ELSEVIER

Speech Communication 34 (2001) 213–225

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Distortion-class modeling for robust speech recognition under GSM RPE-LTP coding

Juan M. Huerta *, Richard M. Stern

Department of Electrical Computer Engineering and School of Computer Science, Carnegie Mellon Univeristy, Pittsbugh, PA 15213, USA

Abstract

We present a method to reduce the degradation in recognition accuracy introduced by full-rate GSM RPE-LTP coding by combining sets of acoustic models trained under different distortion conditions. During recognition, the a posteriori probabilities of an utterance are calculated as a weighted sum of the posteriors corresponding to the individual models. The phonemes used by the system's word pronunciations are grouped into classes according to amount of distortion they undergo in coding. The acoustic model used in the decoding process is a weighted combination of models derived from clean speech and models derived from speech that had been degraded by GSM coding (the source models), with the relative combination of the two sources depending on the extent to which each class of phonemes is degraded by the coding process. To determine the distortion class membership, and hence the weights, we measure the spectral distortion introduced to the quantized long-term residual by the RPE-LTP codec. We discuss how this distortion varies according to phonetic class. The method described reduces the degradation in recognition accuracy introduced by GSM coding of sentences in the TIMIT database by more than 70% relative to the baseline accuracy obtained in matched training and testing conditions with respect to a system using the source acoustic models, and up to 60% relative to the best baseline systems regardless of the number of Gaussians.

1. Introduction

Recent progress in automatic speech recognition technology along with the increase in use of wireless telephony have produced an increased level of interest in voice-activated applications using wireless communication channels. While functionally similar to that of their public-switched telephone network (PSTN) counterparts, speech recognition systems and applications that include a wireless link with coding such as GSM confront the end user with the additional potential problems of wireless-channel noise, data dropouts, and signal degradation originated by the speech codec. Mobile applications provide the additional problem of potentially more varied and more intense environmental noise (Mokbel et al., 1996; Haeb-Umbach, 1997). These phenomena result in error rates for speech recognition systems which are substantially higher than their counterparts in non-mobile environment (e.g., Das et al., 1999; Delphin-Poulat and Mokbel, 1997; Elvira and Torrecilla, 1998; Mokbel et al., 1996).

Even though the degradation in the signal due to the speech codec amounts only to a fraction of the overall degradation in wireless channels compared to the effects of additive noise, the error rate does depend on the coding scheme used and

* Corresponding author. Tel.: +1-412-2687109.
E-mail address: juan@speech.cs.cmu.edu (J.M. Huerta).

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00055-8

the codec bit rate, with accuracy decreasing as bit rate decreases (cfr., Euler and Zinke, 1994; Lilly and Paliwal, 1996; Digalakis et al., 1998). Reduction of the sensitivity of a recognition system to the overall presence of speech coding in a communication link, will result in more robust systems whose performance is more independent of the type and bit-rate of the codec that happens to be used in a particular communications channel.

It has been observed that under mismatched conditions (such as when the data used to test a system are coded but the data used to train it are not), a typical increase in the error rate of a system due to coding can be as high as 20% (Euler and Zinke, 1994; Lilly and Paliwal, 1996; Huerta and Stern, 1998). As the demand for mobile telephony increases one should expect further reductions in bit rates, so the importance of systems which are codec bit-rate insensitive will only become more evident.

Research into the problem of robustness in mobile applications has addressed some specific issues such as codec tandeming (Salonidis and Digalakis, 1998), and hole detection and rejection (Paping and Fahnle, 1997; Fissore et al., 1999; Karray et al., 1998). Other groups (Gallardo-Antolin et al., 1998; Huerta and Stern, 1998; Gallardo-Antolin et al., 1999) have focussed on the assumption of the availability of the codec parameters during recognition. Gupta et al. (1996) have focused on robustness to environmental noise encountered in mobile applications. Also, there is work regarding the application and tailoring of robustness techniques to the problem, such as spectral subtraction, adaptive filtering and model adaptation (Mokbel et al., 1996), robust front ends (Dufour et al., 1996), bias removal and equalization (Delphin-Poulat and Mokbel 1997). Some work has focused on the use of robust acoustic modeling (Puel and Obrecht 1997) and model adaptation (Soulas et al., 1997).

While all these approaches can potentially reduce the absolute word error rate of a speech recognition application, none of them attempt to identify the source and nature of the distortion from the operating model of the codec with this same purpose.

In this work we specifically analyze the effect of a full-rate GSM codec on the spectrum produced by the signal and on recognition accuracy. Based on this analysis we propose a method to alleviate the degradation in recognition accuracy introduced by the GSM distortion. GSM coding begins with an LPC analysis which produces an all-pole representation of the spectrum and a residual signal that represents the excitation. GSM coding represents the spectrum in the form of log area ratio (LAR) coefficients. The residual signal is processed by the RPE-LTP codec. The distortion of the speech signal introduced by the GSM codec can be traced to the quantization of the lot-area ratio coefficients LAR and to the quantization and downsampling of the residual signal performed in the RPE-LTP process. The distortion of the residual signal affects recognition to a greater extent than the quantization undergone by the LAR coefficients. Huerta and Stern (1998) presented an analysis of the impact of the coding and quantization of these codec coefficients on recognition, together with a method to combine selectively the information contained in these parameters to minimize this performance degradation. In the present work we make no assumptions about the availability of the codec parameters, focussing instead on the development of better acoustic models of the reconstructed speech signal that minimize the degradation recognition accuracy produced by GSM coding.

