PAPER Special Section on Statistical Modeling for Speech Processing

Speech Analysis Based on Modeling the Effective Voice Source

M. Shahidur RAHMAN[†a)], ***Student Member and*** **Tetsuya SHIMAMURA**[†], ***Member***

SUMMARY A new system identification based method has been proposed for accurate estimation of vocal tract parameters. An often encountered problem in using the conventional linear prediction analysis is due to the harmonic structure of the excitation source of voiced speech. This harmonic characteristic is coupled with the estimation of autoregressive (AR) coefficients that results in difficulties in estimating the vocal tract filter. This paper models the effective voice source from the residual obtained through the covariance analysis in the first-pass which is then used as input to the second-pass least-square analysis. A better source-filter separation is thus achieved. The formant frequencies and corresponding bandwidths obtained using the proposed method for synthetic vowels are found to be accurate up to a factor of more than three (in percent) compared to the conventional method. Since the source characteristic is taken into account, local variations due to the positioning of analysis window are reduced significantly. The validity of the proposed method is also examined by inspecting the spectra obtained from natural vowel sounds uttered by high-pitched female speaker.

key words: *glottal waveform, effective voice source, linear prediction, least square method, system identification*

1. Introduction

According to the source-filter theory of speech production [1], voiced speech is represented as the response of the vocal tract filter to the glottal voice source. The glottal source consists of quasi-periodic pulses which are created by the vibration of vocal cords. To study the acoustic characteristics of either the vocal fold or the vocal tract, the resonance frequencies of the vocal tract system are required to be estimated accurately. The vocal tract parameters estimated in the absence of source effect has direct impacts for producing natural sounding synthetic speech. Linear prediction analysis [2] estimates an all-pole filter to model jointly the glottal, vocal and lip radiation where the driving source is assumed to be white noise. The all-pole filter coefficients are determined by seeking an optimal fit to the envelope of speech spectrum. For voiced speech, the source is a quasi-periodic nature with spiky excitations. However, the structure of the source excitation is not taken into account in the fitting procedure. The periodic excitations are thus coupled with the filter coefficients. For high-pitched speech, estimation of linear prediction spectrum becomes very difficult due to the wide spacing of harmonics. Vocal tract resonant frequencies (i.e. formants) are shifted to the direction of nearest harmonic peak. Bandwidths of the formant peaks can also be grossly overestimated or underestimated depending on the relative position of actual formant with respect to the neighboring peaks. If the source characteristics can be taken into account for the estimation of filter coefficients, a better source-filter separation can be achieved.

Several attempts [3]–[9] have been taken considering the source characteristics within the estimation process. Among them Miyanaga et al. [3] proposed a system identification based technique which estimates a sequence of pulses from the prediction residuals. The pulse train is then used to characterize the input of the all-pole filter (to be estimated) together with speech waveform as output. To decouple the source effects from estimation, Yanagida et al. [4], Miyoshi et al. [5], and Lee [6] proposed techniques by weighting the prediction residuals. Results presented on the synthetic speech perform quite well. El-Jaroudi [7] and Kabal [9] proposed two iterative methods for eliminating aliasing effects from the autocorrelation function due to periodicity which produce improved all-pole modeling. Recently, Arima proposed another system identification based approach [8] which is conceptually similar with that in [3]. This method, however, employed simple threshold logic on the prediction residuals which made it computationally more effective. Additionally, the method in [8] is shown to be more accurate in estimating spectrum than that in [4] and [6]. This paper presents a modified system identification based approach that produces more accurate estimate of the vocal tract filter than that produced by the method in [8].

A system identification method requires an estimate of the input together with the output for evaluating the underlying model. In case of voiced speech, if an estimate of the glottal source is available, the least square analysis can produce all-pole filter which is expected to be independent to the variation of the fundamental frequency (F_0). The input source assumed in [3], [8] is a pulse train. For natural voiced speech, however, due to glottal and radiation activities the input to the vocal tract system is no longer a plain pulse train [10]–[12]. To estimate the vocal tract characteristics accurately, it is thus important that the input resembles the actual voice source. Unfortunately, accurate estimation of the glottal waveform from the speech waveform is a computationally very expensive process. Numerous methods [10]–[12] have already been proposed, which are, in general, iterative and requires the detection of a good number of parameters at every iteration. Usually, closed phase analysis is performed prior to the estimation of glottal wave and the

Manuscript received July 11, 2005.
Manuscript revised September 30, 2005.

[†] The authors are with the Department of Information and Computer Sciences, Saitama University, Saitama-shi, 338–8570 Japan.
a) E-mail: rahmanms@sie.ics.saitama-u.ac.jp

DOI: 10.1093/ietisy/e89–d.3.1107

methods produce accurate estimates if the closed phase interval is around 3 ms [13]. The closed phase interval is, however, very short for high-pitched female and children speech. Several methods [14]–[16] have also been reported for estimating the voice source and vocal tract parameters simultaneously. These methods also have been devised pitch synchronously (which is troublesome to implement) and require generation and updating of parameters iteratively. A good number of iterations are needed to achieve a satisfactory convergence of the estimation. For real time speech analysis systems, such expensive algorithms are not appropriate.

