DISTORTION-CLASS WEIGHTED ACOUSTIC MODELING FOR ROBUST SPEECH RECOGNITION UNDER GSM RPE-LTP CODING

Juan M. Huerta and Richard M. Stern

Carnegie Mellon University
Department of Electrical and Computer Engineering
Email: {juan, rms}@speech.cs.cmu.edu

ABSTRACT

We present a method to reduce the effect of full-rate GSM RPE-LTP coding by combining two sets of acoustic models during recognition, one set trained on GSM-distorted speech and the other trained on clean uncoded speech. During recognition, the *a posteriori* probabilities of an utterance are calculated as a sum of the posteriors of the individual models, weighted according to the distortion class each state in the model represents. We analyze the origin of spectral distortion to the quantized long-term residual introduced by the RPE-LTP codec and discuss how this distortion varies according to phonetic class. For the Research Management corpus, the proposed method reduces the degradation in frame-by-frame phonetic recognition accuracy introduced by GSM coding by more than 20 percent.

1. INTRODUCTION

Speech coding reduces the accuracy of speech recognition systems [4]. As speech recognition application in cellular and mobile environments become more ubiquitous, robust recognition in these conditions becomes crucial. Even though the speech codec is only one of the several factors that contribute to the degradation in recognition accuracy, it is necessary to understand and compensate for this degradation in order to achieve a system's full potential. We focus on the acoustic distortion introduced by the full-rate GSM codec [1]. This distortion can be traced to the quantization of the log-area ratio (LAR) and to the quantization and downsampling performed in the RPE-LTP process. The distortion introduced in the residual signal affects recognition to a larger extent than the quantization that the LAR coefficients undergo [2].

In Section 2 of this paper, we discuss the origin of the distortion in the RPE-LTP codec. We observe that based on the "predictability" of the short-term residual signal, the RPE-LTP will be able to minimize the error in the quantized long-term residual. This predictability is later shown to be related to phonetic characteristics of the signal. In Section 3 we show that the relative spectral distortion introduced in the quantized long-term residual tends to be concentrated around three levels, and that this amount of relative spectral distortion can be loosely related to the relative degradation introduced to the frame recognition accuracy by GSM coding. In Section 4 we separate the set of phonemes into clusters according to their relative spectral distortion. In Section 5 we propose a method to weight two sets of acoustic models based on the distortion categories introduced in Section 4. We describe the results of recognition experiments using this techniques in Section 6.

2. THE RPE-LTP CODEC AS A SOURCE OF ACOUSTIC DEGRADATION

The full-rate GSM codec decomposes the speech signal into a set of LAR coefficients and a short-term residual signal [1]. The LAR coefficients are quantized and transmitted to the decoder while the short-term residual is segmented into subframes and coded by an RPE-LTP coder [3]. In this section we discuss the sources of degradation that exist in the RPE-LTP coder. We refer to Figures 1 and 2, a simplified version of an ideal RPE-LTP codec and a simplified version of a real RPE-LTP codec respectively.

Figure 1 depicts an ideal codec able to reconstruct a signal that is identical to the original input sequence. Even though this codec would not achieve any reduction of the bit rate of the input sequence it helps illustrate the sources and nature of the distortion of the degradation introduced by the actual RPE-LTP codec.

The ideal RPE-LTP codec works as follows: a time sequence *x[n]* (the short-term residual) enters the codec and is compared to a predicted sequence (the reconstructed short-term residual) produced by the predictor block (the LTP section). The difference between these two signals is what the predictor block is unable to predict, which we will refer to as the "innovation sequence". In the absence of a source of acoustic degradation, our ideal codec will be able to generate a reconstructed time sequence by adding together the innovation sequence and the predicted sequence. This reconstructed sequence will be identical to the original time sequence *x[n]* by definition. In reality, the RPE-LTP coder, does not transmit the necessary information for the decoder to obtain the innovation sequence exactly. It sends subsampled and quantized information that will result in a reconstructed innovation sequence (called the reconstructed short-term residual) that will approximate the original innovation sequence to the extent possible. The amount of degradation in the reconstructed signal will be related to the energy of the innovation signal which in turn depends on how good the predictor module (RPE) is able to "follow" or predict the next subframe of the time sequence based on previous reconstructed subframes.

