SPEECH RECOGNITION FROM GSM CODEC PARAMETERS

Juan M. Huerta and Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

ABSTRACT

Speech coding affects speech recognition performance, with recognition accuracy deteriorating as the coded bit rate decreases. Virtually all systems that recognize coded speech reconstruct the speech waveform from the coded parameters, and then perform recognition (after possible noise and/or channel compensation) using conventional techniques. In this paper we compare the recognition accuracy of coded speech obtained by reconstructing the speech waveform with the speech recognition accuracy obtained when using cepstral features derived from the coding parameters. We focus our efforts on speech that has been coded using the 13-kbps full-rate GSM codec, a Regular Pulse Excited Long Term Prediction (RPE-LTP) codec. The GSM codec develops separate representations for the linear prediction (LPC) filter and the residual signal components of the coded speech. We measure the effects of quantization and coding on the accuracy with which these parameters are represented, and present two different methods for recombining them for speech recognition purposes. We observe that by selectively combining the cepstral streams representing the LPC parameters and the residual signal it is possible to obtain recognition accuracy directly from the coded parameters that equals or exceeds the recognition accuracy obtained from the reconstructed waveforms.

1. INTRODUCTION

Speech coding affects speech recognition accuracy, with word accuracy deteriorating as the coded bit rate decreases [4, 6]. Due to the increase of speech communication applications employing coding algorithms and the interaction of these speech communications systems with automatic speech recognition applications, coding of speech can become a significant problem that limits the performance of such applications [3, 6, 7]. Several approaches that deal with this problem have been proposed (*e.g.* [3, 7]). These approaches involve the regeneration of the speech signal prior to applying compensation and adaptation techniques. The degradation in recognition accuracy is greater when the speech used to train the recognizer had not undergone the identical coding process (*i.e.,* "mismatched conditions"). Nevertheless, using similarly-coded speech for both training and testing reduces but does not eliminate the degradation in recognition accuracy compared to the accuracy obtained with uncoded speech [7].

Using the 13-kbps full-rate GSM codec, we consider in this paper the effects of speech coding on parameter representation accuracy and on speech recognition accuracy. GSM is a Regular Pulse Excited Long Term Prediction (RPE-LTP) coding process [2]. We assume that the speech recognition system has access to the transmitted GSM parameters of the coded speech signal. We analyze the effects of lossy compression and quantization on the cepstra derived from quantized Log Area Ratios (LAR), and from the residual signal reconstructed from the RPE-LTP parameters, by comparing them to corresponding cepstra derived from uncoded and unquantized versions of these signals.

We will demonstrate that the effects of quantization and coding affect the individual coefficients cepstral representations of the LPC filter and residual excitation signal in differing amounts. We will use these observations to guide us in combining the cepstral representations of the LPC filter and the residual signal to minimize speech recognition error rate.

In Section 2 we discuss briefly the characteristics of the GSM codec. We discuss the effect of GSM coding and quantization on speech on cepstral features in Section 3, and we present recognition results employing those features. In Section 4 we discuss methods for recombining the coefficients extracted from these cepstral features in order to minimize the recognition error rate of GSM-coded speech signals.

2. THE FULL-RATE GSM SPEECH CODEC

The full-rate GSM speech codec [2] is a lossy speech coding-decoding algorithm based on a regular pulse excited long term prediction scheme [5]. GSM converts 13-bit digital signals sampled at 8 kHz into blocks of 260 bits for every 160 original samples. Hence, the GSM coding algorithm produces a gross bit rate of 13.0 kbps, although the actual GSM transmitted bit rate is higher due to added error recovery and packet information. The RPE-LTP coding algorithm is a member of the linear predictive analysis-by-synthesis (LPAS) family of coding algorithms [4].

As is the case with all LPAS algorithms, the GSM codec represents the speech signal using two sets of parameters: information about the LPC filter (in the form of quantized log area ratios, or Q-LARS) and information about the coded residual signal (in the form of quantized RPE-LTP parameters). The compression of the residual signal is a lossy process which introduces distortion into the residual signal. During decoding, the residual signal is first reconstructed from the RPE-LTP information, and then filtered by the short-term synthesis filter, whose parameters are derived from the received LARs.