The purpose of this paper is to devise a method for accurate estimation of vocal tract parameters that exploits a fast and robust technique for approximating the effective voice source. Rather than estimating either a pulse train or a complete glottal waveform, we propose a non-iterative technique for approximating the principal parts of the effective voice source which uses a time-efficient threshold logic of prediction residual without requiring any parametric information of the glottal cycle. The whole procedure requires only a fraction of computation needed in [10]–[12], [14]–[16]. The approximated voice source is observed to provide a significant source separation while estimating the vocal tract parameters. The formant frequencies and bandwidths estimated from the AR coefficients are found to be accurate up to a factor of more than three in percent compared with those estimated using the covariance method.

In Sect. 2, we define and formulate the method of estimating the effective voice source. The proposed method for extracting the vocal tract parameters is described in Sect. 3. Sections 4 and 5 present the analyzing results of synthetic and natural speech, respectively.

2. Analysis Method

In this section, we define the effective voice source and propose a technique for estimating the significant parts of the effective voice source in case of pitch asynchronous linear prediction analysis.

2.1 Defining the Effective Voice Source

The linear model of speech production [1] defines the speech model $S(z)$ as

$$S(z) = E(z)G(z)V(z)R(z) \tag{1}$$

where $E(z)$ is the driving function to the glottal shaping model $G(z)$, $R(z)$ is the lip radiation model, and $V(z)$ is the vocal tract model. The linear prediction model of speech analysis is based on the model of Eq. (1) where $V(z)$ is all-pole and the composite spectrum effects of $G(z)$, $V(z)$, and $R(z)$ are represented by a filter:

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{2}$$

where p is the prediction order and a_ks are the filter coefficients. Equations (1) and (2) together give an analysis model $E(z) = S(z)A(z)$, which can be equivalently expressed in the sampled data domain as

$$e_n = s_n + \sum_{k=1}^{p} a_k s_{n-k} \tag{3}$$

for $n = 0, 1, 2, \ldots, N-1$ where N is the length of the analysis frame. The sequence e_n is the prediction error of the sampled speech s_n which is also called the residual. Ideally, the residual for voiced speech should consist of impulses separated by pitch periods. However, such an output is not observed except for vowel sounds synthesized using impulses as the excitation function. The residual signal can rather be related with the glottal source [17].

Let u_n be the quasi periodic sequence of glottal pulses representing the transfer function $E(z)G(z)$ in Eq. (1). If u_n is equivalently considered as the output of a double integrator excited by the second derivative of u_n, $u_n^{(2)}$, then Eq. (1) can be rewritten as

$$S(z) = U^{(2)}(z)\hat{V}(z) \tag{4}$$

where $U^{(2)}$is the z-transformation of $u_n^{(2)}$ and $\hat{V}(z)$ is the cascade of the double integrator, $V(z)$ and $R(z)$. It is well known that the digital inverse filter $A(z)$ in linear prediction analysis is in general an optimum spectral whitening filter. Since $u_n^{(2)}$ is known to possess usually flat spectral characteristics [18], it is reasonable to assume that $A(z)$ flattens the spectrum of $\hat{U}^{(2)}(z)$ only; i.e.

$$\hat{V}(z)A(z) \simeq 1. \tag{5}$$

The prediction residual e_n, which is the output of $A(z)$, is therefore given by

$$e_n \simeq u_n^{(2)}. \tag{6}$$

This analysis brings out that the prediction residual is an approximation of the second derivative of glottal pulses. Glottal waveform and its derivatives generated using the Liljencrants-Fant (LF) [19] model at $F_0 = 200$ Hz are shown in Fig. 1. The LF model is used here because this glottal model has been shown to be more suitable for description and for producing natural sounding synthetic speech [11], [20]. First derivative of the glottal waveform in Fig. 1 (b) corresponds to the glottal waveform inclusive of lip radiation characteristics and the second derivative in Fig. 1 (c) includes additionally the preemphasis filter. An estimate of the second derivative of glottal waveform can be determined by first preemphasizing the speech signal by the filter $1 - z^{-1}$ and then applying the inverse filter $A(z)$ to the preemphasized speech signal. Application of the inverse filter $A(z)$ (obtained from the preemphasized speech) to the unpreemphasized speech signal gives an estimate of the first derivative of the glottal waveform. The first case is common in speech analysis applications using linear prediction. In this paper, we call the second derivative of the glottal waveform as the effective voice source. A representation of the

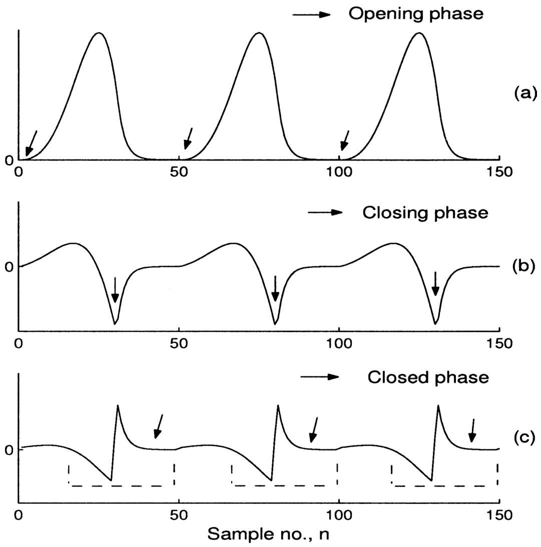

Fig. 1 (a) LF model glottal waveform; (b) First derivative of the glottal waveform; (c) Second derivative of the glottal waveform.

