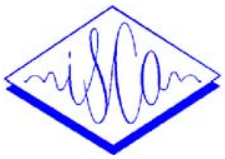
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Soft Decision MMSE Amplitude Estimator as a Noise Preprocessor to Speech Coders Using a Glottal Sensor

C. Demiroglu and D. V. Anderson

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

10.21437/Interspeech.2004-314

Abstract

A soft-decision Ephraim-Malah suppression rule based speech enhancement algorithm is proposed for intelligibility enhancement in parametric speech coders. A glottal sensor is used to improve the intelligibility of a baseline system that uses only the acoustic microphone. The objective measure test shows that the proposed system decreases the spectral distortion by 2-3 dB for most phonetic classes. Moreover, significant improvements in DRT scores of nasality and sibilation features are obtained compared to the baseline system when the noise suppression systems are concatenated with a MELP based speech coder.

1. Introduction

Parametric speech coders, such as MELP, produce high quality and intelligible speech in quiet environments. However, with increasing background noise, intelligibility drops substantially. MELPe uses an Ephraim-Malah suppression rule (EMSR) based noise suppressor to reduce the effects of noise. Although, the MELPe noise suppressor helps improve the speech quality in noisy environments, except for a few noise conditions, the intelligibility does not increase and sometimes even decreases compared to the MELP coder [1]. Therefore, designing a noise suppression algorithm that can increase the intelligibility of parametric speech coders is an important research problem.

In this work, a previously proposed soft-decision EMSR algorithm that is shown to produce high quality speech [2] is used as the baseline system. The overall DRT score of the baseline system is higher than the MELPe score. The baseline system uses a signal-to-noise ratio (SNR) measure to assign weights to each fft bin of the speech spectrum. The idea is to assign higher weights to higher SNR bins. The weights are optimally incorporated into the EMSR algorithm such that low weight frequencies are suppressed more aggressively compared to high weight frequencies. The problem with this approach is that especially in nonstationary noise environments SNR estimation is not accurate. Moreover, the weights are heuristically set, and the constants in the heuristic formula might work better in one condition compared to the other. Thus, tuning is necessary for each SNR and noise condition to have optimum quality, which is not practical. These problems are addressed by incorporating a recently developed glottal radar sensor, namely the glottal electromagnetic sensor (GEMS) device, into the baseline system.

The proposed system uses the GEMS device to detect voiced speech frames and high energy harmonics in those frames. Frequency weights calculated from the audio signal are combined with the weights that are calculated from the GEMS signal. The proposed system is tested with an objective distance measure and is shown to decrease the spectral distortion by 2 dB to 3 dB in most phonetic classes. Moreover, the in-house DRT scores using a MELP based parametric speech coder developed at Georgia Tech [3] indicate significant improvement for nasality and sibilation features over the baseline system.

2. Description of the GEMS Device

The glottal electromagnetic sensor, now *general* electromagnetic sensor (GEMS), is a micro-power 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. A fully developed, commercial version based on these principles is currently available from 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 speech is a signal that resembles an ideal excitation waveform. The exact physical structures whose motion are detected are currently not completely understood. The signal, however, is often very stable and as such would be very useful in further processing. Additionally, the signal obtained is robust to external acoustic influences, such as noise. The GEMS device responds to vocal fold vibration at the larynx. Other devices such as the EGG do this by measur-

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.

ing changes in conductivity at the throat, but it is considered too cumbersome for everyday use.

The GEMS device can be deployed in handheld devices for speech applications in a less intrusive manner. It can be used for applications such as noise robust pitch detection [4] and speech enhancement [1]. In this work, the GEMS device is used for speech enhancement as described in Section 5.

3. Using the GEMS Device for Voicing Detection

A comparison of the spectrums of the GEMS signal and the corresponding speech signal are shown in Fig. 1. The GEMS signal clearly has a higher energy for the voiced speech segments compared to the unvoiced segments. Moreover, its spectrum resembles the excitation signal.

Fig. 1b shows the speech signal in Fig. 1a contaminated with the synthetically added M2 tank noise. The additive noise considerably masks the voicing bar. Therefore, reliably detecting the voicing bar from the noisy speech becomes a challenging task. The radar signal, however, is not disturbed by the acoustic noise, and it clearly has a voicing bar when the speech signal is voiced. Therefore, it can provide accurate voicing information for all acoustic environments, which proves to be beneficial for speech enhancement.

