obtained in Step 1 are shown in Figures 1(b) and 1(c), respectively. One can observe, from Figure 1(b), that $\widehat{u}(k-\tau)$ approximates a quasi-periodic positive pulse train amazingly well with pitch period $P = 67$ and the fluctuation (i.e., variance) of $\widehat{u}(k-\tau)$ between any two consecutive positive pulses is quite small. These results also indicate that HOS based inverse filters of large order can be high-resolution speech deconvolution filters.

In Step 2, the optimum $\widehat{\tau}$ was found to be 49. The obtained optimum ARMA parameters $\widehat{a}_i$ and $\widehat{b}_i$ for $p = q = 16$ are shown in Table 1. The associated vocal-tract filter $\widehat{h}(k)$, the amplitude response $|\widehat{H}(\omega)| = |\widehat{H}(z{=}exp\{j\omega\})|$ and locations of poles (×'s) and zeros (circles) of $\widehat{H}(z)$ are shown in Figures 1(d), 1(e) and 1(f), respectively. Obviously, $\widehat{h}(k)$ is nonminimum-phase because eight zeros of $\widehat{H}(z)$ are located outside the unit circle. Note that the length of $\widehat{h}(k)$ is much longer than the pitch period $P = 67$. One can see, from Figure 1(a), that the reconstructed speech signal $\widehat{x}(k)$ (bottom part) is very close to $x(k)$ (top part). To show its usefulness, one more set of experimental results with real speech data obtained through the same 2-step procedure is presented below.

The speech data (shown in the top part of Figure 2(a)) of sound /n/ uttered by a man were obtained by the same way as we obtained the speech data shown in the top part of Figure 1(a). The order of the inverse filter used was also equal to 80. The optimum $\widehat{\tau}$ was equal to 49. The results corresponding to those shown in Figures 1(a)–1(d) and 1(f) are shown in Figure 2. Again, from Figure 2, one can observe that the estimate $\widehat{u}(k-\tau)$ approximates a pseudo-periodic positive pulse train well with pitch period $P=60$ (see Figure 2(b)), that the estimated vocal-tract filter $\widehat{h}(k)$ is nonminimum-phase with length much longer than the pitch period (see Figures 2(d) and 2(e)) and that the reconstructed speech signal $\widehat{x}(k)$ is a very good approximation to the the original speech data (see Figure 2(a)). The capability of recovering the phase of $h(k)$ is because the estimate $\widehat{u}(k)$ is the output of the inverse filter $1/H(z)$ instead of the output of $1/H_{MP}(z)$ as in spectral estimators based on input-output identification algorithms. These experimental results justify that the proposed two-step deconvolution and vocal-tract parameter estimation method works well.

4. DISCUSSION AND CONCLUSIONS

In this paper, we presented a two-step deconvolution and vocal-tract parameter estimation method, which is an integration of HOS based inverse filters and input-output system identification algorithms, for voiced speech. In the presented tests to the proposed method with real speech data, the inverse filter obtained by Wiggins' algorithm, which, by our experience, outperforms other inverse filters reported in [9-12] besides computational efficiency, was used to estimate the non-Gaussian pseudo-periodic positive pulse train input to the vocal-tract filter in Step 1. Experimental results showed that the proposed method works well because HOS based inverse filters with high order are able to accurately estimate the pseudo-periodic positive pulse train. Moreover, the high-resolution deconvolution results presented in Section 3 (see Figures 1(b) and 2(b)) indicate that HOS based signal processing algorithms are very promising in high-quality speech processing. In addition, accurate estimation of vocal-tract filter including both amplitude and phase by HOS based signal processing algorithms opens the avenue of using phase related new features in the applications mentioned in the introduction section.

ACKNOWLEDGEMENTS

The research presented in this paper was supported by the National Science Council under Grant NSC82-0404-E-007-219 and the Telecommunication Laboratories under Grant TL-82-5209.

REFERENCES

[1] M. V. Mathews, J. E. Miller and E. E. David, Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Am.*, vol. 33, pp. 179-186, 1961.

[2] J. D. Markel and A. H. Gray, Jr., *Linear prediction of speech*, Springer-Verlag, New York, 1976.

