represents the assumed transfer function variation. Here the algorithm gain, interpreted in the frequency domain, is represented by $\sqrt{\hat{r}_1(\omega)/\hat{r}_2}$, and we see that the first and second terms in (37) are proportional and inversely proportional, respectively, to this quantity.

IV. Conclusions

We have derived approximate expressions for the mean-square error of transfer function estimates for time-varying systems with finite impulse response when the system output is affected by correlated disturbances. The approximate MSE expressions, (34), (36), and (37), illustrate the tradeoff between tracking capability and disturbance rejection. They furthermore show the influence of design variables, model order, system variation, and signal properties on the model quality.

References

[1] S. Gunnarsson and L. Ljung, "Frequency domain tracking characteristics of adaptive algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1072–1089, 1989.

[2] A. Benveniste, "Design of adaptive algorithms for the tracking of time-varying systems," *Int. J. Adaptive Contr. Signal Processing*, vol. 1, pp. 3–29, 1987.

[3] L. Ljung and S. Gunnarsson, "Adaptation and tracking in system identification—a survey," *Automatica*, vol. 26, pp. 7–21, 1990.

[4] S. Gunnarsson, "Frequency domain description of the quality of recursively identified FIR models in the presence of correlated disturbances," Tech. Rep. LiTH-ISY-I-1057, Dep. Elec. Eng., Linköping University, Linköping, Sweden, 1990.

Glottal Source Estimation Using a Sum-of-Exponentials Model

Ashok K. Krishnamurthy

***Abstract*—This correspondence describes an algorithm for estimating the glottal source waveform in voiced speech. The glottal source waveform is described using the LF model proposed by Fant *et al.* The vocal tract filter is modeled as a pole-zero system. The analysis of vowel sounds from several talkers shows that the analysis procedure leads to an accurate estimate of the glottal source.**

I. Introduction

There is clear evidence that the glottal source waveform has a significant impact on voice quality [1], [2]. Differences between male and female voices [2], and between different phonation types (e.g., normal, breathy, pressed, etc.) appear to be related to systematic differences in the characteristics of the glottal source waveform [1]. To achieve natural sounding synthetic speech, it is essential that we understand and model these glottal source variations.

One of the major difficulties in using realistic glottal source models in synthesis is the lack of suitable analysis methods to estimate the glottal waveform parameters from natural speech. The most common technique for examining the glottal source signal is inverse filtering [3], [4]. But inverse filtering is subject to many restrictive assumptions, and is frequently impossible. To overcome the limitations of inverse filtering, several researchers have developed analysis methods that simultaneously estimate the parameters of the vocal tract filter and the glottal source waveform [5]–[7]. However, these results assume that the overall glottal source/vocal tract system is time invariant during the glottal cycle. The source/tract system, though, is in fact time varying due to coupling between the subglottal and supraglottal systems [8]. The most predominant effects of this coupling is that the first formant frequency and bandwidth are different in the closed and open glottal phases. Parthasarathy and Tufts [9] allow for this effect by using different models for the vocal tract filter in the closed and open phases. However, they do not estimate the glottal source, but instead model the glottal excitation by means of initial conditions.

This correspondence develops an analysis procedure for estimating the glottal source signal that overcomes these limitations. The effective voice source (which is the derivative of the glottal source waveform, see below) is modeled by the so-called LF model proposed by Fant *et al.* [10]. The source-tract interaction effect is modeled by allowing the vocal tract filter parameters to be different in the closed and open phases. The voiced speech model is described in the next section, followed by a description of the analysis procedure. Results from the analysis of several vowel sounds are then presented.

II. Glottal Source and Vocal Tract Models

Consider the speech production model shown in Fig. 1. The vocal tract filter $V(z)$ models the transfer function relating the volume velocity at the lips to the glottal volume velocity. The input to the vocal tract filter in this model is the differentiated glottal source signal, and not the glottal volume velocity. The differentiated glottal source signal, $q(n)$, called here the effective voice source signal, models the combined effects of the glottal volume velocity and the radiation at the lips (which is typically modeled as a differentiator). The output of the system is thus the radiated sound pressure signal $s(n)$.

