AN APPROACH TO FORMANT FREQUENCY ESTIMATION AT LOW SIGNAL-TO-NOISE RATIO

S. A. Fattah, Student Member, IEEE, W.-P. Zhu, Senior Member, IEEE, and M. O. Ahmad, Fellow, IEEE

Centre for Signal Processing and Communications, Dept. of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8

ABSTRACT

A new approach for the formant frequency estimation of the voiced speech segments in the presence of noise is presented in this paper. A correlation model for the voiced speech is proposed considering the vocal-tract system as an autoregressive moving average (ARMA) model with a periodic impulse-train excitation. It is shown that the formant frequencies can be directly obtained from the model parameters. An adaptive residue-based least-squares optimization algorithm is proposed to estimate the model parameters, which overcomes the failure of conventional correlation based techniques in estimating formant frequencies at a low signal-to-noise ratio (SNR). The proposed algorithm has been tested on synthetic and natural vowels as well as voiced segments of some naturally spoken sentences from TIMIT database in presence of white Gaussian or babble noises. The experimental results show that the proposed method is more robust to noise than some existing methods even at a low SNR of 0 dB.

Index Terms— Speech analysis, autoregressive moving average processes, formant frequency, correlation model.

1. INTRODUCTION

Formant frequency is one of the most useful speech parameters. It has fundamental importance in many speech processing applications such as speech synthesis, compression, and recognition. Free resonances of a vocal-tract (VT) system are called formants. Formants are associated with peaks in the smoothed power spectrum of speech [1]. Among the different formant estimation methods, the linear predictive coding (LPC) based methods have received considerable attention where the VT system is considered as an autoregressive (AR) model [2]. Most of the formant frequency estimation methods so far reported deal only with noise-free environments [1]-[2]. However, formant estimation from noisy speech is very difficult but essential as far as practical applications are concerned. In [3], a peak-picking algorithm is used on a segmented spectrum to estimate the formant frequencies in noisy environments. The multi-cyclic covariance method, reported in [4], can determine the formant frequency from noisy speech at a relatively high SNR. Recently in [5], based on an adaptive bandpass filterbank (AFB), a formant frequency estimation method for noisy speech has been proposed where the estimation accuracy depends on the initial estimates.

In this paper, the formant frequency estimation problem under noisy conditions is addressed. Within a short duration of time, voiced speech can be considered as the output of an ARMA system excited by a periodic impulse-train. Recently in [6], we have proposed a correlation model for the output of an ARMA system excited by the white noise. In order to estimate the formant frequencies, in the current paper, we propose a correlation model for the output of an impulse-train excited ARMA system. Since the proposed correlation model provides a direct relationship between the formant frequencies and the model parameters, the main task is now to estimate accurately the model parameters. Unlike the conventional correlation based methods, a correlation-fitting approach is proposed where an adaptive residue-based least-square (ARBLS) optimization technique is introduced in order to obtain an accurate estimate of the model parameters even in the presence of significant noise.

2. PROPOSED CORRELATION MODEL

For the formant frequency estimation from the observed speech signal, it is sufficient to restrict the analyses only for voiced speech, where the excitation of the VT system can be modeled as the output of a glottal filter whose input is a periodic impulse-train. The spectral shaping effects of the glottis and the VT are combined into one filter $H(z)$, which can be represented by an ARMA system transfer function as

$$H(z)=\frac{\prod_{j=1}^{Q}(1-z_j z^{-1})}{\prod_{k=1}^{P}(1-p_k z^{-1})}=\sum_{k=1}^{P}\frac{\eta_k}{1-p_k z^{-1}} \qquad (1)$$

where p_k and z_j denote, respectively, the pole and the zero of the system with no pole-zero cancellation, P and Q the model orders with $P>Q$ and η_k the partial fraction coefficient corresponding to $p_k = r_k e^{j\omega_k}$. In order to model each formant, a pair of complex conjugate poles is required. For a sampling frequency F_S in samples/sec, Formant frequency (F_k) and

bandwidth (B_k) are given by

$$F_k = \frac{F_S}{2\pi}\omega_k \; ; \quad B_k = -\frac{F_S}{\pi}\ln(r_k). \tag{2}$$

