Robust Speaker Recognition in the Presence of Speech Coding Distortion for Remote Access Applications

Robert W. Mudrowsky[1], Ravi P. Ramachandran[1], Umashanger Thayasivam[2] and Sachin S. Shetty[3]
[1] Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ, USA
[2] Department of Mathematics, Rowan University, Glassboro, USA
[3] Department of Electrical and Computer Engineering, Tennessee State University, Nashville, TN, USA

***Abstract*— For wireless remote access security, forensics, border control and surveillance applications, there is an emerging need for biometric speaker recognition systems to be robust to speech coding distortion. This paper examines the robustness issue for three codecs, namely, the ITU-T 6.3 kilobits per second (kb/s) G.723.1, the ITU-T 8 kb/s G.729 and the 12.2 kb/s 3GPP GSM-AMR coder. Both speaker identification and speaker verification systems are considered and use the Gaussian mixture model classifier. The systems are trained on clean speech and tested on the decoded speech. To mitigate the performance loss due to mismatched training and testing conditions, four robust features, two enhancement approaches and fusion strategies are used. A detailed two-way ANOVA analysis is performed.**

Keywords: speaker recognition, robust system, fusion, ANOVA

I. Introduction and Motivation

Robust speech processing for wireless applications is an active area of research especially in terms of gaining remote access to voice activated services [1] (like performing a bank account transaction). This includes remote access using mobile telephony and the internet. Although most practical applications are aimed at a wireless internet connection, the technical approach described in this paper can be applied to a wired connection. In this context, the particular applications of speaker recognition [1][2][3][4][5][6][7][8] and speech recognition have been investigated [9][10]. The challenge is the diminished performance due to background noise, speech coding distortion and communication channel errors. In this paper, biometric speaker recognition in the presence of speech coding distortion is addressed. Both speaker identification (SI) (determining the most likely speaker among a set of candidates) and speaker verification (SV) (accepts or rejects a claimed identity) systems are described. The systems are both text-independent and language-independent.

The operation of the proposed system is as follows. A person seeking remote access speaks into the microphone of his/her wired node (desktop PC) or mobile/wireless node (laptop or smart phone). This speech signal is communicated to a remote server via the internet and is subject to speech coding distortion. The remote server uses a biometric based speaker recognition algorithm. The algorithm can either accomplish (1) speaker identification (SI) in which a person is identified among a set of candidates or (2) speaker verification (SV) that accepts or rejects the claimed identity of a person. The database acts as the central storage for all biometric data. The communication between the database and the server is secure and encrypted.

The speech coders [11] investigated in this paper include the ITU-T 6.3 kilobits per second (kb/s) G.723.1 coder [12][13], the ITU-T 8 kb/s G.729 coder [14] and the 12.2 kb/s 3GPP GSM-AMR coder [15]. The three coders are based on the Code Excited Linear Prediction (CELP) format. The G.723.1 and the G.729 are used in voice over internet protocol (VoIP) applications. The GSM-AMR is widely used in cellular telephony. The ITU implementations are used for the G.729 and GSM-AMR coders. The implementation in [13] is used for the G.723.1 coder.

Training of the SI and SV systems is done on clean speech. Testing (performance evaluation) is done on the decoded speech which is the clean speech passed through the speech coder and then, decoded. The systems are based on a Gaussian mixture model (GMM) classifier. A speaker independent universal background model (UBM) is first configured [8][16] using the expectation maximization (EM) algorithm [18][19]. Individual speaker models are obtained by maximum a posteriori (MAP) estimation of the UBM parameters using the training speech of the speaker [17]. Four robust features, namely, the linear predictive cepstrum (CEP), the adaptive component weighted (ACW) cepstrum [20], the postfilter (PFL) cepstrum [20] and the mel frequency cepstrum (MFCC) [3][4] are compared.

