Reducing Speech Coding Distortion for Speaker Identification

Alan McCree

MIT Lincoln Laboratory
Lexington, MA 02420
E-mail: mccree@ll.mit.edu

Abstract

In this paper, we investigate the degradation of speaker identification performance due to speech coding algorithms used in digital telephone networks, cellular telephony, and voice over IP. By analyzing the difference between front-end feature vectors derived from coded and uncoded speech in terms of spectral distortion, we are able to quantify this coding degradation. This leads to two novel methods for distortion compensation: codebook and LPC compensation. Both are shown to significantly reduce front-end mismatch, with the second approach providing the most encouraging results. Full experiments using a GMM-UBM speaker ID system confirm the usefulness of both the front-end distortion analysis and the LPC compensation technique.

Index Terms: speech coding, speaker identification.

1. Introduction

With the widespread use of digital speech communication, for example in cellular telephony and voice over IP (VoIP), speech coding has become commonplace. The high user acceptance of modern speech coding algorithms results from their extensive optimization to minimize perceived distortion in listening tests; however, these optimizations may not be appropriate for systems relying on automated analysis, such as speaker identification. In this work, we first examine the impact of widely-used speech coding standards on the mel-cepstral filterbank front-end commonly used in speech recognition or speaker ID. Based on this analysis, we then propose and test two methods of compensating for speech coder distortion: codebook and LPC compensation. Finally, we perform speaker ID experiments confirming the improved performance of the LPC compensation technique.

Previous work on speaker ID of coded speech has considered using featured vectors derived from the speech coding bitstream or the decoded output speech [1, 2, 3]. In particular [1] concluded that performance was better using the coded speech for feature vector analysis. These results were extended in [4], which quantified the speaker ID degradation for a number of speech coders and showed a slight loss in performance for matched training and test conditions, becoming more significant for mismatched conditions. In this work, we continue to use the coded speech for feature vector analysis, both for performance reasons and also for the added convenience of using the same speaker ID modeling approach for both coded and uncoded speech. However, our goal is to develop waveform compensation techniques to improve the speaker ID performance.

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. Impact of Speech Coding on Speaker ID Front-End

We set up an experimental framework to directly measure speaker ID front-end distortion from speech coding. We ran the speaker ID front-end on both uncoded and coded speech, and computed the Spectral Distortion (SD): the RMS error between the two sets of log filterbank energies. The front-end [5] uses a 19-dimensional mel-cepstral vector extracted every 10 ms and bandlimited to 300-3138 Hz. Delta cepstra are computed using a 5 frame span and appended to the cepstral vector to produce a 38-dimensional feature vector. An adaptive energy-based speech detector discards low-energy vectors.

For initial experiments, we used the testing partition of the 'si' sentences of the TIMIT [1] database, resulting in approximately 130,000 test frames. The speech coders tested cover a wide range of bit rates from 64 to 5.3 kb/s, including:

- 64 kb/s ITU G.711 mu-law PCM
- 32 kb/s ITU G.726 ADPCM
- 12.2 kb/s 3GPP GSM-AMR (same as GSM-EFR)
- 6.7 kb/s GSM-AMR
- 5.3 kb/s ITU G.723.1

The first two of these are traditional speech waveform coders: 64 kb/s G.711 Pulse-coded Modulation (PCM) and 32 kb/s G.726 Adaptive Differential PCM (ADPCM). These are toll-quality standards defined by the International Telecommunications Union (ITU) and widely deployed throughout the conventional telephone network. The next three coders are based on Code Excited Linear Prediction (CELP), a more sophisticated speech-specific waveform coding technology providing near-toll quality at medium bit rates. Two of these are widely used in the cellular telephony systems: 12.2 kb/s GSM-EFR for worldwide TDMA systems and 6.7 kb/s GSM-AMR for Japanese TDMA and newer GSM systems. The lower rate 5.3 kb/s G.723.1 CELP coder is used in VoIP applications.