In Section 2 of this paper, we discuss the origin and nature 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 general 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 two regions, and that this amount of relative spectral distortion can be loosely associated with the relative degradation in phone recognition accuracy introduced by GSM coding. In Section 4 we sort the set of phonemes into clusters according to their relative log spectral distortion distributions. In Section 5 we propose a method to

weigh two sets of acoustic models based on the distortion categories introduced in Section 4 with the intention of allowing phones that undergo a moderate amount code distortion to be modelled predominantly by clean models, and of allowing highly distorted phones to be modeled by "noisy" models. We describe the results of recognition experiments using these techniques in Section 6.

2. The RPE-LTP codec as a source of acoustic degradation

The full-rate GSM codec is a linear predictive RPE-LTP based codec with a bit rate of 13 kbps (ETSI, 1994). The 8-kHz speech signals enter the codec where they are analyzed in frames of 160 samples, and the 8th-order LPC parameters are obtained. The LPC parameters are represented as LAR coefficients, and they are quantized and transmitted. The residual signal from the LPC analysis (the short-term residual) is subdivided into subframes of 40 samples each and coded by a regular pulse excited-long-term prediction coder whose quantized parameters are also transmitted. In this section we briefly describe the RPE-LTP coding process (Kroon et al., 1986; Vary et al., 1988) of the short-term residual signal in the GSM full-rate codec. For the work of this paper we refer to and use the publicly-available implementation of GSM in C by Degener and Bormann (1992).

The RPE-LTP codec can be described in simplified form as a two-part process: a Long Term Predictor (the LTP block) process that produces an estimate of the Short-term residual signal, and the regular pulse excitation block (RPE block) which is responsible for representing the "unpredicted" part of the short-term residual signal (called the long-term residual signal) using a reduced number of bits. Under normal conditions, the LTP block will try to capture the long-term periodicity of the signal associated principally with voiced speech segments based on a cross correlation analysis. We now explain these concepts in more detail.

For the purpose of illustration we present two diagrams representing simplified versions of the RPE-LTP codec that process the short-term residual signal that comes out of the LPC analysis. The two diagrams we present correspond to two versions of the RPE-LTP codec: an ideal codec and a real codec. By comparing and contrasting these simplified codecs we can identify the source and nature of the distortion introduced to the reconstructed version of the residual signal. In the following sections we will relate the behavior of the signal distortion to degradation in recognition accuracy.

Fig. 1 is a simplified block diagram of an ideal RPE-LTP codec. The primary difference between the ideal codec and a real RPE-LTP codec is that the ideal codec does not produce quantized versions of its signals or parameters. For this reason, the ideal codec does not achieve any reduction in bit rate. The short-term residual signal $e[n]$ enters the ideal codec and is compared to the short-term residual estimate $\bar{e}[n]$ produced by the LTP block. The difference between these two signals corresponds to the part of the residual signal which the LTP block was unable to predict. This signal is called the long-term residual signal $r[n]$, and it represents what needs to be added to the short-term residual estimate to obtain the reconstructed short-term residual signal. In other words, this signal represents a sort of "innovation" or unpredictable part of the short-term residual signal. The decoder section of the codec contains an identical LTP block which generates a short-term residual estimate, based on the received LTP parameters and the previously reconstructed version of the short-term residual. After the short-term residual estimate is generated, the ideal codec adds the received innovation part of the signal (i.e., the

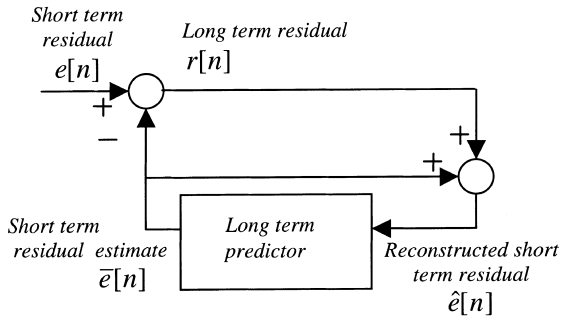

Fig. 1. A simplified block diagram of an ideal REP-LTP codec.

long-term residual) to it. Although the sum of the long-term residual and the short-term residual estimate signal results in exactly the residual sequence, the ideal codec produces no loss or distortion in the restored signal. However, the ideal codec must transmit an exact copy of the long-term residual signal to achieve this, so its bit rate is no less than the bit rate of the original short-term residual sequence.

In reality, the RPE-LTP coder transmits a subsampled and quantized approximation of the long-term residual sequence and the LTP information in order to achieve bit-rate reduction. Generally, the coder does not provide all the information that is needed to obtain a perfect reconstruction. The reconstructed representation of the long-term residual obtained from the transmitted information (called the quantized long-term residual $\hat{r}[n]$) is only an approximation to the original innovation sequence. Fig. 2 illustrates this process by adding to the codec the block labeled *RPE coding*. The amount of degradation in the reconstructed signal will be related to the energy of the original LTR signal which in turn depends on how well the LTP module in the coder is able to "follow" or predict the next subframe of the time sequence based on previous reconstructed subframes.

The RPE codec introduces distortion to the quantized long-term residual that is proportional to the energy present in it. From the analysis of the operation of the RPE-LTP codec above, we suggest that the energy of the long-term residual can be associated with the predictability of the short-term residual. Although the different phones of any given language can be associated with a certain

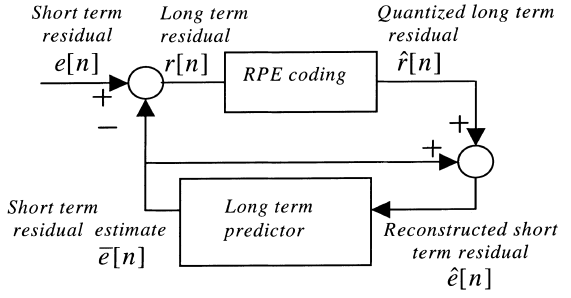

Fig. 2. A simplified block diagram of an ideal REP-LTP codec.

level of periodicity, or predictability (for example, vowels are likely to be more predictable than consonants), we can expect to find certain patterns in the distribution of the amount of distortion introduced by the RPE-LTP coding process. We will illustrate this point in the following sections.