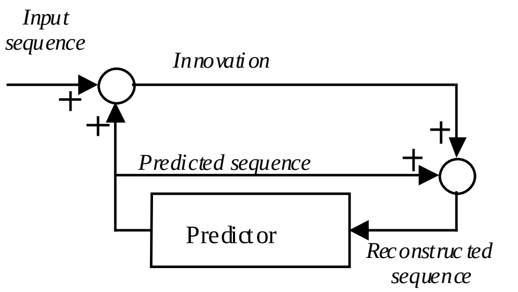

Figure 1. Simplified block diagram of an ideal RPE-LTP coder.

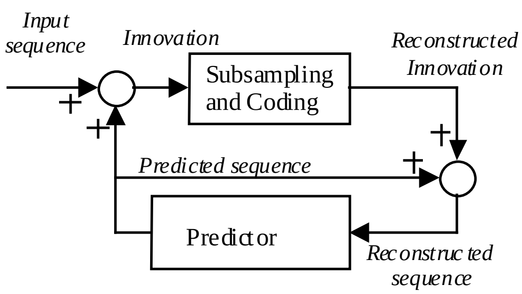

Figure 2. Simplified block diagram of an actual RPE-LTP coder.

3. THE RPE-LTP INDUCED SPECTRAL DISTORTION

3.1. Relative log spectral distortion introduced by the RPE-LTP coder

We use the relative log spectral distance below to measure the dissimilarity between the reconstructed and the original innovation sequence. $E(\omega)$ is the power spectrum of the innovation sequence and $E_R(\omega)$ is the power spectrum of the quantized innovation sequence. We integrate the absolute value of the difference between both log power spectra and normalize it by the integral of the log of the power spectrum of the original innovation signal.

$$L_1 = \int_0^{\pi} \left|\log(E(\omega)) - \log(E_R(\omega))\right| d\omega \Big/ \int_0^{\pi} \left|\log(E(\omega))\right| d\omega$$

3.2. Distribution of the relative log spectral distortion

We computed the relative log spectral distortion introduced by the RPE-LTP codec to the RM corpus. Figure 3 is a histogram that shows the log of the relative frequency of observing various levels of log spectral distortion. The horizontal axis is the amount of relative distortion observed in a frame. We observe that the counts are roughly clustered in 3 regions: low, medium and high distortion, separated by breakpoints of roughly 33 and 67 percent. We also observe that the majority of the frames suffer only a small amount of distortion. Thus, most of the time the LTP section of the codec is able to do a reasonably good job of predicting the short-term residual signal.

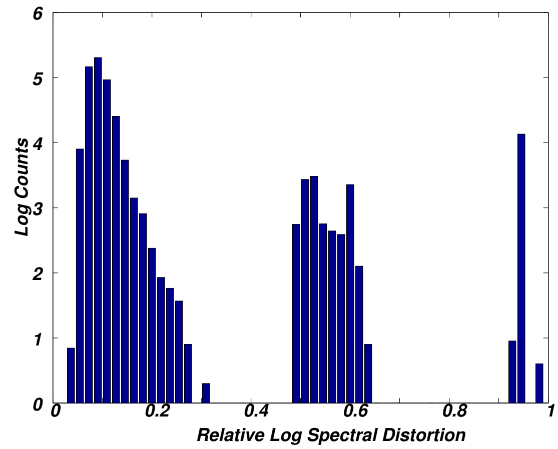

Figure 3. Histogram of the log distribution of relative spectral distortion introduced to a large sample of speech

3.3. Impact of relative log spectral distortion on phonetic recognition

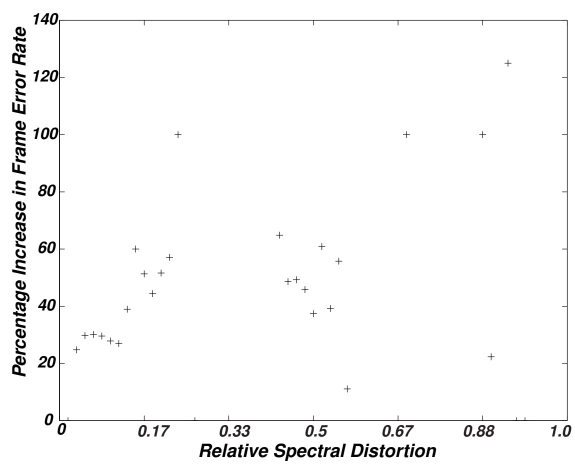

Figure 4. Relative degradation in frame-based phonetic accuracy as a function of the relative log spectral distortion introduced by the RTE-LTP coder.

