Segmentation-Based Speech Enhancement for Intelligibility Improvement in MELP Coders Using Auxiliary Sensors

Cenk Demiroglu, Sunil D. Kamath, and David V. Anderson

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

Abstract

Intelligibility of spoken words in noisy environments is an important problem of speech coders particularly for military applications. The intelligibility problem of MELP speech encoder at noisy environments is addressed by using a novel speech enhancement algorithm at the front end. The speech signal is segmented into broad phonetic classes using auxiliary sensors in addition to the acoustic microphone. Each phoneme class is enhanced by suppressing maximum noise while minimally distorting perceptually important cues using the acoustic-phonetic knowledge about the class. The DRT scores in an M2 tank noise environment show substantial improvement over the MELPe coder.

1. Introduction

Speech enhancement is one of the main strategies for improving the intelligibility of speech coders. Although some progress has been made in enhancing the intelligibility of encoded speech in noise, the intelligibility gap between the uncoded and coded speech is still a major issue [1] [2]. In this work, auxiliary sensors in addition to the acoustic microphone are used to design a novel enhancement system that substantially improves the intelligibility of MELP encoded speech.

Auxiliary sensors, such as the general electromagnetic sensor (GEMS) device [3], have been used for improving the intelligibility of noisy speech [4], [5], [6]. In this work, we have extended our previous speech enhancement algorithm [4] for explicitly incorporating the acoustic-phonetic knowledge of intelligibility into the enhancement algorithm. The speech signal is segmented into broad phonetic classes using the auxiliary sensors with the motivation of adjusting the gain of the enhancement algorithm in each subband based on the sound class information. The relevant cues for a sound class are mildly suppressed while the irrelevant cues are severely suppressed. Although the work in [7] uses a similar approach, it uses hand-segmented data for segmentation, and it is based on a glottal correlation filter while our work is based on a modified Ephraim-Malah Suppression Rule (EMSR) filter.

The Diagnostic Rhyme Test (DRT) scores of the proposed system for an M2 tank noise environment show significant improvement over the NATO standard MELPe coder. The M2 tank noise is known to be a particularly difficult environment [8], and the improvement in speech intelligibility in this environment is promising for next generation coders.

This paper is organized as follows. In Section 2, the segmentation algorithm is described. In Section 3, the speech enhancement algorithm is described. Finally, experimental setup and results are presented in Section 4.

2. Segmentation Algorithm

Speech is segmented using a coarse-grained segmentation method that classifies a speech frame into seven classes. The algorithm has been designed to distinguish between five phonemic categories, i.e., vocalic (vowels, liquids and glides), unvoiced fricative, voiced fricatives, unvoiced plosives (including affricatives) and voiced plosives. The algorithm classifies changeover regions (for example: the aspiration following plosives, the voice-bar preceding voiced plosives, regions just preceding vocalic segments) as 'Transition' regions and silence or noise-only regions between utterances as 'Non-speech' regions.

The algorithm makes hard segmentation decisions based on activity of acoustic and non-acoustic sensor signals using adaptive thresholding of the signal energy. The GEMS signal provides reliable information about the presence of voicing in the speech data. The GEMS sensor is also able to detect the presence of voice-bars (the unspoken periodic vibrations present just prior to the release of voiced plosives). When used with the acoustic signal during these periods, this property provides a fairly good indicator of the presence of voiced plosives. The P-mic sensor does not have the excellent noise immunity property exhibited by the GEMS device and was primarily used to complement the information from the GEMS sensor. The acoustic signal, obtained from a noise-cancelling acoustic microphone, is split into low and high frequency sub-bands with a cutoff at around 3 KHz. This enables the use of reliable high-frequency information for the identification of consonants and plosives when used with information from other sensors.

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.

2.1. Algorithm Details

The algorithm operates on 11.25 ms segments of 16 KHz sampled sensor signals. The GEMS signal is low-passed at 1 KHz to remove noise. The sub-band acoustic signal and the non-acoustic signals are buffered into non-overlapping 11.25 ms frames. The energy estimates of the magnitude spectrums of these signal segments are then computed and normalized to the respective levels of maximum energy regions in a 500 ms window around the present speech segment with a look-ahead of 150 ms.