Other existing coding schemes in which the error minimization block consists of a predictive component (i.e., closed loop prediction-based coders) (Kroon and Kleijn, 1995; Kleijn and Paliwal, 1995) can be thought to operate in a similar fashion as the basic system of Fig. 2, with the main differences between codecs being the way the long-term prediction is performed and how the long-term residual gets represented and the effects of this quantized representation in the reconstructed long-term residual.

3. RPE-LTP-induced spectral distortion

In order to establish the relation between phonetic identity and amount of distortion introduced by coding we must specify a metric that will reflect the degradation between the long-term residual and its quantized approximation. In this section we present such a metric and describe the distributions we obtained when applying it to the TIMIT database (LDC, 1993).

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

We use the relative log spectral distortion (RLSD) distance to measure the dissimilarity between the reconstructed and the original innovation sequence (i.e., between the long-term residual and the quantized long-term residual) at each frequency ω. Let $S(\omega)$ represent the power spectrum of an innovation sequence subframe and let $S_R(\omega)$ represent the power spectrum of the corresponding quantized innovation sequence subframe. The RLSD is then defined to be

$$\text{RLSD} = \frac{1}{\pi}\int_0^{\pi}\left|\frac{\log(S(\omega)) - \log(S_R(\omega))}{\log(S(\omega))}\right| d\omega. \qquad (1)$$

As can be seen, this metric reflects the ratio of the differences between the distortion introduced to the log-power spectra of the long-term residual, normalized for each frequency by the magnitude of the log power spectra. When no distortion is introduced, both power spectra are equal and the RLSD is equal to zero. When a relatively large amount of distortion is present, the RLSD becomes large. The RLSD can be thought of as a type of average inverse SNR.

3.2. Distribution of the relative log spectral distortion

We computed the relative log spectral distortion introduced by the RPE-LTP codec on a subset of the training utterances of the TIMIT corpus that were lowpass filtered with a cutoff of 3.2 kHz and downsampled by a factor of 2 to an 8-kHz sampling rate. We modified the GSM RPE-LTP codec to produce output files containing the samples corresponding to the long-term residual and the quantized long-term residual. We then computed the log power spectrum for the two types of subframes and computed the relative log spectral distortion as in Eq. (1). Fig. 3 is a histogram that shows the logarithm of the frequency of the values of relative log spectral distortion. The horizontal axis represents the amount of relative distortion observed per codec subframe (i.e., 40 samples). We can observe that the RLSD ranges from 0 to 3000%. The log-counts are roughly clustered in two regions separated at approximately the value of 1000%. It should be noted that many of the frames with the greatest RLSD are silence or similar frames, for which $S(\omega)$ can be relatively small in magnitude compared to the amount of distortion introduced. The majority of the frames suffer only a relatively small amount of distortion, so 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.

Even though a large portion of the frames incur only a moderate amount of RLSD (i.e., below 1000%), the histogram shown in Fig. 3 has a relatively low level of resolution for the region where most of the frames occur. Fig. 4 is a similar histogram of RLSD, but with a logarithmic abscissa. The bimodal pattern observed in Fig. 3 is preserved, with the two modes centered around common logarithmic values of approximately 1 and 3, respectively. It is also clear from this figure that the majority of the frames incur only a moderate amount of degradation.

3.3. Impact of relative log spectral distortion on phonetic recognition

In order to analyze the relation that exists between the degradation in recognition accuracy

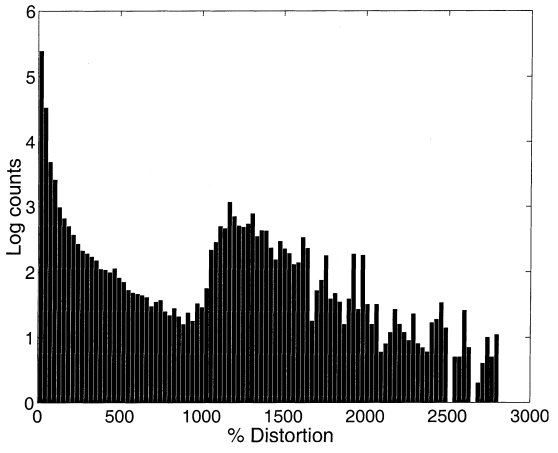

Fig. 3. Log histogram of the RLSD observed in a portion of the training part of the TIMIT corpus.

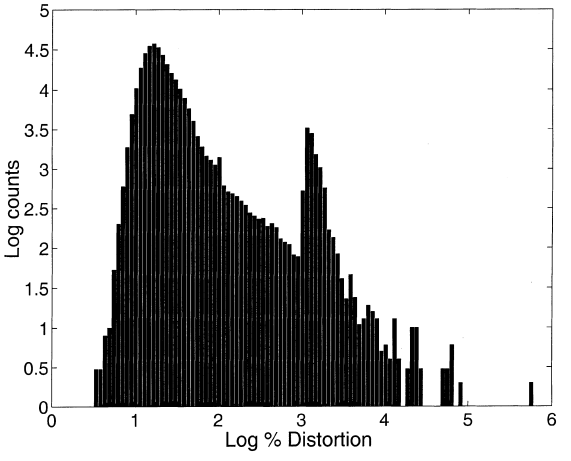

Fig. 4. Log histogram of the RLSD observed in a portion of the training part of the TIMIT corpus as in Fig. 3.

and the amount of RLSD introduced by the RPE-LTP block, we performed two phonetic recognition experiments: testing using clean (i.e., non-GSM coded) speech data and testing using speech that underwent GSM coding. Details of the system's configuration are in Section 6. In both cases the acoustic models were trained using clean speech. We computed the phonetic error rate of both experiments when GSM coding is present and when it is not. From this information we were able to compute the percentage of increase in error rate for each phone due to GSM coding.

