SPEAKER AND LANGUAGE RECOGNITION USING SPEECH CODEC PARAMETERS*

T.F. Quatieri[1], *E. Singer*[1], *R.B. Dunn*[1], *D.A. Reynolds*[1], *and J.P. Campbell*[2]

[1]MIT Lincoln Laboratory, Lexington, MA, USA
[2]Department of Defense
Email: quatieri@ll.mit.edu

ABSTRACT

In this paper, we investigate the effect of speech coding on speaker and language recognition tasks. Three coders were selected to cover a wide range of quality and bit rates: GSM at 12.2 kb/s, G.729 at 8 kb/s, and G.723.1 at 5.3 kb/s. Our objective is to measure recognition performance from either the synthesized speech or directly from the coder parameters themselves. We show that using speech synthesized from the three codecs, GMM-based speaker verification and phone-based language recognition performance generally degrades with coder bit rate, i.e., from GSM to G.729 to G.723.1, relative to an uncoded baseline. In addition, speaker verification for all codecs shows a performance decrease as the degree of mismatch between training and testing conditions increases, while language recognition exhibited no decrease in performance. We also present initial results in determining the relative importance of codec system components in their direct use for recognition tasks. For the G.729 codec, it is shown that removal of the postfilter in the decoder helps speaker verification performance under the mismatched condition. On the other hand, with use of G.729 LSF-based mel-cepstra, performance decreases under all conditions, indicating the need for a residual contribution to the feature representation.

1 INTRODUCTION

Due to the widespread use of digital speech communication systems, there has been increasing interest in the performance of recognition systems from resynthesized coded speech. The question arises as to whether the speech coding and quantization can effect the resulting mel-cepstral representations which are the basis for most recognition systems including speech, speaker, and language recognition. There is evidence, for example, that speech recognition performance can deteriorate when speech used to train and test the recognizer are "mismatched" in the sense that one data set is coded and the other is uncoded; for the GSM codec, this performance degradation can be reduced when both training and testing are performed with coded speech [2, 6]. There is also interest in performing recognition directly using codec parameters rather than from the resynthesized coded speech. For speech recognition, it has been found, again for the GSM codec and matched training and testing, that performance loss is eliminated using mel-cepstra that combine codec parameter representations of all-pole and residual signal components [2, 6].

In this paper, we investigate the effect of speech coding on speaker and language recognition tasks. Speaker recognition experiments are performed using a Gaussian mixture model-universal background model (GMM-UBM) speaker verification system [7] applied to a subset of the NIST eval98 Switchboard database. Experiments with speaker recognition have shown sensitivity in performance with respect to mismatch in UBM training, claimant training, and testing data. For example, a large error increase can be incurred when background (UBM) and claimant training data are from uncoded speech, while the test data is from coded speech. On the other hand, a fully matched condition, i.e., coding across the three data sets, yields a marked improvement, but still not a performance equal to the fully uncoded case. In addition, experiments show a decrease in performance with decreasing coder rate, i.e., from GSM (12.2 kb/s) to G.729 (8 kb/s) to G.723.1 (5.3 kb/s), and a significant difference across gender, i.e., coded female speech exhibits higher error rates relative to an uncoded baseline in both the matched and mismatched conditions. Our language identification system consists of six phone recognizers, each followed by 12 language models [8]. Here we also see a decline in performance with coding rate, but, unlike in speaker verification, no decrease in performance from the matched to mismatched conditions.

We also present preliminary results in understanding the importance of codec system components, in particular G.729, in speaker recognition. An important observation in the mismatch condition is that the presence of the postfilter in G.729 can contribute to loss in performance due to the time-varying nature of this filter; when the postfilter in the mismatch condition is removed, the performance loss is reduced. In the matched condition, changes in performance with and without the postfilter appear to be negligible. Preliminary experiments with G.729 have also shown significant loss in performance with mel-cepstra derived from LSFs, reflecting perhaps an "over smoothing" of the speech spectrum relative to standard mel-cepstral coefficients.

