

Speech Enhancement Using Non-Acoustic Sensors

Rongqiang Hu, Sunil D. Kamath, David V. Anderson

Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA, USA

(rqhu, skamath, dva)@ece.gatech.edu

Abstract

This paper describes a speech enhancement system that significantly improves speech intelligibility of noisy speech in the context of a speech coder in low SNR conditions. The system uses two state-of-the-art non-acoustic sensors, a general electromagnetic motion sensor (GEMS) that detects the internal motions of glottis, and a physiological microphone (P-mic) that measures vibrations of the skin associated with speech. Both sensors are relatively immune to ambient acoustic noise, but provide incomplete information of speech. In the proposed system, the strengths of two algorithms, a perceptually motivated constant-Q (CQ) algorithm and an enhanced glottal correlation (GCORR) algorithm, are combined. The CQ algorithm employs a perceptually inspired signal detection technique to estimate the presence of speech cues in low SNR conditions. To reduce annoying artifacts, a state-dependent mechanism discriminating the distinct acoustic properties of each phoneme, and a psychoacoustic masking model are used to control enhancement gains. The enhanced glottal correlation algorithm extracts the desired speech signal from the noisy mixture, using a modified speech-GEMS correlation estimation of the speech signal with the glottal waveform supplied by GEMS. Both subjective and objective experiments were performed in a variety of noise conditions to indicate the improvement relative to the EMSR algorithm.

1. Introduction

In the presence of high ambient noise, the intelligibility of a voice communication system is significantly degraded. Therefore, speech enhancement has been an active research area for decades and continues to be an important problem.

One of the main problems for noise suppression algorithms is that of detecting or estimating the signal of interest in the presence of noise. Although many speech enhancement algorithms have been developed in minimum mean square error (MMSE) sense [1], these algorithms are based on statistical assumptions of speech and additive noise that don't always hold. Therefore, they may suffer from perceptually annoying residual noise and spectral distortion.

In this paper we present an algorithm inspired by human audio signal detection, referred to as "CQ" in reference to the constant-Q filterbank used at the front-end. The idea is to use an analysis filter bank having passbands similar to those in the cochlea. Coherence between bands and other perceptually motivated processes are employed to detect signals that may be below an SNR threshold. Feedback is employed between bands and between levels to provide *a priori* estimates of signal activity and thus set signal detection thresholds accordingly. Once the signal has been estimated, the perceptual cues gathered are used to drive a state-dependent noise suppressor which is adaptive for different phoneme classes. In the perception of speech,

it has been shown that noise does not impact equally on all phonemes because of the distinct acoustic properties of each phoneme [2]. But, in general speech enhancement algorithms, the noise subtracting procedure is usually blindly applied over the entire speech signal, or the suppression is adapted based only on the probable presence of speech in a frame. Phoneme-adaptive strategies adapt to be much more specific to the current signal, enhancing speech cues while suppressing simultaneous interferers.

Another opportunity in the area of speech processing is the exploitation of alternative non-air conductive sensors. A general electromagnetic motion sensor (GEMS) detects the internal motions of glottis, and a physiological microphone (P-mic) measures vibrations of the skin associated with speech. Both signals exhibit significant attenuations of ambient noise, but provide incomplete information about speech signal. In our system, a segmentation algorithm and an enhanced glottal correlation (GCORR) algorithm are proposed using these information.

The diagram of the proposed system, referred to as "CQ-GCORR", is shown in Figure 1.

