A Missing-Data Approach to Noise-Robust LPC Extraction for Voiced Speech Using Auxiliary Sensors

C. Demiroglu and T. Barnwell

Department of Electrical and Computer Engineering
Georgia Institute of Technology, USA
demirogc,tom@ece.gatech.edu

Abstract

Noise robust LPC extraction from the voiced speech signal is addressed with a missing-data approach. Harmonics in the voiced speech spectrum are detected using a General Electromagnetic Motion Sensor (GEMS) that is immune to acoustic background noise. Non-harmonic frequencies are treated as missing-data and severely suppressed while no processing is done on the harmonic frequencies since they are assumed to have high SNRs. Objective measure tests using the log likelihood ratio (LLR) show significant improvement over the noisy case for severely noisy environments.

1. Introduction

Linear Predictive Coding (LPC) is a widely used tool for modeling the envelope of the speech spectrum. It has been used in many speech applications such as parametric speech coders [1] and automatic speech recognition (ASR) [2].

LPC creates a perceptually attractive model of the spectral envelope since it models the perceptually important spectral peaks more accurately than the spectral valleys [1]. However, in additive noise environments formant peaks are smoothed out and/or shifted in the spectrum, which significantly reduces the quality of the coded speech [3]. Moreover, performance of the LPC-based ASR systems drop substantially with increased background noise [2]. Therefore, it is important to develop new LPC extraction algorithms that are immune to background noise. Such algorithms can find applications in all speech applications where the LPC method is used.

In this work, a new noise robust LPC extraction system is proposed that relies on the fact that LPC can still model the spectral envelope of voiced speech when a significant portion of the spectrum is missing. For example, if pitch and LPC order are low enough, LPC can generate an accurate spectral envelope using only the information at the harmonic frequencies. In a noisy environment, suppressing the signal at the non-harmonic frequencies can substantially increase the SNR and the accuracy of the LPC estimator.

Detecting voicing and harmonic locations accurately in a noisy environment are challenging tasks. We propose using the General Electromagnetic Motion Sensor (GEMS) device in addition to the acoustic microphone for these tasks. The GEMS device is an electromagnetic sensor, and therefore it is relatively immune to acoustic background noise. Moreover, it provides information about the excitation signal when a voiced sound is articulated. Another advantage of the GEMS device is that it is not cumbersome for daily use and can be employed in commercial speech applications.

This paper is organized as follows. A description of the GEMS device is given in Section 2. The idea behind noise robust LPC extraction from voiced speech is described in Section 3. The proposed system is discussed in Section 4. Experiment results are presented in Section 5, and the paper is concluded in Section 6.

2. The Auxiliary Sensor

One of the primary tasks associated with the missing-data approach is determining parts of the signal that exhibit high SNR and parts that have a low SNR. In this work, an additional sensor is used to provide this information. Several sensors exist which would work well in this category including throat accelerometers, physiological microphones (p-mics), bone-conduction microphones, or electromagnetic glottal or vibration sensors. All of these have a low-pass characteristic and most do not do a good job of reproducing vocal-tract modulation of the glottal spectrum. However, all of them can be used to identify voicing and pitch/harmonic. For this paper, we report on results generated using the GEMS device from Aliph, Inc.

The general electromagnetic sensor (GEMS), is a micropower device that can be used, among other things, to detect motion in the region of glottis. The GEMS device consists of a penetrating radar whose principles have been studied extensively both at the Lawrence-Livermore Laboratory and Aliph, Inc. Descriptions of its properties can be found in [4].

When positioned correctly on the exterior of the throat adjacent to the glottis, the output of the radar during voiced

The GEMS speech coding work is sponsored by the Defense Advanced Research Projects Agency under Contract N00024-02-C-6339, and this paper has been designated "Approved for public release, distribution unlimited." Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Government.

speech is a signal that resembles an ideal excitation waveform. The GEMS device responds to vocal fold vibration at the larynx. The signal obtained is robust to external acoustic influences, such as noise, and it can be used for applications such as noise robust pitch detection and speech enhancement [5], [6]. In this work, the GEMS device is used for noise robust harmonic tracking as described in section 4.2.

3. Noise-Robust LPC Extraction

The LPC spectrum approximates the smooth speech spectrum with an all-pole model. Thus, given the LPC parameters $a_{i,k}$ for frame k, one can model the magnitude spectrum of frame k as

$$\hat{H}_k = \frac{\sigma_k}{\prod_{i=1}^{N}(1 - a_{i,k}z^{-1})} \quad (1)$$

for an N^{th} order LPC model, where σ_k is the spectral gain for frame k. The LPC parameters $a_{i,k}$ are derived using a minimum mean square error (MMSE) estimation method.