2 CODERS AND CONDITIONS

We investigated speaker verification performance from coded speech for the G.729 (8 kb/s), GSM (12.2 kb/s), and G.723.1 (5.3 kb/s) codecs. All three coders are based on a residual/LSF/postfilter analysis/synthesis, with the primary difference being the manner of coding the residual. The G.729 codec is a fixed point codec at 8 kb/s standardized by ITU-T for personal communication and satellite systems, and is based on a conjugate-structure algebraic CELP residual coding scheme [3]. The GSM codec is the ETSI Pan-European standard fixed-point enhanced GSM at full rate 12.2 kb/s and is based on a regular multi-pulse residual coding scheme [1]. Finally, the G.723.1 codec is the floating point[1] CELP-based ITU-T multi-media standard codec at 5.3 kb/s [4]. For each codec, there are three conditions that we tested against a baseline condition in which no coding is performed in training or testing. We describe the conditions in order of decreasing degree of matching between training and testing.

*THIS WORK WAS SPONSORED BY THE DEPARTMENT OF DEFENSE UNDER AIR FORCE CONTRACT F19628-95-C-0002. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

[1]Speaker recognition performance with the G.723.1 fixed point version, which is also available, was determined to be equivalent to the G.723.1 floating point version under a variety of coding scenarios.

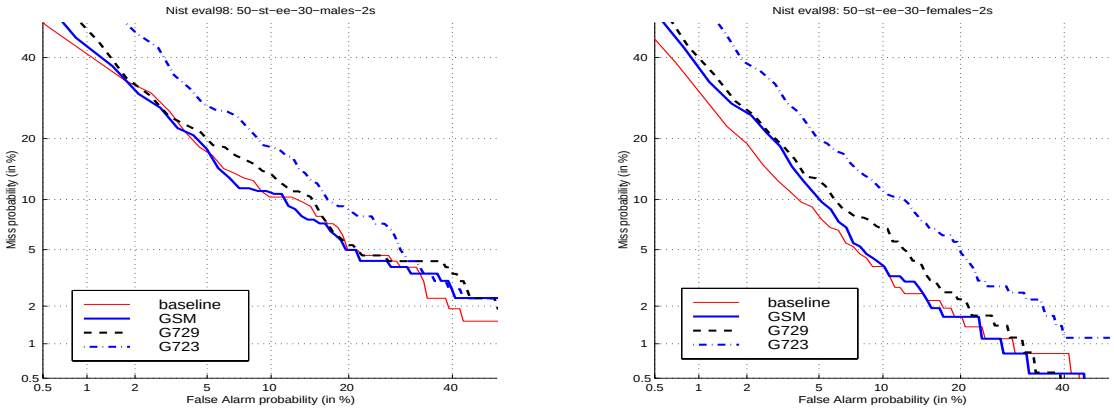

Figure 1: DET performance on male (left panel) and female (right panel) speakers under matched condition (A) for the three codecs GSM (12.2 kb/s), G.729 (8 kb/s), and G.723.1 (5.3 kb/s).

Condition A: This is the "fully matched" case where background and target models are derived from coded speech and the test data is also coded.
Condition B: This is the "partially mismatched" case where the background and target models are derived from uncoded speech and the test data is from coded speech. Since the coded test messages are scored against two uncoded models, we expect performance to decrease relative to condition A.
Condition C: This is the "fully mismatched" case where the background models are derived from uncoded speech and the target models and test data are from coded speech. The test data is thus scored against models derived from both coded and uncoded speech; as such, we expect worse performance for this case.

3 SPEAKER RECOGNITION

The speaker recognition system is a Gaussian mixture model-universal background model (GMM-UBM) speaker verification system [7]. The system consists of a speaker-independent universal background model (UBM) and a claimant model derived from the UBM via Bayesian adaptation. The features consist of appended cepstra and delta cepstra derived from bandlimted (300-3300 Hz) mel-filterbank spectra. Both cepstral mean subtraction and RASTA filtering are performed on the features prior to training and recognition. During recognition, the verification score for an utterance is the log-likelihood ratio computed by taking the difference between the log-likelihoods of the claimant model and UBM. Performance for the speaker verification task is reported as Detection Error Tradeoff (DET) curves [5], produced by sweeping out a speaker-independent threshold over all verification test scores and plotting the miss and false alarm rates at each point.

In speaker recognition, we experimented with a subset of the NIST eval98 database; the train and test sets both corresponded to electret handsets, but possibly different phone numbers. In this test, 50 target speakers are used for each gender with 262 test utterances for males and 363 for females. In various coding scenarios tested with G.729, this subset produced results equivalent to that obtained with the full NIST data set for the same handset condition, while requiring significantly less processing time. Background models are trained based on the NIST eval96 database and NIST eval96 development database (used in NIST eval98 and eval97).