Figure 1 shows a schematic representation of a general analysis-by-synthesis coder. In the specific case of the full-rate GSM coder the block that minimizes the difference between the actual residual signal and the reconstructed residual signal computes the quantized RPE-LTP representation of this difference. Besides the lossy representation of the residual signal that this algorithm introduces in the RPE-LTP section, quantization of the LAR coefficients plays a role in the degradation observed in speech that has undergone the GSM coding process.

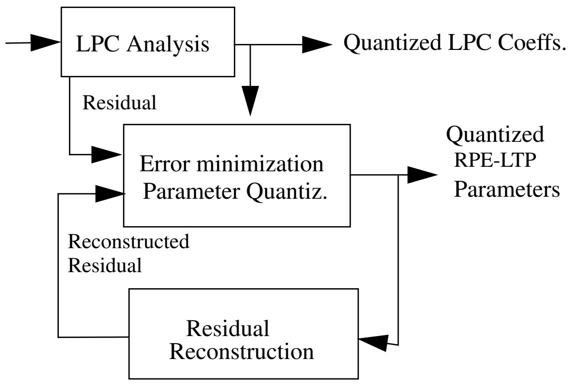

Figure 1: A simplified block diagram of a typical analysis by synthesis coder.

3. THE IMPACT OF PARAMETER QUANTIZATION AND CODING ON CEPSTRA

In this section, we describe the procedure used to develop cepstral features for speech recognition from signals and parameters developed by GSM coding of speech. We consider three sets of cepstral vectors: vectors derived directly from the reconstructed GSM speech signal, vectors derived from the log area ratios representing the LPC filter, and vectors derived from the residual signal. We compare these cepstra with the uncoded and unquantized versions of the signals and parameters listed above to determine the extent to which coding and quantization affects representation accuracy. Finally, we compare the accuracy obtained using these various features in speech recognition systems.

3.1. Recognition using Reconstructed GSM Speech

Most recognition systems operate directly on speech waveforms that are decoded from GSM parameters in conventional fashion. The differences between the GSM-decoded signal and the original speech waveform can cause a degradation in speech recognition. GSM coding affects the various cepstral coefficients used to represent decoded speech in different proportions. In Figure 2 we plot the normalized mean square error (NMSE) between corresponding coefficients of the original and GSM-decoded speech cepstral vectors (normalized by dividing the mean square error by the average squared value of a given coefficient). If we consider the effects of distortion to be an additive noise signal, the NMSE would be roughly proportional to the inverse of the signal-to-noise ratio (SNR). As can be seen in Figure 2, the NMSE introduced by GSM coding generally increases as the coefficient index increases.

3.2. Deriving Cepstra from the LPC Log Area Ratio Parameters

Cepstral coefficients can also be obtained from the quantized log area ratio (LAR) parameters that are developed in the course of GSM coding. The LAR parameters are transformed into the corresponding LPC coefficients, from which cepstral coefficients are generated directly using the approach described in [1]. The GSM standard specifies that 8 coefficients are generated using an eighth-order LPC analysis.

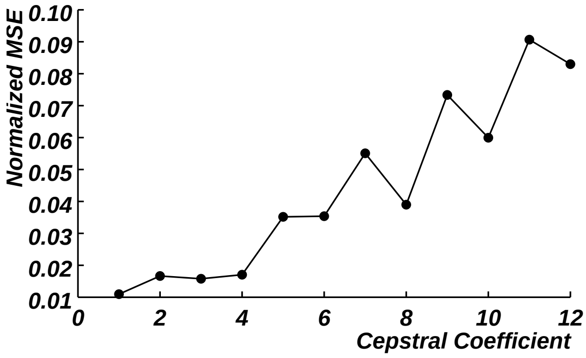

Figure 2: Normalized mean square error (NMSE) of the cepstra of GSM-reconstructed speech waveforms using the cepstra of the original waveforms as the standard. Normalization is with respect to the average energy of each cepstral coefficient.

