AN ALGORITHM FOR FORMANT FREQUENCY ESTIMATION FROM NOISE-CORRUPTED SPEECH SIGNALS

Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad
Centre for Signal Processing and Communications, Dept. of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8
{shaik_fa, weiping, omair}@ece.concordia.ca

1. INTRODUCTION

Formant frequency estimation of speech signals plays an important role in speech synthesis, compression, and recognition. For example, formant information serves as a significant acoustic feature and offers a phonetic reduction in speech recognition. It plays a vital role in the design of some hearing aids [1]. Free resonances of the vocal-tract (VT) system are called formants. Formants are associated with peaks in the smoothed power spectrum of speech. Among different formant estimation techniques, linear predictive coding (LPC) based methods have received considerable attention [2]. In this case, formant frequencies are computed from the autoregressive (AR) parameters of the VT system. Most of the formant frequency estimation methods, so far reported, deal only with noise-free environments. However, formant estimation from noisy speech signals is difficult but an essential task as far as practical applications are concerned. In order to handle noisy environments, recently in [1], a method based on an adaptive band-pass filter-bank, later referred as AFB method, has been proposed where the estimation accuracy depends on initial estimates.

In this paper, we propose a scheme for accurately estimate the formant frequencies of speech signals in a noisy environment. First, a frequency domain noise reduction scheme is proposed. The ACF of the resulting noise-compensated speech signal is then employed in a modified form of least-squares Yule-Walker equations (LSYWE) to estimate poles of the VT system. Finally, a pole selection criterion has been introduced for extracting the desired formants which enables the proposed scheme to avoid errors associated with weak formants. Experimental results on natural and synthetic vowels in the presence of additive white noise are presented.

2. PROPOSED METHOD

The overall vocal-tract (VT) filter can be represented by a P-th order AR system with a transfer function given by

$$H(z)=\frac{G}{A(z)}=\frac{G}{1-\sum_{i=1}^{P}a_i z^{-i}}=\frac{G}{\prod_{i=1}^{P}(1-p_i z^{-1})} \tag{1}$$

where G is the gain factor, $\{a_i\}$ the AR parameters of the filter $A(z)$, and the system pole $p_k = r_k e^{j\omega_k}$ with a pole magnitude r_k and angle ω_k. During a short duration of time, a speech signal, the output of the VT filter, is generally assumed to be stationary. A pair of complex conjugate poles is required to model the formant frequency (F_k) and bandwidth (B_k), respectively, defined as

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

where F_S is the sampling frequency. In the autocorrelation method, formants are estimated from the AR parameters using a least-squares (LS) solution of the following relation [2]

$$r_x(\tau)=\sum_{k=1}^{P}a_k r_x(\tau-k),\ \tau=1,2,\ldots\ldots,P \tag{3}$$

where $r_x(\tau)$ is the autocorrelation function (ACF) of an N length sequence $x(n)$ and its estimate can be computed as

$$r_x(\tau)=\frac{1}{N}\sum_{n=0}^{N-1-|\tau|}x(n)x(n+|\tau|),\quad \tau=0,1,\ldots\ldots,L'-1;L'\le N \tag{4}$$

In order to reduce the estimation variance, a combination of more than P equations can be used in the LS estimation.

In the presence of noise $v(n)$, the observed signal is given by

$$y(n)=x(n)+v(n) \tag{5}$$

The ACF of $y(n)$ can be expressed as

$$r_y(\tau)=r_x(\tau)+r_w(\tau),\ r_w(\tau)=r_v(\tau)+r_{vx}(\tau)+r_{xv}(\tau) \tag{6}$$

It can be observed that in the presence of $r_w(\tau)$, the ACF introduced by the noise, it is difficult to obtain an accurate estimate of $r_x(\tau)$. Even in the case of white Gaussian noise, at a heavy noisy condition, the effect of $r_w(\tau)$ on $r_x(\tau)$ cannot be completely neglected just after the zero lag. In order to reduce the noise effect, instead of directly using the noisy signal, first, we proposed to obtain a noise-compensated signal as

$$\tilde{y}(n)=F^{-1}\left[|Y_C(\omega)|e^{j\angle Y(\omega)}\right],\quad Y_C(\omega)=|Y(\omega)|-\lambda|\hat{Y}_{HF}(\omega)| \tag{7}$$
$$\text{If } Y_C(\omega)\le 0,\ \text{set } Y_C(\omega)=0$$