We also computed the average value of the RLSD associated with each of the recognizer's feature frames (i.e., 100 ms of speech) of every phone. This was performed based on the phone segmentations of TIMIT. In Fig. 5 we present a scatter plot in which the horizontal axis represents the average of the log of the RLSD, while the vertical axis describes the relative increase in WER due to GSM coding. We can see that phones that incur an average log RLSD values of about 2.6 or below have a degradation in error rate of 20% or less. These phones are mostly vowels (**ae, eh, ah, aw, aa, ay, uh,** etc). There are phones with log RLSD values between 2.6 and 2.8 whose relative degradation goes above 20%. Finally, we can see that the consonants **f, z, v** and **dh** fall in the region above 2.8 and suffer a degradation of over 40% in error rate. Informally, from Fig. 5 we can observe a modest relation or trend between the mean RLSD observed by a phone and its increase in phonetic error rate due to GSM coding, as well as between phonetic classes and mean observed RLSD. The most notable exception to this trend is the cluster of nasals (**em, en, n, m** and **ng**). This group shows a relatively high average RLSD but little degradation in recognition due to GSM coding. The average RLSD is clearly only one of several different sources of deterioration in accuracy due to GSM coding. (Other sources include the quantization of the LAR coefficients, the quantization of the LTP coding, etc.)

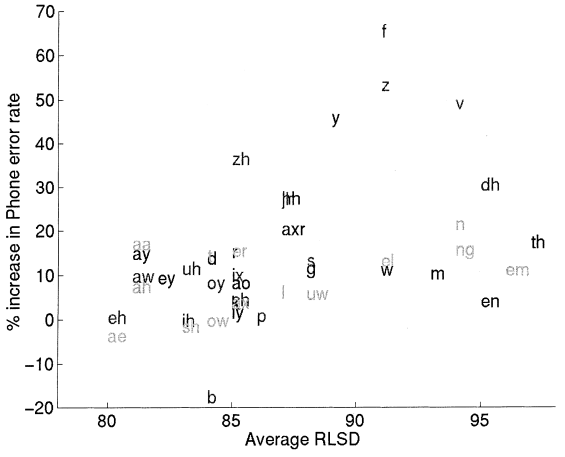

Fig. 5. Scatter plot of the phonetic units of the TIMIT corpus, according to their average RLSD and their relative increase in eror rate due to GSM coding.

4. Relating RLSD patterns to phonetic classes

In Section 3 we described the RLSD metric and suggested that there appears to be a relation between the mean RLSD observed at the subframes corresponding to a phone and the phonetic identity (or properties) of the phones. In this section we will extend this analysis of phonetic accuracy based on histograms of RLSD.

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

We constructed histograms of the log counts of the logarithm of the values of the relative log spectral distortion for each of the 61 phonetic units of the TIMIT database (LDC, 1993). Having obtained such histograms we normalized their areas in order to account for the differences in frequency of occurrence of the phonetic units. We grouped these units into phonetic clusters by incrementally clustering the closest normalized histograms, using as a distance the sum of the square difference between the values of each bin. The clustering process was started by calling each phonetic element a class of its own, and finding the closest two histograms. After finding them, the normalized histograms are added and renormalized, which is equivalent to computing the geometric mean of the bins of both histograms because we are working with the logarithms of the

Table 1
Phonetic classes generated by automatically clustering phone distortion histograms and the corresponding class phone error rates

Class	Class members	Class phone error rate (no GSM)	Class phone error rate (GSM)	Percent degradation
1	**hh jh dx b d g k ch p t**	32.4%	36.7%	13.27%
2	**ng m n em en v eng**	29.3%	42.1%	43.7%
3	**h# bcl dcl gcl kcl pcl tcl pau**	NA	NA	NA
4	**hv aa ae ah eh aw ay ey**	41.9%	44.7%	6.68%
5	**epi**	NA	NA	NA
6	**ix iy ow oy ux zh nx ao ih r er ax sh uh**	35.9%	38.5%	7.24%
7	**dh th q**	47.3%	64.3%	35.94%
8	**axr uw f l ax-h el s w y z**	30.9	39.5	27.83%

counts. Table 1 shows the clusters obtained when the process was stopped at 8 classes. It also shows the average phonetic error rate per class with and without GSM coding and the corresponding relative degradation in error rate. In describing the clustering we use the categorization and description of the phones included in the TIMIT documentation (LDC, 1993).

4.2. Phonetic properties of distortion pattern-derived classes

From Table 1 we can see that Class 5 corresponds to the segments labeled as epenthetic silence, described in the TIMIT documentation as generally found between a fricative and a semivowel or nasal. As this symbol does not appear in the phonetic dictionary used for recognition, no phone error rate is associated to it. Similarly, Class 3 grouped all the closures for the stops **b, d, g, k, p** and **t**, as well as the begin and end markers and the pauses. Since the pronunciations in the dictionary do not explicitly have the closures indicated, no phone error rate is associated with them either.