In Fig. 1, we see that the RMS error varies between 0.5 and 3.5 dB, with increasing error as the speech coding bit rate decreases. In spectral quantization for speech coding, the standard rule of thumb is that 1 dB of spectral distortion is transparent to a

[1] Available from http://www.ldc.upenn.edu

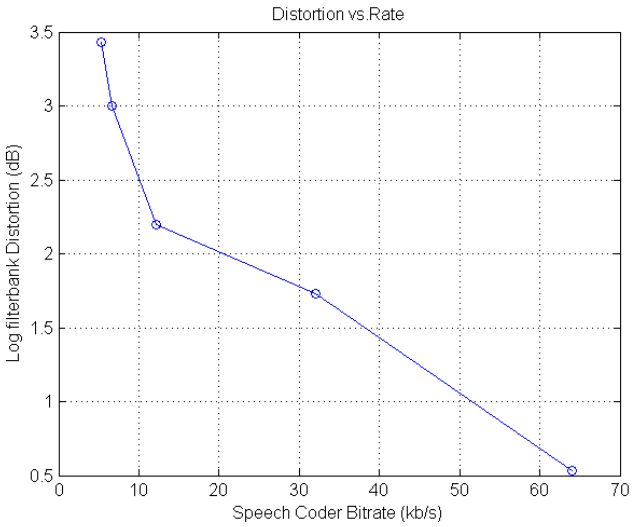

Figure 1: Filterbank distortion vs. speech coder bitrate.

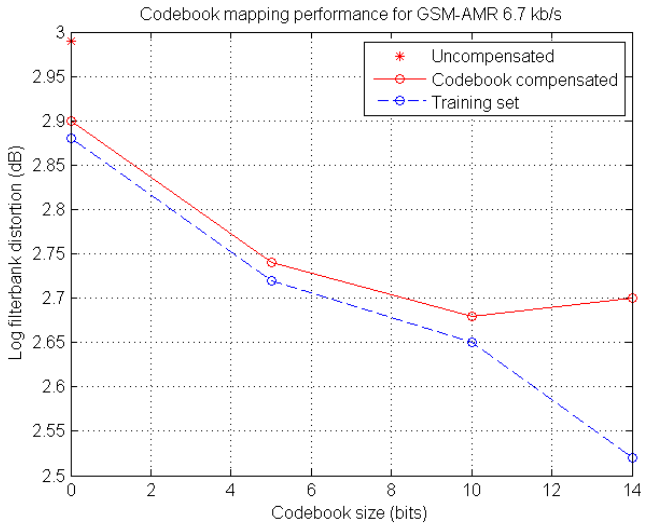

Figure 2: Codebook compensation performance for 6.7 kb/s GSM-AMR speech coder.

human listener [6]. Notice that G.711 is the only coder to achieve this distortion threshold.

3. Coded Speech Compensation Techniques

Our goal is to generate a speech coder postprocessor to optimally reduce the speaker recognition front-end mismatch between coded and uncoded speech, using as much information from the speech coder as possible, not just the decoded output speech, to improve the compensation performance. This postprocessor should clearly help to compensate for the speech coder distortion in mismatched training and test conditions, and we hope that by exploiting internal coder information we can get improvement even for matched conditions.

At first glance it might seem that a well-designed speech decoder cannot be improved. However, we see two reasons for optimism. First, in this application we can use additional delay and lookahead of decoded speech; traditional speech decoders do not have this luxury. More importantly, we are interested in a different goal: to improve the speaker ID performance rather than the perceived speech quality. In particular, we do not need to recover the entire speech waveform (which is clearly not possible since information has been lost), but only the filterbank spectral magnitudes for the front-end. In theory, uncorrelated quantization noise should simply be additive in the power spectral domain, so that the spectral distortion due to coding could in fact be deterministic and removable.