[3] G. E. Kopec, A. V. Oppenheim, and J. M. Tribolet, "Speech analysis by homomorphic prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 40-49, Feb. 1977.

[4] Ira S. Konvalinka and Miroslav R. Atausek, "Simultaneous estimation of poles and zeros in speech analysis and ITIF-iterative inverse filtering algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 5, pp. 485-492, June 1979.

[5] Y. Miyanaga, N. Miki, N. Nagai, and K. Hatori, "A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 88-96, Feb. 1982.

[6] H. Morikawa and H. Fujisaki, "System identification of the speech production process based on a state-space representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 252-262, Apr. 1984.

[7] C.-Y. Chi and W.-T. Chen, "A novel adaptive maximum-likelihood deconvolution algorithm for estimating positive sparse spike trains and its application to speech analysis," *Proc. IEEE International Workshop on Intelligent Signal Processing and Communication Systems*, pp. 388-402, Taipei, Taiwan, Republic of China, March 1992.

[8] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol. 16, pp. 21-35, 1978.

[9] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. Infor. Theory*, vol. IT-36, pp. 312-321, March 1990.

[10] J. K. Tugnait, "Inverse filter criteria for estimation of linear parametric models using higher order statistics," *Proc. IEEE 1991 Intern. Conf. Acoustics, Speech, and Signal Processing*, pp. 3101-3104, Toronto, Canada, May 1991.

[11] C.-Y. Chi and J.-Y. Kung, "A new cumulant based inverse filtering algorithm for identification and deconvolution of nonminimum-phase systems," *Proc. Sixth IEEE SP Workshop on Statistical Signal and Array Processing*, pp. 144-147, Victoria, B.C., Canada, Oct. 1992.

[12] W.-T. Chen and C.-Y. Chi, "New inverse filter criteria for identification and deconvolution of nonminimum-phase systems by single cumulant slice," *Proc. IEEE 1993 International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, Minnesota, April 1993.

[13] J. Durbin, "The fitting of time-series models," *Revue Inst. Int. Statis.* vol. 28, pp. 233-243, 1960.

[14] D. Q. Mayne and F. Firoozan, "An efficient, multistage, linear identification method for ARMA processes," *Proc. IEEE Conf. Decision and Control*, New Orleans, vol. 1, pp. 435-438, Dec. 1977.

[15] C.-Y. Chi and J.-Y. Kung, "A new identification algorithm for allpass systems by higher-order statistics," to appear in *Signal Processing*, vol. 34, no. 1, Jan. 1994.