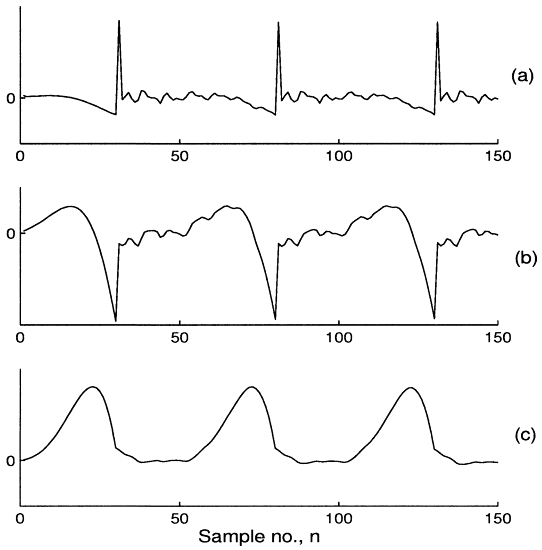

Fig. 3 (a) Residual of synthetic vowel sound /a/; (b) First order integral of (a); (c) Second order integral of (a).

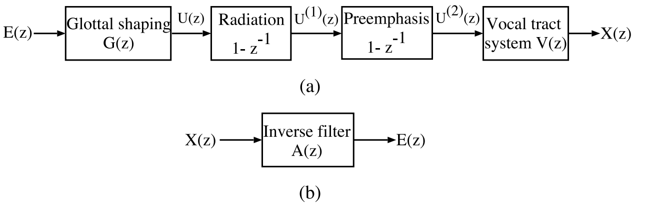

Fig. 2 (a) Representation of the preemphasized speech; (b) Linear prediction inverse filter.

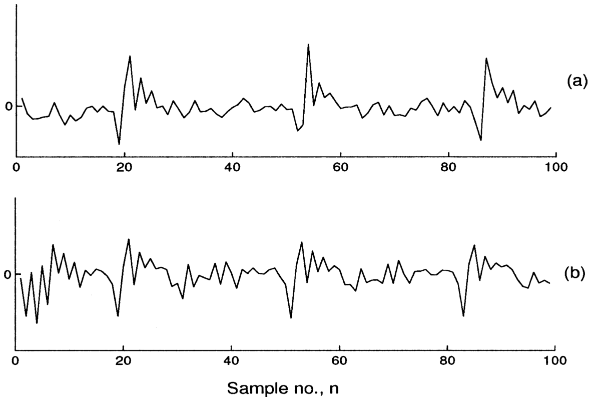

Fig. 4 Residual of (a) natural vowel sound /a/; (b) natural vowel sound /u/.

preemphasized speech and linear prediction inverse filter are summarized in Fig. 2 (a) and (b), respectively. According to Fig. 2 (b), Eq. (3) is rewritten as follows:

$$e_n = x_n + \sum_{k=1}^{p} a_k x_{n-k} \tag{7}$$

where x_n corresponds to the preemphasized speech signal. A residual signal obtained by applying the covariance method on preemphasized synthetic vowel /a/ is shown in Fig. 3 along with its first and second order integrals. The first and second order integrals of the residual correspond to the first order derivative of glottal waveform and the glottal waveform itself, respectively. The synthetic vowel is generated using the LF glottal model as excitation. It is evident that other than the small fluctuations the residual in Fig. 3 (a) produces back a close estimate of the glottal source in Fig. 3 (c) which establishes the validity of Eq. (6). The likeness of the residual with the effective voice source for natural speech signal is shown in Fig. 4. From Figs. 3 and 4 it is obvious that an estimate of the effective voice source can be obtained from the prediction residual.

2.2 Modeling the Effective Voice Source

Previous researches [3], [8] attempted to model pulse train from the residual to account for the excitation effect in the estimation process of $A(z)$ using the covariance matrix constructed from the speech and resulting pulse sequence (where the least square method is applied). The resulting filter $A(z)$ was thus thought to be insensitive to the variations of F_0. Equation (6), however, implies that the contribution of the effective voice source must be decoupled from the speech signal in order to obtain the spectrum of pure vocal tract system.

The inverse filtering methods for glottal source estimation [10]–[12] and the methods for simultaneous estimation of source-tract parameters [14]–[16] require pitch synchronous (or closed phase) analysis. These methods when applied on a segment of multiple pitch periods can not produce a complete estimate of the glottal source. However, it is always possible to achieve an estimate of the significant instants of the effective voice source, which, in fact, can result in sufficient source-filter separation when applied as input to the least square method with speech waveform as output.

One way to justify this concept is that though the true excitation for voiced speech is a sequence of quasi-periodic glottal pulses, the significant excitation of the vocal tract system usually coincides with the glottal closure. Referring to the second derivative of glottal wave in Fig. 1 (c), the major changes are seen to occur around the closing phase. Except the beginning of the opening phase, the rest parts of the effective voice source in Fig. 1 (c) (the region marked with dashed line) can be closely approximated from the residual.