where $Y(\omega)$ is the Fourier transform (FT) of $y(n)$, $\hat{Y}_{HF}(\omega)$ is the average value computed from the high frequency region of $Y(\omega)$, and λ is a scaling factor with $\lambda \leq 1$ to avoid overcompensation. Since, in practical applications it is sufficient to estimate first few formants and speech spectrum shows a decaying nature, only the high frequency region is considered to obtain $\hat{Y}_{HF}(\omega)$. Note that, in the inverse FT operation, we have utilized the phase information of the noisy signal and negative values of $Y_C(\omega)$ are forced to zero to obtain a better result in a noisy environment. In addition, very low frequency region (<100 *Hz*) is excluded which is not in our interest. Thus, the effect of additive white Gaussian noise $v(n)$ on $x(n)$ is significantly reduced and from $\tilde{r}_y(\tau)$, the ACF of $\tilde{y}(n)$, one may obtain an estimate of $r_x(\tau)$ as

$$\hat{r}_x(\tau) = \begin{cases} \tilde{r}_y(\tau) - r_v(\tau), & \text{for } \tau = 0 \\ \tilde{r}_y(\tau), & \text{for } \tau \neq 0 \end{cases} \quad (8)$$

Since, the zero lag value causes a significant error, considering $\tau = P + 1, \ldots, P + S$ in (4), a modified form of LSYW equations excluding zero lag can be obtained as

$$\begin{bmatrix} \tilde{r}_y(P) & \tilde{r}_y(P-1) & \ldots & \tilde{r}_y(1) \\ \tilde{r}_y(P+1) & \tilde{r}_y(P) & \ldots & \tilde{r}_y(2) \\ \vdots & \vdots & & \vdots \\ \tilde{r}_y(P+S-1) & \ldots & \ldots & \tilde{r}_y(S) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = - \begin{bmatrix} \tilde{r}_y(P+1) \\ \tilde{r}_y(P+2) \\ \vdots \\ \tilde{r}_y(P+S) \end{bmatrix} \quad (9)$$

Here S governs the number of equations. An estimate of AR parameters can be obtained from the LS solution of (9) as

$$\hat{\boldsymbol{a}} = (\acute{\mathbf{R}}^{\mathrm{T}} \acute{\mathbf{R}})^{-1} \acute{\mathbf{R}}^{\mathrm{T}} \acute{\mathbf{r}} \quad (10)$$

Thus, system poles are estimated and an estimate of formant frequencies can be obtained using (2). In our implementation, we perform the formant estimation every 10 ms with a 20 ms window applied to overlapping voiced speech segments. In literature, the region of formant frequencies has been well studied [2]. There exists a high degree of overlaps between the formant frequency regions. First, we estimate F1 from a pole with the highest magnitude inside the F1-region. The pole, outside F1 but within F2-region, having the highest magnitude is chosen for F2. In this way the three formants are estimated. A spectral peak-picking operation is performed to estimate a formant if no pole is obtained in a particular region.

3. SIMULATION RESULTS

The proposed formant frequency estimation algorithm has been tested using different synthetic vowels synthesized using the Klatt synthesizer [2] and natural vowels from the North Texas speech database [3], and some natural sentences from the TIMIT speech database with their reported formant references [4]. For the performance comparison, the 12th order LPC [2] and the AFB methods [1] are considered and the percentage root-mean-square error (RMSE) at different noise levels are computed with $S = 4P$ where each noise level consists of 20 independent trials of noisy environments.

Table 1. %RMSE (Hz) for Synthetic Vowels

Vowels			0 dB			5 dB		
			Prop.	LPC	AFB	Prop.	LPC	AFB
Male	/a/	F1	12.63	23.63	31.29	8.93	15.87	11.74
		F2	13.57	27.78	34.82	4.62	15.53	9.51
		F3	11.65	19.28	19.34	7.28	13.19	8.23
	/i/	F1	23.54	28.53	28.16	15.93	19.28	16.25
		F2	4.02	9.68	4.27	2.94	7.54	3.33
		F3	5.78	13.29	7.75	3.81	7.82	3.97
Female	/a/	F1	11.51	17.76	16.76	5.71	9.76	15.67
		F2	9.52	19.43	14.39	5.35	8.68	7.49
		F3	3.84	9.26	4.57	2.11	3.18	2.34
	/i/	F1	30.29	39.81	32.27	17.75	28.27	19.14
		F2	11.92	26.21	13.78	5.89	12.43	6.19
		F3	3.04	15.83	3.78	2.31	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	12.14	14.33	13.93	7.34	9.57	8.54
		F2	19.29	44.93	28.76	14.62	28.27	16.29
		F3	14.93	38.01	23.34	12.28	23.67	18.31
Female	/i/	F1	10.72	21.19	16.84	5.72	8.61	5.95
		F2	13.17	23.78	19.49	6.81	11.28	10.21
		F3	15.48	31.29	24.82	5.28	21.92	14.58

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 is able to provide lower %RMSE at low SNRs 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.

4. CONCLUSION

The proposed formant frequency estimation algorithm is capable of handling noisy environments, since, a noise-compensated speech signal is used in a modified form of LSYW equations along with an effective formant selection criterion. Experimental results on natural and synthetic speech signals show the efficacy of the proposed method in estimating formant frequencies at a moderate to low levels of SNR.

REFERENCES

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

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

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

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