In order to analyze the relation that exists between degradation in recognition performance and the amount of relative log spectral distortion introduced by the RPE-LTP block, we performed two phonetic recognition experiments: one training and testing using clean (*i.e.*, non-GSM coded) speech data and a second experiment using speech that underwent GSM coding. We computed the frame accuracy for each amount of relative spectral distortion both for when GSM was present and when it was absent. Figure 4 depicts how much the frame-based phonetic recognition error rate increased due to the introduction of the GSM coding as a function of the relative spectral distortion. We observe that, generally speaking, the phonemes that produced a greater amount of distortion due to GSM coding also suffered greater amounts of frame error rate.

4. RELATING RELATIVE LOG SPECTRAL DISTORSION PATTERNS TO PHONETIC CLASSES

4.1. Clustering Phonetic-classes using the relative log-spectral distortion

We grouped the 52 phonetic units used by our recognition system into phonetic clusters by incrementally clustering the closest histograms of the counts of the relative log spectral distortion for the frames associated with each phonetic unit in the corpus. The distance between each pair of histograms was calculated using normalized-area histograms. The clustering yielded five classes, as shown in the table below.

Class	Phonemes in class	Frame Acc.	Frame Acc., GSM	% Degra-dation
1	PD KD	46.76%	44.10%	5%
2	IX B BD DD DH M N JH NG V W Y DX ZH AX SH	68.28%	58.34%	31.34%
3	D G K P T TD CH	70.88%	65.94%	16.96%
4	F HH S Z TH TS	72.39%	57.63%	53.43%
5	IY OW OY UW L R AA AE AH AXR EH AO IH ER AW AY UH EY	74.37%	67.90%	25.24%

Table 1. Phonetic classes generated by automatically clustering the phonemes distortion histograms

Classes 1, 3, and 4 encompass the majority of the consonants, Class 4 being mostly fricatives. Class 2 includes the remaining consonants and a couple of vowels, while Class 5 encompass the vowels, diphthongs and semivowels. Silence units were omitted from this analysis. We can see that even though the classes do not exactly correspond to phonetic classes, they achieve a reasonable partition between broad classes of sounds. The pattern of distributions of relative log spectral distortion introduced by the RPE-LTP is strongly related to phonetic properties of the signal.

A plot of the normalized histograms of the five classes is shown below. These five classes exhibit a different pattern of distortion in each of the three distortion regions defined in Section 3.2. Classes 1 and 3 produce a large percentage of counts in the high distortion area and a presence in the middle area, while Classes 3 and 4 have more counts in the middle area. Class 5's counts are almost completely concentrated in the low distortion region.

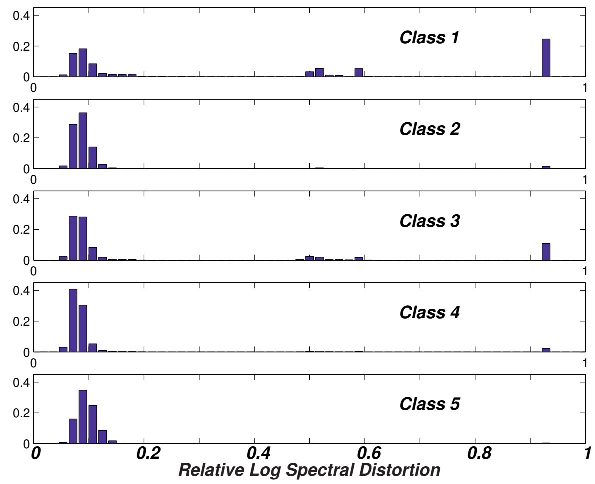

Figure 5. Histograms of the relative spectral distortion for the five phonetic classes of Table 1.

5. ACOUSTIC MODEL WEIGHTING

Acoustic modeling for HMM-based speech recognition commonly makes use of mixtures of Gaussian distribution representing a set of tied states. The posterior probability that an observed vector has been emitted by a certain state is thus expressed by

$$b_j(o_t) = \sum_{k=1}^{M_j} c[j,k] N(o_t, \mu_{jk}, C_{jk})$$

The term *c[j, k]* expresses the prior probability of the k^{th} Gaussian component of the j^{th} HMM. For a given state *j* the sum of *c[j,k]* over all *k* is equal to 1.