We develop a method to obtain an estimate of the above mentioned regions of the effective voice source based on detecting the residual peaks. The peaks are identified using a threshold logic. Rather than using one threshold (as used in [8]), we introduce the use of two thresholds, one for identifying the positive peak and the other for the negative peak. The use of two thresholds provides robust detection of the residual peaks corresponding to the excitation instants. This is because the information of a single peak can be misleading by the presence of false residual peaks. Referring to Eq. (7) the first p samples of e_n are more erroneous, which is, in fact, the inherent property of the covariance method. Due to this reason, e_n is analyzed from right to left direction and the thresholds are computed using the rest of the samples (i.e. from e_{N-1} to e_{p+1}). The magnitude of the positive and negative thresholds can be 60% of the maximum and minimum of e_n (excluding the first p samples), respectively. The samples exceeding the value of positive and negative threshold are considered tentatively as the positive and negative peaks, respectively. Finally, a positive peak immediately followed by a negative peak (considering right to left direction of analysis) is treated as true excitation peak. Once the positive peak is decided, samples are traversed starting from its position until a switch from negative to positive value (upward zero crossing) occurs. This is repeated for every excitation peak. The rest of the samples are set to zero. The knowledge of pitch period can further ease the process of peak detection which can be approximated as the difference between two successive positive or negative peaks. In Fig. 5, the approximated version of the effective voice source (solid line) estimated from the residuals of synthetic vowel sound /a/ and natural vowel sound /u/ are shown together with the residuals (dashed line). The approximated signal is used as input for least square identification of the AR model. It is verified that an estimate of the input signal obtained by capturing only the instants of positive and negative peaks of the residual at every pitch period can result in significant separation of the excitation effect from the estimated spectrum.

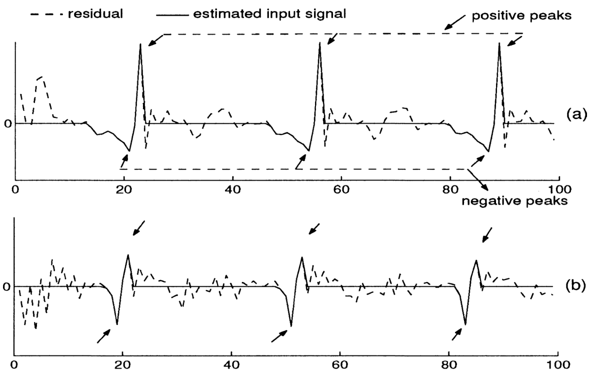

Fig. 5 Residual and the input signal estimated (a) from synthetic vowel sound /a/; (b) from natural vowel sound /u/.

3. Estimation of Vocal Tract Parameters

The overall method of estimating the vocal tract parameters is outlined in Fig. 6. The first-pass covariance parameters a_k estimated from the preemphasized speech x_n are used to derive the residual which is then operated to obtain an estimate of the simplified equivalent, $\hat{e}_n$, of the effective voice source. With an estimate of the input available, the second-pass least square parameters can be realized in the absence of source effect. Since the second-pass analysis method is not exactly similar to the first-pass covariance analysis, it is referred to as least square analysis in the block diagram. The modified error for the second-pass least square method can then be defined as:

$$\varepsilon_n = (x_n - \hat{e}_n) + \sum_{k=1}^{p} a_k x_{n-k} \tag{8}$$

In this case, the error criterion

$$J = \sum_{n=p}^{N-1} [x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} - \hat{e}_n]^2 \tag{9}$$

is minimized. The gain factor is not considered here because its role is not significant for the purpose of formants/bandwidths estimation. After the vector

$$\mathbf{x} = [x_p, x_{p+1}, \ldots, x_{N-1}]^T$$

and the matrix

$$\mathbf{\Omega} = \begin{bmatrix} x_{p-1} & \ldots & x_0 & \hat{e}_p \\ x_p & \ldots & x_1 & \hat{e}_{p+1} \\ x_{p+1} & \ldots & x_2 & \hat{e}_{p+2} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{N-2} & \ldots & x_{N-p-1} & \hat{e}_{N-1} \end{bmatrix}$$

are constructed, J can be derived as

$$J = (\mathbf{x} - \mathbf{\Omega}\theta)^T (\mathbf{x} - \mathbf{\Omega}\theta)$$

where T denotes transpose and θ corresponds to the coefficient vector:

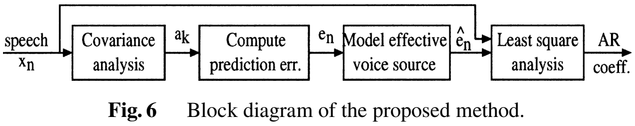

Fig. 6 Block diagram of the proposed method.

$$\theta = [-a_1, -a_2, \ldots, -a_p]^T.$$

Minimizing J in the conventional manner by setting the partial derivatives with respect to the unknown parameters to zero leads to an estimate of the coefficient vector as

$$\hat{\theta} = [\mathbf{\Omega}^T\mathbf{\Omega}]^{-1}\mathbf{\Omega}^T\mathbf{x}.$$

The solution $\hat{\theta}$ gives the estimate of all-pole vocal-tract filter.

A generalized example is illustrated in Fig. 7 for synthetic vowel sound /a/ synthesized at eight different F_0 values. From the figure, it is obvious that using a simplified version of the effective voice source as input to the identification model results in more accurate spectra. As seen in the figure, NRSA (Noise Robust Speech Analysis) method [8] produces much better estimates of the formants than the conventional covariance method. The consistency of estimation at higher values of F_0, however, is not as good as the current method when comparing with the 'true' spectrum (thick solid line in the figure). Accuracy of the formants and bandwidths is mirrored in the similarities of all the spectra estimated using the proposed method. The 'true' spectrum is obtained from the vocal tract impulse response.

