A FORMANT FREQUENCY ESTIMATOR FOR NOISY SPEECH BASED ON CORRELATION AND CEPSTRUM

Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad
Dept. of Electrical and Computer Engineering
Concordia University, 1455 De Maisonneuve Blvd. W., Montreal, Quebec, Canada H3G 1M8

1. INTRODUCTION

Formant is one of the most informative speech features which helps in interpreting the mechanisms of human speech production. Formant frequency estimation of speech has numerous applications, such as, speech recognition, synthesis, and compression. As an acoustic feature, formant offers phonetic reduction in speech recognition [1]. Formants are associated with peaks in the smoothed power spectrum of speech. Among different formant estimation techniques, linear predictive coding (LPC) based methods are most commonly used [1]. Cepstrum features are also used in the formant estimation. Most of the formant frequency estimation methods, so far reported, deal only with the noise-free environments. The estimation performance of correlation based formant estimators deteriorates noticeably in the presence of noise. Formant estimation from noisy speech is a difficult but essential task as far as practical applications are concerned. Recently in [2] and [3], methods have been proposed in order to handle noisy environments. The method proposed in [2] is based on an adaptive band-pass filter bank (AFB) where the estimation accuracy depends on initial estimates. The formant frequency estimation approach, proposed in [3], is based on a correlation model of voiced speech signals.

The objective of this paper is to develop a formant estimation scheme combining the advantageous features of correlation and cepstral domains, which is capable of handling the adverse effect of observation noise. We propose a ramp cepstrum model of a once-repeated autocorrelation function (ORACF) of voiced speech signals in terms of formant parameters. It has been shown that in comparison to the conventional ACF, the ORACF can drastically reduce the effect of additive noise. A residue based least-squares optimization technique based on a model-fitting approach is introduced in order to obtain formant frequencies from noisy observations. Simulations are carried out to estimate formant frequencies from synthetic and natural speech signals under noisy conditions.

2. PROPOSED METHOD

2.1 A Ramp Cepstrum Model of ORACF of Speech

The human vocal-tract system can be represented by a P-th order AR system with a transfer function given by

$$H(z) = \frac{G}{\prod_{k=1}^{P}(1 - p_k z^{-1})} \tag{1}$$

where G is the gain factor and p_k is the system pole. In order to model each formant, a pair of complex conjugate poles is required. Formant frequency (F_k) and bandwidth (B_k) can be computed from the pole magnitude r_k, angle ω_k, and sampling frequency F_S as

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

The complex-cepstrum of the impulse-response $h(n)$ of the vocal-tract filter is defined as

$$c_h(n) = F^{-1}\{\ln(H(e^{j\omega}))\} \tag{3}$$

where $F^{-1}\{\cdot\}$ denotes the inverse Fourier transform (FT). From (1) and (3), assuming a minimum phase system, $c_h(n)$ can be expressed in terms of system poles as

$$c_h(n) = \sum_{k=1}^{P} \frac{p_k^n}{n}, \quad n > 0 \tag{4}$$

During a short duration of time, an observed speech $x(n)$ can be assumed to be stationary. The FT of the ACF of $x(n)$, $r_x(\tau)$, can be written as

$$R_x(e^{j\omega}) = \left|H(e^{j\omega})\right|^2 R_u(e^{j\omega}) \tag{5}$$

where $R_u(e^{j\omega})$ is the FT of the ACF of the input excitation $u(n)$. In the cepstum-based speech analysis, generally, cepstral coefficients are computed from observed speech or the estimate of its non-parametric power spectral density (PSD). The cepstral coefficients thus estimated, especially in the presence of observation noise, provide poor estimation accuracy. Hence, we propose to utilize a once-repeated ACF (ORACF) $\varphi_x(\tau)$ that exhibits a higher noise immunity. The FT of $\varphi_x(\tau)$ can be expressed as

$$\Phi_x(e^{j\omega}) = \left|R_x(e^{j\omega})\right|^2 \tag{6}$$

Using (5) and (6), cepstrum corresponding to $\Phi_x(e^{j\omega})$ can be expressed as

$$c_{\Phi_x}(n) = F^{-1}\{\ln(\Phi_x(e^{j\omega}))\} = c_h'(n) + c_u'(n) \tag{7}$$