A primary binary-level activity detection is performed on the GEMS and P-mic signals (*G* and *P* respectively) using an adaptive energy threshold. These thresholds are slowly adapted with respect to the noise floor of previously identified non-speech regions. The combination of the GEMS - P-mic activity decisions (*GP*) basically identifies all regions with voiced speech activity including voiced plosives. A second detector makes activity decisions on the low and high frequency energy contours. It also provides activity information of the first derivative of the high-frequency energy contour. The low frequency activity detector (*LP*) gives information of regions of vocalic speech. The high frequency decision (*HP*) identifies consonants and high-energy plosive regions, while the first-derivative (acceleration) activity (*HPacc*) identifies high-energy plosives only. Vocalic decision (*V*) is made by combining the detected regions of *GP* and *LP* with the detected regions of *G* and *LP* and that of *P* and *LP*. Decisions of unvoiced plosive (*UVP*) is done by combining detected regions of the *HP* and *HPacc*) with regions where voicing was not detected (NOT(*V*)). Table 1 gives the combinations that were used to detect each category.

Table 1: Categorization of phonemes based of detected activity of sensor signals.

Category	Decision Rule
V	((*GP* AND *LP*) OR (*P* AND *LP*) OR (*G* AND *LP*))
UVP	(*HP* AND *HPacc*) AND NOT(*V*)
UVF	(*HP* AND NOT(*HPacc*)) AND NOT(*V*) AND NOT(*GP*)
VF	*HP* AND NOT(*HPacc*) AND (*GP*) AND NOT(*LP*)
T	(*GP* AND NOT(*UVP*)) AND NOT(*UVF*) AND NOT(*LP*)

The individual decisions are combined to get a final decision vector, where each phoneme category is represented by a unique amplitude level. The decision vector is then run through a heuristic rule-checking algorithm that checks for inconsistencies and spurious decisions. The algorithm was tested on the isolated-word DRT sentences of the DARPA /ARCON speech database (created for the DARPA Advanced Speech Encoding (ASE) program). The algorithm gives a classification accuracy of around 70% in high-noise conditions (around 0dB). However, it was observed that the segmentation performance degrades for continuous speech. This can be attributed to the fact that the behavior of the GEMS signal is is different in the isolated-word versus continuous speech case and the heavy dependence of the algorithm on the GEMS signal. Currently, work is in progress to develop a soft-decision speech segmentation and classification framework. The non-acoustic sensor information will be used to sustain performance in adverse noise conditions rather than be the basis for segmentation/classification process itself.

3. The Speech Enhancement Algorithm

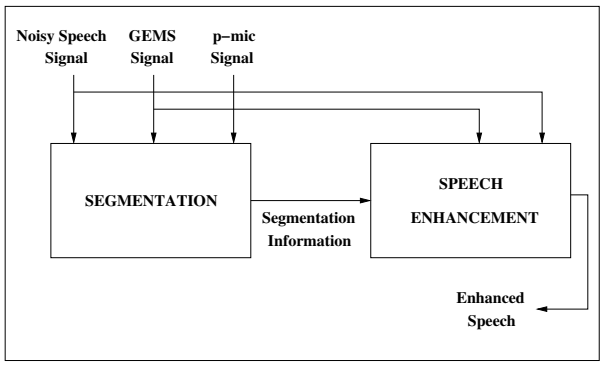

Figure 1: An overview of the proposed system. The speech signal is segmented using the noisy acoustic signal, the GEMS signal, and the P-mic signal. The segmentation information is used in the speech enhancement block.

The general overview of the proposed system is shown in Fig. 1. The proposed speech enhancement algorithm receives segmentation information for each frame from the segmentation algorithm. Speech frames are enhanced differently based on the phoneme class to which they belong. The key idea in this system is to maximally suppress the noise in the signal without distorting or deleting the perceptual cues that are vital for identifying the speech sound.

The operation of the system is fundamentally based on the Ephraim-Malah Suppression Rule (EMSR) that incorporates signal uncertainty to the enhancement system. The gain function of the EMSR algorithm is modified by taking into account the signal presence probability as proposed in [9]. The modified gain function G_m is given as

$$G_m = G_{emsr}^{P_s} G_{min}^{1-P_s} \quad (1)$$

where G_{min} is set to -20 dB, and P_s is the signal presence probability. The bands that have high chance of carrying useful speech information are mildly suppressed while low chance bands are suppressed severely. The problem in this approach is reliably estimating P_s. Initial consonants, particularly the unvoiced ones, have relatively weak energy and misestimation of P_s can significantly reduce intelligibility. In this work, combinations of three methods are used to detect signal presence as described below.