Class 1 includes all the stops except **q,** both the affricates **jh** and **ch,** and the semivowel **hh.** Class 2 encompass all the nasals except the nasal flap **nx,** but includes the fricative **v**. Classes 4 and 6 split the vowels, Class 4 including the voiced **h: hv,** and Class 6 the fricatives **sh** and **zh,** the nasal **nx,** and the glide **r**. Class 7 includes the fricatives **dh** and **th,** as well as the stop **q**. Class 8 is the most heterogeneous, and includes fricatives, semivowels and vowels.

Class 7 has the highest absolute class phone error rate without GSM coding, and classes 1, 2, 6 and 8 have the lowest. Classes 2 and 7 are the classes that suffer the greatest amount of relative degradation when GSM coding is introduced, and Classes 1, 4 and 6 are the most robust to GSM coding. We can see that the use of the distribution of the RLSD introduced by the RPE-LTP in the form of normalized histograms for the purpose of clustering the phones produces classes with some phonetic homogeneity. We also note that classes dominated by vowels suffer the least from GSM coding, while groups dominated by nasals, fricatives, and some other consonants suffer substantially larger relative degradation in their class phonetic error rates due to GSM coding.

Fig. 6(a,b) shows the normalized histograms for Classes 3 and 4. We can see that for Class 3 the number of counts in the region of high distortion (i.e., the rightmost mode of the histogram) has a substantial number of counts, while Class 4 does not show such mode. Another visible difference is the location of the leftmost mode. We can see that the mode for Class 4 is closer to the vertical axis, meaning that the subframes of the phone realizations of this class suffer in average a considerably smaller amount of relative log spectral distortion. This confirms our observation that vowels are less sensitive to the effects of coding than (in this case) stop closures, or more generally consonants and silences.

In Section 6 we will be performing recognition experiments based on phonetic categories that have been derived in the way described above. We

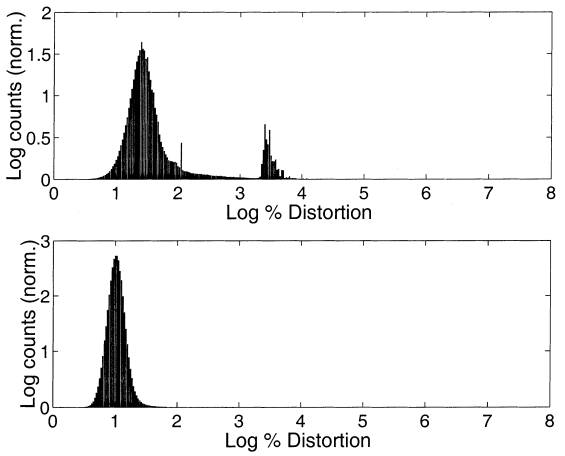

Fig. 6. Normalized log histogram of the log of the RLSD for Class 3 (plosive closures) and Class 4 (vowels), classified as in Table 1.

Table 2
Phonetic classes generated by automatically clustering phone distortion histograms, 15 classes

Class	Class members
1	**b d**
2	**ng m n en v**
3	**bcl dcl gcl pau**
4	**eng**
5	**hv aa ae ah eh aw ay ey**
6	**hh jh dx g k ch p t**
7	**epi**
8	**ix iy ow oy ux zh nx ao ih r er ax sh uh**
9	**f el w z**
10	**dh th**
11	**ax-h**
12	**h# kcl pcl tcl**
13	**em**
14	**axr uw I s y**
15	**q**

will be doing experiments based on 15 phonetic classes. This number of classes result in a small number of parameters to estimate, yet allows the phonetic clusters to be distinct. In Table 2, we present the results of following the clustering procedure described above, but stopping at 15 phonetic classes. We can observe from this figure that the difference between the 8 and 15 classes are principally due to some phones separating into their own classes.

5. Weighted acoustic modeling of phonetic categories

We saw in Section 4 that not all the phones in an utterance undergo the same amount of RLSD. Due to this, we can expect to obtain better recognition accuracy for those phones that undergo a small amount of degradation in coding by using models trained from speech that had not undergone GSM coding (which we refer to as "clean" speech). Similarly, for those phones for GSM coding produces a larger average distortion, one can expect greater accuracy if we employ models that reflect a higher amount of distortion during decoding. Possible ways to achieve this type of modeling are to use two acoustic models during decoding: one derived from clean speech and the other from GSM speech. In this section we will describe different strategies of combining both models during decoding.

It has been shown by other authors (e.g., Ming et al., 1999; Beyerlein 1998) that a recognition system's performance can be improved by combining several different acoustic models during recognition. In this kind of approach, three issues are of relevance: the nature of the individual models, the way in which the models/scores are obtained, and the way in which the weighting factors are determined. As mentioned above, our aim is to combine two sets of models: one derived from undistorted data and one from distorted data. In the rest of this section we describe the method used to combine such models and the way we obtain the weighting coefficients.

5.1. Combining acoustic models by means of mixture weighting

One way to consider several models when decoding an utterance is by combining the posterior probabilities obtained from this model into a log linear posterior probability distribution, (e.g., Beyerlein 1998). Eq. (2) describes the probability of a string w given the observation O is expressed in term of the posterior log probabilities of the N different models. The term $C(\Lambda)$ is a normalization factor. These models are weighted by the terms λ_j

and incorporated into a log linear score expression,

$$p(w|O) = \exp\left\{ \log C(\Lambda) + \sum_{j=1}^{N} \lambda_j \log\ p_j(w|O) \right\}. \tag{2}$$

A different approach to achieving this model combination is by merging the distributions of both model distributions into a single set of model distributions. 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

$$b_j(o_t) = \sum_{k=1}^{M_j} c[j,k] N(o_t, \mu_{jk}, C_{jk}). \tag{3}$$

The term $c[j,k]$ in Eq. (3) expresses the prior probability of the kth Gaussian component of the jth HMM. In a tied state-based system, this expression represents the jth tied state. For a given state j the sum of $c[j,k]$ over all k is equal to 1. The parameters of the Gaussian distribution $N(\ )$ are μ_{jk} and C_{jk}, respectively.