During a short duration of time (frame), a given speech signal is generally assumed to be stationary. Hence, $H(z)$ can be modeled with constant coefficients within a frame. An impulse-train excitation with period T can be expressed as

$$u_{imp}(n) = \sum_{i=0}^{\lambda-1}\delta(n-iT) \tag{3}$$

where λ is the total number of impulses within the duration N. For a duration m, the value of λ can be computed as

$$\lambda_m = \left\lceil \frac{m+1}{T} \right\rceil \tag{4}$$

where $\lceil \xi \rceil$ represents the nearest integer greater than or equal to ξ. For the input $u_{imp}(n)$, the output $x(n)$ can be written as

$$x(n) = u_{imp}(n) * h(n). \tag{5}$$

Using (1) and (3), for an initially relaxed ARMA system, $x(n)$ can be obtained as

$$x(n) = \sum_{k=1}^{P}\sum_{i=0}^{\lambda_n-1}\eta_k p_k^{n-iT}. \tag{6}$$

The autocorrelation function (ACF) of $x(n)$ with data length N can be estimated as [6]

$$r_x(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|}x(n)x(n+|\tau|), \quad \tau = 0,1,\ldots\ldots,M-1. \tag{7}$$

In periodic impulse-train excited systems, all the necessary information required in order to estimate the system poles lies within the first T lags of $r_x(\tau)$ and hence, in (7), it is sufficient to consider $M = T/2$ number of ACF lags. Due to the complicated form of the signal model (6), unlike the case of white noise excitation proposed in [6], we compute and simplify the correlation function for each lag separately and finally obtain the following form

$$r_x(\tau) = \sum_{k=1}^{P}\psi_{Tk}p_k^{\tau}, \;\; \tau = 0, 1, ..., M-1; \; M < T/2; \tag{8}$$

where

$$\psi_{Tk} = \frac{1}{N}\sum_{j=1}^{P}\eta_k\eta_j\Theta_{Tk}. \tag{9}$$

Here, Θ_{TK} can be expressed in terms of the poles. The correlation model, derived in (8), is expressed explicitly in terms of the poles of the ARMA system. Letting $p_k = r_k e^{j\omega_k}$, $\psi_{Tk} = \zeta_k e^{j\upsilon_k}$, and θ = number of real poles + the number of complex conjugate pole pairs, correlation model reduces to

$$r_x(\tau) = \sum_{l=1}^{\theta} r_l^{\tau}[\alpha_l\cos(\omega_l\tau) + \beta_l\sin(\omega_l\tau)], \;\; \tau = 0, 1, ..., M-1 \tag{10}$$

where $\alpha_l = \zeta_l\cos\upsilon_l$, $\beta_l = -\zeta_l\sin\upsilon_l$. Each of the θ components in (10) for $0 < \omega_k < \pi$, corresponds to a particular formant.

3. FORMANT ESTIMATION IN NOISE

In the presence of additive noise $v(n)$, the observed noisy speech $y(n)$ can be written as

$$y(n) = x(n) + v(n) \tag{11}$$

and assuming $v(n)$ is uncorrelated with the input $u_{imp}(n)$, $r_y(\tau)$, the ACF of $y(n)$, can be expressed as

$$r_y(\tau) = r_x(\tau) + r_v(\tau). \tag{12}$$

In the case of additive white Gaussian noise, one can obtain

$$r_x(\tau) = \begin{cases} r_y(\tau) - \sigma_v^2, & \tau = 0 \\ r_y(\tau), & \tau \neq 0. \end{cases} \tag{13}$$

However, under the heavy noisy conditions, estimation of $r_x(\tau)$ using (12) or (13) may cause significant error at all lags resulting in poor pole estimates for conventional correlation based methods. To alleviate this problem, we propose a correlation-fitting approach where an ARBLS error minimization algorithm is used to estimate the model parameters. Note that to reduce the noise effect in correlation-fitting, $\tau > 0$ is considered.

