Using a low-bit Speech Enhancement Adaptive Filter to Improve GSM Speech Recognition Performance

N.S. Mahlanyane and D.J. Mashao

Department of Electrical Engineering, University of Cape Town,
Rondebosch, Cape Town, South Africa

mhlnku002@mail.uct.ac.za, daniel@eng.uct.ac.za

Abstract

Robustness of speech recognition systems to speech that has been transmitted through GSM voice communication channels is generally very poor. The two main causes of this poor performance are speech coding and GSM channel noises. In this paper we propose the use of a low-bit speech enhancement adaptive filter as a speech recognition system pre-filter to improve robustness of recognition systems to GSM channel speech. Results of experiments performed with this filter indicated that the use of the filter improved GSM speech recognition performance by over 6%.

1 Introduction

Performance of speech recognition systems when they are used to recognize speech that has been transmitted through GSM voice communication channels (GSM speech) is generally poor. This poor performance limits the use of speech recognition systems over GSM networks.

There are two main sources of the degradation of performance and these are GSM speech coding and GSM channel noise. Previous research has shown that recognition of speech degrades as its bit rate is lowered [1, 2]. However, it has also been found that GSM channel noise contributes more to this degradation than speech coding. The aim of this work is to improve performance of speech recognition systems with GSM speech.

Different channel normalization techniques, which attempt to alleviate the effects of channels on channel modified speech signals in speech recognition, have been developed. These techniques can be classified into three broad categories, namely, model modification, signal pre-processing and feature processing techniques. Model modifying methods attempt to reduce effects of channel noise during the recognition process. Signal pre-processing and feature processing techniques attempt to reduce effects of channel noise prior to the recognition process. These usually involves the use of filters. Some of the well known of these are the RelAtive SpecTrA (RASTA) processing [3] and Cepstral Mean Subtraction (CMS) [4]. In most cases, techniques from each of these three categories are used together in one recognizer.

In this work we propose the use of an adaptive filter [5] to improve robustness of recognition systems to GSM speech. The aim of this filter is to reduce the effects of GSM speech coding. This filter, in combination with channel normalization techniques improves performance much more than when channel normalization techniques are used alone. As outlined in Section 3 this filter was tested in combination with the CMS channel normalization method. Details of this filter are outlined in Section 2.

Performance of systems in this paper is measured in terms of percentage word error rates (WER). The formula used for calculating performance is,

\[
WER = \frac{I + D + S}{N} \times 100\%
\]

where \(I\) is the total number of insertion errors, \(D\) the total number of deletion errors, \(S\) the total number of substitution errors made by the evaluated system and \(N\) is the total number of words there are in the given test set.

2 Adaptive Filter

In this Section the adaptive filter proposed to improve robustness of recognition systems to GSM speech is described. The filter has been used previously as a coded speech enhancement post-filter in low-bit (≤ 8kbps) speech coding telecommunications. It was so successful in its use in telecommunications that different variations of it have been incorporated into several national and international speech coding standards [6]. Among these standards are the U.S. Federal Standard 4.8kbps CELP (FS1016) [7] and the North American digital radio standard 8kbps VSELP (IS-54) [8]. In this paper we propose its use as pre-filter for a GSM channel speech.

In low-bit speech coding telecommunications it is used to enhance the perceptual quality of coded speech by reducing perceived noise introduced by low-bit speech.
In this work we show that it also improves speech recognition performance by countering the effects of GSM speech coding.

The transfer function of the filter is defined as,

\[ F(z) = \frac{A(z/\beta)}{A(z/\alpha)} \]

where \( A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \) (an LPC analysis filter), \( \alpha \geq 0 \) and \( \beta \leq 1 \). The choice of parameters \( \alpha \) and \( \beta \) depends on the speech coding scheme for which the filter is used and the bit rate of that speech codec [3]. In this work the LPC order \( p \) was set to 10. The frame size over which the LPC coefficients \( a_k \) are computed, was set to 200 samples on 8kHz sampled speech signals.