The optimization criterion for LPC extraction can be viewed in the spectrum domain where the mean square error is given as

$$\epsilon_k = \int_{-\pi}^{\pi} \frac{|H_k(f)|^2}{|\hat{H}_k(f)|^2} df, \quad (2)$$

where $H_k(f)$ is the spectrum the original speech frame, and $\hat{H}_k(f)$ is the spectral envelope of the LPC model [1].

An interesting property of the LPC spectrum is that complete spectrum is not necessary for generating a smooth spectral estimate [1]. For instance, in voiced speech the smooth vocal tract spectrum is naturally sampled by the excitation signal at the harmonic frequencies, and the spectral values are unknown at the non-harmonic frequencies. Still, the LPC method generally works well with voiced speech. In this work, such spectral representations with missing information are called sparse spectral representations.

One way of understanding the way LPC handles the sparse spectrum is by analyzing the error function given in Eq. 2. The speech spectral envelope $H_k(f)$ at the missing frequencies can be assumed 0 for this intuitive analysis. In this case, the ratio in the error term at that frequency is automatically 0 independent of the estimated $\hat{H}_k(f)$. Therefore, the estimation becomes insensitive to the missing spectrum, and LPC fits a model on the rest of the spectrum. As long as the LPC order is low enough, LPC does not follow the fine structure of the spectrum. The resulting LPC spectrum is a good approximation to the actual smooth spectrum with the complete data.

In the next section, the use of sparse spectral representation for noise robust extraction of the LPC parameters from the voiced speech is described.

3.1. LPC Extraction for Voiced Speech

A voiced speech signal is typically modeled with a smooth all-pole transfer function $h(n)$ driven by a quasi-stationary source signal $e(n)$. The spectrum of $e(n)$ is similar to an impulse train where each impulse occurs at the integer multiple of the fundamental frequency F_0. Thus, the voiced speech spectrum $S(k)$ is a sampled version of $H(k)$ at the harmonic frequencies. Although there is still some energy at the non-harmonic locations, it is negligible compared to the energy at the harmonic locations as seen in Fig. 2.

While the lack of spectral information at the non-harmonic frequencies typically does not affect the operation of LPC, the effect of relatively low power at those frequencies should be taken into account in noisy environments. The non-harmonic frequencies can have very low SNRs in a noisy environment, which can significantly distort the LPC spectrum. However, this problem can be an advantage if one can detect the harmonic frequencies and suppress the signal at the non-harmonic frequencies. LPC can model sparse spectral representations smoothly, and does not track the fine harmonic structure as long as the pitch is not too high as discussed in the previous section. The ability of LPC to handle such sparse spectral representations can be used for better spectral modeling in noisy environments. Suppression of noise in non-harmonic frequencies can significantly increase the frame SNR while keeping the perceptually important harmonics. SNR of the noisy signal at the harmonic locations are relatively high since speech spectrum has high power at those frequencies. This assumption is true especially if the noise is concentrated in the low bands where most of the speech signal resides.

4. The Proposed LPC Extraction Algorithm for Voiced Speech

The diagram of the proposed system is shown in Fig. 1. The speech signal spectrum $S(k)$ is labeled voiced or unvoiced using a voicing detector that utilizes the GEMS signal spectrum $R(k)$. If the frame is labeled as voiced, then a harmonic detector module detects the harmonic locations, and the non-harmonic locations are suppressed with a suppression factor of G_{min}. The LPC parameters are extracted from the sparse spectral representation. In the next two sections, voicing and harmonic detection modules are described respectively.

4.1. Using the GEMS Device for Voicing Detection

The GEMS signal can be a very reliable indicator of voiced speech when the sensor is directed at the glottis because it has significantly high energy for the voiced speech segments compared to the unvoiced segments. Moreover, the sensor signal is not affected by the acoustic noise, and it can provide accurate voicing information for all acoustic noise environments which proves to be useful for speech enhancement [5].