We choose the NRSA method for making comparison because this method is the immediate predecessor of the current method. To our knowledge, the NRSA method, in addition, is a recent non-iterative method that employed least square identifications for estimating speech spectrum.

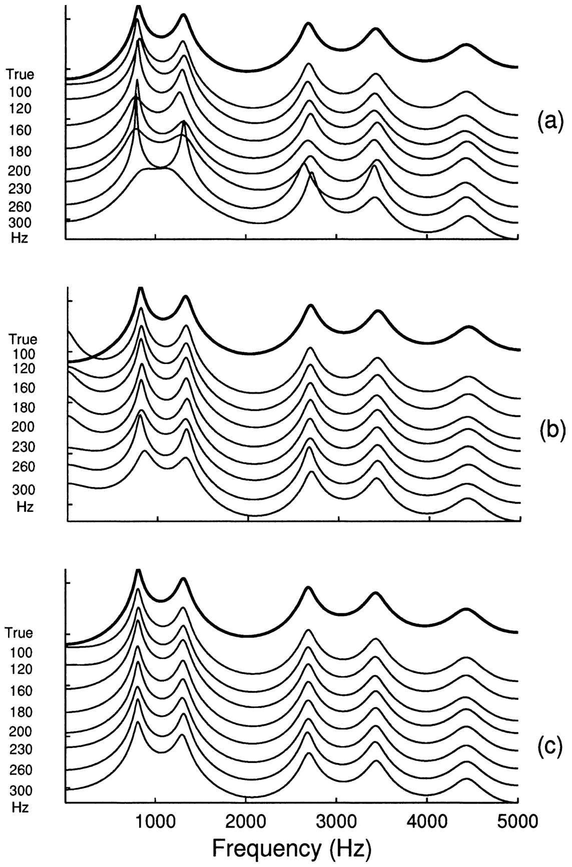

Fig. 7 Spectra estimated from synthetic vowel sound /a/ at different F_0 values (a) using covariance method; (b) using NRSA method; (c) using proposed method.

We note that the results presented here are based on analyzing a single frame. Detail results concerning various positions of the analysis window are described in Sects. 4 and 5.

4. Simulation Results Using Synthetic Speech

The modeling of the effective voice source has been applied to the estimation of formant frequencies and corresponding bandwidths of synthetic vowels. Five Japanese vowels are synthesized by exiting formant resonators by the glottal waveform generated using LF model. Lip radiation is simulated by the filter $1 - z^{-1}$. Speech at different F_0 values is synthesized by varying the F_0 values of the glottal source while keeping the values of other parameters fixed. The formant frequencies used to synthesize vowels are shown in Table 1. Bandwidths of the five formants are set to 60, 100, 120, 175, and 281 Hz, respectively, for all the five vowels. Synthetic speech is preemphasized before analysis. Frame size is set to three pitch period and the predictor order is 12. Formant and bandwidth values are obtained from the AR coefficients using the root-solving method.

4.1 Estimation of Formant Frequency

Every formant value is determined as the arithmetic mean of the formants obtained from twenty individual analysis frames where the frame overlap is set to half the frame length. The relative estimation error (REE) of the ith formant frequency is obtained as

$$EF_i = \frac{1}{5}\sum_{j=1}^{5} \mid \hat{F}_{ij} - F_{ij} \mid / F_{ij} \qquad (10)$$

where F_{ij} denotes the ith formant frequency of the jth vowel and $\hat{F}_{ij}$ is its estimated value. The REEs of the first, second, third, and fourth formant frequencies are shown in Fig. 8 (a), (b), (c), and (d), respectively, using the covariance, NRSA, and proposed methods over a wide range of F_0 values. As seen in the figures, the proposed method produces better estimates of all the formants at all F_0 values used in this example. In contrast, the covariance and NRSA method fail to produce equal accuracy at all F_0 values.

We combine the REEs of all the five formants as

$$E = \frac{1}{25}\sum_{j=1}^{5}\sum_{i=1}^{5} \mid \hat{F}_{ij} - F_{ij} \mid / F_{ij}. \qquad (11)$$

Table 1 Formant frequencies used to synthesize vowels.

vowel	F_1	F_2	F_3	F_4	F_5 Hz
/a/	813	1313	2688	3438	4438
/i/	375	2188	2938	3438	4438
/u/	375	1063	2188	3438	4438
/e/	438	1813	2688	3438	4438
/o/	438	1063	2688	3438	4438