The NMSE of cepstral coefficients developed from the LPC analysis of GSM-encoded speech signals are plotted in Figure 3, in the same fashion as in Figure 2. The general effect of GSM coding for these coefficients appears to be similar to that of the NMSE of the coefficients representing the original waveform in that the NMSE generally increases as the coefficient order increases.

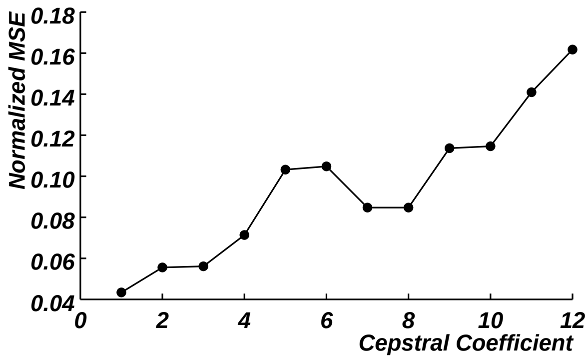

Figure 3: Normalized mean square error of cepstra derived from the quantized LARs of GSM-encoded speech waveforms with respect to the corresponding cepstra of the original waveforms (without quantization).

3.3. Deriving Cepstra from the Residual Signal

Cepstral coefficients can also be generated from the RPE-LTP parameters that represent the residual excitation signal. The RPE-LTP coefficients are obtained from conventional cepstral analysis of time functions. While the residual signal is generally assumed to contain primarily information that is less relevant to the speaker independent speech recognition task such as pitch, periodicity, and glottal waveform information [8]. However, because only an eighth-order LPC analysis is used in LPC coding, the residual signal still carries information that is useful for speech recognition.

We generated cepstral coefficients from the residual obtained from the RPE-LTP parameters of the GSM codec (*i.e.,* the reconstructed GSM residual) and compared their values to the corresponding coefficients for the original uncoded speech signal. Figure 4 shows the NMSE of the cepstral coefficients representing GSM-encoded speech, with respect to the corresponding coefficients of the original uncoded speech. In contrast to the NMSE of the reconstructed waveform and the Q-LARs shown in Figs. 2 and 3, the NMSE of the cepstral coefficients representing the residual signal tends to *decrease* as the coefficient order increases. We also note that the magnitude of the NMSE of the residual is much greater than that of the cepstra of both the Q-LARs and the reconstructed speech waveform.

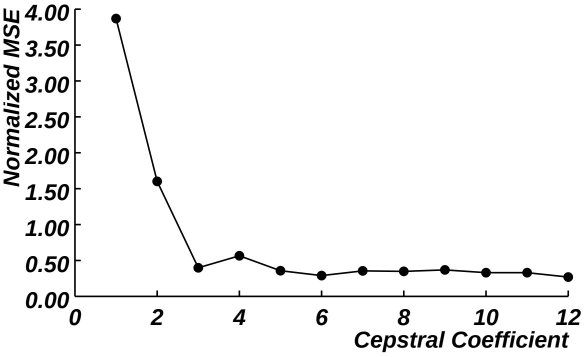

Figure 4: Normalized mean square error of cepstra derived from the residual signal of GSM-encoded speech waveforms with respect to the corresponding cepstra of the original waveforms (without quantization).

4. EFFECT OF GSM CODING ON SPEECH RECOGNITION ACCURACY

In this section we describe the results of a series of speech recognition experiments using cepstral features derived from the reconstructed waveforms and from the GSM parameters themselves. Recognition experiments were performed using a reduced-bandwidth and downsampled version of the speaker independent component of the Resource Management RM1 corpus [9] under clean and noisy conditions. In all cases the speech signal was low-pass filtered to 3.5 kHz and downsampled to 8 kHz. For noisy conditions, stationary additive lowpass colored noise was added to yield a resulting SNR of approximately 18 dB. The colored noise was generated by filtering white gaussian noise through a simple 2-pole filter with a resonance of approximately 650 Hz and a half-power bandwidth of approximately 400 Hz. The acoustic models employed consisted of a set of senonically-tied continuous density HMMs, modeled by approximately 2500 senones and 2 gaussians per mixture.

