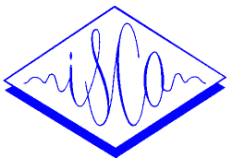

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

INSTANTANEOUS-DISTORTION BASED WEIGHTED ACOUSTIC MODELING FOR ROBUST RECOGNITION OF CODED SPEECH

Juan M. Huerta, Richard M. Stern

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

ABSTRACT

In this paper we apply the Weighted Acoustic Modeling (WAM) technique to the recognition of speech coded by the full-rate GSM codec or the FS-1016 CELP codec employing various estimates of instantaneous distortion. In the WAM method, separate hidden Markov models are developed for regions of speech that exhibit low levels of codec-induced distortion and for regions with higher levels of such distortion. At recognition time, the contributions of these models are mixed together with a weighting that is determined by estimating the instantaneous distortion. In this paper instantaneous distortion was estimated from the instantaneous cepstral distortion, the long-term gain parameter of the codec, the long-term predictability of the reconstructed signal, and measurements of recoding sensitivity. We observe that the use of the long-term gain parameter produces results that are similar to those obtained by use of cepstral distortion (which can only be obtained if the original cepstra are transmitted along with the speech signal) for the GSM codec. Overall, the effect of the degradation in error rate introduced by coding can be reduced by up to 55% with these techniques for GSM coding, and by up to 38% for the CELP coding.

1. INTRODUCTION

One of the principal sources of degradation in the performance of speech recognition applications deployed in mobile environments is the presence of coding-decoding processes in the communication link. This performance degradation is due to the distortion introduced to the reconstructed speech signal by the quantization and bit rate reduction that takes place in the speech codec. It has been shown [2, 8] that, in general, the effect of coding on recognition accuracy tends to increase as the bit rate of the compressed signal decreases. As more speech applications are deployed into wireless, mobile, and cellular environments, the problem of distortion due to coding will become more widespread.

The distortion introduced by the codec to the signal can be thought of as an additive non-stationary noise which is a function of the speech signal itself. However, because most modern speech codecs are fairly complex, their effect on the original speech signal is hard to model analytically. Previous approaches to the problem have included robust features, robust acoustic modeling, and methods of acoustic model compensation and adaptation.

In this work we focus on reducing the effect of this distortion on recognition accuracy by making use of the Weighted Acoustic Modeling method employing different distortion estimates. In the WAM technique [4, 5] a set of N acoustic models, each representing a certain distortion condition, is employed during decoding and its contribution to the overall likelihood is weighted by a running estimate of the distortion observed by each observation frame. In this work we estimate the instantaneous distortion in four different ways: using measured cepstral distortion, the long-term gain (adaptive codebook gain), the long-term predictability, and recoding sensitivity.

In Section 2 we describe the basic idea behind linear prediction-based coding of speech as well as some specifics about the GSM and the FS-1016 codecs. Section 3 is a brief overview of the WAM method. In Section 4, we describe the instantaneous estimates of the distortion that we use in the paper. Section 5 contains the experiments and results obtained using the distortion estimates and the TIMIT corpus. And finally we present some conclusions and discussion of the results in Section 6.

2. LINEAR PREDICTION-BASED CODING

Linear-prediction based speech codecs assume that the speech signal is the output of an all-pole filter excited by an input signal (called the short-term residual, or excitation signal). Using z-transform notation we can use this model to express the resulting speech signal $S(z)$ as the product of an excitation signal $E(z)$ and a spectral shaping filter $\frac{1}{B(z)}$.

$$S(z) = \frac{E(z)}{B(z)} \tag{1}$$

To separate the spectral envelope of the speech signal from the excitation component, speech codecs first perform an LPC analysis on the speech signal and obtain the LPC polynomial $B(z)$. Then, the excitation component is obtained by inverse filtering the speech signal. The LPC filter is an all-pole filter whose coefficients can be found by minimizing the mean squared error between the predicted and true value of the original speech signal $s[n]$.

The speech signal has, in its voiced segments, strong long-term correlation components due to the quasi-periodic vibration of the vocal chords. While the broad spectral properties are modeled by the analysis filter, the long-term correlation information is retained in the residual signal $e[n]$. In order to be able to represent the excitation signal with a reduced num-

ber of bits, speech codecs remove the redundancy left in the residual signal by exploiting this long-term correlation. The residual in any subframe frequently resembles itself in adjacent or close subframes. Based on this observation, speech codecs approximate the residual in a given subframe according to its resemblance to the residual in previously reconstructed subframes.

