Noise Robust Digit Recognition Using a Glottal Radar Sensor for Voicing Detection

C. Demiroglu and D. V. Anderson

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

Abstract

A voicing feature is used in concatenation to MFCC features to increase the performance of digit recognition at both low and high SNRs. The problem of noise robust extraction of the voicing feature is solved by using the glottal electromagnetic sensor (GEMS). The GEMS device provides reliable voicing information at all SNRs and noise environments. It is shown that although the voicing feature increases the performance for the clean speech case, the relative improvement for the noisy case is significantly higher for a digit recognition task. Our results indicate that the GEMS device can solve the fundamental problem of extracting reliable voicing information in noisy environments.

1. Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems commonly use Mel Frequency Cepstrum Coefficients (MFCCs) as spectral features. MFCCs are shown to achieve high performance in quite environments, but they are not robust to acoustical noise. Therefore, there has been considerable amount of research done for finding noise robust acoustic features [1]. One of the possibilities in this direction is to find complementary features that are robust to noise and increase the performance when fused with the MFCC features. Voicing is an example of such features that is recently shown to increase the performance when fused with the MFCC features [2, 3, 4].

The previous work on voicing mostly focus on clean speech. In [2], three different voicing detection algorithms are compared, and a voicing feature is used for small and large vocabulary tasks. In [3], standard MFCC features are fused with a voicing feature, and its first and second order derivatives. An autocorrelation based voicing measure is used. In [4], fundamental frequency and voicing are used with the MFCC features using Linear Discriminant Analysis. The fundamental problem in the proposed systems is the difficulty in extracting the features reliably at low SNRs. Moreover, the problem becomes even more severe under non-stationary noise conditions. Furthermore, misclassifications in voicing detection can decrease ASR performance rather than increase it at low SNRs.

In this work, a voicing feature is used at low SNRs to improve noise robustness of the baseline ASR system. The glottal electromagnetic sensor (GEMS) device is used to reliably extract the voicing information. The GEMS device uses micro-power radar to sense glottal movement in the throat. Therefore, it is immune to acoustic noise. Moreover, these sensors are designed to be easily worn and are not cumbersome for the user. Thus, as opposed to sensors such as electroglottograph (EGG), it is possible to use the GEMS device in daily life.

Since the GEMS device is used for voicing detection, the voicing feature is robust to any type of noise at any SNR, which is shown to be valuable for a noise robust digit recognition task. In fact, although the voicing feature helps in the clean case, the relative improvement it provides is shown to be significantly higher for the noisy case. The results indicate that the GEMS device can help solve the fundamental problem of extracting reliable and useful speech information in noisy environments.

This paper is organized as follows. In section 2, a brief description of the GEMS device is done, and the voicing detection algorithm using the GEMS device is explained. The GEMS data is not available for any standard speech recognition database. Therefore, the GEMS signal is simulated using the clean speech signal to run the speech recognition experiments. In section 3, the simulation methodology is described. In section 4, the experimental results using the Aurora2 small vocabulary task is presented, and finally a conclusion is done in section 5.

2. Proposed ASR System with the Voicing Feature

Fig. 1 shows an overview of the proposed system. The MFCC features x are extracted from the speech signal $s(n)$ while the voicing feature v is extracted from the radar signal $r(n)$. The two feature vectors are fused, and the resulting vector y is fed into the ASR engine.

A description of the GEMS device and its use for detecting voicing in speech signals are given below.

This 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.

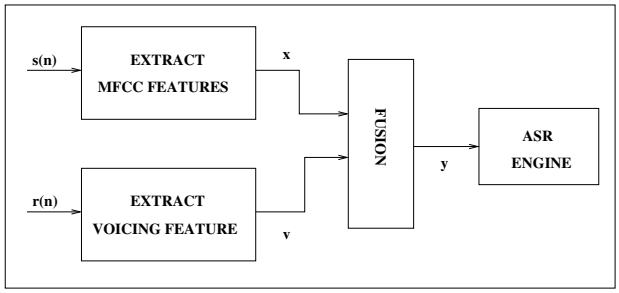

Figure 1: Overview of the proposed system. $s(n)$ is the input speech signal and $r(n)$ is the radar signal.