3.1. Codebook Compensation

If the spectral distortion introduced by speech coding is deterministic, i.e. if a given coded spectrum always corresponds to a given uncoded spectrum, then it can be learned from a training set by an algorithm such as generalized vector quantization (also called codebook mapping) [7]. In this approach, the training set of coded speech filterbank vectors is partitioned into regions, and for each region the optimal minimum-mean-square-error compensation vector is computed as the mean difference between uncoded and coded filterbank vectors for this region. While it is tempting to map each coded vector directly to a corresponding uncoded vector, we achieved better performance by learning the error for each partition, since then the space of possible output vectors is continuous rather than discrete. Codebook mapping assumes no particular functional relationship between uncoded and coded speech, and in the limit of large training set and codebook sizes provides theoretically optimal performance.

To test this method, compensation codebooks were trained over the standard training partition of the 'si' sentences in TIMIT. Processing of uncoded and coded speech files was the same as described above; this training set contains approximately 340,000 non-silent frames. Fig. 2 shows the performance of this method for the 6.7 kb/s GSM-AMR coder, as a function of increasing codebook size. Note that even a 0 bit codebook provides improvement over the uncompensated case; this is simply compensation of the difference in mean feature vector. Also, for large codebooks the degradation in performance from the training data to the independent test set is quite significant, indicating that the training data was not sufficient for this codebook size or that the training algorithm is not sufficiently robust. A codebook size of 10 bits (1024 vectors) provides a reasonable performance/complexity tradeoff.

Therefore, 10-bit codebooks were trained for all speech coders listed in the previous section. Test set results for the coders of primary interest (5.3 - 12.2 kb/s) are shown in Fig. 3 (labeled as "CB 1014"), along with the uncompensated results ("None"). Codebook compensation provides noticeable performance improvement for these coders, as well as for 32 kb/s G.726, but not for 64 kb/s G.711.

3.2. LPC Compensation

In most medium and low bit-rate speech coders, the transmitted bitstream contains a representation of the original speech spectrum using some form of linear prediction coefficients (LPC). Since the

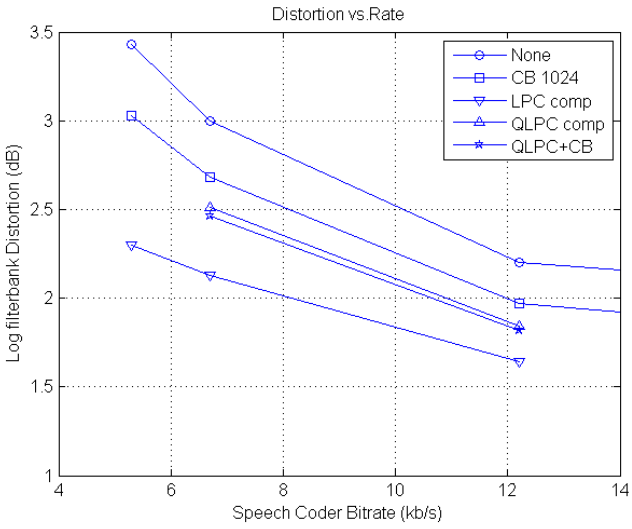

Figure 3: SD performance of compensation techniques vs. speech coder bitrate. Techniques include baseline (None), codebook compensation (CB 1024), ideal LPC compensation (LPC comp), quantized LPC compensation (QLPC comp), quantized LPC + codebook compensation (QLPC+CB).

decoder has access to this information directly, it could be used in the compensation process. Earlier work [1] has shown that performing speaker ID using the LPC spectrum directly does not improve performance because it contains no excitation information. However, we can assume that the transmitted LPC represents the correct (undistorted) spectral envelope for the synthesized speech, and postfilter the output appropriately. The steps involved in this process are:

- Compute the LPC of the synthesized speech.
- Inverse filter the synthesized speech with this filter to remove the distorted spectral envelope.
- Perform synthesis filtering with the transmitted LPC of the original speech to restore the correct spectral envelope.

To get an initial feel for the potential of this approach, we used a separate LPC analysis of the uncoded speech for compensation. While this would not be possible in practice since only the quantized LPC would be available, it does provide an upper performance bound. As shown by the curve labeled "LPC comp" in Fig. 3, this technique provides significantly more improvement than the codebook mapping approach for low bit rates, since it exploits an additional information source.

