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 cluster their relative spectral distortion. In Sec a method to weight two sets of acoustic mo the distortion categories introduced in S describe the results of recognition experi techniques in Section 6.

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

The full-rate GSM codec decomposes the sp into a set of LAR coefficients and a shor signal [1]. The LAR coefficients are qu transmitted to the decoder while the short- segmented into subframes and coded by an coder [3]. In this section we discuss degradation that exist in the RPE-LTP code Figures 1 and 2, a simplified version of LTP codec and a simplified version of a r codec respectively.

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

The ideal RPE-LTP codec works as follow sequence x[n] (the short-term residual) enter and is compared to a predicted sequ reconstructed short-term residual) prod predictor block (the LTP section). The between these two signals is what the pre unable to predict, which we will refe "innovation sequence". In the absence of acoustic degradation, our ideal codec wi generate a reconstructed time sequence together the innovation sequence and th sequence. This reconstructed sequence wil to the original time sequence [1]. by definition. Reality, the RPE-LTP coder, does not t necessary information for the decoder t innovation sequence exactly. Instead quantized information that will result in innovation sequence (called the reconstruc residual) that will approximate the origi sequence to the extent possible. The degradation in the reconstructed signal wi the energy of the innovation signal which on how good the predictor module (RPE)

"follow" or predict the next subframe of the sequence based on previous reconstructed subframes.

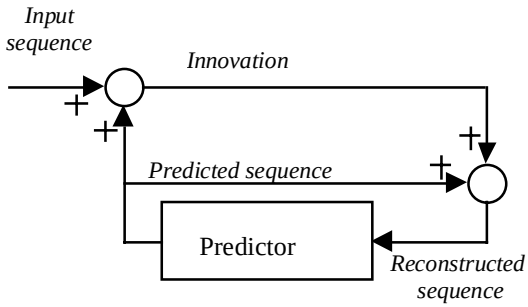

Figure 1. Simplified block diagram of an ideal RPE

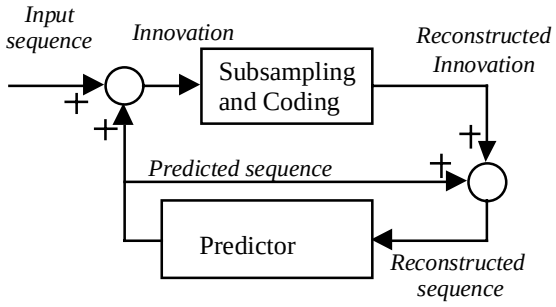

Figure 2. Simplified block diagram of an actual RPE-LTP coder.

3.THE RPE-LTP INDUCED SPECTRAL DISTORTION

3.1 Relative log spectral distortion introduc by the RPE-LTP coder

We use the relative log spectral distanc measure the dissimilarity between the recon 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 We integrate the absolute value of the diffe both log power spectra and normalize it b of the log of the power spectrum of th innovation signal.

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

1.2 Distribution of the relative log spectral distortion

We computed the relative log spectral introduced by the RPE-LTP codec to the RM 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 observ frames are roughly clustered in 3 regions: and high distortion, separated by breakpoint 33 and 67 percent. We also observe that the the frames suffer only a small amount of dist most of the time the LTP section of the code do a reasonably good job of predicting the residual signal.

.

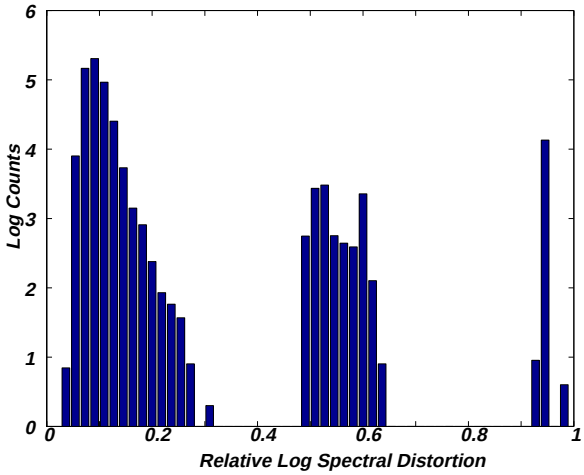

Figure 3 Histogram of the log distribution o spectral distortion introduced to a large samp