2.1. 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 [5]. 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 signal responds to vocal fold vibration at the larynx. Other devices such as the EGG do this by measuring changes in conductivity at the throat, but it is considered too cumbersome for everyday use.

The GEMS device can be employed in handheld devices for speech applications in a less intrusive manner. It can be used for applications such as noise robust pitch detection [5] and speech enhancement [6]. In this work, the GEMS device is used for a noise robust ASR application. The GEMS signal can be used as a reliable voicing detector at all SNRs and for all noise environments. Therefore, it is used to provide accurate voicing information that is fused with MFCC features.

2.2. Using the GEMS Device for Voicing Detection

A comparison of the spectrums of the GEMS device output and the corresponding speech signal is shown in Fig. 2. The GEMS device clearly has high energy during the voiced speech segments. Moreover, its spectrum resembles the excitation signal although this behavior is not well understood. Fig. 2b shows the spectrum of the same speech signal with the M2 tank noise artificially added on it. The additive noise masks the voicing bar significantly. Therefore, reliably detecting the voicing bar from the noisy speech becomes a difficult task. However, the radar signal is not influenced from the acoustic noise, and it clearly shows the voicing bar at the correct time instants.

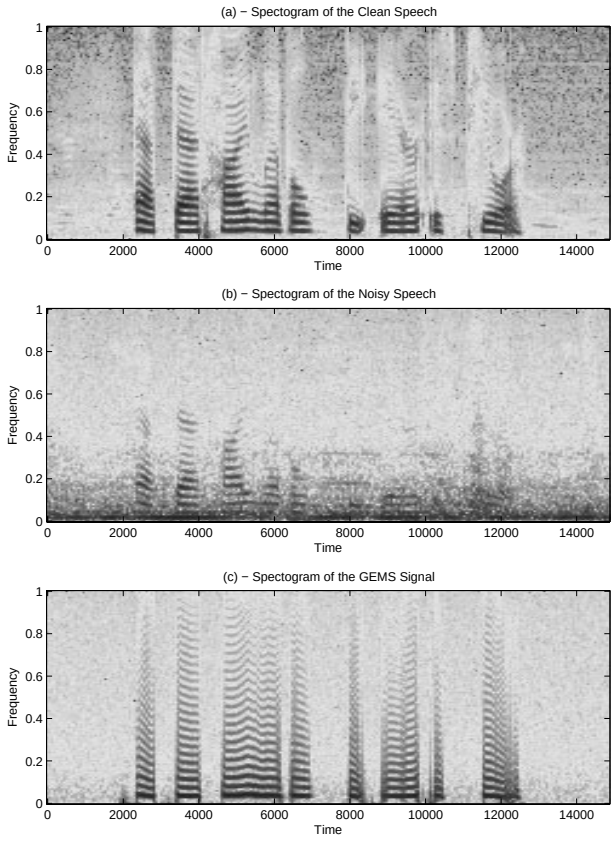

Figure 2: The spectrograms of clean speech, noisy speech and GEMS signal are shown in (a), (b) and (c) respectively. All signals are sampled at 8 kHz and the M2 tank noise is added to the clean speech at a global SNR of -10 dB.

A hard decision energy based algorithm is used to detect the voiced segments using the radar data. Each 20 msec of speech segment i is windowed with a Hanning window. Energy ζ_i in the 100-400 Hz range is calculated for each frame i. This range generally contains at least one harmonic in the case of voicing. Thus, significant energy rise in this band compared to the silence level indicates existence of voicing. A minimum energy level $\zeta_{min,r}$ is set to the median of energy levels of the 10 lowest energy frames in the radar signal, and a threshold energy level $\zeta_{th,r}$ is set to 6 dB above the $\zeta_{min,r}$. The voicing feature V_i for frame i is set as

$$V_i = \begin{cases} 1 & \text{if } \zeta > \zeta_{th,r}, \\ 0 & \text{if } \zeta \leq \zeta_{th,r}. \end{cases}$$

3. Simulation of the GEMS Device

The GEMS system is being investigated for ultra low bitrate speech coding ($\sim$ 300 bps) in noise [6]. As a part of this project, an extensive database was created by ARCON Corporation that has simultaneous speech, GEMS, EGG, and other sensor data for various noise conditions [7]. This database is designed for the assessment of speech coders using the DRT (diagnostic rhyme test) and the CVC (consonant vowel consonant) words lists as well as the Harvard sentences.