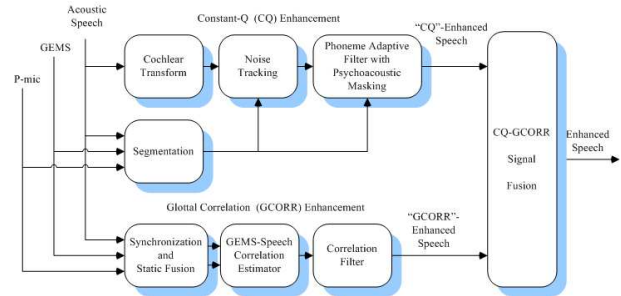


Figure 1: Block diagram of the CQ-GCORR speech enhancement system

2. Non-Acoustic Sensors

2.1. GEMS

The general electromagnetic motion sensor (GEMS) captures motion from the subglottal region of the trachea [3] [4]. The output is a signal that resembles an ideal excitation function in near real time. When the sensor is attached at proper facial locations, typically placed on the throat at the laryngeal notch, the measured signal is often very stable and as such can be very useful in further speech processing. Additionally, because the signal collected from a GEMS device depends on the tissue movement in the speech production anatomy, it is relatively immune from external acoustic noise. The GEMS output provides better side information in pitch estimation, pitch epoch localization, and approximated excitation.

⁰The GEMS speech coding work is sponsored by the Defense Advanced Research Projects Agency under Contract N00024-02-C-6339. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the US Government.

2.2. P-mic

The physiological microphone (P-mic) is composed of a gel-filled chamber and a piezoelectric sensor behind the chamber [6]. P-mic measures the signal in response to applied forces that are generated by the movement of corresponding tissue where the sensor is placed. Because of the poor coupling properties between ambient noise and the fluid-filled pad, the output exhibits large attenuation of background noise. The P-mic signal at the throat contains clearer low-pass vocal tract formants than those of normal microphone.

3. Segmentation

Based on acoustic and non-acoustic sensor signal activity, the speech segmentation algorithm makes hard segmentation decisions using an adaptive threshold of the signal energy. The algorithm is a coarse-grained segmentation algorithm that classifies a speech frame into broad phonemic categories, viz., vocalic (vowels, liquids and glides), unvoiced fricative, voiced fricatives, unvoiced plosives (including affricatives) and voiced plosives. Changeover regions are classified as ‘‘Transition’’ regions and noise-only regions between utterances as ‘‘Non-speech’’ regions. More detail is described in [7].

4. CQ–GCORR System

The proposed CQ–GCORR speech enhancement system is a hybrid system that combines the strengths of a perceptually motivated CQ algorithm and a GCORR algorithm, which is a statistical based filter using outputs from GEMS and P-mic. At the back-end, a signal fusion is derived for intelligibility improvement.

4.1. CQ Algorithm

It is generally believed that the perception of speech is performed throughout the auditory system. The CQ algorithm uses perceptually inspired techniques for frequency separation, noise estimation, signal detection and estimation, and anticipatory speech cue estimation [8, 9].

Consider a speech signal $s(k)$ corrupted by statistically independent background noise $n(k)$, the noisy mixture $x(k)$ can be represented as:

$$x(k) = s(k) + n(k) \quad (1)$$

The algorithm does not use block processing; instead, the signal is filtered into sub-bands ($x(m, k)$) with 1/3-octave filters modeling the human peripheral auditory system. The envelope of the signal in each sub-band is then calculated and it is the envelope that is used in most of the subsequent processing.

A sub-band filter is proposed as:

$$\hat{s}_c(k) = \sum_m \frac{\nu(m, k) \cdot \varepsilon^2(m, k)}{\varepsilon^2(m, k) + \zeta(m, k)} x(m, k) \quad (2)$$

Where $\varepsilon(m, k)$ is the sub-band SNR at time instant k in the m -th sub-band, $\nu(m, k)$ is the level of boosting to strengthen the attenuated speech cues, which is indicated from the temporal masking properties, and $\zeta(m, k)$ represents the enhancement gain.

The sub-band SNR $\varepsilon(m, k)$ is estimated using a recursive noise envelope averaging. The noise envelope is tracked during non-speech periods and characterized according to its mean and variance. Approximately 50 msec of non-causality is used to control the adaptation in the vicinity of speech to avoid contamination by speech onsets.

At the second level, a perceptually inspired speech detector is employed by evaluating the coherence of all sub-bands based on the perceptual weights [8]. The speech presence probability $\gamma(m, k)$ is then computed.