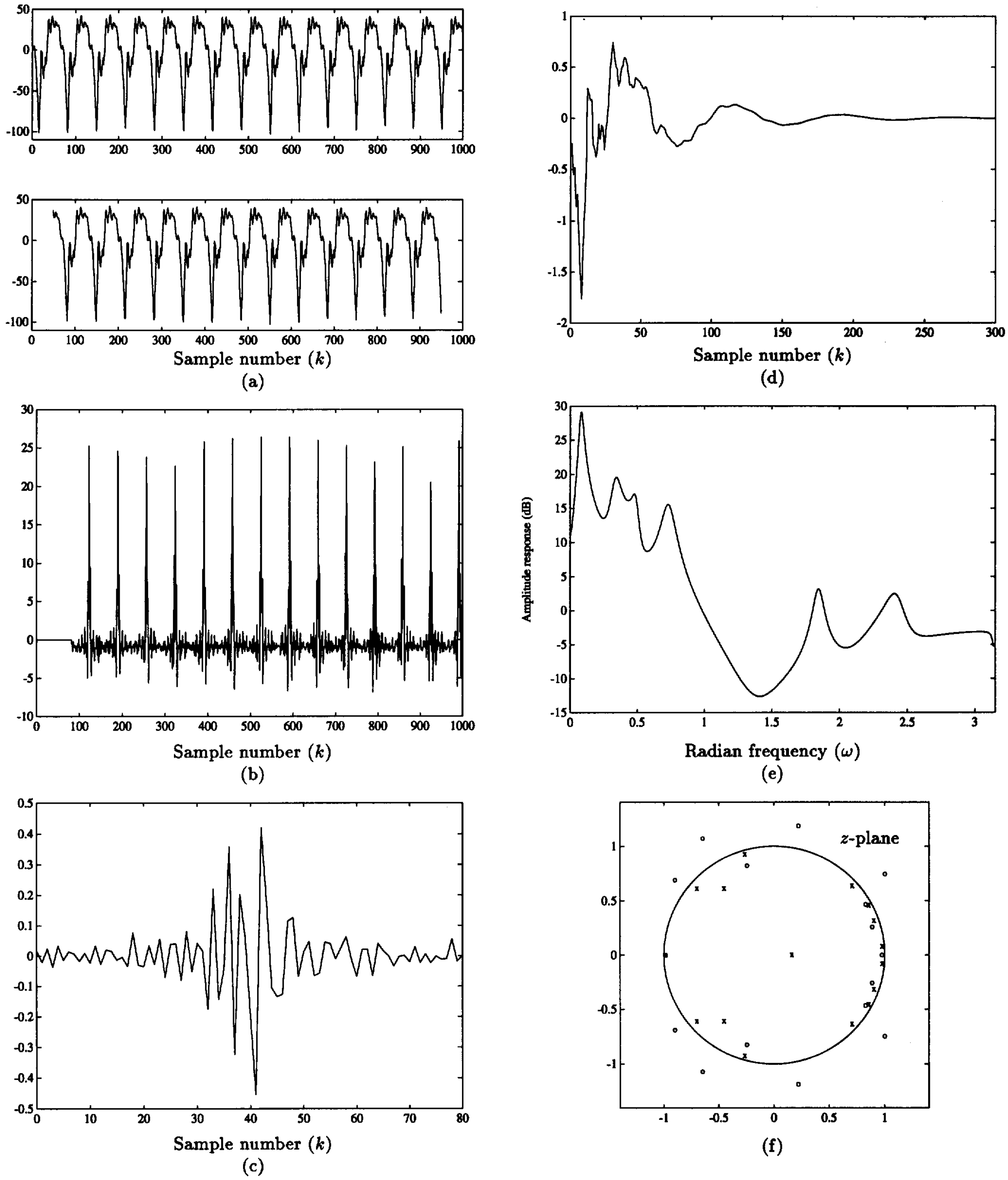
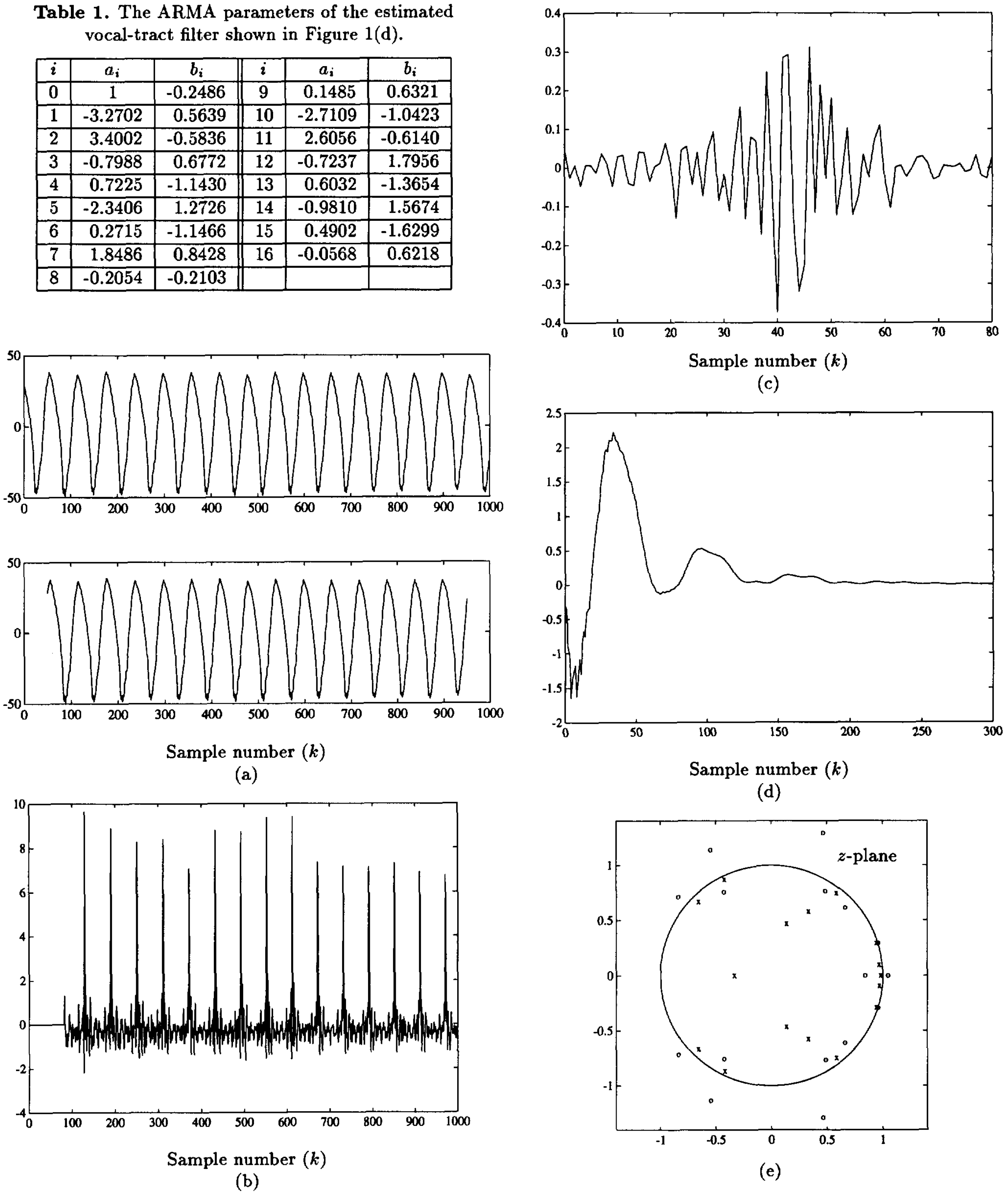