A reasonable first approximation to the excitation of the current subframe, then, is the excitation signal of adjacent subframes. Speech codecs implement this approximation by finding the gain and lag that minimize the difference between the reconstructed excitation signal in previous subframes and the excitation in the current subframe [7], By performing the long-term prediction (LTP) of the excitation signal (or short-term residual), the predictable or periodic part of the residual is captured. Speech codecs utilize a second set of parameters (*e.g.*, multi-pulses, regular-pulses or fixed codebooks) to model the non-predictable or non-periodic portion of the excitation signal.

The reconstructed version of the excitation or reconstructed short-term residual is produced by the weighted sum of a predictable component and an unpredictable component. These two contributions can be interpreted as being based on two codebooks, one adaptive and one fixed.

The energy of the quantization error is proportional to the energy of the unpredictable part of the signal (the long term residual), so the frame being processed will incur a smaller amount of distortion if the long-term predictor produces a good estimate of the short-term residual than if it produces a poor prediction. Because the long-term prediction is based on subframe level prediction, a given frame will incur less distortion if the corresponding segment of speech has higher cross-subframe periodicity (and thus is easier to predict using subframes in its proximity). In this paper we exploit this idea in order to achieve robust acoustic modeling.

In this paper we analyze the effect of two specific Linear-prediction based speech codecs: the full rate GSM codec and the FS-1016 codec, which we briefly describe here.

2.1 The full-rate GSM codec

The full-rate GSM codec is a linear predictive regular-pulse excited-long-term predictive (RPE-LTP) based codec operating with a bit rate of 13 kbps [3]. The 8-kHz speech signals enter the codec where they are analyzed in frames of 160 samples from which 8^{th}-order LPC parameters are obtained every 20 ms, producing an LPC analysis rate of 50 frames per second. The LPC parameters are represented as log area ratio (LAR) coefficients which are quantized and then transmitted. The residual signal from the LPC analysis (*i.e.*, the short-term residual) is subdivided into subframes of 40 samples and coded by a regular pulse excited-long-term prediction codec whose parameters are quantized. The long-term part of the residual codec is responsible for performing the prediction of the current subframe in terms of the adjacent subframes, producing a lag and a gain.

2.2 The FS-1016 CELP codec

The FS-1016 standard is a 4.8-kpbs codec, based on a 10^{th} order LPC analysis, followed by a CELP representation of the short-term residual signal [1]. The long-term periodicity is modeled by an adaptive codebook (which is an equivalent process to the long-term prediction block of the GSM codec). The adaptive codebook is generated from previous subframes of the reconstructed short-term residual. The long-term residual signal, or the difference between the short-term residual and the short-term residual signal estimate, is coded by means of a fixed ternary stochastic codebook. This codec uses, as does the FR-GSM counterpart, 8 kHz as sampling rate. Its frame size is 30 ms long with four 7.5-ms subframes per frame.

3. THE WEIGHTED ACOUSTIC MODELING METHOD

We describe the use of the weighted acoustic modeling (WAM) technique with which we employ instantaneous distortion information or some estimate of it for robust recognition of speech that is degraded by GSM or FS-1016 CELP coding. The basic idea of the WAM method [4, 5] is related to the observation described in Section 2 that not all segments of speech in a coded corpus are distorted to the same extent. As noted above, when the speech codec performs a short-term and long-term analysis of the speech signal, the level of distortion introduced by the long-term predictive analysis of the short-term residual can be associated with the predictability of the speech signal. Thus, the predictability of the speech signal is directly related to the detrimental effect of coding in recognition [5]. Therefore, it is reasonable to expect that decoding the segments of speech that suffered less distortion yields better recognition accuracy when evaluating their likelihoods using HMM models that were trained using undistorted speech than when recognizing coded speech using a system that was trained on generic coded speech without characterizing the particular coding distortion involved.

The weighed acoustic method, combines the contributions from M constituent models to the overall likelihood at the state level. Specifically, let $p(a_i|q_i)$ be the probability that state q_i emits the observation a_i at time i, then:

$$p(a_i|q_i) = \sum_{m=1}^{M} \left(\sum_{k=1}^{K} f\left(p_{q_{i,m}}(k), d_i, q_{i,m}\right) N\left(a_i, \mu_{q_{i,m},k}, C_{q_{i,m},k}\right) \right) \quad (2)$$

where the term $N\left(a_i, \mu_{q_{i,m},k}, C_{q_{i,m},k}\right)$ represents the k^{th} Gaussian density of state q_i belonging to the m^{th} constituent environment and $f\left(p_{q_{i,m}}(k), d_i, q_{i,m}\right)$ is a weighting function that incorporates the original mixing weights of the constituent state models $p_{q_{i,m}}(k)$, the identity of the constituent states $q_{i,m}$, and d_i which is the estimate of the amount of distortion observed at time i. In this work we use four different methods to estimate appropriate values for d_i given the observed reconstructed speech signal for every frame i.