A two-stage algorithm is used for detecting voicing from the sensor spectrum $R(k)$. A hard decision energy based algorithm is used to roughly detect the voicing segments in the first stage. An energy threshold of ε_{th} is used for detecting

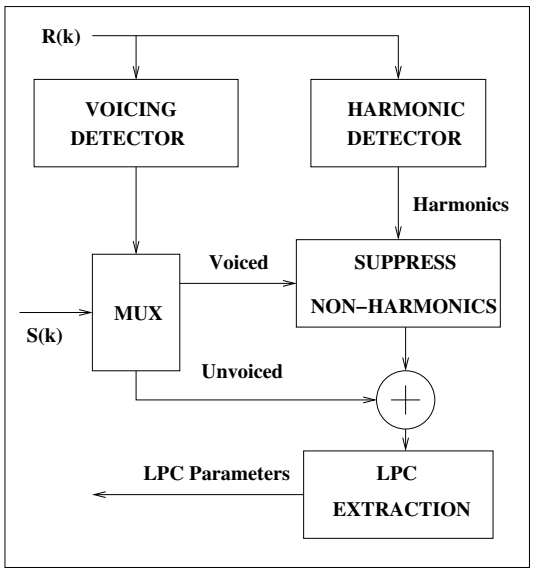

Figure 1: The diagram of the proposed system. $S(k)$ is the spectrum of the speech signal; and $R(k)$ is the spectrum of the radar signal.

voicing at this stage. The voicing feature V_i for frame i is set as

$$V_i = \begin{cases} 1 & \text{if } \varepsilon > \varepsilon_{th}, \\ 0 & \text{if } \varepsilon \leq \varepsilon_{th}, \end{cases}$$

where ε is the energy of the sensor frame.

In the second stage, the rough detection is refined by using a correlation-based approach. The frames that are labeled as unvoiced are not considered in this stage. The frames that are labeled voiced are refined as follows. The lag with the highest value in the autocorrelation function is calculated for lags 2.5 msec to 12 msec. The maximum correlation value ρ is compared with a threshold of ρ_{th}. Finally, the voicing feature V_i for frame is set as

$$V_i = \begin{cases} 1 & \text{if } \rho > \rho_{th}, \\ 0 & \text{if } \rho \leq \rho_{th}, \end{cases}$$

where ρ is the autocorrelation function of the sensor signal.

4.2. Harmonic Detection in Voiced Speech Using the GEMS Device

In addition to voicing information, the GEMS signal is used for detecting the harmonic locations. The GEMS signal has a harmonic structure that is very similar to the acoustic signal as shown in Fig. 2. Thus, if both signals are windowed using the same window, the GEMS device can accurately detect the high signal power (HSP) locations in the voiced speech spectrum. The HSP locations are within the neighborhood of the exact harmonic locations. The exact harmonic locations cannot be detected since the resolution in the frequency domain is limited with the sampling rate. Therefore, harmonic tracking and HSP tracking are used interchangeably in this work.

Table 1: Parameters of the proposed system.

Parameter	Value
ε_{th}	4 dB
G_{min}	-20 dB
ρ_{th}	0.4
F_s	8 kHz

A hard-decision thresholding algorithm is used for detecting the HSP locations; and the binary decisions are stored in the vector P_s.

The GEMS spectrum is divided into 18 subbands with equal bandwidth. The algorithm for detecting the HSP locations in each subband can be described as follows. Three types of high signal power cues are identified in the spectrum:

1. $P_s(k)$ is set to 1 at the highest energy frequency bin in the subband.
2. $P_s(k)$ is set to 1 if the signal power ζ_k is greater than $\zeta_{k-1} + \zeta_{th}$ where ζ_{th} is a power threshold and k is the frequency bin index.
3. Similarly, $P_s(k)$ is set to 1 if the signal power ζ_k is greater than $\zeta_{k+1} + \zeta_{th}$ where ζ_{th} is the same constant used in case 1.

The algorithm attemps to find at least three HSP locations in each subband. An iterative technique is used for tracking the HSP locations. ζ_{th} is initialized to 3 dB. If the number of HSP locations is less than three after the first iteration then ζ_{th} is decreased with a step size of 0.2 dB, and the procedure is repeated until at least three HSP locations are detected, or ζ_{th} is less than 1 dB. P_s is a binary vector, and the elements that are not explicitly set to 1 by the algorithm are by default 0.

5. Experiments

An extensive database was created by ARCON Corporation having simultaneous speech, GEMS, EGG, and other sensor data for various noise conditions. The ARCON database is used for objective testing in this work. Twenty minutes of 8 kHz speech files have been hand-labeled into five phonetic classes that represent voiced phonemes. The log-likelihood ratio (LLR) objective measure is used to compare the performance of the proposed system with the noisy system for each class. The parameters of the proposed system are given in Table 1.

Noise is artificially added to clean speech. Segmental SNRs for both environments are approximately 0 dB in these experiments.