Table 1: Procedure of CQ algorithm

| |
|---------------------------------------------------------------------------------------------------|
| Input time-domain noisy signal $x(k)$ and segmentation $se(k)$ |
| For all time instant k |
| • Apply lowpass filters to get the sub-band signal $x(m, k)$ |
| • For all subband m |
| ◦ Compute the signal envelope $X(m, k)$ using a set of frequency-dependent smoothing windows |
| ◦ Compute noise envelope $\hat{N}(m, k)$ recursively according to the estimated mean and variance |
| ◦ Compute the subband SNR $\varepsilon(m, k)$ |
| ◦ Compute speech presence probability $\gamma(m, k)$ using the perceptual inspired parameters [8] |
| ◦ Compute phoneme saliency $\phi(m, k)$ according to segmentation input $se(k)$ [8] |
| ◦ Compute the masking compensation $\theta(m, k)$ and $\nu(m, k)$ [9] |
| ◦ Compute enhancement gain $\zeta(m, k)$ using equation (3) |
| ◦ Compute CQ output $\hat{s}_c(k)$ using equation (2) |

To reduce the annoying residuals, phoneme adaptation and psychoacoustic masking are two useful factors to be incorporated. It is generally believed that the effects of background noise on conspicuousness of ‘‘significant’’ signals are different for each class. Therefore, it is important to drive parameters selectively adjusted to speech signals based on phoneme content. In the proposed algorithm, speech sounds are classified from the time-varying spectral characteristics revealed through the spectrogram, including the distribution and energy. The examples of saliency functions ($\phi(m, k)$) are shown in [8]. Additionally, a psychoacoustic masking model is used as the post-processing [9]. From the model, the spectral audibility level $\theta(m, k)$ is estimated from spectral masking model to reduce the annoying artifacts, the temporal audibility $\nu(m, k)$ is derived from the temporal masking model to boost the attenuated speech cues.

After the estimation of those parameters, the enhancement gain $\zeta(m, k)$ is expressed as:

$$\zeta(m, k) = F(\gamma(m, k), \phi(m, k), \theta(m, k)) \quad (3)$$

Where function $F(\cdot)$ is introduced in [9].

4.2. GCORR Algorithm

Studies on speech production mechanisms show that the acoustic speech signal is generated by the frequency shaping of the glottal excitation signal by the vocal tract. Those internal motions of glottis, which are measured by the GEMS device, are independent of ambient noise. The idea of the glottal correlation filter is to extract acoustic speech signal, that statistically correlates with glottal excitation, from noisy mixture [10].

As observed, the P-mic signal contains clearer low-pass vocal tract formants than those of normal microphone. Therefore, the low-frequency components (<300Hz) of acoustic signals are replaced by corresponding P-mic signals to produce more intelligible speech ($\hat{X}(f)$).

The enhanced output of GCORR is constructed by evaluating the statistical properties:

$$S_G(f) = \frac{P_{G\hat{X}}(f)}{P_{GG}(f)} \cdot G(f) \quad (4)$$

Where, ($P_{G\hat{X}}(f)$) represents the cross-correlation of the GEMS signals and the modified acoustic signals [10].

The detail implementation is described in [10]. This approach can also be applied to the input of other sensors, where the acquired signal is highly correlated to clean speech and immune to background noise, such as bone-conduction mic.

Table 2: Pseudocode of GCORR algorithm

| |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Input time-domain signals from acoustic microphone ($x(k)$), GEMS ($g(k)$), and P-mic ($x_p(k)$) <ul style="list-style-type: none"> • Synchronize signals by maximizing the cross-correlation • Estimate the pitch and pitch epoches • Construct adaptive frames using the length of two epoches • For all time frame k <ul style="list-style-type: none"> ◦ Compute modified speech signal $\hat{X}(f)$ using low-pass fusion with P-mic signal in equation (??) ◦ For all frequency bin f <ul style="list-style-type: none"> · Compute GEMS-Speech correlation $P_{G\hat{x}}(f)$ [10] · Compute the GCORR output $S_G(f)$ using equation (4) |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