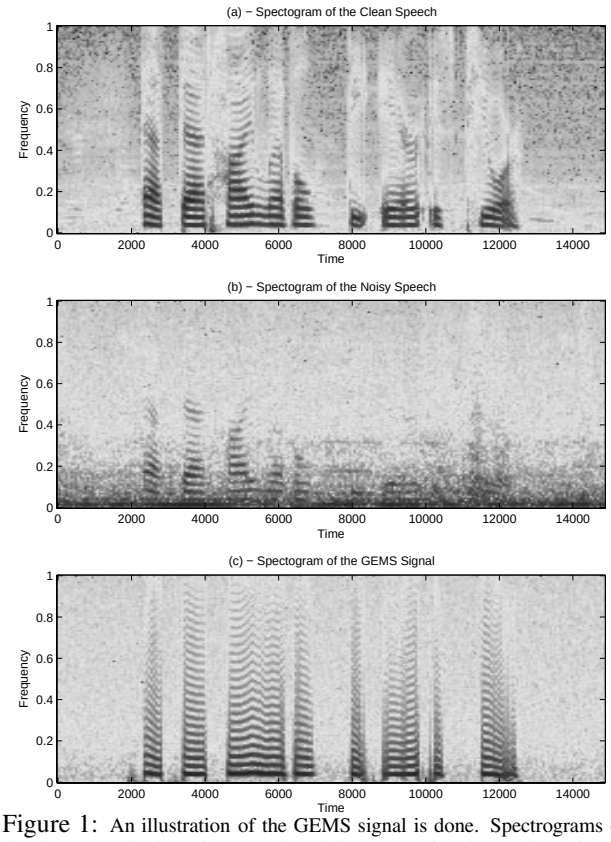


Figure 1: An illustration of the GEMS signal is done. Spectrograms of the clean speech, the noisy speech and the GEMS signals are shown in (a), (b) and (c) respectively. All signals are sampled at 8 kHz and the M2 tank noise is added to the clean speech at a global SNR of -10 dB.

A hard decision energy based algorithm that uses the GEMS signal to detect voicing in the speech signal is developed. Each 20 msec of speech frame i is windowed with a Hamming window. Energy ε_i in the 100-400 Hz range is calculated for frame i. This frequency range typically contains at least one harmonic in the case of voiced speech. Thus, a significant energy rise in this band compared to the silence level indicates existence of voicing in the speech frame. A minimum energy level ε_{min} is set to the median of energy levels of the 10 lowest energy frames, and a threshold level ε_{th} is set to 6 dB above ε_{min}. Finally, the voicing feature V_i for frame i is found by

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

4. The harmonic Tracking Problem

The speech production mechanism is typically modeled as a linear filter driven by a source signal. The source is modeled with a random noise generator for unvoiced speech and a periodic pulse generator for voiced speech. The linear filter excited by the periodic pulse generator has high energy harmonics in the spectrum at every kF_f Hz where k is an integer number and F_f is the frequency of the pulses. These high energy harmonics are clearly visible in the spectrograms shown in Fig. 1.

Formants are important for quality and intelligibility of voiced speech. Harmonic locations and their relative powers determine the locations and the bandwidths of formants. Unfortunately, there are several problems in detecting the harmonics in the voiced speech spectrum. One of the biggest problems is that the vocal cords generate quasistationary signals, and even if the pitch period is reliably estimated, which itself can be a difficult problem, the harmonics do not necessarily appear at every integer multiple of the pitch period.

The problem becomes more dramatic for noisy speech signals. Fig. 2 shows the case where a noisy speech spectrum is contaminated with the M2 tank noise. One of the effects of noise is masking the harmonics. In Fig. 2, the first harmonic is an example to the situation where the harmonic is masked by noise.

5. The Proposed System

Both the GEMS signal and the speech signal are used for tracking the harmonic frequencies. The speech spectrum is utilized in a way similar to the method described in [2]. The idea can be briefly described as follows. SNR ζ at frequency j in the power spectrum of the noisy speech $Y(k)$ can be approximated as