A. LF Model for the Effective Voice Source

We use the so-called LF model, proposed by Fant *et al.* [10] for the effective voice source $q(n)$. The LF model in continuous time is given by

$$e(t) = \begin{cases} E_0 e^{\alpha t} \sin(\Omega_g t), & 0 \leq t < t_e \\ -\dfrac{E_0}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c - t_e)}], & t_e \leq t < t_c \end{cases} \tag{1}$$

where $e(t)$ is the differentiated glottal volume velocity.

A slightly modified version of the above equation is used to define the effective voice source in the discrete-time domain. 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 $q(n)$ is defined as

$$q(n) = \begin{cases} A_{go} e^{\alpha_{go} n} \sin(\omega_{go} n + \phi_{go}), & n = 0, \cdots, N-1 \\ -A_{gc} e^{-\alpha_{gc}(n-N)}, & n = N, \cdots, M-1. \end{cases} \tag{2}$$

Manuscript received March 19, 1990; revised March 17, 1991.

The author is with the Department of Electrical Engineering, Ohio State University, Columbus, OH 43210.

IEEE Log Number 9105676.

1053-587X/92$03.00 © 1992 IEEE

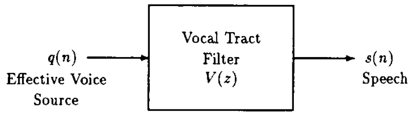

Fig. 1. Speech production model.

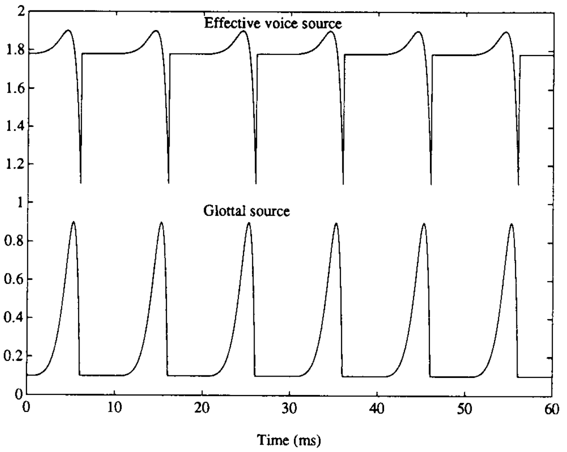

Fig. 2. Effective voice source $q(n)$, and the corresponding glottal source signal.

The above equation can be rewritten using complex exponentials as

$$q(n) = \begin{cases} C_o z_{go}^n + C_o^*(z_{go}^*)^n, & n = 0, \cdots, N-1 \\ C_c z_{gc}^{n-N}, & n = N, \cdots, M-1 \end{cases} \tag{3}$$

where * is the complex conjugate, and

$$C_o = 0.5\, A_{go}\, e^{j(\phi_{go} - \pi/2)}, \quad z_{go} = e^{\alpha_{go} + j\omega_{go}} \quad C_c = -A_{gc}, \quad \text{and}$$
$$z_{gc} = e^{-\alpha_{gc}}. \tag{4}$$

The effective voice source is modeled as a growing sinusoid in the open phase (since α_{go} is usually positive), and a decaying exponential in the closed phase. Also, ϕ_{go} is an additional parameter used in the discrete time model. In the ideal case, ϕ_{go} should be 0; however, leaving it as a free parameter in (2) allows for the case where $q(n)$ in the open phase does not start at 0. This can occur either because of an error in determining the glottal opening instant or because the opening instant does not coincide with a sampling instant. In any case, the use of ϕ_{go} allows for the offset between the actual and estimated glottal opening instant to be more than one sample point.

An example of a typical effective voice source waveform and the corresponding glottal source signal, synthesized using this model, are shown in Fig. 2.