4.3. CQ-GCORR Fusion

In the signal fusion, the observed signals include the output of CQ ($\hat{s}_c(k)$), the output of GCORR ($\hat{s}_g(k)$), and the noisy signal ($x(k)$). We propose the final output of CQ-GCORR system is constructed as the linear combination of the observed signals in all sub-bands.

$$\hat{s}(k) = \sum_{m=1}^M \left[\hat{h}_c(m) \hat{s}_c(m, k) + (1 - \hat{h}_c(m)) \hat{s}_g(m, k) \right] \quad (5)$$

where $\hat{s}_c(m, k)$ and $\hat{s}_g(m, k)$ represent the sub-band signal of CQ and GCORR output respectively. The minimization of mean square error (MSE) can be described as:

$$\min_k \sum_k (s(k) - \hat{s}(k))^2 = \arg \min_{h_c(m)} \sum_k \sum_m (s(m, k) - \hat{s}(m, k))^2 \quad (6)$$

There is no explicit solution for this equation. In this implementation, we look for a solution with lower computation complexity in order to process the signal in real-time applications. We note that the minimization of equation (6) equals to the maximization of the segmental SNR improvements in each critical sub-band, as indicated in equation (7).

$$SNR_{imp} = \frac{1}{L} \sum_{m=0}^{L-1} 10 \cdot \log_{10} \frac{\frac{1}{N} \sum_{k=0}^{N-1} n^2(m, k)}{\frac{1}{N} \sum_{k=0}^{N-1} [s(m, k) - \hat{s}(m, k)]^2} \quad (7)$$

where L represents the number of frames in the voiced signal and N is the number of samples in m th frame.

We set the observation as the ratio of the smoothed signal envelope of CQ output ($\hat{S}_c(m, k)$) and GCORR output ($\hat{S}_g(m, k)$).

$$\lambda(m, k) = 10 \cdot \log_{10} \frac{\hat{S}_c(m, k)}{\hat{S}_g(m, k)} \quad (8)$$

The maximization of equation (7) can be approximated as:

$$\arg \max_{h_c(m)} (SNR_{imp} | \lambda(m, k)) \quad (9)$$

The parameters $h_c(m)$ equals to the conditional probability of CQ outperforming GCORR given the observation $\lambda(m, k)$.

$$h_c(m) = p(DSNR_{imp} < 0 | \lambda(m, k)) \quad (10)$$

where $DSNR_{imp}$ is the SNR gains of GCORR over CQ.

$$DSNR_{imp} = SNR_{imp}(GCORR) - SNR_{imp}(CQ) \quad (11)$$

The conditional probabilities were determined in off-line training by the clean speech with additive noise. In the implementation, interpolation and smoothing were used to avoid the transitional variations.

5. Evaluation

Speech data was selected from a speech corpus created by ARCON for the DARPA Advanced Speech Encoding project. This corpus is an extensive multi-sensor speech corpus collected from ten male and ten female talkers in nine different acoustic noise environments. In the experiments, we used speech data from six speakers (3 females and 3 males) in three types of noise: M2 Bradley Fighting Vehicle, military gun shoot operations in urban terrain, and Blackhawk helicopter.

5.1. Illustration of CQ-GCORR behavior

Figure 2 shows the behavior of the CQ-GCORR system in M2 fighting vehicle noise environment. The GEMS signal exhibits large attenuation of background noise. But it can not capture the speech in unvoiced section (phoneme "t"). This factor gives the same problem in the output of GCORR algorithm. The combination of CQ and GCORR indicates better performance.