We 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 express this concept by introducing a function f that weights the kth posterior probability of the pth 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 considered to be a weighted version of $c_p[j,k]$. The resulting expression for the posterior probability becomes

$$b_{j,d}(o_t) = \sum_{p=1}^{N} \sum_{k=1}^{M} f(c_p[j,k], d_t, j) N(o_t, \mu_{pjk}, C_{pjk}). \tag{4}$$

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.

For the case of two sets of models (i.e., clean and GSM models), by making the weighting function dependent only on j (i.e., the state in the model), Eq. (4) becomes

$$b_j(o_t) = \lambda_j \sum_{k=1}^{M} c_1[j,k] \mathrm{N}(o_t, \mu_{1,j,k}, C_{1,j,k}) + (1-\lambda_j) \sum_{k=1}^{M} c_2[j,k] N(o_t, \mu_{2,j,k}, C_{2,j,k}). \tag{5}$$

The function $f(c_p[j,k], \mathrm{d}_t, j)$ was separated into the mixing weights terms $c_1[j,k]$ and $c_2[j,k]$, and the state weighting factors λ_j which are factored out of the sum. In the next subsection we describe how we found these weights.

5.2. Estimating the weighting factors from RLSD histograms

For a given certain set of phonetic classes, we want to associate a set of weights such that each class i will combine the two sets of models with weights λ_j and $1-\lambda_j$, respectively. These weights should be made proportional to the amount of distortion observed per category.

We determined the values of the weights from the normalized log histograms of the RLSD. We first clustered the phones into phonetic categories using the method described in Section 4.1. We then obtained the normalized log-histograms of each class, and from these histograms we computed for each class the value of the bin for which 50% of the counts had been accumulated (i.e., a value close to the median of the distribution of the RLSD), and then divided this value by the value of the highest bin (this bounds the value of lambda between 0 and 1, which is desirable). A value close to zero indicates that 50% of the counts are close to the low distortion area and the associated weight λ_j will be small and $1-\lambda_j$ will be high (i.e., close to

one), indicating that $1-\lambda_j$ is the weight that should be associated to the clean models.

6. Speech recognition experiments

6.1. Recognition system setup and baseline experiments

Recognition experiments were performed using the TIMIT corpus and the Carnegie Mellon University SPHINX-3 system. The TIMIT corpus was reduced in bandwidth down to 4 kHz and downsampled to 8000 samples per second. A GSM-coded version of the TIMIT corpus was generated by passing the bandwidth-reduced speech signal through the codec. Our baseline system consisted of a cross-word triphon-based, continuous-density HMM recognition system, modeled by approximately 600 multigaussian distributions with diagonal covariance matrices, and 8 Gaussian densities per mixture. The acoustic features consisted of Mel frequency cepstral coefficients including a power coefficient, along with delta and double-delta coefficients. Models were trained for both GSM-coded speech and non-GSM (clean) speech separately using the same model definitions (i.e., state tyings) to ensure tied-state compatibility between both sets of models. Our dictionary consisted of the 6329 words found in the TIMIT corpus, plus silence. The language model (LM) consisted of a bigram model trained on the transcription of the training utterances. The language weight was set to a conservative value to achieve a balance between maximizing the decoding speed while preventing the LM score from dominating the overall score at the expense of evidence from the acoustic models (which represent the effects of GSM coding). We employed an LM weight equal to 9.5.

Lines 1 and 2 in Table 3 summarize recognition error rates for clean and GSM speech using clean models, while lines 3 and 4 describe results when models are trained using GSM-coded speech or a combination of clean and GSM-coded speech, respectively. Given the number of words in the TIMIT test set the interval in the WER that can be considered to be statistically reliable is approximately 0.5% (Gillick and Cox, 1989).

Table 3
Baseline error rates for the reduced-bandwidth TIMIT database under diverse train and test coding conditions

Testing data	Training data	Word error rate
Clean	Clean	11.5%
GSM	Clean	13.0%
GSM	GSM	12.2%
GSM	GSM + clean multistyle	11.9%

From Table 3 we can see that the absolute word error rate increase by 1.5% from 11.5–13% due to the presence of coding in the testing data using clean speech models, which corresponds to a 13% relative increase in word error rate. By recognizing using models trained in matched conditions (i.e., GSM coding) the absolute increase in word error rate is 0.7%, corresponding to a relative increase of 6.1%. When both clean and GSM data are used during training in "multistyle" fashion, there is a further reduction of word error rate of 0.3%. Some authors have noticed improvement when performing multistyle training in cellular and telephone communications or by judicious use of both types of models (Puel and Obrecht, 1997; Das et al., 1999; Mokbel et al., 1996). In our case, the same utterances that were used for training the clean models were coded and reused in the multistyle training. In other words, our multistyle training procedure uses the same data only twice, once before coding and once after. Other authors (cfr. Haavisto 1999) have observed larger effects of GSM coding on ASR even when recognizing using matched acoustic model conditions. The smaller effect of GSM coding on our task can be associated with the task's limited acoustic confusability, and the relatively low perplexity of the LM for the lexicon size.