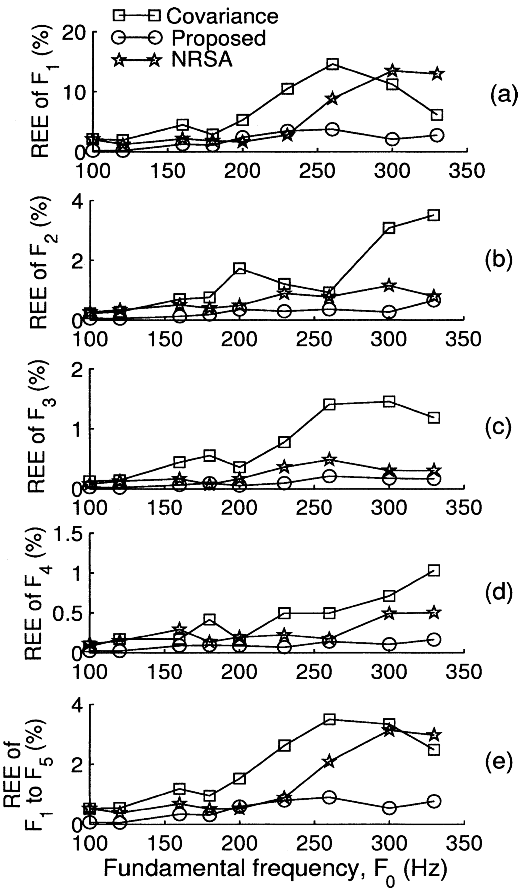

Fig. 8 Relative estimation error (REE) of (a) first formant; (b) second formant; (c) third formant; (d) fourth formant; (e) all five formants.

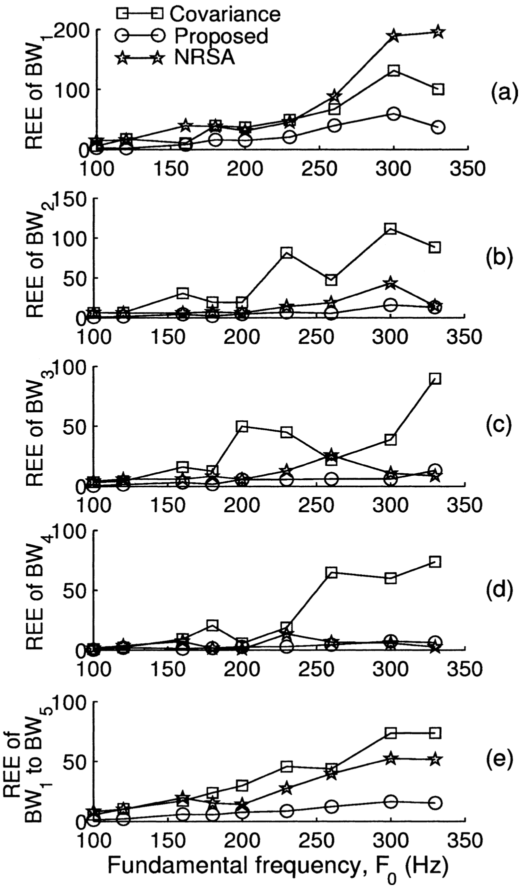

Fig. 9 Relative estimation error (REE) of (a) first bandwidth; (b) second bandwidth; (c) third bandwidth; (d) fourth bandwidth; (e) all five bandwidths.

Figure 8 (e) shows the REE corresponding to Eq. (11). An improvement up to a factor of more than three over the covariance method is achieved. The greatest improvement, as can be predicted, is obtained at the higher values of F_0 where linear prediction is more erroneous. The results suggested that the proposed method can be used for analyzing female and children speech as well as typical male speech with significant improvement in accuracy.

4.2 Estimation of Bandwidth

Bandwidth of a formant is determined similarly as formant by taking the mean of the bandwidths obtained from twenty individual frames. The REEs of the first, second, third, fourth, and all the five bandwidths are shown in Fig. 9, which are computed in a similar fashion of Eqs. (10) and (11). The term BW_i used in the figure implies the ith bandwidth. Using the NRSA method the first bandwidth is observed to be expanded at higher F_0 values with lower first formant. The proposed method, in contrast, produces estimates with smaller errors in all cases. Smaller estimation error of formants and bandwidths implies that the estimated spectrum is more close to the 'true' spectrum. This is, in fact, the outcome of successful source-filter separation in the estimation process.

5. Results Using Real Speech

Experiments have also been conducted on natural vowel sounds. Spectra obtained from Japanese vowels and CV sounds uttered by a skilled Japanese female speaker are inspected for evaluating the performance of the current method. The voiced segments are extracted from vowels /a/, /o/, /u/ and from the vowel parts of the CV sounds /bo/ and /bu/. The F_0 values of speech segments range from 250 to 300 Hz. Speech is recorded in an almost sound-proof room using a SONY ECM-G3M superdirectional microphone. The recorded speech is then processed by an ONKYO digital audio processor (SE-U55X) and sampled primarily at 44 kHz rate which is down sampled to 10 kHz rate before analysis. The predictor order is set to 12 and speech is preemphasized by the same filter $1 - z^{-1}$ as for the synthetic speech. Spectra obtained using the proposed method are compared with those obtained using the covariance and NRSA methods in Fig. 10 (a) and (b), respectively. The examples illustrate some typical situations produced by linear prediction when analyzing higher pitched speech. In the spectra obtained using the covariance method in Fig. 10 (a), it is seen that formants are not resolved accurately (F_1 and