The parameters $\{r_l, \omega_l, \alpha_l, \beta_l\}$ of each component $G_l(\tau)$ of (10) with $\tau > 0$ are determined such that the total squared error between the (l–1)th residual function and $G_l(\tau)$ is minimized. The lth residual function can be defined as

$$\Re_l(\tau) = \Re_{l-1}(\tau) - G_l(\tau), \; \Re_0(\tau) = r_y(\tau), l = 1,2,..., \theta-1 \tag{14}$$

and the objective function can be formulated as

$$J_l = \sum_{\tau=1}^{M-1}\left|\Re_{l-1}(\tau) - G_l(\tau)\right|^2, \; l = 1, 2, ..., \theta. \tag{15}$$

For each set of chosen values of r_l and ω_l, the optimum values of α_l and β_l can be obtained by making zero the partial derivatives of J_l with respect to r_l and ω_l. The values of $\hat{r}_l$ and $\hat{\omega}_l$ corresponding to the global minimum of J_l, are selected as an estimate of formant frequencies if $0 < \omega_l < \pi$.

In order to suppress the natural spectral tilt of the signal, a first order pre-emphasis filter is used. We perform a frame by frame analysis considering an overlapping window. Although in the previous section rectangular window is considered implicitly, the performance remains almost same even if the hamming window is used.

In the ARBLS method, unlike the conventional harmonic retrieval methods, K formant frequencies are sequentially determined from K number of stages. The possible extreme ranges of the formants (ROF) (both frequency and bandwidth) are available in literature and are utilized to restrict the search space [1], [7]. In each stage of the ARBLS algorithm, an updated initial frequency estimate is used. At the first step, the frequency candidate for the first formant (F1) is estimated from the smoothed spectral peaks of noisy observations. The candidate (with the largest peak) inside the desired region specified by the ROF is taken as an

Table I: RMSE for synthetic vowels

Vowels			SNR = 0 dB			SNR = 5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	43.0	128.1	244.7	36.7	104.9	98.5
		F2	75.2	358.6	485.8	33.4	177.4	140.5
		F3	198.2	403.3	461.9	128.9	305.6	182.1
	/u/	F1	93.0	294.2	792.1	50.6	96.2	143.4
		F2	205.4	711.2	720.1	57.4	232.2	204.1
		F3	252.1	774.7	620.2	176.0	376.8	317.0
	/i/	F1	129.8	254.8	982.3	57.7	127.1	549.6
		F2	77.1	156.2	53.8	38.2	130.1	47.2
		F3	100.2	268.5	164.8	71.8	144.5	71.6
Female	/a/	F1	98.6	185.2	265.5	70.0	122.9	278.3
		F2	77.4	186.0	205.0	60.3	124.3	116.2
		F3	80.0	92.16	80.7	73.8	81.3	72.9
	/u/	F1	134.9	396.1	379.2	93.3	117.9	132.0
		F2	140.3	599.6	356.1	125.5	241.5	201.1
		F3	163.8	519.5	259.4	115.8	287.1	152.9
	/i/	F1	153.4	629.8	490.2	57.7	203.0	233.8
		F2	179.3	537.4	319.2	107.3	119.2	142.6
		F3	51.7	244.3	47.5	34.2	75.7	25.4

initial estimate and the frequency search is performed in its neighborhood. For the initial estimates at the remaining stages of the ARBLS algorithm, smoothed spectral peaks of the residue functions obtained for the corresponding stages are used where similar to the case of first formant, we utilize the practical knowledge of the ROF for different formants [1], [7].

4. SIMULATION RESULTS AND DISCUSSION

The proposed formant frequency estimation algorithm has been tested using some synthetic and natural vowels, and some natural sentences taken from the well-known TIMIT speech database. Recently, a reference database for the VT resonances (VTR) of a large number of TIMIT sentences is reported in [8]. The VTR database is carefully used (keeping in mind the differences between VTR and formant frequencies, [1]) as a reference for the TIMIT sentences. For the performance comparison, we consider the LPC (14th order) and the adaptive filter-bank (AFB) methods [5].