3.1 Experiments with resynthesized speech

The speaker verification performance from resynthesized speech for the three codecs corresponds to coder quality: Full-rate GSM has best performance and best quality, G.723.1 has worse performance and worse quality, and G.729 falls between GSM and G.723.1 with respect to performance and quality. Figure 1 shows this relative performance of the three codecs under the matched condition (A). This same relative performance across codecs holds under the two mismatch conditions (B) and (C), with performance generally degrading with increasing mismatch as illustrated by the Equal Error Rate (EER) results in Figure 2. The DET curves for each gender show these same relative performance trends across the three conditions over the range of miss and false alarm probabilities shown in Figure 1. We also see in Figures 1 and 2 that, although the female baseline (i.e., fully uncoded condition) performance is greater than that for males, the relative performance loss with coding is greater for the females.

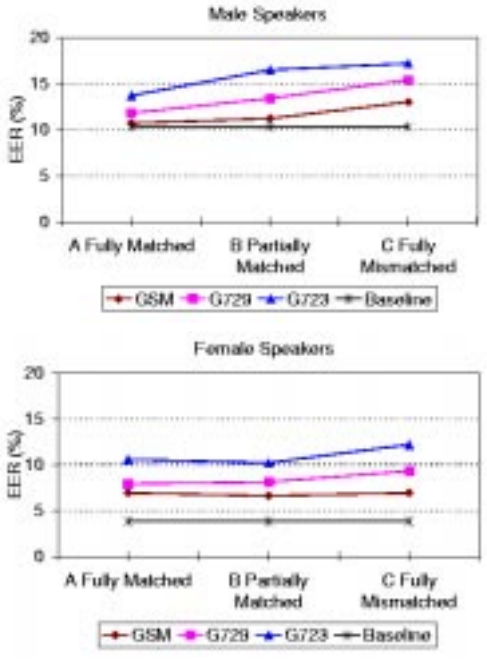

Figure 2: Speaker verification EER performance on male (upper panel) and female (lower panel) speakers with matched and mismatched conditions for GSM, G.729, and G.723.1 codecs.

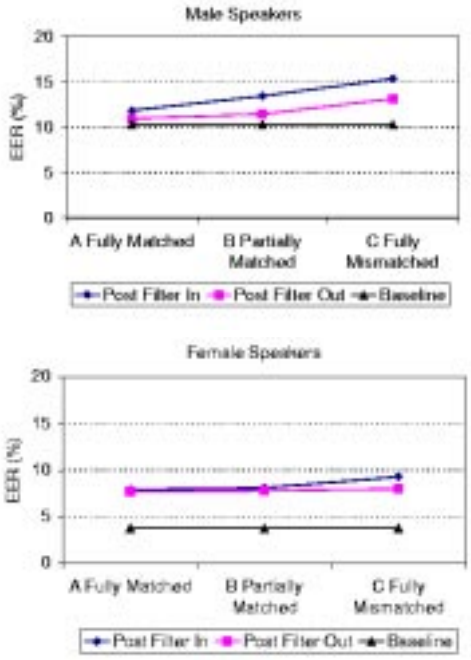

Figure 3: Speaker verification performance on male (upper) and female (lower) speakers for G.729 with and without the postfilter for the three training and testing conditions (A), (B), and (C).

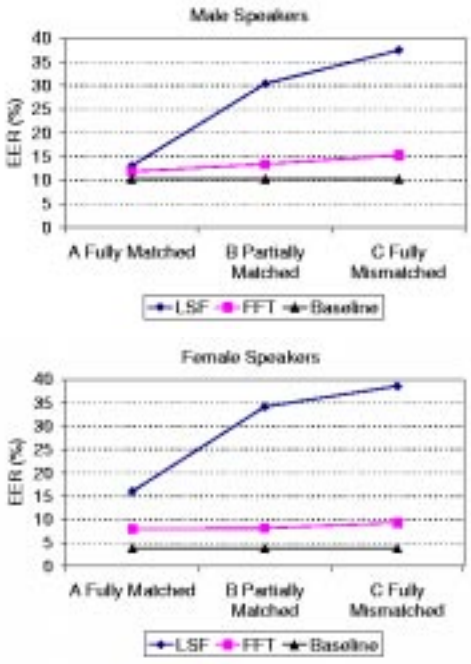

Figure 4: Comparison of speaker verification performance using mel-cepstra derived from G.729 LSF coefficients and FFT calculations for the three training and testing conditions (A), (B), and (C). Upper and lower panels show performance for male and female speakers, respectively.