6.2. Experiments using weighted acoustic models

In this subsection we describe recognition experiments we performed on GSM-coded speech using weighted models, comparing the effects of different phonetic groupings and different weighting schemes. Table 4 summarizes recognition results from these experiments. Experiments were conducted using the automatically clustered

Table 4
Comparison of recognition error rates using several different types of weights for combining acoustic models, as described in Section 6.1

Test data	Number of distortion categories	Number of phonetic categories	Weight search method	Word error rate
GSM	2	1 class (flat weights)	Exhaustive search	11.9%
GSM	2	15 classes	Exhaustive search	11.7%
GSM	2	15 classes	Histogram derived	11.8%
GSM	2	45 classes (1 phone per class)	Histogram derived	11.8%

categories obtained using the method described in Section 4. We considered three types of phonetic clusters: a single phonetic category including all phones (line 1 in Table 4), an automatic clustering that yielded 15 phonetic categories (lines 2 and 3 in Table 4), and assigning each of the phones to its own category (line 4 in Table 4). We compared recognition accuracy using two different types of weights: associating with each class a value of λ which was optimized by searching for the parameter value that maximized recognition accuracy (lines 1 and 2 in Table 4), and associating with each class a λ which was obtained by clustering log-spectral histograms and making λ equal to normalized median value of the corresponding histogram distribution, as described in Section 5.2 (lines 3 and 4 in Table 4).

In order to optimize the weights in the experiment associated with lines 1 and 2, we explored the vicinity of the values obtained by clustering log-spectral histograms for each λ and we adjusted each value independently trying to minimize the work error rate. We associated all the tied states related to a certain basephone j with the weight corresponding to that basephone λ_j. (We refer to this sequential procedure as exhaustive search in the tabular data and discussion that follow.)

The results described in Table 4 indicate that the lowest error rates are obtained by optimizing the weights of each of the 15 classes. The best result, 11.7% is just 0.2% absolute points from the result obtained by training and testing on clean data. In other words, if we use 15 classes and obtain the best weights we will effectively minimize the impact of the codec on word error rate. Results similar to these but based on cepstral frame accuracy and phonetic accuracy using a slightly different log-spectral metric were obtained using the Resource Management database in (Huerta and Stern, 1999). The value of the weight λ when only on single phonetic class was considered (i.e., what we previously referred to as the flat weights condition) was close to 0.6.

The experiments described above were based on the weighted combination of two source acoustic models. For those experiments, the total number of Gaussians used in recognition has been effectively doubled, making the decoding process computationally more expensive and thus slower. An important issue is the extent to which improvement is possible regardless of computational complexity. Fig. 7 shows results of recognition experiments that employ the same phonetic clustering and λs obtained using 15 phonetic classes and exhaustively-searched weights, as a function of the number of Gaussians per mixture. The horizontal axis is labelled according to the number of Gaussians per state in the source models used in the Clean/Clean and GSM/GSM cases (e.g., 8 Gaussians), and according to the number of Gaussians in the resulting weighted acoustic

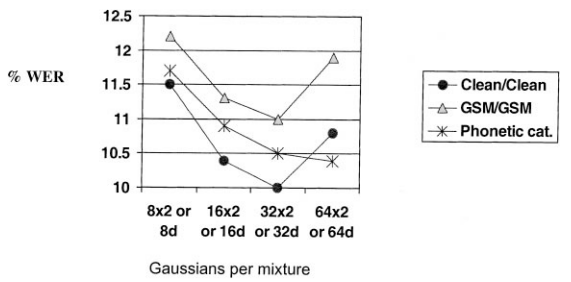

Fig. 7. Error rates using best weights obtained in Table 4 using models with different number of Gaussian densities per mixture. The number of Gaussians indicated in the x-axis refer to the weighted acoustic modeling (top row), and Clean/Clean and GSM/GSM scenarios (bottom row).

models (e.g., 8×2 Gaussians). We can see that lowest word error rate was obtained for the Clean/Clean and GSM/GSM baseline conditions using 32 Gaussians per density. For that case the degradation in recognition accuracy produced by GSM coding under matched conditions is about 1.0% absolute. With a greater number of Gaussians, the base models are overtrained and their error rate increases. We can see that the weighted acoustic modeling based on these models continues to produce an improvement with a greater number of Gaussians (i.e., 64×2d). We believe that this is due to the smoothing effect that this technique has on the likelihood surfaces. Likelihood smoothing techniques have been repeatedly shown to help reduce effects of problems like model overtraining (cfr. Huang et al., 1996). The best absolute performance of the weighted acoustic modeling, then, reduces the gap that exists between the best systems's performance under Clean/Clean and the best system under GSM/GSM conditions by approximately 60% (relative).

These results suggest that once we have established a set of the λ parameters that work with a certain number of Gaussians per state we can apply them with success to other numbers of Gaussians. While we can expect equal or better results when the λs are optimized for each of the particular configurations, this involves more computation. The computational expense associated with the exhaustive weight search is mitigated by three facts: (1) the search can be performed more efficiently using model configurations with small numbers of Gaussians, (2) the phones can be clustered into categories, which will simplify the search and produce results similar to untied conditions, and (3) the search for optimal weights needs to be done only once, after the source models have been trained.

7. Summary

The distribution of the RLSD introduced to the quantized long-term residual by the RPE process of the RPE-LTP codec differs for the various phonetic categories. We presented a way to take advantage of this observation and adjust the acoustic models for each of these phonetic classes by means of weighted models. While a reasonable way to approach the GSM codec problem might be through the use of acoustic models that had been trained exclusively on matched data conditions, the results observed in Section 6 indicate that the degradation in recognition accuracy can be reduced by including both GSM and clean data in the acoustic models. The gap introduced by GSM coding was reduced by 86% for mismatched training and testing conditions, and by 71% for matched conditions. These reductions in error rate were obtained when models were combined using weights derived directly from statistics from the distributions of the RLSD and phonetic clusters derived from these patterns of distortion, using a relatively small number of Gaussians (8 Gaussians per state). When computational expense is not an issue and the number of Gaussians per state are not considered the method provides a reduction of 60% of the degradation gap introduced by GSM coding between the best Clean/Clean conditions system and the best performing GSM/GSM system. A recognition system that processes speech that had undergone GSM coding will greatly benefit from including models based on both clean speech and GSM-encoded speech in the training process. Similar analyses of the effects of coding based on distortion categories can easily be applied to other types of closed-loop predictive coders regardless of their bit rates.