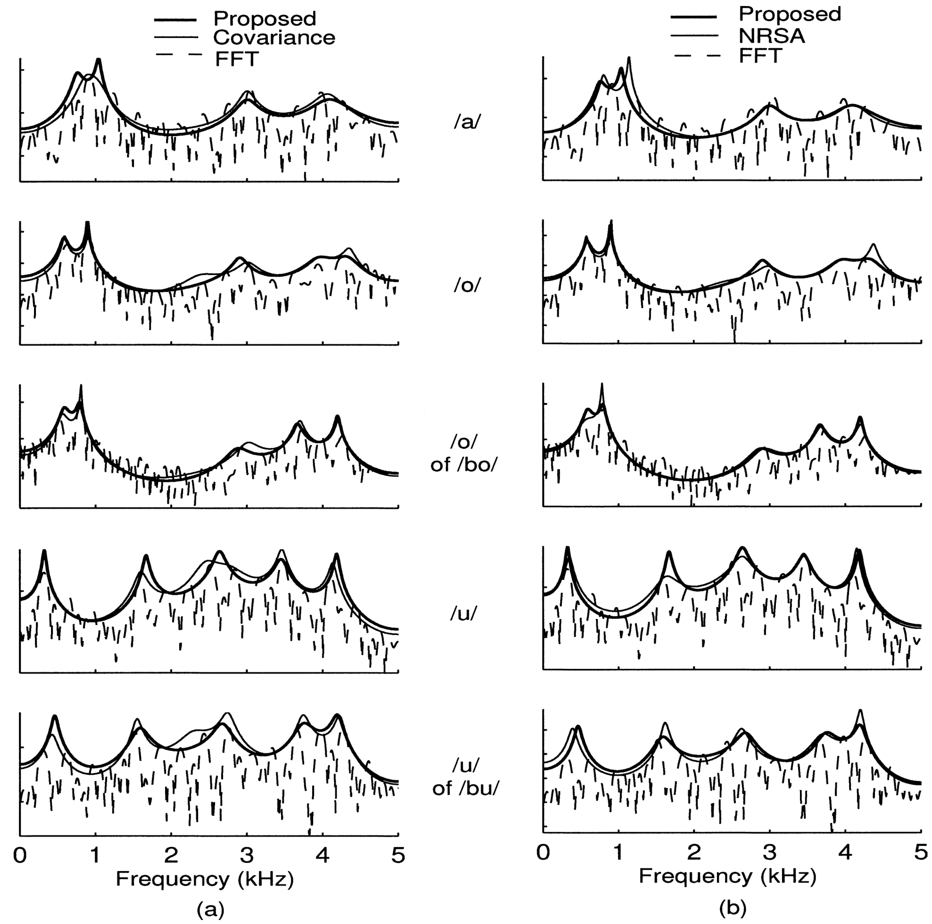

Fig. 10 Comparison of spectra estimated from natural vowel sounds using the proposed method (a) with the covariance method; (b) with the NRSA method.

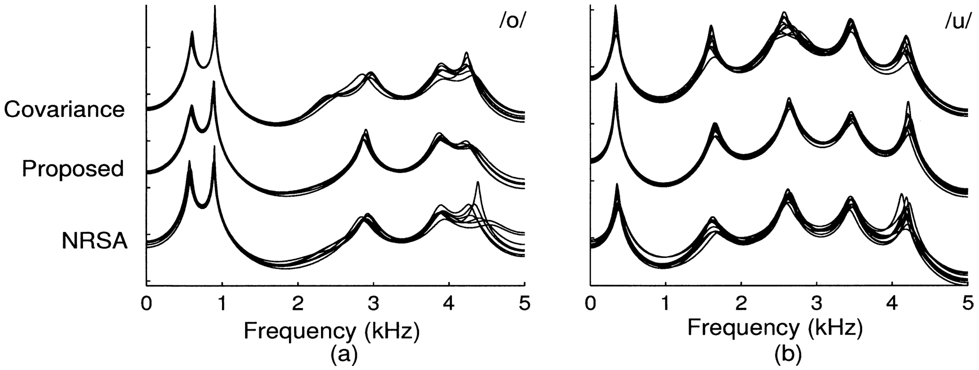

Fig. 11 Comparison of the compactness of several spectra estimated successively (a) from natural vowel sound /o/ at F_0 = 300 Hz; (b) from natural vowel sound /u/ at F_0 = 300 Hz.

F_2 in case of /a/), some spurious peaks other than formants are introduced (in cases of /o/, /u/, and /u/ of /bu/) and bandwidths of some formants (in cases of /u/ and /o/ of /bo/) are also not estimated accurately. These problems arise due to the higher F_0 value of the underlying speech. Since we have accounted the issue of voice source, the above problems are completely absent in the spectra obtained using the proposed method. All the formants and bandwidths are estimated very well. In Fig. 10 (b), on the other hand, the spectra produced by the NRSA method are seen to possess some narrow bandwidth formants (e.g. bandwidths of F_2 of /a/, /o/, and /o/ of /bo/).

The linear prediction method is also known to be affected by the positioning of analysis window. The use of voice source in the estimation process eliminates the problem significantly. Several spectra estimated successively from the vowel sounds /u/ and /o/ by the covariance, NRSA, and proposed methods are shown in Fig. 11. Frame interval used here is 10 ms. It is obvious that deviations of the estimated spectra using the current method in Fig. 11 are much less than those estimated using the covariance and NRSA methods. This indicates that the proposed method produces more compact formant clusters.

6. Conclusion

In an attempt to decouple the source effects from the estimation of vocal tract filter, we model the significant parts of the effective voice source which is then used as input to the least square method together with the speech waveform as output. The current method thus significantly reduces the limitations of the linear prediction analysis. The proposed method improves the robustness of the estimated spectrum to the variation of fundamental frequencies in terms of greater accuracy of both formants and bandwidths. It also greatly reduces the sensitivity of the positioning of analysis window. Unlike many other iterative methods, the current method is a non-iterative one which makes it promising for real time speech analysis applications.