Since the mismatch between training and testing leads to diminished performance, feature enhancement using the affine transform and signal enhancement using the McCree technique [8] are individually investigated and combined. Fusion of the four features [21][22] is also attempted. Statistical methods based on analysis of variance (ANOVA) [23] will compare the performance of the individual features, fusion methods and feature and signal enhancement. The aim is to examine the relative performance of the various approaches as a function of the bit rate. This paper extends the work in [6] by examining three speech coders, implementing both SI and SV and performing two-way ANOVA.

II. System Description and Experimental Protocol

The TIMIT database is used for the experiments. The speech in this database is clean and first downsampled from 16 kHz to 8 kHz. The UBM is first configured using all ten sentences from each of 168 speakers. The k-means algorithm is used to initialize the parameters of the UBM with a diagonal covariance matrix assumed. Ten iterations of the EM algorithm results in the final UBM. Ninety individual

GMM speaker models (different speakers from those used to train the UBM) are obtained by MAP adaptation of the UBM parameters. Both adaptation of the (1) weights, mean vectors and elements of the covariance matrix and (2) only the mean vectors were attempted but no significant difference in performance was obtained. As in [17], only adaptation of the mean vectors is performed. Eight of the ten sentences for each speaker are used to obtain the feature vectors and perform the MAP estimation. Since four different features are investigated, there are four different speaker recognition systems.

For testing the system, the two sentences from each of the 90 speakers not used in training are processed. These sentences are the decoded speech from a particular speech coder. The overall system consists of five components, namely, (1) Signal enhancement using the McCree method, (2) Feature extraction for ensuring speaker discrimination, (3) Feature compensation using the affine transform that maps the test feature vectors to the space reflecting the training condition, (4) GMM classifier and decision logic and (5) Feature fusion.

A. McCree Technique

The basic steps as outlined in [8] were implemented, namely, (1) Perform linear predictive (LP) analysis of the decoded speech to obtain the nonrecursive $A(z)$, (2) Pass the decoded speech through $A(z)$ to remove the distorted spectral envelope and (3) Perform LP synthesis filtering with the transmitted LP information of the input speech to the coder in order to restore the correct spectral envelope.

B. Feature Extraction

The speech is preemphasized by using a nonrecursive filter $1-0.95z^{-1}$ and then divided into frames of 30 ms duration with a 20 ms overlap. A 12th order LP analysis is performed using the autocorrelation method. The LP coefficients are converted into 12 dimensional CEP, ACW and PFL feature vectors. The PFL feature is obtained by weighting the CEP feature by the factor $[1-0.9^n]$ for $n=1$ to 12 [20].

A 12 dimensional MFCC feature vector is computed in each frame. The success of the MFCC is due to the perceptually based filter bank processing of the Fourier transform of the speech followed by cepstral analysis using the DCT [3][4].

For each of the four features, a 12 dimensional delta feature [3] is computed in each frame using a frame span of 5 in order to derive first derivative information. Second derivative information is also obtained. Concatenation of the feature vector, the first derivative and the second derivative results in a 36 dimensional vector that is used for speaker recognition.

Energy thresholding is performed over all frames of an utterance to determine the relatively high energy speech frames. This is the most simple form of voice activity detection [5]. The 36 dimensional vectors are considered only in these high energy frames.

C. Affine Transform

The affine transform accomplishes feature compensation by mapping a feature vector $\mathbf{x}$ derived from test speech to a feature vector $\mathbf{y}$ in the region of the p-dimensional space occupied by the training vectors. This forces a better match between training and testing conditions and in effect, compensates for the distortion of the test speech (in this case, the codec). The transform relating $\mathbf{x}$ and $\mathbf{y}$ is given by Eq. (1) as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{1}$$

where $\mathbf{A}$ is a p by p matrix and $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{b}$ are column vectors of dimension p. Expanding Eq. (1) gives

$$\begin{bmatrix} y(1) \\ y(2) \\ y(3) \\ \vdots \\ y(p) \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \mathbf{a}_3^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(p) \end{bmatrix} + \begin{bmatrix} b(1) \\ b(2) \\ b(3) \\ \vdots \\ b(p) \end{bmatrix} \tag{2}$$

where $\mathbf{a}_m^T$ is the row vector corresponding to the mth row of $\mathbf{A}$.