Importance of G.729 Postfilter

The G.729 postfilter consists of three components: short-term, long-term, and spectral tilt contributions. The short-term postfilter sharpens formants and suppresses noise in formant nulls. The long-term postfilter does the same for harmonics. The spectral tilt modifies the spectral trend. Our objective here is to test the time-varying "channel" effect of the postfilter under both the two mismatch conditions and the fully matched condition. Results for our Conditions A, B, and C for each gender are shown in Figure 3 where we observe that performance without the postfilter degrades with increasing mismatch, as it did with the postfilter present. The removal of the postfilter, however, improved speaker verification performance, especially under the mismatch conditions. The DET curves away from the EER point even more strongly confirm these performance differences with and without the postfilter. One important aberration, however, was observed. For the fully matched condition (A) for male speakers, the removal of the complete postfilter slightly helped performance, while for the females, this removal somewhat hurt performance away from the EER point along the DET.

We see then that under the mismatched conditions, best performance is obtained with removal of the postfilter. Under the matched condition, on the other hand, we have a performance difference with gender; for females, we do not want to remove the postfilter because this hurts the already degraded performance by the coder, while for males, the removal of the postfilter helps performance but the gain is negligible. It seems that intuition is consistent with these results. Under the mismatch conditions, removal of one of the channel effects (i.e., the postfilter interpreted as a channel) helps performance. Under the matched condition, however, performance stays about the same or is hurt because removal of the postfilter also removes the positive effect of reducing noise in formant nulls (short-time) and harmonic nulls (long-time). The greater loss in performance with female speakers is consistent with this interpretation.

3.2 Experiments with G.729 LSFs

We have explored speaker verification performance with mel-cepstra derived from G.729 LSFs [3]. Conversion of the G.729 LSFs to mel-cepstra is given by the following steps: (1) Extract the LSFs and convert to LPC coefficients; (2) Sample the LPC spectral envelope at the DFT frequencies, consistent with the speaker recognition frontend; (3) Apply the mel-cepstral filters and convert to mel-cepstra; (4) Compute frame energy measure (sum of squared sequence values), energy being the basis for speech activity detection.

The EER performance for the matched condition (A) and mismatched conditions (B) and (C) are shown in Figure 4 for each gender. We see that, as compared to previous speaker verification results with G.729 coded speech, performance degrades significantly for both males and females, especially under the mismatched conditions. The overall performance as seen in the DET curves is generally consistent with the EER trends, except for the case of male speakers under the matched condition for which performance further decreased away from the EER point. One possible explanation for the general performance loss with the modified mel-cepstra is that LSF-based features used in these conditions are fundamentally very different from conventional mel-cepstra: we are smoothing an already smooth LPC envelope, rather than a high-resolution short-time FFT as in the conventional scheme. Under the matched condition (A), we have lost spectral resolution while under the mismatch conditions (B) and (C), we add to degradation by using very different feature types in training and testing.

In order to attempt to overcome the "double smoothing" effect of the LSF-based mel-cepstra, we ran speaker verification for the matched condition where mel-cepstra are derived from sampling the LPC envelope at mel-filter center frequencies rather than integrating over these mel-filters.

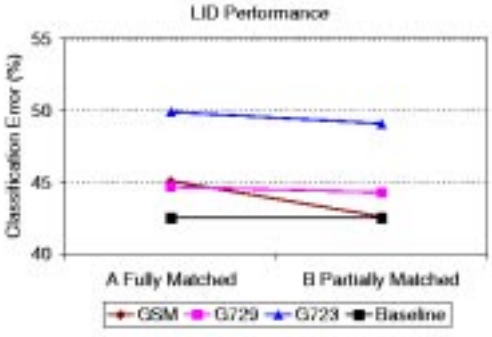

Figure 5: 12-way language classification results (percent error).

The hope is that we will avoid "double smoothing" of the spectrum prior to cepstral computation. For both males and females, however, speaker verification performance degraded slightly in the matched condition (A) with sampling of the LPC envelope.

Finally, the LSFs in all the above experiments are unquantized. To obtain a flavor for the effect of quantization, we looked at speaker verification performance for the matched condition (A) with mel-cepstra derived from quantized LSFs. As expected, there is a small decline in performance with the introduction of quantization on the LSFs.