At first we present results for three synthetic vowels, /*a*/, /*u*/, and /*i*/, corrupted by white Gaussian noise. Vowels with duration of 200 ms are synthesized using the Klatt synthesizer considering the pitch values of 120 Hz and 220 Hz, respectively, for male and female speakers. We perform the formant estimation every 10 ms with a 20 ms window only for the voiced frames. In the ARBLS algorithm, the search range of r_l is chosen as $0.8 \le r_l \le 0.99$ for F3, and $0.85 \le r_l \le 0.99$ for F2 and F1, and ω_l is searched ±10% of π around the initial estimates. An acceptable level of estimation accuracy can be achieved with a search resolution of $\Delta r = 0.01$ for r_l and $\Delta\omega = \pi/100$ for ω_l. The number of lags for the ACF is set to be $M < T/2$.

Table II: Estimated mean and standard deviation for natural vowels

Fi	Male (/*a*/)				Female (/*i*/)			
	Ref.	Prop.	LPC	AFB	Ref.	Prop.	LPC	AFB
F1	754 (33)	778 (38)	815 (58)	826 (51)	435 (3)	443 (12)	422 (17)	451 (28)
F2	1369 (22)	1377 (62)	1416 (85)	1401 (72)	2638 (35)	2616 (51)	1990 (126)	2284 (141)
F3	2402 (29)	2429 (91)	2551 (123)	2516 (101)	3250 (70)	3259 (110)	2831 (128)	2976 (116)

We have computed the root-mean-square errors (RMSE) in the estimation of each formant frequency (Fi) at a particular level of SNR. The mean operation, required in the computation of RMSE, is performed over different voiced frames and 20 independent trials. Table I shows the RMSE values (Hz) obtained by different methods at SNR = 0 and 5 dB. Clearly, the performance of the proposed (Prop.) method is superior to that of the other methods, and at SNR = 0 dB the RMSE values for other methods increase significantly.

Next, four natural vowels /*a*/, /*u*/, /*i*/, and /*e*/ are taken from [9] with the reference formant values. The vowels were contained in the words "hod", "hood", "heed", and "head". For the purpose of analysis, pitch periods (T) are determined from the noise-free speech signals using the autocorrelation method [7]. In Table II, the estimation performance in terms of mean and standard deviation (shown in the parenthesis) for a male vowel /*a*/ and a female vowel /*i*/ corrupted by white Gaussian noise is presented at SNR = 5 dB. It is evident that the proposed method is able to estimate the formant frequencies quite accurately in both cases. Due to the high level of energy in the frequency band of F1, estimation error in F1 is low for all three methods at SNR = 5 dB. However, the low level of energy in the F3 band and the presence of strong background noise make F3-estimation difficult. But the proposed method is capable of overcoming these difficulties by employing the ARBLS algorithm to extract the correlation model parameters which give the formant frequencies. In Fig. 1, the effect of noise on the estimation errors in terms of RMSE is plotted for the same male vowel /*a*/ from SNR = 0 dB to 40 dB. The difference in the RMSE values between the proposed and other methods is quite high for F3. It is to be mentioned that the estimation accuracy of the proposed method for these four vowels is also investigated in the presence of babble noise. In this case, the level of performance achieved by the proposed method in comparison to that of the other methods is almost similar to the case of white noise environment.

Finally, we present the estimation results for a natural male utterance "Rob sat by the pond" which is taken from the TIMIT database (sampling frequency = 16 KHz). In Fig. 2 (a) the reference formant frequencies are plotted on the spectrogram of noise-free speech [8]. The estimated formants by using different methods at SNR = 5 dB in the presence of white noise are plotted on the spectrogram of the

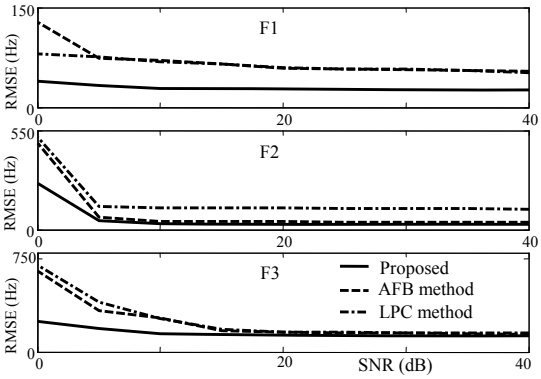

Fig. 1. Effect of SNR on RMSE (Hz).

noise-free signal to clearly show the formant estimation accuracy during the voiced regions. For fare comparison, voicing decisions are taken from the AFB method [5]. Formant frequencies are estimated only in the voiced frames (dark lines in the spectrogram) and the interval between the two voiced frames are just end-point connected (dotted lines in the spectrogram) for Figs. 2(a) to 2(c). Since the AFB method works on sample by sample, it provides values also for those intervals. It is evident that the proposed method provides better estimation accuracy even in the region showing a rise-fall pattern. However, the estimation accuracy degrades for F3 during the last few frames; the reason is same as explained earlier.