One of the problems with the GEMS device is that there is not enough training data for speech recognition experiments. Therefore, the behavior of the GEMS device is simulated using the clean speech signal. Voicing is detected from the clean speech signal using the algorithm described in Section 3. The threshold level $\zeta_{th,s}$ for the speech signal is set to $\zeta_{min,s} + 8(dB)$.

Fig. 3 may be useful for understanding the rationale behind using an energy based voicing detection algorithm. One male speaker's speech from the Harvard sentences database is energy normalized as shown in Fig. 3. Clearly, the energy rises significantly both in the radar and the speech trajectories when voicing occurs. Therefore, energy rise in the 100-400Hz range is used as the indicator of voicing both for speech and radar signals.

The voiced speech sections are segmented with rectangular boxes in Fig. 3. Although there is a very high match between the voicing decisions from the radar and the speech signals, there are also occasional errors. These errors are analyzed as follows. The performance of the simulation is measured using ten minutes of speech from the clean Harvard sentences. Six male and six female speakers are used. The misclassification rates are shown in Fig. 4. The energy is normalized and divided into 1 dB intervals for more accurate simulation of the system. For each energy interval, the probability of misclassification are calculated.

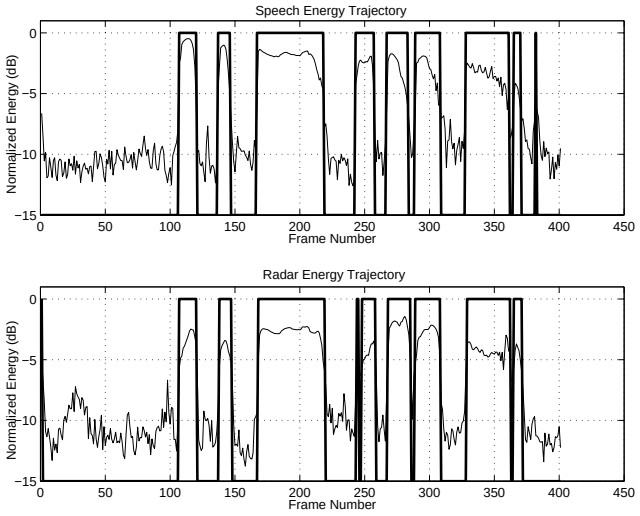

Figure 3: Energy trajectories of the clean speech signal and the radar signal are shown. The voiced segments are shown with rectangular boxes.

4. Experiments

4.1. Experimental Setup

An open source ASR software developed at Mississippi State University is used as the recognition engine [8] . Words are modeled with 20 state left–to–right Hidden Markov Models (HMM) using 16 Gaussian mixtures per state.

The Aurora2 database is used for the experiments. Clean training data is used for training, and the four noisy testing conditions in testa with subway, babble, car, and exhibition noises are used for testing.

GEMS data is not yet available for the Aurora task. The procedure proposed in Section 4 is used for simulating the GEMS signal. The simulation errors shown in Fig. 4 are included both in training and testing. However, it is found that these errors do not significantly affect the performance.

The baseline system contains 12 MFCC coefficients, energy and their first and second derivatives. Thus, there are 39 features in the baseline system. The proposed system fuses these 39 features with the voicing feature and uses a 40 dimensional vector. Energy normalization is performed in both systems.

4.2. Results

The results with the baseline and the proposed systems are shown in Tables 1 and 2 respectively. Adding the voicing feature decreases the word error rate (WER) at all SNRs. It is interesting to note that the performance increase has a Gaussian shape with respect to the SNR. Thus, the gain is highest at the mid SNRs while it is relatively lower at the high and low SNRs. This behavior can be explained using two facts. At high SNRs, the system has a low chance of confusing a voiced region with an unvoiced region particularly due to the energy feature. Therefore, the voicing feature yields relatively more performance gain at medium SNRs where the system has a higher chance of confusing voiced and unvoiced speech. As the SNR keeps dropping, however, the masking effect starts severely affecting the performance. The voicing feature can offer relatively lower gain at very low SNRs since the masking effect becomes the dominant factor, and misclassification of phonemes within the same class (voiced or unvoiced) becomes significantly more probable.