B. Vocal Tract Model

The vocal tract filter, $V(z)$, is modeled as a pole-zero system with transfer function $V_C(z)$ during the closed phase and $V_O(z)$ during the open phase

$$V(z) = \begin{cases} V_C(z) = \dfrac{B_c(z)}{A_C(z)}, & \text{closed phase} \\ V_O(z) = \dfrac{B_O(z)}{A_O(z)}, & \text{open phase.} \end{cases} \tag{5}$$

also, let the poles in the closed and open phases be

$$A_C(z) = \prod_{i=1}^{K_C} [1 - z_C(i)z^{-1}][1 - z_C^*(i)z^{-1}] \tag{6}$$

$$A_O(z) = \prod_{i=1}^{K_O} [1 - z_O(i)z^{-1}][1 - z_O^*(i)z^{-1}]. \tag{7}$$

C. Speech Signal Model

The speech signal $s(n)$ is the output of the vocal tract filter when the input is the effective voice source signal. If we assume that the effective voice source is given by the LF model, and that the vocal tract filter is linear and time invariant in each of the glottal phases, then the output speech signal $\hat{s}(n)$ of the model will be the sum of exponential signals; i.e.,

$$\hat{s}(n) = \begin{cases} B_{go}z_{go}^n + B_{go}^*(z_{go}^*)^n + \sum_{i=1}^{K_O} [B_O(i)\{z_O(i)\}^n + B_O^*(i)\ \{z_O^*(i)\}^n] \\ \qquad n = 0, \cdots, N-1 \\ B_{gc}z_{gc}^{(n-N)} + \sum_{i=1}^{K_C} [B_C(i)\ \{z_C(i)\}^{n-N} + B_C^*(i)\ \{z_C^*(i)\}^{n-N}] \\ \qquad n = N, \cdots, M-1. \end{cases} \tag{8}$$

In the above equation, K_O complex conjugate pairs (in the open phase) and K_C complex conjugate pairs (in the closed phase) of the exponential components correspond to the vocal tract poles, and the remaining components correspond to the effective voice source. For each of the complex conjugate pairs, we need to estimate the complex amplitude and the complex exponent, i.e., 4 real parameters. Thus a total of $4K_O + 4K_C + 6$ real parameters need to be estimated in each pitch period.

An interesting feature of this speech signal model is that, along with the glottal source, the contribution of each formant to the speech waveform can also be estimated. The component due to a formant is the complex conjugate pair of terms in (8) that correspond to the poles of the formant. The model is thus useful, with only small modifications, in parallel formant synthesis [11].

III. Analysis Procedures

The analysis to estimate the model parameters consists of 2 stages. In the first stage, each pitch period of the speech signal is isolated, and separated into the closed and open glottal phases. Then a separate analysis is carried out in each phase to determine the parameters of the model. The two steps are described below. Note that no preemphasis of the speech signal is done prior to the analysis.

A. Locating the Closed and Open Phases

The first step in the analysis procedure is to locate the closed and open glottal phases in each pitch period. We use the electroglottograph (EGG) signal, sampled simultaneously with the speech signal for this purpose. Basically, the closing instants are located as negative peaks, and the opening instances are located as positive peaks in the differentiated EGG (D-EGG) signal. The relevant algorithms are described in [3], [12]. An example of the open and closed glottal phases located using this procedure is shown in Fig. 3. Note that recent work [13]-[15] may lead to methods of determining the closed and open phases from the speech signal alone.

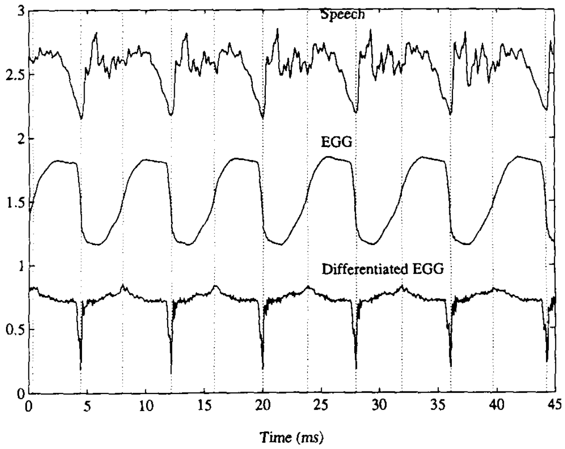

Fig. 3. Locating the closed and open phases using the EGG. The sharp negative peaks in the EGG mark the closing instants, and the smaller positive peaks mark the opening instant.