Based on these encouraging results, we modified the software of one speech coder to correctly implement this compensation using the quantized LPC from the received bitstream. This requires an additional buffering delay of the synthesized speech for the lookahead window of the LPC analysis, since care must be taken to use the same analysis process and window position for this analysis as was used by the encoder. We chose to use the GSM-AMR coder software so that experiments could be run at multiple rates (6.7 and 12.2 kb/s). These results, labeled " QLPC comp" in Fig. 3, confirm that even with LPC quantization this approach provides significant performance improvement for both coders.

3.3. Combination

Since both codebook mapping and LPC compensation provide performance improvement, it is possible that combining the two will provide even better performance. We have tested this approach, where the modified speech coder software described in the previous section is used for processing, and then compensation codebooks are trained based on this version of the coded speech. Unfortunately, Fig. 3 shows that this provides negligible further improvement. Apparently, the LPC compensation already exploits the information available to the codebook mapping process. Therefore, we have concluded that the LPC compensation technique provides the best front-end distortion compensation performance.

4. Speaker ID Performance Characterization and Improvement

Based on these encouraging results for front-end distortion, we have performed full Speaker ID experiments.

4.1. Speaker ID System

We use the MIT/LL Gaussian mixture model with universal background model (GMM-UBM) speaker recognition system [5]. This system is a likelihood ratio detector with target and alternative probability distributions modeled by GMM's. The UBM is used as the alternative hypothesis model, and from this, target models are derived using Bayesian adaptation.

For these experiments, we use a different partitioning of TIMIT than in our earlier experiments, since we now need training and test utterances from the same speakers which are not available with the standard test/train partition. For training each target speaker model, we use two sentences from the 'sa' portion of the corpus, three sentences from the 'si', and three from the 'sx', for a total of eight sentences. For test, we use two different sentences from the 'sx' portion. There are a total of 462 speakers, 136 female and 326 male. Each test sentence is scored using all of the speaker models. To test the impact of speech coder distortion on speaker ID performance, we train all models on uncoded speech and test using utterances processed by the speech coder under test. While this is not the only possible testing configuration, it gives a reasonable and consistent performance measure.

The UBM is a 2048 mixture gender-independent GMM, trained from ten sentences (two from 'sa', three from 'si', and five from 'sx') from each of 168 speakers in the standard test partition of TIMIT; these speaker are all different than the 462 target speakers. The target speaker models are derived from adapting the Gaussian mean vectors only over the eight target training sentences.

4.2. Baseline Results

First, we test the effect of all speech coders on our baseline system. Results for Equal Error Rate (EER), where probabilities and costs of miss and false alarm are equal, are shown in Fig. 4. The resulting curve of distortion vs. speech coder bit rate shows minimal impact at high bit rates vs. the uncoded baseline, but a significant performance decrease for lower rates. This is consistent with the expectation that toll-quality speech coders, which are nearly transparent to human listeners, are also nearly transparent to the speaker

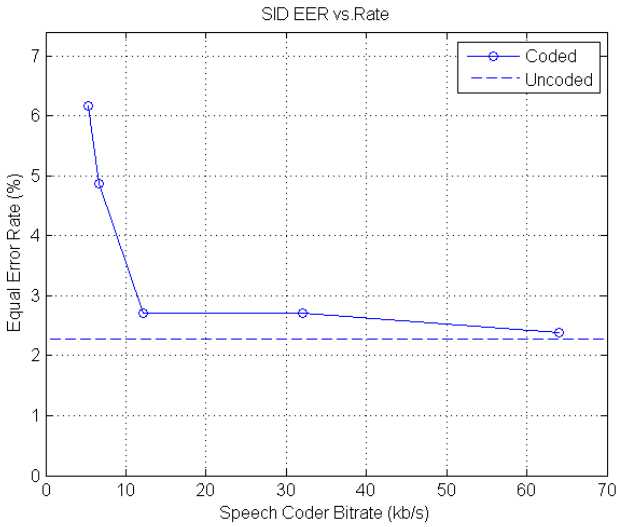

Figure 4: Speaker ID EER performance vs. speech coder bit rate.