The babble noise is a case that deserves particular attention in this context. Since the background noise is human speech-like, it is one of the most challenging noise types. However, the largest improvement is obtained for this noise type for SNRs greater than 5 dB. The reason for this important result is that most of the speech energy is in the lower frequencies (0-2000 Hz) and a speech like noise can easily create a speech-like spectrum at those frequencies. For ex-

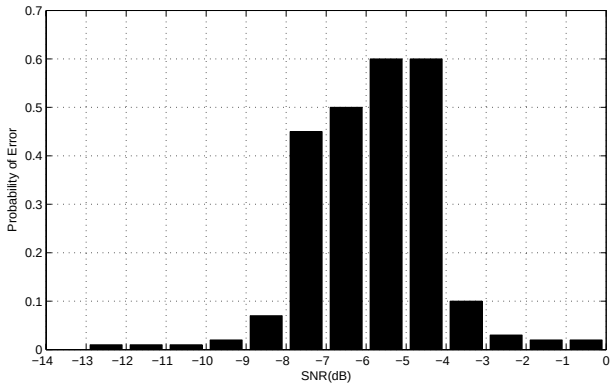

Figure 4: The misclassification rates when the GEMS signal is simulated using the clean speech signal. The normalized energy (dB) is divided into 1 dB intervals, and the error rates are shown for each energy interval.

ample, in a babble noise environment an unvoiced speech frame can be easily confused with a voiced speech frame due to high energy concentration at the lower frequencies and similar spectral content. An accurate voicing detector can eliminate this problem and substantially increase the performance as shown in Tables 1-2.

Table 1: Performance results in terms of WER for all four noise types using the baseline system.

SNR	Subway	Babble	Car	Exhibition	Avg.
-5 dB	88.60	89.70	88.20	88.20	88.68
0 dB	75.10	75.50	76.50	73.90	75.25
5 dB	50.40	53.10	55.70	47.80	51.75
10 dB	26.00	29.40	26.80	21.40	25.90
15 dB	9.50	9.50	7.20	8.10	8.58
20 dB	3.40	2.90	3.30	3.60	3.30
Clean	1.20	1.30	1.70	0.9	1.3

5. Conclusions

It is known that fusing a voicing feature with the MFCC features increases the performance of the ASR systems. The problem is detecting these features accurately in noisy speech, which can be difficult even for clean speech. In this work, the voicing feature is detected using the GEMS device. The GEMS device is immune to acoustic noise, and therefore it is effective for all noise environments. Moreover, they are comfortable for the user, and they can easily be used in commercial applications. A significant performance gain for a digit recognition task at noisy environments is obtained by fusing the voicing feature with the MFCC features.

Table 2: Performance results in terms of WER for all four noise types using the proposed system.

SNR	Subway	Babble	Car	Exhibition	Avg.
-5 dB	83.6	83.3	86.1	81.7	83.7
0 dB	62.9	68.6	70.6	62.6	66.2
5 dB	34.5	37.8	38.1	33.8	36.05
10 dB	16.9	12.5	12.9	15.6	14.5
15 dB	8.3	4.1	5.0	8.8	6.6
20 dB	3.6	2	2.9	4	3.1
Clean	1.1	1.1	1.6	0.8	1.1

Table 3: Percent improvement of WER compared to baseline case averaged over all four noise types.

-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	Clean
5.6	12.1	30.3	42.9	22.8	6.54	9.8

6. References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, 1995.

[2] R. S. A. Zolnay and H. Ney, "Extraction methods of voicing feature for robust speech recognition," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, Sept. 2003.

[3] D. L. Thomson and R. Chengalvarayan, "Extraction methods of voicing feature for robust speech recognition," *Speech Communication*, vol. 37, 2002.

[4] A. Ljolje, "Speech recognition using fundamental frequency and voicing in acoustic modeling," in *ICSLP*, Denver, CO, Sept. 2002.

[5] 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.

[6] 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.

[7] T. F. Quatieri, D. Messing, K. Brady, W. M. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech enhancement," in *Proceedins of the Workshop on Multimodal User Authentication*, Santa Barbara, CA, 11-12 December 2003.

[8] *http://www.isip.msstate.edu*.