3.1. Signal Presence Detection

In order to detect the signal presence probability given the GEMS signal and the segmentation information three methods are employed in the system. The first method utilizes the

GEMS signal for detecting the harmonic frequencies in the voiced speech spectrum. Harmonic frequencies carry perceptually important information in the spectrum. Moreover, they have significantly higher Signal to Noise Ratio (SNR) compared to the other frequencies. In the first method, harmonic frequencies detected by the GEMS signal are assigned signal presence probability (P_s) of 1 while other frequencies are assigned a P_s of 0. The harmonic tracking algorithm is described in Section 3.2

The second method measures the SNR at each frequency and assigns P_s values as proposed in [9]. This method is used for rapidly changing voiced speech segments such as glides and liquids.

The third method uses some of the acoustic phonetic knowledge that is available in the speech literature to suppress the irrelevant parts of the spectrum. For instance, in unvoiced fricatives, such as /s/, perceptually important cues are typically above 2000 Hz. Therefore, P_s for frequencies below 2000 are set to 0. This method not only suppresses a significant amount of noise in the signal, but also increases the DRT scores for sibilation as shown in Section 5.

These three methods and their combinations are used for assigning P_s to spectral bins as follows

- The harmonic tracking method (method 1) is used for all frequencies for the vowel class.

- Acoustic-phonetic knowledge (method 3) is used for unvoiced fricatives. The spectral cues for this class is typically at the higher frequencies. Therefore, a P_s of 1 is used for the $2000 - 4000$ Hz range, and a P_s of 0 is assigned for the $0 - 2000$ Hz.

- P_s is set to 1 for the $400 - 4000$ Hz range of unvoiced plosives. SNR method (method 2) does not perform well for this region since the energy of the burst is low and the SNR estimation is not reliable. At such low SNRs, small SNR estimation errors are found to create significant loss of cues or generate false cues. The $0 - 400$ Hz range is assigned a P_s of 0, since typically the fundamental frequency of voiced sounds is in that region, and residual noise can cause confusion of an unvoiced plosive with a voiced sound.

- The SNR method is used for transitional sounds and transients. Although the GEMS signal detects those regions, we have found that sometimes the harmonic tracker fails to detect perceptually important formant transitions in those regions because of relatively weaker GEMS signal at the onset and offset of the voiced sounds.

- The harmonic tracking method is used for the first 500 Hz of voiced plosives. The GEMS signal is found to be inconsistent and has low energy for frequencies higher than 500 Hz for voiced plosives. As in the unvoiced plosives case, the SNR method is not reliable due to low energy spectral cues, and P_s of 1 is used for all frequencies above 500 Hz.

- In voiced fricatives, both lower and higher frequencies contain important cues. Low frequencies are important since they indicate voicing in the signal. High frequencies are important since they indicate frication in the sound. We have used the harmonic tracker for frequency range of $0 - 2000$ Hz, and the SNR method for $2000 - 4000$ Hz range.

3.2. Harmonic Tracking

In this work, the GEMS signal is used for detecting the harmonic locations in voiced speech spectrum. The GEMS signal has harmonic structure that is very similar to the acoustic signal. Thus, if both signals are windowed using the same window, the GEMS signal can be used to accurately detect the high signal power (HSP) locations in the voiced speech spectrum. [1] A hard-decision thresholding algorithm is used for detecting the HSP locations; and the binary decisions are stored in the vector P_s.

In the initial phase, pitch information is extracted from the GEMS signal. The autocorrelation method is used at this step where the maximum lag in the interval of 2.5 ms and 10 ms is chosen as the pitch. The GEMS spectrum is divided into subbands with a bandwidth of pitch frequency.

The algorithm for detecting the HSP locations in each subband can be described as follows. Four types of high signal power cues are identified in the spectrum:

1. $P_s(k)$ is set to 1 at the two highest energy frequency bins 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 constant 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.

4. Two HSP points can exist consecutively. Therefore, $P_s(k)$ is set to 1 if

   - $|\zeta_k - \zeta_{k+1}| < \zeta_{th,2}$ and $P_s(k+1) = 1$ or
   - $|\zeta_k - \zeta_{k-1}| < \zeta_{th,2}$ and $P_s(k-1) = 1$

where $\zeta_{th,2}$ is a constant representing the second power threshold, and it is set to 0.2 dB in this work. For each subband in a windowed speech frame, a two iteration procedure is followed. In the first iteration type 1, type 2, and type 3 locations are found. The algorithm attempts to find at least three HSP locations in each subband. ζ_{th} is initialized to 3 dB. If the number of HSP locations is less than three, then ζ_{th} is decreased with a step size of 0.5 dB, and the procedure is repeated until at least four 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, set to 0. In the second iteration, type four locations are found.