Figure 1. Experimental results with real speech data of sound /*a:*/ uttered by a man (sampling rate equal to 10 kHz). (a) The original speech data (top part) and the reconstructed speech signal (bottom part), (b) deconvolved results by the Wiggins' inverse filter (a causal FIR filter) $\widehat{f}(k)$ of order $L = 80$, (c) $\widehat{f}(k)$, (d) the estimated vocal-tract filter $\widehat{h}(k)$, (e) the amplitude response $|\widehat{H}(\omega)|$ and (f) locations of poles (×'s) and zeros (circles) of $\widehat{H}(z)$.

Table 1. The ARMA parameters of the estimated vocal-tract filter shown in Figure 1(d).

i	a_i	b_i	i	a_i	b_i
0	1	-0.2486	9	0.1485	0.6321
1	-3.2702	0.5639	10	-2.7109	-1.0423
2	3.4002	-0.5836	11	2.6056	-0.6140
3	-0.7988	0.6772	12	-0.7237	1.7956
4	0.7225	-1.1430	13	0.6032	-1.3654
5	-2.3406	1.2726	14	-0.9810	1.5674
6	0.2715	-1.1466	15	0.4902	-1.6299
7	1.8486	0.8428	16	-0.0568	0.6218
8	-0.2054	-0.2103			



Sample number (k)
(c)

Sample number (k)
(a)

Sample number (k)
(d)

Sample number (k)
(b)

(e)

Figure 2. Experimental results with real speech data of sound /n/ uttered by a man (sampling rate equal to 10 kHz). (a) The original speech data (top part) and the reconstructed speech signal (bottom part), (b) deconvolved results by the Wiggins' inverse filter (a causal FIR filter) $\widehat{f}(k)$ of order $L = 80$, (c) $\widehat{f}(k)$, (d) the estimated vocal-tract filter $\widehat{h}(k)$ and (e) locations of poles (×'s) and zeros (circles) of $\widehat{H}(z)$.