1.3 Impact of relative log spectral distor on phonetic recognition

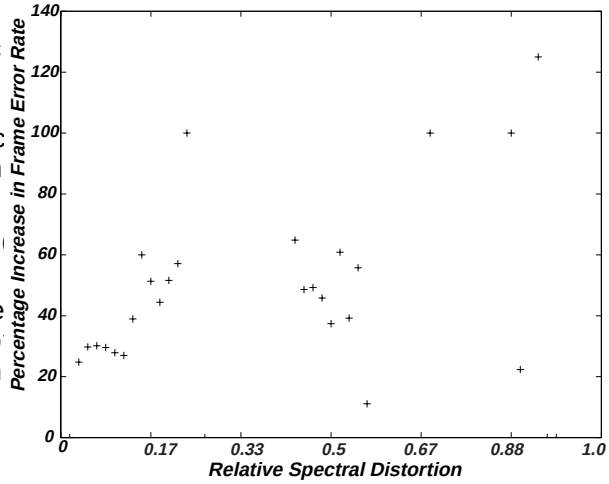

Figure 4. Relative degradation in frame-ba phonetic accuracy as a function of the re spectral distortion introduced by the RTE- coder.

distortion

In order to analyze the relation that exi degradation in recognition performance and t of relative log spectral distortion introduce LTP block, we performed two phonetic recc experiments: one training and testing using

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 system into phonetic clusters by incrementally the closest histograms of the counts of th spectral distortion for the frames associa phonetic unit in the corpus. The distance pair of histograms was calculated using nor histograms. The clustering yielded five classes, as summarized in the table below.

Class	Phonemes in cl	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 computed below. These five classes exhibit distinct spectral distortion in each of the three defined in Section 3.2. Classes 1 and 3 presented most of counts in the high distort region, while Class 2 had most of the counts in the middle area. Class 5 was completely concentrated in the lo region.

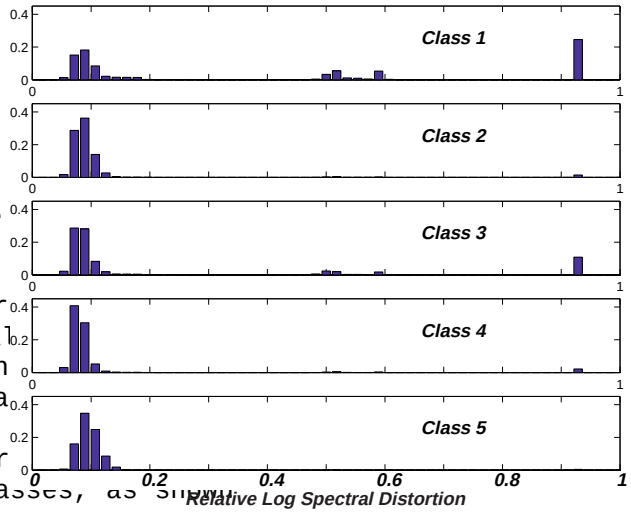

Figure 5. Histograms of the relative spectral dist five phonetic classes of Table 1.

5. ACOUSTIC MODEL WEIGHTING

Acoustic modeling for HMM-based speech re commonly makes use of mixtures of G distribution representing a set of tied sta probability that an observed vector has be 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 probabi k^{th} Gaussian component of the j^{th} HMM. For a given j st the sum of $c[j,k]$ is equal to 1.

We can consider the amount of distortion t a phonetic class undergoes while evaluatin probability using several models that distortion regions. We can express this introducing a function f that weights th probability of the j^{th} model depending on the distortion of the observed frame, which indicat the model representing a given phonetic function also depends on the prior probabi model $c[j,k]$. This function can also be weighted version $[j,k]c_p$ The resulting po 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 d_t of each frame. This way, 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 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 da	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 pho using three automatically-clustered phonet While there is a modest advantage of weig models differently, either of these two weigh provide better results than using multi-s models. The 3-class optimized weights re degradation in frame error rate introduce coding by 27%, from 5.36% to 3.87%. The frame weighting improves recognition ac approximately equally for all three phonetic c

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 models and when using phonetic-class based weightin

7. CONCLUSIONS

In this paper we examined the distortion in the RPE-LTP GSM codec to the short-term r signal. We related this distortion to the re distortion produced by the quantized long-te We observed that this distortion influences performance and that it can be related to properties of the speech signal. We introduc mix two acoustic models that had been traine and distorted speech by taking into consid amount of distortion suffered by the phonetic

8. REFERENCES

[1] European Telecommunication Standards Institute pean digital telecommunications system (Phase rate speech processing functions (GSM 06.01 1994.

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

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

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