For a voiced speech, the input excitation $u(n)$ is treated as a periodic impulse-train with period T and it is found that $c_u'(n)$ exhibits significant values at origin and multiples of the period. Thus $c_u'(n)$ vanishes in the region $0 < n < T$. It can be shown that $c_h'(n) = 2c_h(n)$ in this region and thus (7) reduces to

$$c_{\Phi_x}(n) = 2\sum_{i=1}^{P} \frac{p_i^n}{n}, \qquad 0 < n < T \tag{8}$$

It is clear that the cepstrum decays as $1/n$, which is difficult to handle. Hence, we propose a ramp-cepstrum $\psi_x(n)$, which for real values of $x(n)$ can be expressed as

$$\psi_x(n) = n c_{\Phi_x}(n) = \sum_{i=1}^{K} \eta(\omega_i) r_i^n \cos(\omega_i n), \quad 0 < n < T \tag{9}$$

where K = number of real poles + the number of complex conjugate pole pairs, and $\eta(\omega_i) = 2$ if $\omega_i = 0$ or π, otherwise $\eta(\omega_i) = 4$. Equation (9) is the ramp-cepstrum model of voiced speech signals. Note that each of the K components in (9) for $0 < \omega_k < \pi$, corresponds to a particular formant. Next we will develop a model-fitting algorithm to estimate formants.

2.2 Formant Estimation Algorithm in Noise

The noise-corrupted speech signal is given by

$$y(n) = x(n) + v(n) \tag{10}$$

where, the additive noise $v(n)$ is assumed to be zero mean with variance σ_v^2. The ACF of $y(n)$ can be expressed as

$$\begin{aligned} r_y(\tau) &= r_x(\tau) + r_w(\tau) \\ r_w(\tau) &= r_v(\tau) + r_{vx}(\tau) + r_{xv}(\tau) \end{aligned} \tag{11}$$

The effect of noise term $r_w(\tau)$ on $r_x(\tau)$ is relatively less pronounced since, the crosscorrelation terms in (11) are negligible and $r_v(\tau)$ mostly affects only the zero lag. Hence, in the ORACF of $y(n)$, i.e., the ACF of $r_y(\tau)$, the effect of noise term will be drastically reduced. It can be shown that the ORACF, like the ACF, preserves the poles of the vocal-tract AR system. The ramp-cepstrum computed from the ORACF of $y(n)$ can be expressed as

$$\psi_y(n) = \psi_x(n) + \psi_n(n), \; 0 < n < T \tag{12}$$

Due to the error term $\psi_n(n)$, it is difficult to estimate $\psi_x(n)$ from $\psi_y(n)$ in a noisy condition. In order to overcome this problem, we exclude $r_y(0)$ in computing $\psi_y(n)$ at a low SNR, which significantly reduces the strength of noise term $\psi_n(n)$. Resulting noise-reduced $\psi_y(n)$ is then used to extract the ramp-cepstrum model parameters through a residue-based least-square optimization. The parameters of each component $F_l(n)$ of the ramp-cepstrum model (10) are determined such that the total squared error between the $(l-1)$th residual function and an estimate of $F_l(n)$ is minimized. The lth residual function can be defined as

$$\Re_l(n) = \Re_{l-1}(n) - F_l(n); \; \Re_0(n) = \psi_y(n); \; l = 1, 2, ..., K-1 \tag{13}$$

and the objective function for the minimization is formed as

$$J_l = \sum_{n=1}^{M-1} \left| \Re_{l-1}(n) - F_l(n) \right|^2, \quad l = 1, 2, ..., K; \; M < T \tag{14}$$

Values of $\hat{r}_l$ and $\hat{\omega}_l$, corresponding to the global minimum of J_l, are selected to compute the estimate of the lth formant. Thus different formant frequencies are sequentially determined. In the poposed optimization technique, r_l and ω_l are searched within a certain search space. The region of formants (both frequency and bandwidth) is available in literature and utilized to restrict the search space [1], [3]. Further reduction in the frequency search space is achieved by using an initial estimates obtained through smoothed spectral peaks of the zero lag excluded ACF.