The results are shown for the M2 tank noise and the Blackhawk helicopter noise environments in Tables 2 and 3 respectively. The proposed system reduces LLR significantly for all phoneme classes and both noise types. The distortion

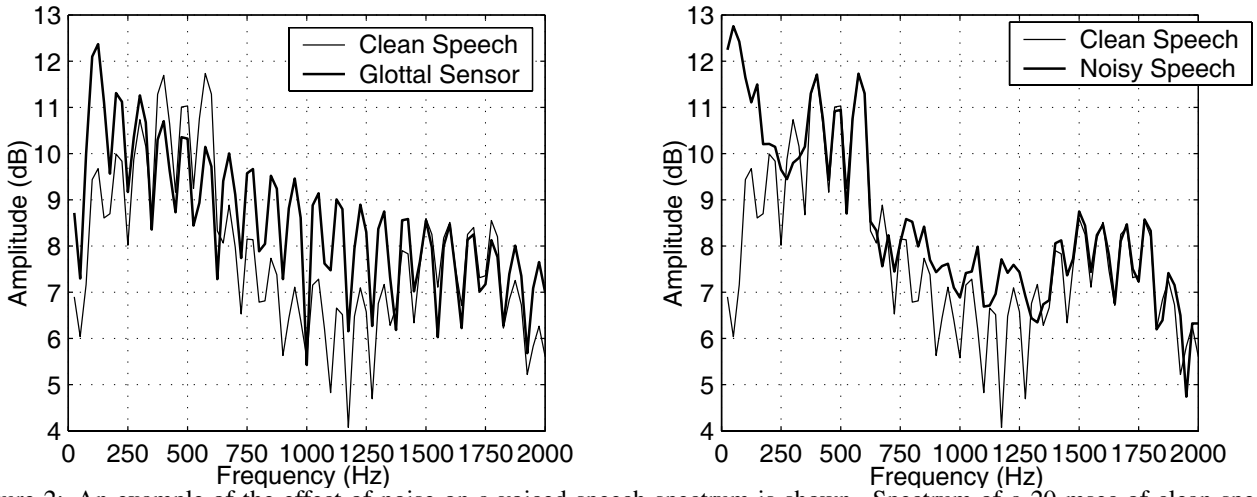

Figure 2: An example of the effect of noise on a voiced speech spectrum is shown. Spectrum of a 20 msec of clean speech segment is shown on the left. On the right, the same speech segment is compared with the case when the M2 tank noise is added on it.

is more severe for the Blackhawk noise case compared to the M2 fighting vehicle noise. Interestingly, the proposed system achieves similar performance for both environments. This behavior is expected since the system makes use of the harmonic frequencies for LPC extraction, and those frequencies typically have high SNRs. The low SNR, non-harmonic frequencies, which severely degrades the performance, are suppressed by the system.

Table 2: Comparison of the proposed system with the noisy system using the log-likelihood distortion measure for the M2 fighting vehicle noise.

Phone Class	Baseline	Proposed System
Voiced Fricative	0.89	0.58
Voiced Plosive	1.11	0.54
Vowel	0.55	0.34
Semivowel	1.47	0.50
Nasal	0.91	0.45

Table 3: Comparison of the proposed system with the noisy system using the log-likelihood distortion measure for the Blackhawk helicopter noise.

Phone Class	Baseline	Proposed System
Voiced Fricative	1.28	0.57
Voiced Plosive	1.89	0.58
Vowel	0.99	0.36
Semivowel	2.35	0.66
Nasal	1.69	0.58

6. Conclusion

A noise robust LPC extraction system for voiced speech is proposed. The GEMS device is used for noise robust voicing detection and harmonic tracking in the voiced speech spectrum. The objective measure test using log-likelihood ratio shows the significant advantage of the proposed system over the noisy case. The proposed system can be used to increase the noise immunity of all speech applications that use the LPC parameters such as ASR and speech coding.

7. References

[1] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*. Elsevier, 1995.

[2] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, Sept. 1993.

[3] S. M. Kay, "The effect of noise on the autoregressive spectral estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, oct 1979.

[4] G. C. Burnett, "The physiological basis of glottal electromagnetic micropower sensors (gems) and their use in defining an excitation function for the human vocal tract," Ph.D. dissertation, University of California Davis, 1999, http://speech.llnl.gov/thesis/.

[5] T. Barnwell, M. A. Clements, D. V. Anderson, E. Moore, M. Lee, A. E. Ertan, V. Krishnan, S. Kamath, W. Choi, J. Hu, C. Demiroglu, P. S. Whitehead, and A. S. Durey, "Low bit rate coding of speech in harsh conditions using non-acoustic auxiliary devices," in *Special Workshop in Maui: Lectures by masters in Speech Processing*, Maui, Hawaii, Jan. 2004.

[6] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for encoding," *submitted to IEEE Transactions on Speech and Audio Processing*, 2004.