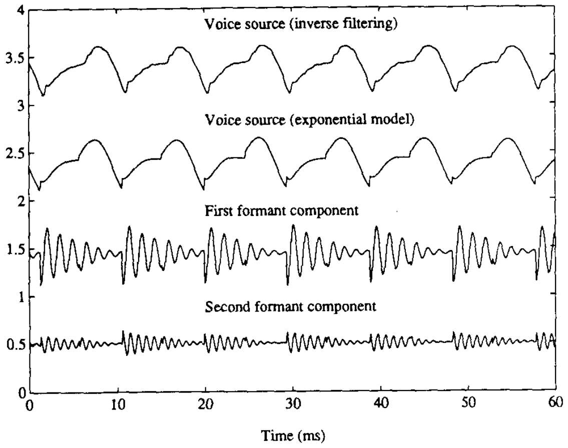

Fig. 4. Effective voice source from inverse filtering (top graph); effective voice source from exponential model (second from top); first formant component (third from top); and second formant component (bottom). (All waveforms are drawn to the same scale.)

B. Estimating the Glottal Source and the Vocal Tract Parameters

Once the open and closed phases have been located, the next step in the analysis is to determine the parameters of the exponential signal model. A separate analysis is done in each phase, leading to two sets of parameters for each pitch period. The errors

$$E_O = \sum_{n=0}^{N-1} [s(n) - \hat{s}(n)]^2 \tag{9}$$

$$E_C = \sum_{n=N}^{M-1} [s(n) - \hat{s}(n)]^2 \tag{10}$$

are minimized in the open and closed phases, respectively, where $\hat{s}(n)$ is the speech signal from the model as given by (8).

It is well known that direct minimization of E_C or E_O leads to a nonlinear problem; therefore, a two step procedure is used to estimate the parameters [9], [16]. In the first step, a backward prediction polynomial of order L is computed. Usually, L is larger than the minimum order required to estimate all the components. Also, a truncated singular value decomposition (SVD) based solution is used to obtain accurate estimates with reduced variance [16]. The complex exponential parameters are related to a subset of the roots of the backward prediction polynomial. In the second step, the complex amplitude parameters are computed by solving a second set of linear equations. The energy of each of the possible complex modes is then computed, and the effective voice source and the vocal tract parameters are identified as the complex modes with the highest energies. The details of this procedure are described in [9], [16], and are omitted here.

We show the results from the analysis of a steady vowel sound spoken by a male talker. The vowel is /a/, and the speech and EGG signals were simultaneously sampled at 10 KHz each using a 16-b A/D converter in a laboratory environment. In both the open and closed phases, the higher order predictor length was 21, and all but the 6 largest singular values were truncated.

Fig. 4 compares the effective voice source estimated by inverse filtering [3] (top graph) with that estimated using the model described in this correspondence (second graph from the top). The two waveforms are remarkably similar, with the LF model source being much smoother. For example, the short step at the start of the closed phase in the inverse filtered signal has its counterpart in the effective voice source. Note however that the effective voice source from the model includes a sinusoidal component in the glottal closed phase. Also shown in this figure are the first formant (third graph from the top) and second formant (bottom graph) components in the speech signal. A small discontinuity can be observed at the opening instant in the first and second formant waveforms. Since the parameters in the two phases are estimated independently, small differences in the waveforms at the points that they fit together give rise to these discontinuities.