$$\zeta(j) = \left(\frac{Y(j) - N(j)}{N(j)} \right) \quad (1)$$

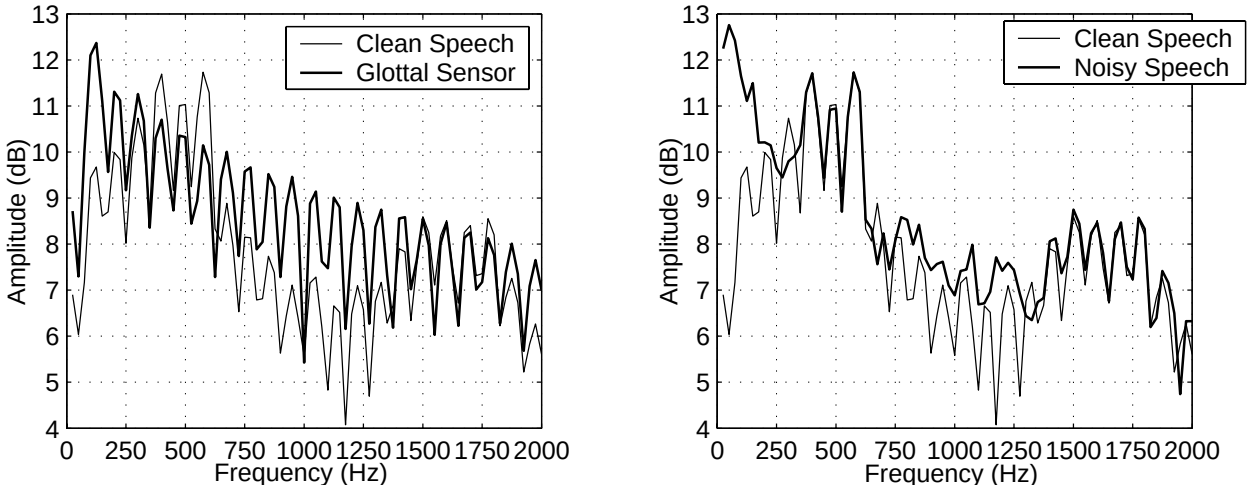


Figure 2: An example of the effect of noise on voiced speech spectrum is shown. A 20 msec clean speech segment is analyzed using a Hamming window. The same speech segment is compared with the case when M2 tank noise is added on it on right figure. The noise level is tuned to match the real-life M2 tank environment. The first 2000 Hz of the spectrums are shown since the GEMS signal is used for this frequency range in the proposed system. The figure on the left shows the high match between radar and speech harmonic locations.

where $N(j)$ is the estimated noise power spectrum. A soft decision weight is assigned to frequency k as follows

$$w_{s_j} = \begin{cases} 1 & \text{if } \zeta_j \geq \zeta_{max}, \\ \frac{log(\zeta_j/\zeta_{min})}{log(\zeta_{max}/\zeta_{min})} & \text{if } \zeta_{max} > \zeta_j > \zeta_{min}, \\ 0 & \text{if } \zeta \leq \zeta_{min} \end{cases}$$

where ζ_{max} and ζ_{min} are constants denoting the maximum and the minimum SNRs and are set to 10 dB and -10 dB respectively. The SNR-based weight estimation is shown to perform well in subjective quality tests. However, the goal in this work is to design a noise preprocessor for an LPC based speech coder. Therefore, objective quality and speech intelligibility after coding are the primary performance criteria.

The GEMS device is utilized for detecting high SNR frequencies of voiced speech spectra that are assigned low weights by the SNR-based algorithm due to inaccurate SNR estimations. High energy harmonics in the voiced speech spectrum have high chances of having high SNRs. Therefore, fft bins of the voiced speech spectrum that are neighbor to exact harmonic frequencies are assumed to have high weights. Fig. 2 clearly shows that in voiced speech the harmonics of the radar signal and the speech signal align in frequency. This fact is used to detect the speech harmonics in the voiced speech spectrum.

The GEMS signal is not guaranteed to span the whole 4000 Hz at all times. Therefore, the GEMS signal is used for harmonic detection in the 0-2000 Hz range while the SNR-based system is used for 2000-4000 Hz for voiced speech.

Detection of harmonics in voiced speech using the GEMS device is done by finding the fundamental frequency (F_f), and starting from the first harmonic searching the harmonics that are approximately one F_f away from each other. F_f is estimated by detecting the highest energy spectral component in the 100-400Hz range in the GEMS signal. The second harmonic is found by finding the highest energy component in the range of $2F_f - \delta F_f$ to $2F_f + \delta F_f$, where δF_f denotes the maximum deviation from the estimated location. It is calculated as

$$\delta F_f = round(F_f/5)Hz. \tag{2}$$

The rest of the harmonics locations are estimated using the same technique. The weights using the GEMS device are assigned as follows

$$w_{r_j} = \begin{cases} 1 & \text{if harmonic is detected,} \\ 0 & \text{if harmonic is not detected.} \end{cases}$$

Weights w_{s_j} are used for the 2000-4000 Hz range. The weights for 0-2000 Hz range are set using both w_{s_j} and w_{rj} as follows

$$w_{c_j} = \begin{cases} 1 & \text{if } w_{r_j} = 1, \\ w_{s_j} & \text{if } w_{r_j} = 0. \end{cases}$$

The weights are used within a soft-decision Ephraim-Malah Suppression Rule (EMSR) framework as proposed in [2]. If the suppression gain calculated by the EMSR algorithm without the weights is G_{em}, then the suppression gain using the proposed method is

$$G_{sd,em} = G_{em}^{w_c} G_{min}^{1-w_c} \tag{3}$$

where G_{min} is set to -20 dB.