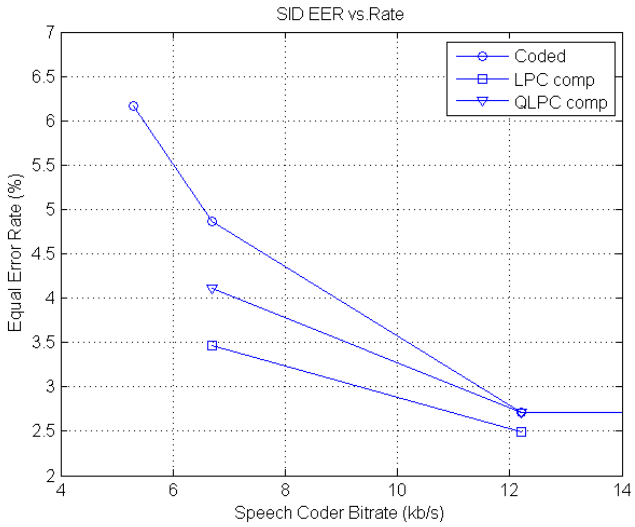

Figure 5: Speaker ID EER improvement with compensation methods.

ID system. Note also that the performance is more a function of bit rate than type of coder; for example, the 12.2 kb/s GSM-EFR CELP coder has similar performance to the 32 kb/s ADPCM waveform coder.

In comparison to the filter bank Spectral Distortion results from Fig. 3, this rate distortion curve is much flatter at high rates, as performance is limited by the minimum error rate achievable for this problem even without speech coding distortion (labeled "uncoded") . This implies that the simpler SD measure can be used to predict speaker ID performance degradation due to speech coding, but there is a non-linear relationship that must be considered as well. SD less than 1 dB has no effect, up to 2 dB has minimal impact, and 3 dB and greater is significant.

4.3. Performance with Coded Speech Compensation

Since the LPC compensation technique appears to be most promising, we have tested it's impact on actual speaker ID performance. For the GMM speaker ID system, we achieve significant performance improvement from the LPC compensation technique, as shown in Fig. 5. While again the ideal (unquantized) LPC compensation provides the largest improvement, the real LPC compensation method is also very effective. The GSM-AMR system using quantized LPC parameters from the bitstream brings the performance of the 6.7 kb/s coder up to the equivalent of about 9 kb/s. Comparing these results to Fig. 3 again confirms the usefulness of the SD metric in predicting speaker ID performance.

5. Conclusion

We have developed a new method for reducing the degradation of speech coding on speaker ID performance. Experiments using the TIMIT corpus have demonstrated reduced speaker ID error rate by compensating the decoded speech with the LPC parameters derived from the speech coder bitstream. An additional advantage of this compensation method is that it is completely transparent to the speaker ID system, since this new LPC postfilter can readily be incorporated inside the speech decoder.

6. Acknowledgements

The author gratefully acknowledges many interesting technical discussions with colleagues Carl Quillen, Alex Solomonoff, Doug Reynolds, Doug Sturim, and Tom Quatieri.

7. References

[1] T. F. Quatieri, E. Singer, R. B. Dunn, D. A. Reynolds, and J. P. Campbell, "Speaker and Language Recognition using Speech Codec Parameters," in *Proc. Eurospeech*, 1999, pp. 787–790.

[2] E. W. M. Yu, M. W. Mak, C. H. Sit, and S. Y. Kung, "Speaker Verification Based on G.729 and G.723.1 Coder Parameters and Handset Mismatch Compensation," in *Proc. Eurospeech*, 2003, pp. 1681–1684.

[3] A. Moreno-Daniel, B. H. Juang, and J. A. Nolazco-Flores, "Robustness of Bit-stream Based Features for Speaker Verification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2005, pp. I749–752.

[4] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, and J. P. Campbell, "Speaker Recognition from Coded Speech and the Effects of Score Normalization," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, 2001, pp. 1562–1567.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[7] A. Rao, D. Miller, K. Rose, and A. Gersho, "A Generalized VQ Method for Combined Compression and Estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, 1996, pp. 2032–2035.