When \( \beta = 1 \) and \( \alpha < 1 \), Equation 2 then becomes,

\[ F(z) = \frac{A(z)}{A(z/\alpha)} \]  

This form of the filter is exactly the same as an error weighting filter [9] used in the GSM speech codec and other Analysis-by-Synthesis speech codecs. Throughout the rest of this paper error weighting filter refers to the adaptive filter in the form of Equation 3. In the GSM speech codec, the error weighting filter is used to mask quantization noise in speech signals.

Figure 1 depicts the operation of the error weighting filter. In the figure the error weighting filter has been segmented into an analysis filter \( A(z) \) and the synthesis filter \( W(z) \) blocks. The synthesis filter is also known as the weighted synthesis filter [10]. The output of the analysis filter is the error signal, \( r(n) \). The error weighting filter attenuates the spectrum of the error signal, \( r(n) \), in formant regions. This way noise in formant regions of speech, which are more important in speech perception, is attenuated. This, however, is achieved with undesirable effects of amplifying the noise in non-formant regions as well. The attenuation of noise in formant regions of speech also attenuate the formants which is also undesirable yet unavoidable.

Figure 2 shows the LPC spectrum and the spectra of the error weighting filter for different values of \( \alpha \). These spectra show the desired attenuation in formant regions and amplification in non-formant regions of the error signal when the error weighting filter is applied. Experiments were conducted with the error weighting filter to determine its effects on the robustness of recognition systems to GSM speech. Results of these experiments are given in Section 4.

3 Experimental Setup and Procedure

No off-the-shelf speech recognition systems were used for experimentation, all of them were built from scratch using the hidden Markov model toolkit version 3.2 (HTK 3.2) [11]. The systems built were all of the same design, differences only come up in front-ends used. They were trained using the TIMIT database training set [12] downsampled to 8kHz. All HMM models for all 49 phones used were 5-state, left-to-right non-skipping models with continuous Gaussian mixture densities. The number of mixture densities was kept as one per state per model until later stages of the building process in which they would be incremented to improve performance.

For all recognizers, feature extraction was applied on 25ms long frames of speech segmented at the rate of
10ms. The hamming window was applied on each frame of speech prior to feature extraction. For all features extracted, delta and acceleration coefficients were computed over a window of 2 frames.

All recognition systems built employ a bigram language model. The best language model settings were worked out to be 30 and 50 for language model (LM) probability weighting and word insertion penalty respectively. These language model settings were used in all the experiments performed.

### 3.1 Experimental Procedure

The development of a speech recognition system for a specific task is experimental in nature. It often starts with a construction of a basic system whose performance on a given task is improved until peak performance is reached. There is often a number of language models that are fine-tuned to reach this peak performance. In this work the same procedure is followed. A basic recognition system with single mixture density HMMs is constructed. The performance of this system is improved by incrementing the mixture densities to determine the number of mixtures that given peak performance. The final number of mixtures for the system is taken as the one at which the system produces peak performance.

### 3.2 GSM Test Set

Because a comprehensive medium vocabulary database of GSM data was not available for use, we had to make our own recordings of such data for testing. We had two Mobile Stations (cell phones), one connected at the transmitting end and the other at the receiving end. The one at the transmitting end was used to transmit speech signals from the TIMIT database core test set [12] over the Vodacom (South African cellular network company) GSM network to the receiver end. The receiving end received the signal from the network and recorded it. The transmitter and the receiver machines were appropriately networked for synchrony purposes.

### 3.3 Base-line System Front-end

Since speech recognition system development is experimental in nature, front-ends usually differ from system to system. There is therefore no front-end standard across all systems. The chosen front-end for base-line results in this work is the one used by the HTK group in the DARPA’02 evaluations [13]. These evaluations involved testing of speech recognition systems developed by different leading research groups in speech recognition. Among different testing speech data was GSM speech. This is the motivation behind the choice of this front-end. This front-end consists of the following:

- Bandwidth reduced to 125 - 3800
- 12 MF-PLP cepstral features + the zeroth cepstral coefficient (C0) and 1st and 2nd order derivatives
- Side-based cepstral mean and variance normalization
- Vocal tract normalization in both training and testing data

The MF-PLP feature-extraction method was first used in the 1996 HTK Broadcast News Transcription System [14]. In the development of that system the MF-PLP feature extraction method was found to be more robust and consistent than the popular MFCC and PLP feature extraction methods. All the systems that the HTK group developed since 1996 implemented used the MF-PLP feature extraction method.