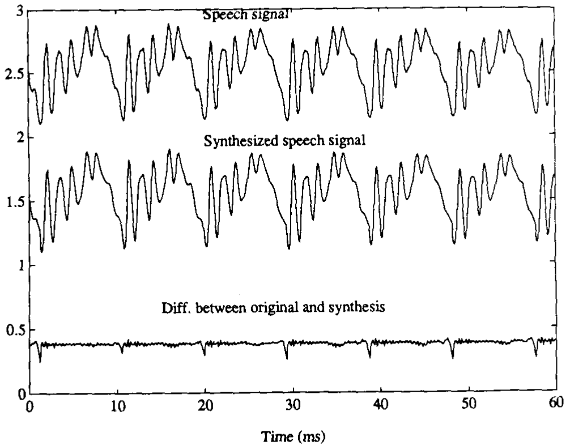

Fig. 5. Original speech signal (top); synthesized speech signal (middle); and difference between the original and synthesis (bottom). (All plots are drawn to the same scale.)

The average first formant frequency and bandwidth for this example are 825 and 74 Hz, respectively, in the closed phase, and 1070 and 241 Hz, respectively, in the open phase.

Fig. 5 shows the original speech signal (top waveform), the synthesized speech signal using (8) (middle waveform), and the difference between the two (bottom waveform). All three waveforms are drawn to the same scale. The similarity between the original and synthesized speech waveforms is remarkable, even though only

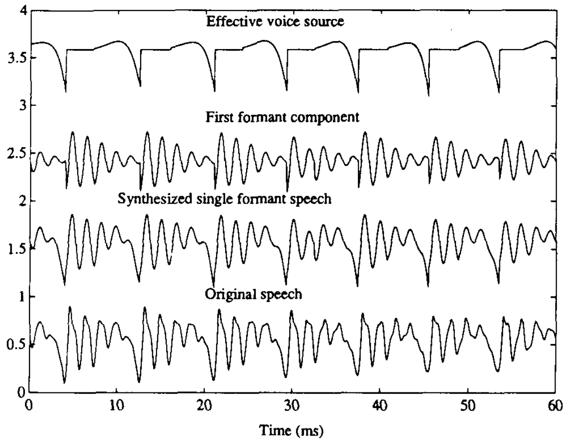

Fig. 6. Effective voice source (top); first formant component (second from top); and original speech signal (bottom) for a second male talker. (All waveforms are drawn to the same scale.)

the first two formant components and the effective voice source signal were used in the synthesis.

C. Improving the Estimation of the Glottal Source

One of the drawbacks of the exponential signal model for the speech signal is the large number of parameters needed to represent each period of the speech signal. For example, if we assume that 3 formants need to be estimated in the closed and open glottal phase, then in each period, 30 real parameters need to be estimated. This is difficult, especially when the pitch period is small. If the primary interest is in the glottal source signal, then one can reduce the number of parameters to be estimated by eliminating the higher formants using selective inverse filtering. Also, data from two or more successive periods can be used in each analysis.

1) Selective Inverse Filtering: Selective inverse filtering is a way of removing selected formant components from the speech waveform. In the procedure that we have adopted, the effect of all formants higher than the first is removed using inverse filtering. Since the primary effects of the subglottal coupling are limited to the first formant, the higher formants are assumed to be time invariant. Covariance linear predictive (LP) analysis [17] is done pitch asynchronously on 20-ms segments of the preemphasized speech signal, and an interactive procedure is used to identify the roots of the LP polynomial that correspond to the second and higher formants. An inverse filter is constructed from the selected roots, and the speech is filtered through it to remove the higher formants.

Once the second and higher formants are removed from the speech signal, only the effective voice source and first formant components need to be estimated. Thus only about 14 parameters need to be estimated in each pitch period.

2) Using Data from Successive Periods: By assuming that the glottal source components and the formant frequencies do not change significantly in two successive periods, the data from the corresponding phase in two periods can be combined and used in the analysis. This also increases the accuracy of parameter estimation. Parthasarathy and Tufts [14] suggest a similar technique. We have found that using data from two successive periods leads to significantly better estimates.