We can consider the amount of distortion that a frame or a phonetic class undergoes while evaluating the posterior probability using several models that reflect these distortion regions. We can express this concept by introducing a function *f* that weights the k^{th} posterior probability of the p^{th} model depending on d_t, the distortion of the observed frame *t*, and *j*, which indicates the model representing a given phonetic unit. The function *f* also depends on the prior probabilities of the model $c_p[j,k]$. This function can also be seen as a weighted version of $c_p[j,k]$. The resulting posterior becomes

$$b_{j,d}(o_t) = \sum_{p=1}^{2} \sum_{k=1}^{M} f(c_p[j,k], d_t, j)) N_p(o_t, \mu_{jk}, C_{jk})$$

The function f can also be dependent on the distortion class the model represents. This way, f will weight more the clean models for states that model phonemes that suffer small average GSM distortion. Alternatively, one can make the function f depend on knowledge of the instantaneous relative distortion of each frame d_t if this information is available. This function should give more weight to the distorted models when the relative frame distortion is greater.

6. SPEECH RECOGNITION EXPERIMENTS

Phonetic recognition experiments were performed using the Resource Management corpus and the SPHINX-3 system. Our basic acoustic models consisted of 2500 tied states, each modeled by a mixture of two Gaussians. One set of models was trained on clean speech while the other set of models was trained with speech that underwent GSM coding. We obtained our reference transcription by performing automatic forced alignment on the 1600 test utterances using their pronunciations. We computed both the phone accuracy and the frame accuracy. Phone accuracy is calculated considering insertions and deletions of phones in the hypothesis. We evaluated the frame accuracy by making a frame-by-frame comparison of the output of the phonetic recognizer with the alignment. Table 2 summarizes recognition accuracy for clean and GSM speech, using clean and GSM models. We also show the results when models are trained using both GSM-coded speech and clean speech (multi-style trained models).

Testing data	Training data	Phone Accuracy	Frame Accuracy
Clean	Clean	63.3%	69.86%
GSM	Clean	55.7%	61.20%
GSM	GSM	60.1%	64.50%
GSM	GSM+Clean Multistyle	60.1%	64.64%

Table 2. Baseline frame accuracy and phoneme recognition accuracy in Resource Management

From Table 2 we can see that the effect of having GSM codec distortion during training and recognition reduces frame accuracy by 5.35 points. Recognition accuracy can improved somewhat by using multi-style trained models.

6.1. Experiments using weighted acoustic models

We performed recognition experiments using weighted models by considering different mixing factors for two acoustic models, trained GSM-coded and clean speech. Table 3 shows results weighting both models equally. We also made these weights dependent of the phonetic class, using three automatically-clustered phonetic classes. While there is a modest advantage of weighting the models differently, either of these two weighting methods provide better results than using multi-style trained models. The 3-class optimized weights reduce the degradation in frame error rate introduced by GSM coding by 27%, from 5.36% to 3.87%. The unequal frame weighting improves recognition accuracy approximately equally for all three phonetic classes.

Weighted modeling	Phone Accuracy	Frame Accuracy
Equal weights	61.2%	65.77%
3-class optimized weights	61.2%	65.99%

Table 3. Frame accuracy obtained when using equally weighted models and when using phonetic-class based weighting.

7. CONCLUSIONS

In this paper we examined the distortion introduced by the RPE-LTP GSM codec to the short-term residual signal. We related this distortion to the relative spectral distortion produced by the quantized long-term residual. We observed that this distortion influences recognition performance, and that it can be related to the phonetic properties of the speech signal. We introduced a way to mix two acoustic models that had been trained on clean and distorted speech by taking into consideration the amount of distortion suffered by the phonetic class.

8. REFERENCES

[1] European Telecommunication Standards Institute, “European digital telecommunications system (Phase 2); Full rate speech processing functions (GSM 06.01)”, ETSI 1994.

[2] Huerta, J. M. and Stern R. M., “Speech Recognition from GSM Codec Parameters” Proc. ICSLP-98, 1998.

[3] Kroon, P., Deprettere, E. F., Sluyter, R. F. “Regular-Pulse Excitation - A Novel Approach to Effective and Efficient Multi-pulse Coding of Speech”, IEEE Trans. on Acoustics, Speech and Signal Processing, **34**:1054-1063, October 1986.

[4] Lilly, B. T., and Paliwal, K. K. “Effect of Speech Coders on Speech Recognition Performance”, Proc. ICSLP-96, 1996.