The affine transform parameters $\mathbf{A}$ and $\mathbf{b}$ are determined from the five (the "sa" and "si") sentences of each of the 90 speakers under consideration. The affine transform is computed only during training and is different for each feature and for each of the three codecs. A total of 12 affine transforms are computed. It is only computed for the 12 dimensional feature vector and not for the first and second derivative information.

Let $\mathbf{y}^{(i)}$ be the feature vector for the ith frame of the training speech utterance. Let $\mathbf{x}^{(i)}$ be the feature vector for the ith frame of the training speech utterance passed through the distortion encountered during testing (in this case, the clean utterance is compressed and decoded by the codec). By using a number of training speech utterances, N sets of vectors are collected, namely, $\mathbf{y}^{(i)}$ and $\mathbf{x}^{(i)}$ for $i=1$ to N. A squared error function is formulated as

$$E(m) = \sum_{i=1}^{N} \left[y^{(i)}(m) - \mathbf{a}_m^T \mathbf{x}^{(i)} - b(m) \right]^2 \tag{3}$$

where $\mathbf{a}_m^T$ is the mth row of $\mathbf{A}$ and $y^{(i)}(m)$ and $b(m)$ are the mth components of $\mathbf{y}^{(i)}$ and $\mathbf{b}$, respectively. Minimization of $E(m)$ with respect to $\mathbf{a}_m^T$ and $b(m)$ results in the system of equations [6]

$$\begin{bmatrix} \sum_{i=1}^{N} \mathbf{x}^{(i)} \mathbf{x}^{(i)T} & \sum_{i=1}^{N} \mathbf{x}^{(i)} \\ \sum_{i=1}^{N} \mathbf{x}^{(i)T} & N \end{bmatrix} \begin{bmatrix} \mathbf{a}_m \\ b(m) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{N} y^{(i)}(m)\mathbf{x}^{(i)} \\ \sum_{i=1}^{N} y^{(i)}(m) \end{bmatrix}. \quad (4)$$

The function $E(m)$ is minimized for $m = 1$ to p. Hence, m different systems of equations of dimension $(p + 1)$ as described by Eq. (4) are solved. The left hand square matrix of Eq. (4) is independent of m and hence, needs to be computed only once.

D. GMM Classifier and Decision Logic

A GMM speaker model λ is completely specified by a conditional probability density expressed as a linear combination of Gaussian densities as given by

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{V} w_i c_i(\mathbf{x}) \quad (5)$$

where $\mathbf{x}$ is a p-dimensional feature vector, $c_i(\mathbf{x})$ is a p-variate Gaussian probability density function and w_i are the mixture weights ($\sum w_i = 1$) for $i = 1$ to V (V is the number of Gaussian mixtures). The Gaussian density $c_i(\mathbf{x})$ is specified by a vector of means μ_i and a covariance matrix Σ_i.

A test speech utterance is processed such that a set of test feature vectors $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_q)$ are computed. In computing these feature vectors, the option of using or not using the McCree technique and the affine transform exists. For the SI system, there are $M = 90$ speakers for which speaker i is represented by GMM λ_i, the identified speaker $M^\star$ is chosen to maximize the a posteriori log-probability as given by

$$M^\star = \arg \max_{1 \le j \le M} \sum_{i=1}^{q} \log p(\mathbf{x}_i|\lambda_j) = \arg \max_{1 \le j \le M} d(j) \quad (6)$$

where $p(\mathbf{x}_i|\lambda)$ is computed as given in Eq. (5) and $d(j)$ is the SI score for speaker j. When many utterances are tested, the identification success rate (ISR) is the number of utterances for which the speaker is identified correctly divided by the total number of utterances tested. For each codec, a total of 180 utterances are tested to get the ISR.