4.1. Recognition Accuracy using Original and Reconstructed Speech Waveforms

Table 1 compares speech recognition accuracy obtained using various cepstral feature sets, with and without the additive noise. For each feature set, acoustic models were trained with features used to test the system, and without the additive colored noise. Results in the first three rows of Table 1 compare the recognition accuracy using Mel-frequency cepstral coefficients (MFCCs) generated from the original speech without GSM coding (Row 1), and GSM-processed speech (Rows 2 and 3). Training is “mismatched” in Row 2 in that the system was trained using uncoded speech; GSM coding is used for both training and testing for the results in Row 3. The effect of GSM coding on recognition error rate was relatively modest for this dataset: the error rate increased by about 20% for clean speech and 6% for noisy speech with mismatched training, and most of the degradation was eliminated when GSM coding was used in training as well as in testing.

Feature Set	Clean	Noisy
MFCC coefficents from original waveform	89.7%	45.0%
MFCC coefficients from GSM-decoded speech (mismatched models)	87.7%	41.5%
MFCC coefficients from GSM-decoded speech (matched models)	89.2%	47.5%
LAR CEPSTRA	87.9%	44.1%
Q-LAR CEPSTRA	87.5%	44.9%
RESIDUAL CEPSTRA	71.1%	1.4%
GSM-RESIDUAL CEPSTRA	67.5%	3.9%

Table 1: Recognition accuracy obtained for speech without and without GSM encoding, and with and without additive noise, using cepstral features derived from the waveform and from the GSM parameters directly. See text.

4.2. Recognition Accuracy using Features Derived from GSM parameters

Rows 4 through 6 of Table 1 compare recognition accuracies obtained using cepstra generated from unquantized and quantized LARs, and from the original residual signal and the GSM-restored residual signal. The accuracy of this pair of features reveals the existence of information relevant to recognition in the residual signal. These results indicate that recognition accuracy obtained from features derived from the LAR and Q-LAR parameters is almost as good as recognition accuracy obtained from the reconstructed waveforms themselves. Features derived from the residual signal are somewhat less effective.

5. COMBINING Q-LAR CEPSTRA WITH GSM-RESIDUAL CEPSTRA

Since in traditional LPC theory, reconstructed speech waveforms are obtained by the convolution of the impulse response of the LPC filter with the residual signal, the cepstrum of the speech waveform can be estimated by adding the cepstra of the LPC filter and of the residual. As discussed in Section 3, however, the NMSE of these two sets of cepstral coefficients behave differently. In this section we show that we can improve recognition accuracy by *selectively* combining Q-LAR cepstral coefficients with cepstral coefficients derived from the GSM-restored residual signal.

We consider two ways of combining the cepstra representing the LPC filter and the residual filter: (1) direct addition of the two sets of cepstra (which indeed corresponds to convolving the impulse response of the LPC filter with the residual signal), and (2) assembling a *13*-dimensional composite cepstral vector by concatenating a subset of the cepstral coefficients representing the LPC filter with a subset of the cepstral coefficients representing the residual waveform. We implemented the latter procedure by combining the first i coefficients of the quantized-LAR Cepstra and the last 13 *minus* i coefficients of the GSM-restored residual cepstra. These subsets of coefficients were chosen because the NMSE of the residual cepstra is smaller for the higher order coefficients, as shown in Figure 4. In further experiments we confirmed that good recognition accuracy for the concatenated vector could be obtained provided using other combinations of specific coefficient, provided that the first two cepstral coefficients from the residual signal were excluded. (These coefficients exhibit the greatest NMSE.)