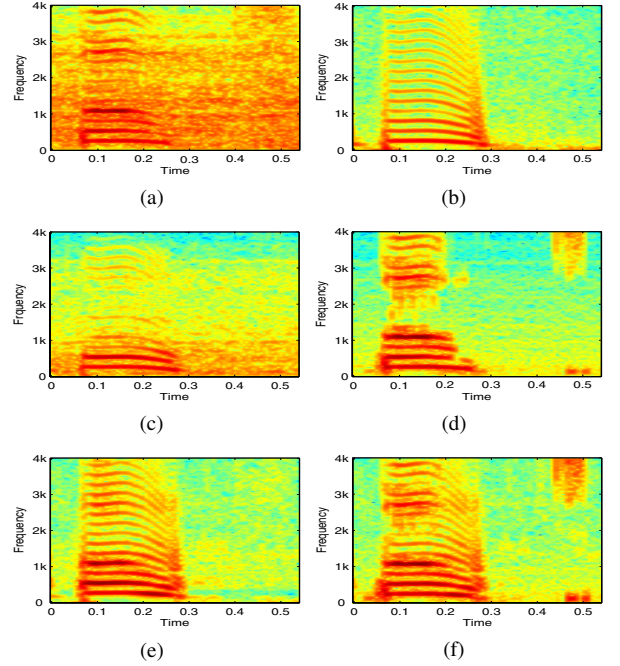


Figure 2: Results of a speech clip (word "boot") in M2 fighting vehicle noise environment ($SNR = 5dB$), (a) noisy speech, (b) the GEMS signal, (c) the P-mic signal, (d) the CQ output, (e) the GCORR output, (f) the CQ-GCORR output.

5.2. Subjective Intelligibility Assessment

A diagnostic rhyme test (DRT) was performed to give a quantitative indication about speech intelligibility. The results are provided in Table 3 and 4. The largest gain is achieved in voicing feature set, where both CQ and GCORR are shown to be effective. Although some small degradation in three feature sets, in average, the CQ-GCORR system gives significant improvement in speech intelligibility for low-bit-rate speech coding in acoustic harsh environments.

Table 3: DRT scores of the enhanced speech signals in harsh environments (SNR<5dB) in low-bit-rate (MELP@2400bps) coding.

| Noise Type | Speaker | EMSR | CQ-GCORR | Gain |
|--------------------|---------|-------|----------|-------------|
| Blackhawk | Male | 77.82 | 78.52 | 0.70 |
| Helicopter | Female | 80.62 | 85.51 | 4.89 |
| M2 Bradley vehicle | Male | 67.36 | 70.53 | 3.17 |
| | Female | 76.43 | 82.42 | 5.99 |
| Gun Shoot | Male | 80.99 | 85.20 | 4.21 |
| | Female | 87.67 | 90.02 | 2.35 |

Table 4: DRT scores of different feature sets

| Feature Set | EMSR | CQ-GCORR | Gain |
|--------------|-------|----------|--------------|
| Voicing | 74.45 | 87.52 | 13.07 |
| Nasality | 80.62 | 86.41 | 5.79 |
| Sustentation | 64.36 | 66.53 | 2.17 |
| Sibilant | 79.83 | 78.92 | -0.91 |
| Graviness | 75.52 | 75.20 | -0.32 |
| Compactness | 88.43 | 85.12 | -3.31 |
| EXP | 77.07 | 80.02 | 2.95 |

5.3. Objective Quality Measure

The log spectral distance (LSD) is plotted as cumulative distribution functions of the magnitude of speech distortion. In this experiment, noise has been added to the clean speech signal with a varying SNR. From overall results shown in Fig.3, the enhanced output of CQ-GCORR system suffers less speech distortion. The amount of noise suppression is also demonstrated by