For the test utterance for the SV system, let the claimed identity be speaker k. The a posteriori log-probability, as in Eq. (6) is calculated for the speaker model λ_k and for the UBM model. These two quantities are subtracted to yield the SV score. The SV score is compared to a threshold to either accept of reject the claimed identity. Different thresholds are attempted to yield a receiver operating characteristic (ROC) from which the equal error rate (EER) is the performance measure. For each codec, there will be 180 genuine attempts and $(180)(89) = 16,020$ impostor attempts.

For each speaker, the 2 "sa" and three "si" sentences are always used in training the GMM speaker model as they are used in computing the affine transform. Of the remaining five "sx" sentences, three are rotated for training purposes and the other two are used for testing. This enables 10 trials for each experimental condition.

E. Feature Fusion

Since four different features are used, each with a separate GMM classifier, an ensemble system [21] results which naturally leads to the investigation of fusion. For SI systems, decision level fusion is the simplest technique and involves taking a majority vote of the four different features to get a final decision. The second fusion method is to use Borda count based on the SI scores $d(i)$ for $i = 1$ to $M = 90$ (see Eq. (6)) obtained for the four features. For the SV system, the genuine and impostor SV scores are normalized to be in the range [0,1] for each feature separately. Fusion of these normalized SV scores for the four features are used to obtain the EER. The three methods of fusion considered are the sum of the scores, product of the scores and taking only the maximum score.

III. General Results

The first aspect was to examine how many mixtures to use for the UBM and the individual speaker models. The performance of the four features for all conditions (clean, codec only and using the McCree method and/or the affine transform) was compared when the number of mixtures was varied from 16 to 2048, each time doubling the number of mixtures. The use of 256 and 512 mixtures yielded the best and statistically comparable performance. Using more than 512 mixtures did not necessarily result in higher performance. The results continue for the case of using 256 mixtures.

Tables I and II show the speaker identification and verification results respectively. The average performance over 10 trials is given.

IV. Two-way ANOVA

The balanced two-way ANOVA model (two factors) is considered. A model is considered for each codec separately and within each codec, the SI and SV scenarios separately. There is a continuous response, Y, and two factors, A and B. In this case, Y refers to either the ISR or EER. Factor A are the methods used and has $a = 4$ levels, namely, coder distortion only, coder distortion mitigated by the McCree method, coder distortion mitigated by the affine transform and coder distortion mitigated by both. Factor B are the features and the fusion methods and has b levels. With two fusion methods for SI, $b = 6$. With three fusion methods for SV, $b = 7$. The interests lie in determining whether or not the response (performance) differs due to the two factors and the interaction between the factors. For the SI and SV systems involving each codec, ANOVA revealed that there is a significant interaction between the methods and the features.

Experiments with multiple factors will generally look at a factorial treatment structure. This implies 'treatments' are combinations of the levels of different factors. The experiments in this study are **complete** in that there are

Condition	CEP	ACW	PFL	MFCC	Decision Level Fusion	Borda Count
Clean	93.1	93.2	92.1	95.4	95.2	95.2
G.723.1	64.6	62.5	65.3	79.3	69.0	72.3
G.723.1, McCree	70.4	67.6	71.2	83.0	75.3	77.3
G.723.1, Affine	77.7	74.2	77.7	86.3	83.4	84.3
G.723.1, McCree and Affine	82.8	78.9	80.7	85.8	86.5	87.9
G.729	65.7	61.4	64.6	78.5	69.9	70.2
G.729, McCree	85.0	83.5	83.6	91.1	88.3	89.3
G.729, Affine	84.3	80.9	82.1	89.3	87.8	88.9
G.729, McCree and Affine	86.8	85.5	86.7	90.3	90.2	91.1
GSM-AMR	75.9	73.8	75.3	78.9	78.9	76.3
GSM-AMR, McCree	86.1	83.7	84.2	84.2	87.7	84.4
GSM-AMR, Affine	86.5	86.6	86.2	84.0	89.8	85.3
GSM-AMR, McCree and Affine	85.3	84.2	83.8	83.6	88.2	83.8