5. CONCLUSION

A correlation model for the voiced speech is proposed for the formant frequency estimation in the presence of noise. It is shown that even at a low SNR, the use of an adaptive residue-based least-square optimization algorithm can provide accurate estimation of the correlation model parameters, which are used to calculate the formant frequencies. Analyses on some synthetic and natural speech segments have been performed under noisy condition in order to evaluate the accuracy of the proposed formant frequency estimator. Simulation results show that the proposed method can provide fairly accurate formant frequency estimates at moderate to even low SNR.

6. REFERENCES

[1] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 2, pp. 425–434, Mar. 2006.

[2] M. Lee, J. V. Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 741–750, Sept. 2005.

[3] B. Chen and P. C. Loizou, "Formant frequency estimation in noise," in *Proc. ICASSP'04*, May 2004, vol. 1, pp. 581–584.

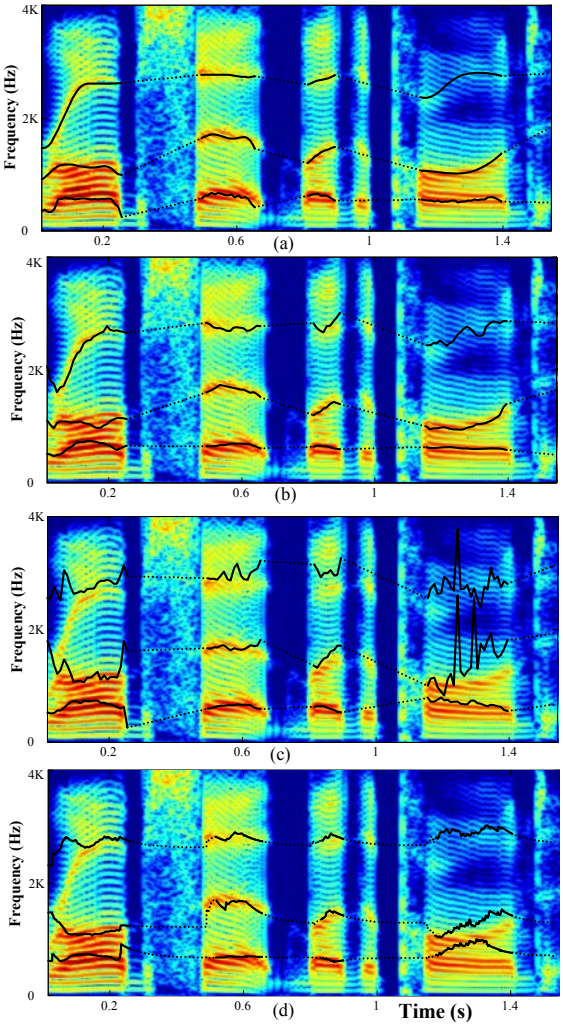

Fig. 2. Formant estimation results for a male utterance "Rob sat by the pond" at SNR = 5 dB plotted on clean speech spectrogram. (a) Reference, (b) Proposed, (c) LPC, and (d) AFB methods.

[4] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 4, pp. 313–327, July 1998.

[5] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 2, pp. 435–444, Mar. 2006.

[6] S. A. Fattah, W.–P, Zhu, and M. O. Ahmad, "An approach to ARMA system identification at a very low signal-to-noise ratio," in *Proc.* ICASSP'05, Mar. 2005, vol. 4, pp. 113–116.

[7] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, NY, second ed., 2000.

[8] L. Deng; X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP'06*, May 2006, vol. 1, pp. 369–372.

[9] J. M. Hillenbrand, L.A. Getty, M. J. Clark, and K. Wheeler, ''Acoustic characteristics of American English vowels,'' *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, May 1995.