Table 2 compares recognition results for a set of values of the parameter *i*, which we refer to as "cutoff values", ranging from *i=5* to *i=10*. We note that in this table a cutoff of zero is equivalent to using a 13-element GSM-residual cepstral vector; a cutoff of 13 is equivalent to using Q-LAR cepstra. From Table 2 it appears that best results are obtained when approximately 8 cepstral coefficients representing the LPC filter are combined with 5 coefficients representing the residual signal.

Cutoff	Clean	Noisy
0	67.5%	3.9%
5	88.8%	40.5%
6	89.2%	42.7%
7	89.7%	46.4%
8	89.7%	49.4%
9	89.6%	50.2%
10	88.7%	50.2%
13	87.5%	44.9%

Table 2: Recognition accuracy obtained by combining Q-LAR and GSM-residual cepstral using various cutoff values. (See text.)

.

Feature	Clean	Noisy
Concatenation of Q-LAR and residual cepstra (cutoff value equals 8)	89.7%	49.4%
Sum of Q-LAR and residual cepstra	89.1%	47.1%

Table 3: Performance of Cepstral features resulting from the concatenation and addition of LAR and residual cepstral streams.

Table 3 compares recognition accuracy obtained by concatenating the Q-LAR and the GSM-residual cepstral vectors, as discussed above, with simply adding them together as would be suggested by LPC theory. As can be seen, the concatenated feature vector is more effective than simple addition for both conditions considered. Even more interesting is the fact that recognition accuracy obtained using the concatenated GSM feature vector is greater than both the accuracy obtained using reconstructed waveforms, and the accuracy obtained with the original uncoded speech waveform.

6. DISCUSSION AND SUMMARY

The degrading effect of GSM coding on speech recognition accuracy has been associated with the distortion introduced to cepstral representations of the log area ratios and the restored residual signal, after quantization and lossy coding. Of the representations of GSM parameters considered, we observed greatest normalized mean-square error for the *highest*-order cepstral coefficients representing the LARs (and hence the LPC filter), and for the *lowest*-order cepstral coefficients representing the residual excitation signal. In order to obtain best speech recognition accuracy, it is necessary to concatenate lower-order coefficients that represent the LPC filter with higher-order coefficients representing the residual signal. Speech recognition accuracy for the NIST RM1 database was greater when the concatenated feature vector derived directly from the GSM parameters was used than when features were extracted from speech waveforms reconstructed by the GSM decoder.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

7. REFERENCES

1. Atal, B. S. "Effectiveness of Linear Prediction Characteristics of the Speech Wave", *J. Acoust. Soc. Am.* Vol. 55, No.6: 1304-1312, June 1974

2. European Telecommunication Standards Institute, "European digital telecommunications system (Phase 2); Full rate speech processing functions (GSM 06.01)", ETSI 1994

3. Haeb-Umback, R. "Robust Speech Recognition for Wireless Networks and Mobile Telephony" *Proc. EUROSPEECH 97* Vol.5: 2427-2430., 1997.

4. Kleijn, W. B., and Paliwal, K. K. (editors), *Speech Coding and Synthesis*, Elsevier Science B.V., Amsterdam 1995

5. Kroon, P., Deprettere, E. F., Sluyter, R. F. "Regular-Pulse Excitation - A Novel Approach to Effective and Efficient Multipulse Coding of Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-34 No. 5:1054-1063, October 1986

6. Lilly, B. T., and Paliwal, K. K. "Effect of Speech Coders on Speech Recognition Performance", *Proc. ICSLP 96* Vol. 4: 2344-2347, 1996.

7. Mokbel, C., Mauuary, L., Jouvet, D., Monne, J., Sorin, C., Simonin, J., and Bartkova, K. "Towards Improving ASR Robustness for PSN & GSM Telephone Applications" *2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications(IVTTA1994)* Vol 1: 73-76, 1996.

8. Rabiner, L. and Juang, *B.-H. Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs 1993.

9. NIST, The Resource Management Corpus (RM1), Distributed by NIST, November 1989