Though the proposed method is primarily intended for spectral estimation in case of clean speech, it can also be applied for noisy speech with relatively higher SNR like the NRSA method. In case of lower SNR, the task of peak picking of the residuals will be difficult and the accuracy will thus be decreased. However, the peak picking technique employed in the current paper is more robust than that used in the NRSA method. By limiting the current peak picking algorithm to only the positive residual samples, the NRSA method can also be derived from the proposed one. It is therefore safe to state that the proposed method is at least equivalently applicable for noisy speech as the NRSA method.

Since the proposed method identifies the input signal (which is an approximation to the effective voice source), it can also be extended as an ARMA method similar to that in [3] but with improved accuracy. A general model for estimating the ARMA parameters based on the current method is under development.

Acknowledgment

The authors are grateful to the IEICE reviewers for their insightful comments on the manuscript.

References

[1] J.L. Flanagan, Speech Analysis, Synthesis, and Perceptions, 2nd ed., Springer-Verlag, New York, 1976.

[2] B.S. Atal and S.L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol.50, no.2, pp.637–655, 1971.

[3] 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 Process., vol.30, no.1, pp.1870–1887, 1982.

[4] M. Yanagida and O. Kakusho, “A weighted linear prediction analysis of speech signals by using the given’s reduction,” (M.H. Hamza, Ed.), IASTED Int. Symp. Appl. Signal Processing and Digital Filtering, Paris, pp.129–132, 1985.

[5] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida, and O. Kakusho, “Analysis of speech signal of short pitch period by a sample-selective linear prediction,” IEEE Trans. Acoust. Speech Signal Process., vol.35, no.9, pp.1233–1240, 1987.

[6] C.H. Lee, “On robust linear prediction of speech,” IEEE Trans. Acoust. Speech Signal Process., vol.36, no.5, pp.642–650, 1988.

[7] A. El-Jaroudi and J. Makhoul, “Discrete all-pole modeling,” IEEE Trans. Signal Process., vol.39, no.2, pp.411–423, 1991.

[8] Y. Arima and T. Shimamura, “Noise robust speech analysis using system identification methods,” IEICE Trans. Fundamentals (Japanese Edition), vol.J83-A, no.12, pp.1455–1466, Dec. 2000.

[9] P. Kabal and W. Kleijn, “All-pole modeling of mixed excitation signals,” IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.97–101, Salt Lake City, 2001.

[10] D.Y. Wong, J.D. Markel, and A.H. Gray, Jr., “Least square glottal inverse filtering from the acoustic speech waveform,” IEEE Trans. Acoust. Speech Signal Process., vol.27, no.4, pp.350–355, 1979.

[11] H. Fujisaki and M. Ljungqvist, “Proposal and evaluation of model for the glottal source waveform,” Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing, vol.4, pp.1605–1608, 1986.

[12] A.K. Krishnamurthy, “Glottal source estimation using a sum-of-exponentials model,” IEEE Trans. Signal Process., vol.40, no.3, pp.682–686, 1992.

[13] A.K. Krishnamurthy and D.G. Childers, “Two-channel speech analysis,” IEEE Trans. Acoust. Speech Signal Process., vol.34, no.4, pp.730–743, 1986.

[14] H. Fujisaki and M. Ljungqvist, “Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform,” IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.12, pp.637–640, 1987.

[15] W. Ding, N. Campbell, N. Higuchi, and H. Kasuya, “Fast and robust joint optimization of vocal tract and voice source parameters,” IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol.2, pp.1291–1294, 1997.

[16] K. Funaki, Y. Miyanaga, and K. Tochinai, “A time varying ARMAX speech modeling with phase compensation using glottal source model,” IEEE Int. Conf. Acoustics, Speech, and Signal Processing, pp.1299–1302, 1997.

[17] T.V. Ananthapadmanabha and B. Yegnarayana, “Epoch extraction from linear prediction residual for identification of closed glottis interval,” IEEE Trans. Acoust. Speech Signal Process., vol.27, no.4, pp.309–319, 1979.

[18] J.N. Holmes, “Formant excitation before and after glottal closure,” Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing, pp.39–42, 1975.

[19] G. Fant, J. Liljencrants, and Q.G. Lin, “A four parameter model of glottal flow,” Quart. Progress and Status Rep., Speech Transmission Lab, Royal Inst. Technol., Oct.-Dec., pp.1–13, 1985.

[20] H. Strik, “Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses,” J. Acoust. Soc. Am., vol.103, no.5, pp.2659–2669, 1998.

M. Shahidur Rahman received the B.Sc. (Hons) and M.Sc. degree in electronics and computer science from Shah Jalal University of Science and Technology, Sylhet, Bangladesh, in 1995 and 1997, respectively. In 1997, he joined Shah Jalal University as a junior faculty. Since October 2003, he has been with Saitama University, Saitama City, Japan, to pursue Ph.D. degree in mathematical information systems. His current research interests include speech analysis, speech synthesis, and digital signal processing. He is a student member of IEEE.

Tetsuya Shimamura received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1986, 1988, and 1991, respectively. In 1991, he joined Saitama University, Saitama City, Japan, where he is currently as Associate Professor. His research interests are in digital signal processing and applications to speech and communication systems. He is a member of IEEE and EURASIP.