TABLE I
IDENTIFICATION SUCCESS RATE (%) FOR VARIOUS CONDITIONS (AVERAGE OVER 10 TRIALS)

Condition	CEP	ACW	PFL	MFCC	Sum Fusion	Product Fusion	Maximum Score Fusion
Clean	3.61	3.35	3.39	3.13	2.78	2.77	3.40
G.723.1	8.43	8.79	8.87	5.98	6.48	6.65	6.62
G.723.1, McCree	7.87	8.09	7.67	5.43	5.75	5.88	6.01
G.723.1, Affine	5.75	6.61	6.22	4.59	4.60	4.56	5.27
G.723.1, McCree and Affine	4.95	5.73	5.51	4.29	4.10	4.13	4.74
G.729	8.11	8.59	8.44	6.69	6.44	6.57	6.63
G.729, McCree	4.82	4.85	4.80	3.90	3.74	3.67	4.12
G.729, Affine	5.29	4.85	4.79	4.07	3.79	3.77	4.38
G.729, McCree and Affine	4.05	4.04	3.93	3.51	3.13	3.19	3.77
GSM-AMR	6.63	6.18	6.25	4.61	4.90	4.90	5.77
GSM-AMR, McCree	5.38	4.86	4.94	3.29	3.51	3.47	4.55
GSM-AMR, Affine	4.65	4.58	4.55	3.44	3.26	3.24	4.37
GSM-AMR, McCree and Affine	5.34	4.96	4.96	3.39	3.66	3.58	4.89

TABLE II
EQUAL ERROR RATE (%) FOR VARIOUS CONDITIONS (AVERAGE OVER 10 TRIALS)

observations at each level combination. The parameterization of the Two-way ANOVA model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk} \quad (7)$$

for $i = 1, ..., a$, $j = 1, ..., b$, and $k = 1, ..., n_{ij}$. Normality is assumed for the errors. The parameter constraints sum to zero in order to get a unique solution. The variable Y_{ijk} is the k^{th} observation on the ij^{th} treatment, where $k = 1, ..., n_{ij}$ and n_{ij} is the size of the sample drawn from each treatment (this is a balanced design so all treatment samples are of size n). The variable μ denotes the global mean. The quantity α_i is the differential effect of the ith level of factor A on the mean response (the mean of the dependent variable) such that $\sum_i(\alpha_i) = 0$. The quantity β_j is the differential effect of the jth level of factor B on the mean response (the mean of the dependent variable) such that $\sum_j(\beta_j) = 0$. The quantity $(\alpha\beta)_{ij}$ is the differential effect of the ijth treatment on the mean response such that $\sum_i(\alpha\beta)_{ij} = 0$ and $\sum_j(\alpha\beta)_{ij} = 0$. The error is denoted by E_{ijk}.

A. Speaker Identification

A discussion of the comparison among the methods and features individually is given. Considering the interaction between the methods and features, the best approaches are also mentioned.

1) G.723.1: Figure 1 and 2 show the 95% confidence interval for the methods and features respectively. It is clear that combining the McCree technique and the affine transform is the best method. The features (includes decision level fusion and Borda count) are similarly compared and the best feature is the MFCC. Due to the interaction of the feature and method, the best performance (average ISR of 87.9%) is obtained using the McCree technique and the affine transform in conjunction with Borda count fusion.