We have implemented the WAM technique in our recognition system by appending the densities of the constituent models, and associating subsets of the Gaussian densities that belong to a given state to a distortion environment (*e.g.*, clean versus distorted) and then rescaling the mixing weights according to distortion information either provided externally to the recognizer or based on an estimate of the instantaneous distortion affecting each segment of speech.

4. INSTANTANEOUS DISTORTION INFORMATION AND ESTIMATES

In this section we suggest four ways of calculating the parameters that serve as rescaling factors or weighting factors for the Gaussian components in the Weighted Acoustic Modeling method described in Section 3. Using this information, the M constituent models get assigned a multiplicative constant (*i.e.*, they are rescaled) according to how much evidence there is that the observed frame belongs to a certain environment. If the system has access to both the distorted and the undistorted signals (or their cepstral representations), this distortion information d_i can be computed for at every instant i at the front-end level. Otherwise, the amount of distortion can be estimated from the reconstructed speech signal. In the following subsections we describe four different measurements and estimates of the instantaneous distortion introduced by the speech codec.

4.1 Instantaneous cepstral distortion

The instantaneous cepstral distortion measure is obtained by computing the normalized value of the norm of the difference between the cepstral vector of the original speech signal $c_i[j]$ at frame i and the corresponding cepstral vector of the reconstructed speech signal $\hat{c}_i[j]$:

$$d_i = \sum_{j=0}^{12} \frac{(c[j] - \hat{c}[j])^2}{c[j]^2} \qquad (3)$$

This measure is proportional to the exact value of the magnitude of the distortion introduced to the recognition features by the speech codec. Since it is necessary to have access to the original (uncoded) speech signal to compute this measure, this parameter needs to be computed at the terminal device and should be transmitted along with the codec parameters to the server where the ASR application is located.

4.2 Long-term gain of the short-term residual codec

The use of the long-term gain (or adaptive codebook) as a feature for ASR of coded speech has been shown to be beneficial to the recognition of coded speech [6]. In this work we use it as an estimate of the distortion introduced by the codec. The basic idea is that the speech codec introduces larger distortion to regions of speech which are harder to predict. The value of the long-term gain serves as a measurement of the predictability of the subframe. Thus, access to this parameter (or an estimate of it), can provide the recognizer with information related to the amount of distortion present in the reconstructed speech signal. Because of the different configurations of the various existing codecs, however, the way this measure is applied must be codec dependent.

4.3 Speech long-term predictability

The measured long-term predictability of the reconstructed speech signal (which is similar to the long-term gain of the short term residual as described previously) should constitute, using the same rationale as in the subsection above, a reasonable estimate of the distortion introduced by the codec. Essentially, the difference between the long-term gain of the short-term residual and this estimate, is that the long-term predictability of the speech signal can be made independently of the speech codec. Specifically, we obtain this estimate by computing the maximum value found in the cross-correlation between the current subframe $s[n]$ and the adjacent subframe $s[n+N]$ normalized by the energy of the current subframe.

$$LTPM[n] = \log\left\{ \frac{\max_{0<p<TN}\left(\sum_k s[n]s[n+k+p+N]\right)}{\sum s[n+k]s[n+k]} \right\} \qquad (4)$$

Because we are trying to make this estimate independent of any particular speech coding, the values of *N*, the size of the subframes, and *T*, the number of adjacent subframes considered must be chosen such that they fall close to or within the actual range of values of these parameters employed by the codecs.

4.4 Recoding sensitivity

We also can derive an estimate of the amount instantaneous distortion introduced to the speech signal by means of a second coding pass on the already-coded speech. The second coding pass is introduced in order to enable a comparison between the received speech and the speech resulting from the second decoding pass. We note that the received speech will be employed for recognition; the second coding is only applied in order to estimate the distortion weights. The basic motivation for this method is that if $s[n]$ is the original speech signal and $\hat{s}[n]$ the reconstructed speech signal, then $\hat{s}[n] = s[n] + \hat{h_1}[n] * \Upsilon(r_1[n])$ where $\hat{h_1}[n]$ is the impulse response of the quantized LPC synthesis filter, and $\Upsilon(r_1[n])$ is the error introduced by the coding of the lon-term residual $r_1[n]$. Hence, if $\hat{s}_2[n]$ is the output of a second coding pass then

$$\hat{s}_2[n] = s[n] + \hat{h}_1[n] * \Upsilon(r_1[n]) + \hat{h}_2[n] * \Upsilon(r_2[n]) \qquad (5)$$

If the reconstructed speech signal is dominated by the original speech signal, then the amount of quantization introduced in the second decoding pass for a given speech segment will be similar to the distortion produced by the first decoding pass. This estimate is codec dependent, like the estimate based on the long-term gain of the short-term residual codec.