In our implementation, we perform the formant estimation every 10 ms with a 20 ms window applied to overlapping speech segments. In addition to signal pre-emphasis a FFT pre-filtering is performed to remove very low frequencies (<100 Hz) which are not in our interest.

3. SIMULATION RESULTS

The proposed formant frequency estimation algorithm has been tested using various synthetic vowels synthesized using the Klatt synthesizer [1] and some natural vowels extracted from the TIMIT and North-Texas standard databases with corresponding reference values [1], [4]. For the performance comparison, the 12th order LPC [1] and the AFB methods [2] are considered and the percentage root-mean-square error (RMSE) at different noise levels are computed where each noise level consists of 20 independent trials of noisy environments. In the proposed optimization algorithm, the search range of r_l is chosen as $0.8 \le r_l \le 0.99$, and that for ω_l is $\pm 0.1\pi$ near initial estimates [3]. In Table 1, the estimated %RMSE (Hz) is shown for two synthesized vowels at SNR = 0 dB and 5 dB. It is clearly observed that the proposed method provides lower %RMSE for both male and female speakers. In Table 2, estimation accuracy in terms of %RMSE (Hz) for natural vowels /a/ and /i/ (contained in the words "hod" and "heed") are shown. It is found that the proposed method provides better estimation accuracy at both conditions.

Table 1. %RMSE (Hz) for Synthetic Vowels

Vowels			0 dB			5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	9.78	23.63	31.29	7.02	15.87	11.74
		F2	9.65	27.78	34.82	3.12	15.53	9.51
		F3	8.39	19.28	19.34	5.43	13.19	8.23
	/i/	F1	19.27	28.53	28.16	12.51	19.28	16.25
		F2	3.95	9.68	4.27	2.72	7.54	3.33
		F3	4.93	13.29	7.75	3.71	7.82	3.97
Female	/a/	F1	11.15	17.76	16.76	4.84	9.76	15.67
		F2	9.01	19.43	14.39	4.51	8.68	7.49
		F3	4.54	9.26	4.57	2.08	3.18	2.34
	/i/	F1	21.34	39.81	32.27	13.59	28.27	19.14
		F2	9.56	26.21	13.78	4.81	12.43	6.19
		F3	2.89	15.83	3.78	2.12	7.63	2.78

Table 2. %RMSE (Hz) for Natural Vowels

Vowels			0 dB			5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	9.85	14.33	13.93	6.72	9.57	8.54
		F2	14.52	44.93	28.76	11.26	28.27	16.29
		F3	12.25	38.01	23.34	10.39	23.67	18.31
Female	/i/	F1	9.84	21.19	16.84	5.03	8.61	5.95
		F2	10.25	23.78	19.49	5.96	11.28	10.21
		F3	10.11	31.29	24.82	5.05	21.92	14.58

4. CONCLUSION

A formant frequency estimation scheme based on a new ramp cepstrum model has been developed which is capable of efficiently handling the noisy environment. In the development of the ramp cepstrum model, the once-repeated ACF is employed which can significantly reduce the effect of noise in the correlation domain. It has been shown that even at a low level of SNR, the proposed residue based least-squares optimization algorithm for the model-fitting provides an accurate estimation of the formant frequencies. From experimental results on synthetic and natural speech signals under a noisy condition, it has been found that the proposed method provides an accurate formant frequency estimate even at a low level of SNR.

REFERENCES

[1] D. O'Shaughnessy, (2000). Speech Communications: Human and Machine (2nd ed.). *IEEE Press, NY*.

[2] K. Mustafa and I. C. Bruce, (2006). Robust formant tracking for continuous speech with speaker variability. *IEEE Trans. Audio Speech Lang. Processing*, 14, 435–444.

[3] S. A. Fattah, W. -P. Zhu, and M. O. Ahmad, (2007). An approach to formant frequency estimation at low signal-to-noise ratio. *ICASSP '07*, 4, 469–472.

[4] J. M. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97, 3099–3111.