2) G.729: Figures 3 and 4 show the 95% for the methods and features respectively. As in the case of G.723.1, (1) the best method is to combine the McCree technique and the

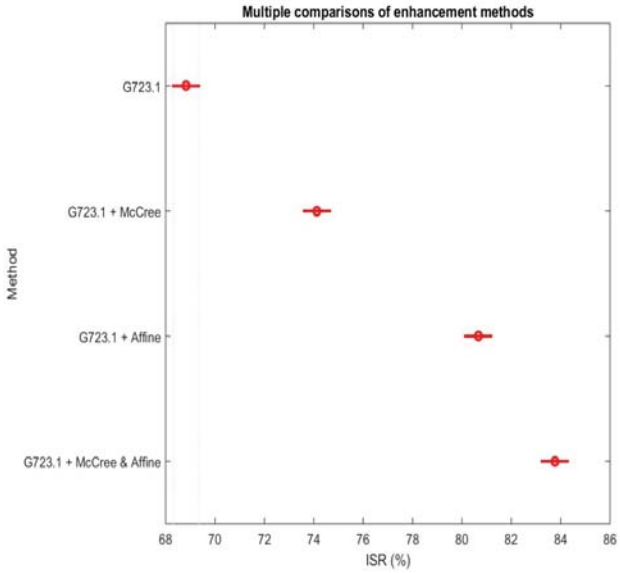

Fig. 1. Speaker identification: Comparison of the methods (G.723.1).

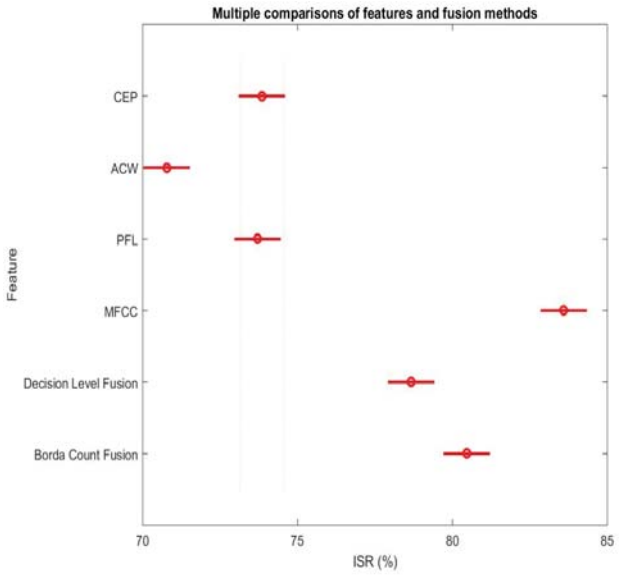

Fig. 2. Speaker identification: Comparison of the features (G.723.1).

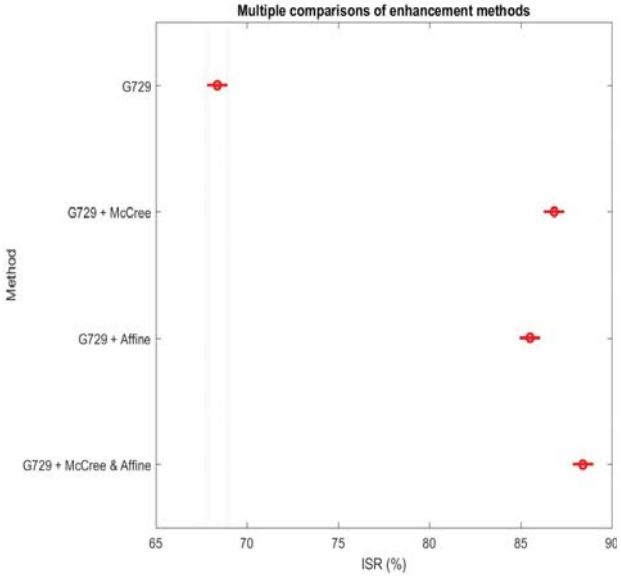

Fig. 3. Speaker identification: Comparison of the methods (G.729).