5. EXPERIMENTAL RESULTS

We performed experiments using the TIMIT database and the SPHINX-III recognition system, using both the full-rate GSM codec and the CELP codec. The weighted acoustic modeling for these experiments was implemented using two source environments, clean speech and matching conditions coded speech. No retraining of the acoustic models was performed. A bigram language model was employed, and the HMM models shared 600 tied states.

Table 1 summarizes the results of baseline experiments for the three coding conditions: no coding, GSM coding and CELP coding. These experiments were performed using either a single set of HMMs with 16 densities (Table 1) or two models comprised of 16 Gaussian per mixture densities were combined (Table 2). Table 2 compares results obtained using WAM with two constituent environments (clean and coded), the GSM and FS-106 codecs, and the various instantaneous distortion estimates described in Section 4. The numbers in parenthesis in Table 2 represent the percentage improvement provided by WAM relative to the size of the gap in WER between results obtained using clean speech for training and testing and coded speech for training and testing (Table 1).

Train	Test	%WER
Clean	Clean	10.4%
GSM	GSM	11.3%
FS-1016	FS-1016	12.5%

Table 1: Baseline Recognition experiments using the TIMIT corpus and GSM and FS-1016 coding techniques.

For the GSM codec, our results indicate that it is possible to reduce the effect of GSM and coding on recognition by up to 55% when instantaneous cepstral distortion information is available or when the codec's long-term gain parameter is used. It is also possible to reduce the effects of GSM coding by 11% and 33% when estimates of the long term predictability and the recoding sensitivity are employed respectively.

Instantaneous distortion information	% WER GSM	% WER FS-1016
Cepstral distortion	10.8% (55%)	11.7% (38%)
LTP gain	10.8% (55%)	12.2% (14%)
LT predictability	11.2% (11%)	11.8% (33%)
Recoding sensitivity	11.0% (33%)	12.2% (14%)

Table 2: Weighted Acoustic modeling recognition experiments using the TIMIT corpus and different instantaneous distortion information.

For the FS-1016 codec, however, the biggest reductions in the effects of coding are observed using instantaneous cepstral distortion and the long term predictability information, which provide 38% and 33% relative reduction, respectively. The LTP gain gives a more modest reduction for the FS-1016 codec than for the GSM codec, which can be attributed to the differences in the complexity and properties of the codecs (*e.g.*, length of the subframes).

6. CONCLUSIONS

We demonstrated the effectiveness of weighted acoustic modeling (WAM) in reducing the effects of codec-induced distortion on speech recognition by as much as 55%. A key attribute of this algorithm is the use of an estimate of instantaneous distortion, and we compared the performance of four approaches to accomplish this estimation. If direct measures of instantaneous cepstral distortion are available, their use can produce a 55% decrease in the degradation of WER introduced by GSM codecs, and a corresponding 38% decrease for the FS-1016 CELP codec. Otherwise, "blind" estimates of instantaneous distortion can reduce the degradation introduced by GSM codecs by 55% (using the LTP gain metric) and the corresponding degradation introduced by FS-1016 codecs by 33% (using the LT predictability metric). These differences indicate that the different codecs have substantially different properties that are relevant to the introduction of distortion, so compensation for the effects of new codecs will have to be developed in a codec-by-codec basis.

ACKNOWLEDGEMENTS

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

REFERENCES

1. Campbell Jr. J.P., Tremain T.E., Vanoy C. Welch, "The DoD 4.8 KBPS Standard (Proposed Federal Standard 1016)", in *Advances in Speech Coding* edited by Atal B. Cuperman V. and Gersho A. Kluwer Academic Publishers 1989.

2. Euler S., Zinke J., "The Influence of Speech Coding Algorithms on Automatic Speech Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol ASSP- No. 1994.

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

4. Huerta, Juan M. and Stern, Richard M. "Distortion-class weighted acoustic modeling for robust speech recognition under GSM RPE-LTP coding", *Proceedings of the Workshop on Robust Method for Speech Recognition in Adverse Conditions*, Tampere Finland, 1999.

5. Huerta, Juan M. *Speech Recognition in Mobile Environments*, Ph.D. Thesis Department of Electrical and Computer Engineering, Carnegie Mellon University. May 2000.

6. Kim, H-K., and Cox R.V., "Bitstream-based feature extraction for wireless speech recognition", *Proc. ICASSP 2000*.

7. Kleijn W.B., Paliwal K.K, "An introduction to Speech coding", in *Speech Coding and Synthesis*, Elsevier Science B.V., W.B. Kleijn and K.K.Paliwal (editors), Amsterdam 1995.

8. Lilly B. T., Paliwal K. K., "Effect of Speech Coders on Speech Recognition Performance", *Proc. ICSLP-96*, 1996.