*Side-based cepstral mean and variance normalization* is termed side-based because the cepstral means and variances are computed before from a somehow clustered subset of data. In our implementation, side-based cepstral means and variances were computed per speaker for all data types.

Base-line results obtained using this system are outlined and discussed in subsection 4.1.

### 4 Results and Discussions

In this section results of experiments conducted are mentioned and discussed. As mentioned in Section 2, there were two forms of the filter considered, each characterized by the values assigned to its variable parameters. The parameters are the error weighting filter ($\beta = 1$ and $\alpha$ varied) and the adaptive filter ($\alpha$ fixed and $\beta$ varied). The filters were applied on both the training data of the system and the GSM testing data. Results of each of these, as well as of the base-line system, are listed and discussed separately in the sub-sections following. Since
the best settings for these parameters ($\alpha$ and $\beta$) depend on the bit rate of the speech signals on which the filter is applied [5], they had to be varied to determine the best combination for GSM speech whose bit rate is 13kbps.

4.1 Base-line Results

Base-line results upon which improvements are to be made are shown on Figure 4. This system was tested with both the clean test data and the GSM test data. As may be seen on the figure, the system produced the best %WER of 7.2% at 4 mixtures when tested with clean speech and 33.97% at the same number of mixtures when tested with GSM speech. This increase in word error rate served as proof of the poor robustness of speech recognition systems to GSM speech.

4.2 Error Weighting Filter Results

As mentioned in Section 2, the error weighting filter is the same as the adaptive filter with $\beta = 1$ and $\alpha < 1$. In [5], $\alpha = 0.8$ was found to give the best speech enhancement results. We chose this value as the initial value when the error weighting filter was applied. Other values in the vicinity of this value were also tested to see if they would bring about better performance. $\alpha = 0.8$, of all the values tested, was found to produce the best recognition results for GSM speech. It can be seen on Figure 5 that the error weighting filter with $\alpha = 0.8$ also brought about an improvement in the recognition rate of GSM speech compared to that of the base-line system. At 4 mixtures the error weighting filter system matched the base-line system with a %WER of 33.97. Beyond 4 mixtures the error weighting filter system proved the better of the two systems.

At mixture numbers 7, 8 and 9 the error weighting filter system (i.e. without the adaptive filter, $\beta = 1.0$) produced word error rates lower than the best produced by the base-line system. The best % WER produced by this system was 32.76 at 8 mixtures. This is about 4% drop in the error rate produced by the base-line system. These improvements are attributed to the speech coding noise masking ability of this filter for which it was considered in the first place.

4.3 Adaptive Filtering Results

Results of applying the adaptive filter in its complete form showed that the $\beta$ value (with $\alpha$ fixed at 0.8) that produces the best results is 0.4. At this value the system shows a steady decrease in the % WER with the increase in the number of mixtures. As can be seen on Figure 5, the system’s performance is lower than that of the base-line system at 4 mixtures where the base-line system produced the best performance (i.e. 33.97%). Beyond 4 mixtures the system produced error rates lower than that of the base-line system. The best of these is 31.93% at 7 mixtures. This is a 6% reduction of the WER produced by the base-line system.

These results also show that the application of this filter brings about further improvements to those brought about by the error weighting filter. For all mixtures its error rate is below that of the error weighting filter except at 4 mixtures. This improvement over the error weighting filter system is attributed to the fact that the adaptive filter eases the spectral tilt of the error weighting filter when values of $\beta$ less than 1 are used bringing about further improvements in the system’s performance on GSM speech.
5 Conclusions

From the results obtained, it is concluded that the adaptive filter does improve robustness of speech recognition systems to GSM speech especially at $\alpha = 0.8$ and $\beta = 0.4$. The error weighting filter used in the GSM speech codec also improves robustness of clean speech recognition systems to GSM speech on its own through noise masking.

References