Figs. 6 and 7 show the results of the analysis of a vowel sound from a male talker and a female talker, respectively. In both cases, the speech and EGG signals were simultaneously digitized at 12 KHz each in an anechoic chamber. Each sample was of 16 b precision. The speech signal was selectively inverse filtered to remove all formants except the first (and the effective source). Data from two successive periods was used in each analysis. The order of the predictor was 21 in the open phase and 17 in the closed phase for the male talker, and 17 in the open phase and 14 in the closed phase for the female talker. Only the 4 largest singular values were retained.

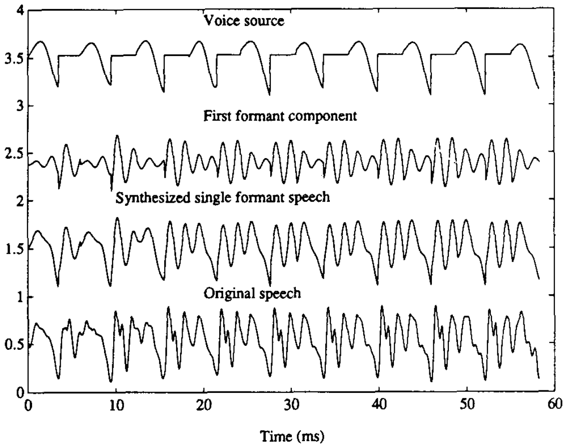

Fig. 7. Effective voice source (top); first formant component (second from top); synthesized first formant only speech (third from top); and original speech signal (bottom) for a female talker. The average fundamental frequency for this data is 164 Hz. (All plots are drawn to the same scale.)

Figs. 6 and 7 shows the estimated effective voice source (top graph), followed by the first formant component (second from top), and the synthesized speech signal (third from top) for these two talkers. Note that only the effective voice source and the first formant component are used to form the synthesized signal. Finally, the bottom graph shows the original speech signal (before selective inverse filtering). Comparing the last two graphs, it appears that the voice source and the first formant account for most of the speech waveform shape. In these example, the first formant frequency and bandwidth are 563 and 61 Hz, respectively, in the closed phase and 590 and 129 Hz, respectively, in the open phase for the male talker. The corresponding formant frequency and bandwidth values for the female talker are 580 and 38 Hz in the closed phase, and 596 and 146 Hz in the open phase.

D. Discussion

We have examined data for vowels from four talkers, three male and one female, using this analysis procedure. The results obtained are consistent with the examples shown here, once the closed and open glottal phases have been accurately located. The greatest problem appears to be with the use of the EGG for locating the glottal opening instant. While the EGG provides very precise information about the closing instant [13], the error in locating the opening instant is sometimes large. We are presently pursuing methods of using the speech signal itself to locate the opening instant [9].

Our results indicate that there are significant differences in the first formant bandwidth in the closed and open phases, with the open phase value being much larger. The first formant frequency is also usually higher in the open phase, but the difference is much less significant. This is in keeping with predictions from models of speech production that model source-tract interaction [8]. A systematic study to quantify these differences, including the effects (if any) of phonetic context, is presently underway.

IV. Conclusions

We have described a method of simultaneously estimating the parameters of a glottal source model and a vocal tract model from the speech signal. The effective voice source is modeled 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. Analysis results suggest that the method has the potential for providing accurate estimates of the glottal source signal, and the formant frequencies and bandwidths.

References

[1] J. Gauffin and J. Sundberg, "Spectral correlates of glottal voice source waveform characteristics," *J. Speech Hearing Res.*, vol. 32, pp. 556-565, Sept. 1989.

[2] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among male and female talkers," *J. Acoust. Soc. Amer.*, vol. 87, pp. 820-857, Feb. 1990.

[3] A. K. Krishnamurthy and D. G. Childers, "Two channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 730-743, Aug. 1986.

[4] D. J. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech wave," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350-355, Aug. 1979.

[5] P. Hedelin, "A glottal LPC-vocoder," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (San Diego, CA), Mar. 1984, pp. 1.6.1-1.6.4.

[6] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Tokyo, Japan), 1986, pp. 1605-1608.

[7] H. Fujisaki and M. Ljungqvist, "Estimation of the voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Dallas, TX), 1987, pp. 637-640.