References

Beyerlein, P., 1998. Discriminative model combination. In: Proceedings ICASSP'98, 12–15 May 1998. Seattle, Washington, DC USA, Vol. 1, pp. 481–484.

Das, S., Lubensky, D., Wu, C., 1999. Towards robust speech recognition in the telephony network environment-cellular and landline conditions. In: Proc. EUROSPEECH 1999.

Degener, J., Bormann, C., 1992. GSM speech compression software implementation. ftp://ftp.cs.tu-berline.de/pub/local/kbs/tubmik/gsm/.

Delphin-Poulat, L., Mokbel, C., 1997. Frame-synchronous adaptation of cepstrum by linear Regression. In: IEEE Proc. ASRU 1997.

Digalakis, V., Neumeyer, L., Perakakis, M., 1998. Quantization of cepstral parameters for speech recognition over the www. In: Proc. ICASSP 1998.

Dufour, S., Glorion, C., Lockwood, P., 1996. Evaluation of the root-normalised front-end (RM_LFCC) for speech recognition in wireless GSM network environments. In: Proc. ICASSP 1996.

Elvira, J.M., Torrecilla, J.C., 1998. Name dialing using final user defined vocabularies in mobile (GSM & TACS) and fixed telephone networks. In: Proc. ICASSP 1998.

Euler, S., Zinke, J., 1994. The influence of speech coding algorithms on automatic speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing.

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

Fissore, L., Ravera, F., Vair, C., 1999. Speech recognition over GSM: specific features and performance evaluation. In: Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999.

Gallardo-Antolin, A., Diaz-de-Maria, F., Valverde-Albacete, F., 1998. Recognition from GSM digital speech. In: Proc. ICSLP 1998.

Gallardo-Antolin, A., Diaz-de-Maria, F., Valverde-Albacete, F., 1999. Avoding distortions due to speech coding and transmission errors in GSM ASR tasks. In: Proc. ICASSP 1999.

Gillick, L., Cox, S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: Proc. ICASSP 1989.

Gupta, S.K., Soong, F., Haimi-Cohen, R., 1996. High accuracy connected digit recognition for mobile applications. In: Proc. ICASSP 1996.

Haavisto, P., 1999. Speech recognition for mobile communications. In: Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, 1999.

Haeb-Umbach, R., 1997. Robust speech recognition for wireless networks and mobile telephony. In: Proc. EUROSPEECH 97.

Huang, X.D., Hwang, M., Jiang, L., Mahajan, M., 1996. Deleted interpolation and density sharing for continuous hidden Markov models.In: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference.

Huerta, J.M., Stern, R.M., 1998. Speech recognition form GSM codec parameters. In: Proc. ICSLP-98.

Huerta, J.M., Stern, R.M., 1999. Distortion-class weighted acoustic modeling for robust speech recognition under GSM RPE-LTP coding. In: Proceedings of the Robust Methods for Speech Recognition in Adverse Conditions, Tampere Finland, 1999.

Karray, L., Ben Jelloun, A., Mokbel, C., 1998. Solution for robust speech recognition over the GSM cellular network. In: Proc. ICASSP 1998.

Kleijn, W.B., Paliwal, K.K. (Eds.), 1995. Speech Coding and Synthesis. Elsevier, Amsterdam.

Kroon, P., Kleijn, W.B., 1995. Linear-prediction base analysis-by-synthesis coding. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis, Elsevier, Amsterdam.

Kroon, P., Deprettere, E.F., Sluyter, R.F., 1986. Regular-pulse excitation-a novel approach to effecitve and efficient multipulse coding of speech. IEEE Trans. Acoust Speech Signal Process. 34, 1054–1063.

LDC, The Linguistic Data Consortium (Distributors), 1993. TIMIT Acoustic-Phonetic Continuous Speech Corpus. LDC Catalog Number: LDC93S1.

Lilly, B.T., Paliwal, K.K., 1996. Effect of speech coders on speech recognition performacne. In: Proc. ICSLP-96.

Ming, J., Hanna, P., Stewart, D., Owens, M., Smith, J., 1999. Improving speech recognition performance by using multi-model approaches. In: Proc. ICASSP 1999.

Mokbel, C., Mauuary, L., Jouvet, D., Monne, J., Sorin, C., Simonin, J., Bartkova, K., 1996. Towards improving ASR robustness for PSN & GSM telephone applications. In: Proceedings of the Second IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA1996).

Paping, M., Fahnle, T., 1997. Automatic detection of disturbing robot voice and ping pong effects in GSM transmitted speech. In: Proc. Eurospeech 1997.

Puel, J.B., André-Obrecht, 1997. Cellular phone speech recognition: noise compensation vs. robust architectures. In: Proc. Eurospeech 1997.

Salonidis, T., Digalakis, B., 1998. Robust speech recognition for multiple topological scenarios of the GSM mobile phone system. In: Proc. ICASSP 1998.

Soulas, T., Mokbel, C., Jouvet, D., Monné, J., 1997. Adapting PSN recognition models to the GSM environment by using spectral transformation. In: Proc. ICASSP 1997.

Vary, P., Hofmann, R., Hellwig, K., Sluyter, R.J., 1988. A regular-pulse excited linear predictive codec. Speech Communication 7 (2), 209–215.