Any frame that is not labeled as voiced by the proposed voicing detection algorithm is suppressed less aggressively compared to the voiced speech frames. The noise power spectrum estimate is multiplied by 0.5 to reduce the aggressiveness of the enhancement algorithm in those regions. The idea is to keep the perceptually important spectral cues in the unvoiced speech intact. Since the GEMS signal does not

provide any consistent information for the unvoiced speech frames, the algorithm relies on the acoustic microphone in those regions.

6. Experiments

Performance of the proposed system is compared with the SNR-based baseline system. Comparison of performance is done for an objective measure test and an in-house Diagnostic Rhyme Test (DRT). Subjective quality of the proposed system is similar to the baseline system. It could, however, correct the occassional harmonic losses that can create unnatural speech. Nevertheless, these situations are not frequent and mostly happen when the speaker's speech energy drops such as the cases that might happen at the end of an utterance.

The objective distance measure used is a log spectral similarity measure that is defined as

$$D(i,j) = \left[\frac{1}{N}\sum_{k=1}^{N}\left(20log_{10}(\frac{A_c(e^{(jw_k)})}{A_e(e^{(jw_k)})})\right)^2\right]^{1/2} \quad (4)$$

where $A_c(e^{(jw_k)})$ and $A_e(e^{(jw_k)})$ are the LPC spectrums of the clean speech and the enhanced speech respectively. The results for the M2 tank noise are given in Table 1. The M2 tank noise is particularly important since it has been found that the performance of the MELP coders drop significantly in that environment [1]. The M2 tank noise is synthetically added to the DRT test sets in the ARCON database that includes acoustic and GEMS data for various quality and intelligibility tasks [1]. The noise level is perceptually adjusted to match the actual speech samples recorded in an M2 tank. These recordings are also available in the ARCON database. The results are given in Table 1. The proposed system using the GEMS device provides around 2 dB improvement over the baseline system for most phonetic classes. The reason for the almost equal performance on vowels is the high energy harmonics in vowels that can be detected using the noisy speech signal.

The DRT scores are given for speech coded with the Melpi coder, which is a MELP based speech coder that is developed at Georgia Tech [1]. Six native English listeners performed the tests for three male and three female speakers from the ARCON database. The speech data is recorded in an M2 tank environment with noise canceling microphones. Table 2 shows the results of the DRT test for each distinctive feature. The statistically important differences are in nasality and sibilation where the proposed system outperforms the baseline system.

7. Conclusions

A radar based glottal sensor that can provide robust harmonic information at any noise condition is used for harmonic detection. The harmonic location information is used within a soft-decision MMSE estimation framework. The proposed system outperforms the baseline system in an object measure test. Moreover, the in-house DRT scores indicate that the intelligibility scores for nasality and sibilation features increase significantly compared to the baseline case.

Table 1: Performance results in terms of the log spectral distortion measure. The performance of the proposed system is compared with the baseline system for each phonetic class.

Phone Class	Baseline	Proposed System
Vowels	4.7	4.6
Nasals	9.9	7.2
Semivowels	9.6	8.1
Unvoiced Fricatives	16.2	14.0
Unvoiced Plosives	20.4	18.9
Voiced Fricatives	13.5	11.1
Voiced Plosives	16.9	14.6

Table 2: Performance results in terms of DRT scores for each distinctive feature of consonants.

Feature	MELPe	Baseline	Proposed Sys.
Voicing	63.0	76.7	78.8
Nasality	70.0	60.1	80.4
Sustenation	55.0	46.9	47.9
Sibilation	72.0	77.1	82.1
Graveness	60.5	70.1	67.5
Compactness	79.2	83.0	80.0
Total	66.8	69.0	73.0

8. References

[1] 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.

[2] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Transactions on Speech and Audio Processing*, vol. 9, 2002.

[3] A. E. Ertan, "Pitch synchronous analysis of speech signal for improving the quality of low bit rate speech coders," Ph.D. dissertation, Georgia Institute of Technology, 2003.

[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.