[8] T. V. Anathapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, pp. 167-184, 1982.

[9] S. Parthasarathy and D. Tufts, "Excitation-synchronous modeling of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1241-1249, Sept. 1987.

[10] G. Fant, J. Liljencrants, and Q. Lin, "A four parameter model of glottal flow," STL-QPSR 4/1985, 1985, pp. 1-13; also presented at the French-Swedish Symp., Grenoble, Apr. 22-24, 1985.

[11] J. N. Holmes, "Formant synthesizers—cascade or parallel?," *Speech Commun.*, no. 2, pp. 251-273, 1983.

[12] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography," *CRC Crit. Rev. Bioeng.*, vol. 12, no. 2, pp. 131-161, 1985.

[13] D. Talkin, "Voicing epoch determination with dynamic programming," in *117th Meeting: Acoust. Soc. Amer.*, vol. 85(S1) (Syracuse, NY), p. S149, 1989.

[14] S. Parthasarathy and D. Tufts, "Signal modeling by exponential segments and applications in voiced speech analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Dallas, TX) Apr. 1987, pp. 645-648.

[15] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1805-1815, Dec. 1989.

[16] R. Kumareasan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 833-840, Dec. 1982.

[17] D. O'Shaughnessy, *Speech Communication: Human and Machine.* Reading, MA: Addison-Wesley, 1987.

Performance Degradation of DOA Estimators Due to Unknown Noise Fields

Fu Li and Richard J. Vaccaro

***Abstract*—This correspondence presents a statistical performance analysis of subspace-based directions-of-arrival (DOA) estimation algorithms in the presence of correlated observation noise with unknown covariance. Our analysis of five different estimation algorithms is unified by a single expression for the mean-squared DOA estimation error which is derived using a subspace perturbation expansion. The analysis assumes that only a finite amount of array data is available.**

I. Introduction

In the last decade, subspace processing algorithms have found prominence in the problem of estimating directions of arrival. The most popular algorithms are MUSIC [1], min-norm [2], state-space realization (TAM) [3], ESPRIT [4], and matrix pencil [5]. The increasing popularity of the subspace processing motivates many researchers to analyze the performance of these algorithms, as opposed to the early performance justification based on simulations. Most of these analyses consider finite data effects induced by observation noise (among them are [6]-[13]), while others consider sensor error effects [14]-[17], but few of them consider the effects of spatially correlated noise with unknown covariance, although it is also one of the major error sources in DOA estimation. Also, most of the analysis in the literature assume asymptotically large data records, and only consider an individual algorithm.

The success of subspace-based DOA estimation is based on its ability to perform a complete separation of single and orthogonal subspaces. Observation noise degrades the performance of DOA estimation by causing an incomplete separation of the two subspaces. The degradation due to observation noise can be divided into two aspects: 1) a finite number of noisy observations does not provide accurate knowledge of the signal and orthogonal subspaces—this is a finite sample effect induced by observation noise; and 2) an unknown noise covariance structure prevents a subspace rotation from separating two partially overlapping subspaces caused by a spatially colored noise field—this is called the unknown noise field effect. Correlated noise with known covariance can be easily handled by using a prewhitening filter or solving a generalized eigenvalue problem, which we described above as spatial rotation. However, in practice, the covariance structure is often unknown, and sometimes even time varying.

The presence of correlated noise is generally due to crosstalk between channels, random radiation from distributed sources, reverberation, undesired interference, etc. Typical examples of these occur in sonar where the ocean noise is due to marine life, waves, or distant shipping, and in radar where the background noise is land clutter, sea clutter, and scattering interference.

In our previous work, we developed a unified nonasymptotic

Manuscript received March 26, 1991; revised August 20, 1991. This work was supported by the Office of Naval Research under Grant N00014-90-J1283 and by Portland State University through its Faculty Development Program.

F. Li is with the Department of Electrical Engineering, Portland State University, Portland, OR 97207-0751.

R. J. Vaccaro is with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881-0805.

IEEE Log Number 9105672.

1053-587X/92$03.00 © 1992 IEEE