4 LANGUAGE RECOGNITION

The language recognition system consisted of 6 phone recognizers, 72 language models (12 language models per phone recognizer), and a Gaussian backend classifier [8]. Training of language models was performed using the messages in the "TRAIN" subdivision of the Callfriend database. Long messages were truncated to 7.5 minutes duration to reduce processing time. There are a total of 238 training messages, approximately 20 per language. Test utterances were taken from the 30 second messages in the "lid96d1" development data used in the 1996 NIST Language Recognition Evaluation. The final test set consists of 1184 utterances, approximately 100 per language.

Cross-validation experiments were performed using the lid96d1 messages for backend training and for testing. Baseline results were established by computing 12-language percent error classification using uncoded speech for both training and testing. For each codec (GSM, G.729, and G.723.1) two conditions were evaluated using resynthesized speech, one in which both train and test data were from synthesized speech (Condition A) and the other in which the test data was synthesized but the training data was uncoded (Condition B).

Results, reported as 12-language percent error, are shown in Figure 5. Except for the GSM-match condition, results are consistent with the expectation that performance decreases with decreasing bit rate. Again with the exception of GSM, matching the language models to the coder had no significant effect on performance. These results support the conclusion that the time consuming task of retraining language models for new coders may be unnecessary. In all cases, the phone recognizers used were trained using uncoded data and were not retrained for a specific coder. This mismatch may account for some degree of loss in performance from the baseline to the codecs, but retraining the phone recognizer for every codec variation is computationally expensive and was avoided. Finally, we note that the 42.5% error of the baseline system obtained with 7.5 min messages compares with 34.1% error obtained with 30 min messages.

5 SUMMARY

Speaker verification performance generally degrades with codec bit rate, i.e., from GSM (12.2 kb/s) to G.729 (8 kb/s) to G.723.1 (5.3 kb/s), relative to baseline. This decrease is consistent with decreasing perceptual quality. With language recognition, a similar but less consistent decline in performance was observed. With speaker verification, for all codecs, performance generally decreases as the degree of mismatch between training and testing increases, i.e., in going from Condition A to B to C. Language recognition, on the other hand, exhibited no performance loss under the mismatched condition (B) relative to the matched condition (A) for all three codecs. One explanation of this difference between speaker verification and language recognition behavior is that the speaker verification system operates directly on the acoustic features and so is more affected by changes in these features due to coding conditions. The language recognition system operates on indirect information derived from the acoustic features (i.e., the phoneme sequence) which is less affected by non-severe coder degradations. When the coding degradations begin to affect the consistency of the recognized phones (e.g., random confusions) the language recognition performance will begin to decrease rapidly.

Removal of the G.729 postfilter helps speaker verification performance in the mismatched conditions, but is ineffective (and for females is degrading) in the matched condition. With use of G.729 LSF-based mel-cepstra, we obtain significant performance loss under all conditions. Best performance (but still with a considerable loss) occurs with the matched condition for male speakers. Our attempt to overcome this performance loss by sampling the LPC spectrum at mel-filter center frequencies hurt performance slightly.

Our current direction is to investigate other representations of LSF parameter, residual, and postfilter contributions to the coders to avoid reconstructing the coded speech prior to recognition. The postfilter has helped explain some of the performance loss under specific conditions. It is of interest to investigate other sources of loss that may expose which parameters to avoid when dissecting the codecs.

REFERENCES

[1] European Telecommunication Standards Institute, "European digital telecommunications system(Phase2); Full rate speech processing functions (GSM 06.01)," ETSI, 1994.

[2] J.M. Huerta and R.M. Stern, "Speech recognition from GSM coder parameters," *Proc. 5th Int. Conf. on Spoken Language Processing*, Vol 4, pp 1463-1466, 1998.

[3] ITU-T Recommendation G.729, "Coding of speech at 8 kb/s using conjugate-structure algebraic-code-excited linear prediction," June 1995.

[4] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kb/s," March 1996.

[5] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, Mark Przybocki, "The DET Curve in Assessment of Detection Task Performance," *Proc. Eurospeech 1997*, Vol 4, pp 1895-1898, 1997.

[6] C. Mokbel, L. Mauuary, D. Jouvet, J. Monne, C. Sorin, J. Simonin, and K. Bartkova, "Towards improving ASR robustness for PSN and GSM telephone applications," *Proc. 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Vol 1, pp 73-76, 1996.

[7] D.A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Proc. Eurospeech97*, Vol 1, pp 963-967, 1997.

[8] M.A. Zissman, "Predicting, Diagnosing, and Improving Automatic Language Identification Performance," *Proc. Eurospeech97*, Vol 1, pp 51-54, 1997.