[1] The exact harmonic locations can not be detected since the resolution in the frequency domain is limited with the sampling rate. The HSP locations are within the neighborhood of the exact harmonic location. Therefore, harmonic tracking and HSP tracking are used interchangeably in this work.

The operation of the harmonic tracker is shown in Fig. 2. The system can catch all the high signal power locations in the spectrum. It introduces a false harmonic at $k = 5$. Although these type of errors are occasionally introduced into the system, we have found that they do not distort the perceptual quality significantly.

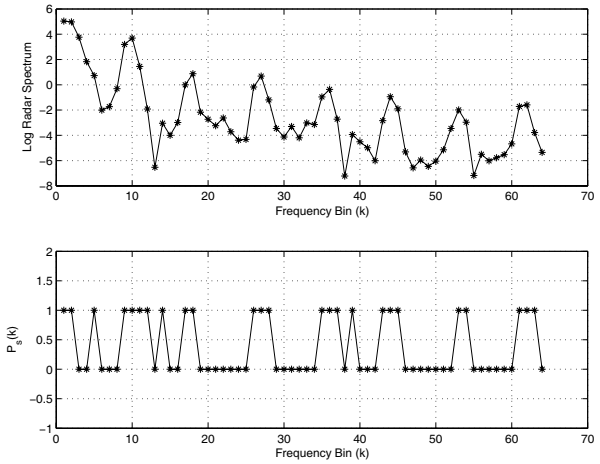

Figure 2: An illustration of the harmonic detection algorithm. An example GEMS spectrum is shown in the top figure. The output of the harmonic tracker is shown in the bottom figure.

4. Experiments

The proposed system is used as a front end to the 2.4 kbps MELP speech coder and compared with the MELPe NATO speech coding standard using in-house diagnostic rhyme tests (DRT). The MELPe coder uses the EMSR algorithm that employs only the SNR based signal uncertainty method. Four native English speakers took the test for 10 minutes of audio. The audio contains the DRT sequences spoken by two male and two female trained native English speakers. The audio files are recorded in the simulated M2 tank noise environment which is found to be one of the most difficult environments for intelligibility of MELP coded speech. The noisy speech data is part of the ARCON speech database that is created as part of the DARPA Advanced Speech Encoding (ASE) program.

Results are shown in Table 1. The proposed system outperforms the MELPe system for all distinctive features. Voicing and nasality are high as expected since the GEMS signal provides harmonic information for those voiced segments. The sustention feature is related with the aspiration after the burst for the plosive sounds. The cues for sustention are very sensitive since the low energy burst and aspiration can be easily masked by noise. The proposed system uses the acoustic-phonetic information for plosives and clearly outperforms the MELPe coder significantly for plosives. Similar observations can be made for graveness and compactness.

Table 2: Performance results in terms of DRT scores for each distinctive feature of consonants for the MELPe coder compared with the proposed system operating at the front end of a MELP coder in M2 high noise environment.

Feature	MELPe	Proposed System
Voicing	74.6	75.26
Nasality	60.21	89.58
Sustention	43.4	58.85
Sibilation	82.6	85.42
Graveness	67.3	70.05
Compactness	75.0	81.25
EXP	88.8	90.42
Total	67.6	76.81

5. References

[1] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Istanbul Turkey, June 2000.

[2] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s melp coder candidate for the new us federal standard," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, May 1996.

[3] 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.

[4] C. Demiroglu and D. V. Anderson, "A soft decision mmse amplitude estimator as a noise preprocessor to speech coders using a glottal sensor," in *ICSLP*, Jeju island, Korea, October 2004.

[5] R. Hu and D. V. Anderson, "Single acoustic-channel speech enhancement based on glottal correlation using non-acoustic sensor," in *ICSLP*, Jeju island, Korea, October 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.

[7] R. Hu and D. V. Anderson, "Audio noise suppression based on neuromorphic saliency and phoneme adaptive filtering," in *DSP Workshop*, Taos, NM, August 2004.

[8] 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.

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