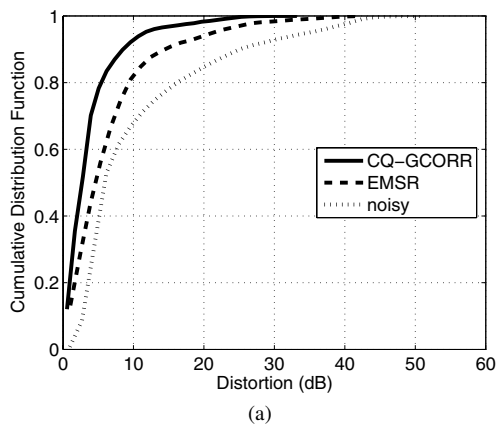


Figure 3: Cumulative log spectral distortion of enhanced outputs in M2 fighting vehicle noise environment. The measured signals are the following: the enhanced outputs of (boldline) CQ-GCORR, (bold - dotline) Ephraim-Malah Suppression Rule (EMSR), and (dotline) noisy speech.

the segmental SNR improvement in the voiced segments. The results are shown in Table 5. The proposed system increases the noise suppression levels in all noise conditions.

Table 5: Segmental SNR improvement for voiced segments in various noise conditions

| Noise Type | Input Seg. SNR (dB) | Seg. SNR Improvement (dB) | | | |
|----------------------|---------------------|---------------------------|------|-------|--------------|
| | | EMSR | CQ | GCORR | C-G |
| Blackhawk Helicopter | -5 | 8.04 | 9.09 | 9.35 | 11.67 |
| | 0 | 7.61 | 8.42 | 8.64 | 10.72 |
| | 5 | 7.08 | 7.91 | 7.65 | 9.47 |
| M2 Bradley Vehicle | -5 | 5.94 | 8.10 | 9.01 | 11.17 |
| | 0 | 5.57 | 7.69 | 8.45 | 10.27 |
| | 5 | 5.24 | 7.11 | 7.42 | 9.13 |
| Gun Shoot | -5 | 3.17 | 5.15 | 6.89 | 7.27 |
| | 0 | 2.81 | 4.75 | 6.60 | 6.89 |
| | 5 | 2.24 | 4.09 | 6.16 | 6.27 |

6. Conclusion

General speech enhancement algorithms utilize the inputs from acoustic sensors only. Although many optimization techniques have been developed, the enhanced outputs may still suffer from considerable speech distortion, especially in low SNR conditions. As a result, the speech intelligibility is degraded. In the proposed system, we exploit the state-of-the-arts non-acoustic sensors and introduce a hybrid speech enhancement system using perceptually inspired techniques. Both subjective and objective experiments were conducted and compared to the EMSR algorithm in various noise types and levels. The proposed system is effective in suppressing background noise. Significant improvement of speech intelligibility for low-bit-rate speech coding in harsh environments is achieved.

7. References

- [1] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," in *Proceedings of ICASSP*, Boston, USA, 1983, pp. 1118-1123.
- [2] M. Deidher and A. Spanias, "HMM-based speech enhancement using harmonic modelling," in *Proceedings of ICASSP*, vol. 2, Munich, Germany, Apr. 1997, pp. 1175-1178.
- [3] L. C. Ng, G. C. Burnett, J. Holzrichter, and T. J. Gable, "Denosing of human speech using combined acoustic and em sensor signal processing," in *Proceedings of ICASSP*, Istanbul, Turkey, June 2000.
- [4] G. C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function," Nov. 1999, presented at the 138th Meeting of the Acoustical Society of America, Columbus, Ohio.
- [5] M. Scanlon, "Acoustic sensor for health status monitoring," in *Proceeding of IRIS Acoustic and Seismic Sensing*, vol. II, 1998, pp. 205-222.
- [6] C. Demiroglu, S. Kamath, and D. V. Anderson, "Segmentation based speech enhancement," in *Proceedings of ICASSP*, Mar. 2005.
- [7] R. Hu and D. V. Anderson, "Audio noise suppression based on neuromorphic saliency and phoneme adaptive filtering," in *IEEE DSP Workshop*, Taos, NM, Aug. 2004.
- [8] —, "Improved perceptually inspired speech enhancement using an auditory model," in *Asilomar*, Nov. 2004.
- [9] —, "Single acoustic channel speech enhancement based on glottal correlation using non-acoustic sensors," in *International Conference on Spoken Language Processing*, Jeju, Korea, Oct. 2004.