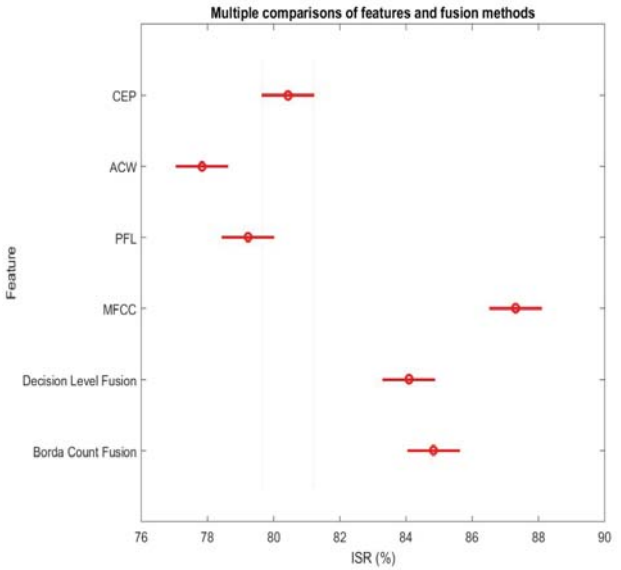

Fig. 4. Speaker identification: Comparison of the features (G.729).

affine transform and (2) the best feature is the MFCC. Due to the interaction of the feature and method, the best performance (average ISR of 91.1%) is obtained using either (1) the McCree technique and the affine transform in conjunction with Borda count fusion or (2) the McCree technique with the MFCC feature.

3) GSM-AMR: Figure 5 and 6 show the results. The best method is using only the affine transform. The best feature is the use of decision level fusion. It is the same two approaches that interact the best achieving an average ISR of 89.8%.

B. Speaker Verification

Results based on the EER are given for the SV system.

1) G.723.1: Figure 7 and 8 show the 95% confidence interval for the methods and features respectively. It is clear that combining the McCree technique and the affine transform is the best method. The features (includes sum, product and maximum score fusion) are similarly compared. Although the best feature is the MFCC, its 95% confidence interval overlaps with that of sum and product fusion. Due to the interaction of the feature and method, the best performance (average EER of 4.1%) is obtained using the McCree technique and the affine transform in conjunction with sum fusion. Using product fusion is statistically comparable and leads to only a slightly higher average EER of 4.13%.

2) G.729: Figures 9 and 10 show the 95% confidence interval for the methods and features respectively. The best method is to combine the McCree technique and the affine transform. Although the sum fusion is the best feature, its

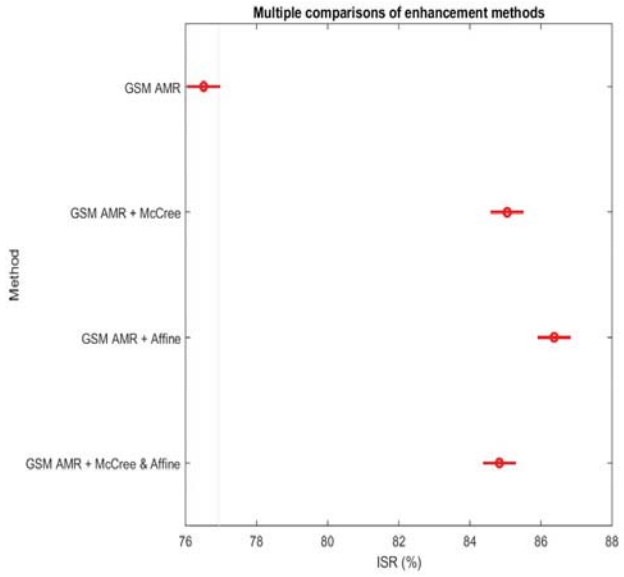

Fig. 5. Speaker identification: Comparison of the methods (GSM-AMR).

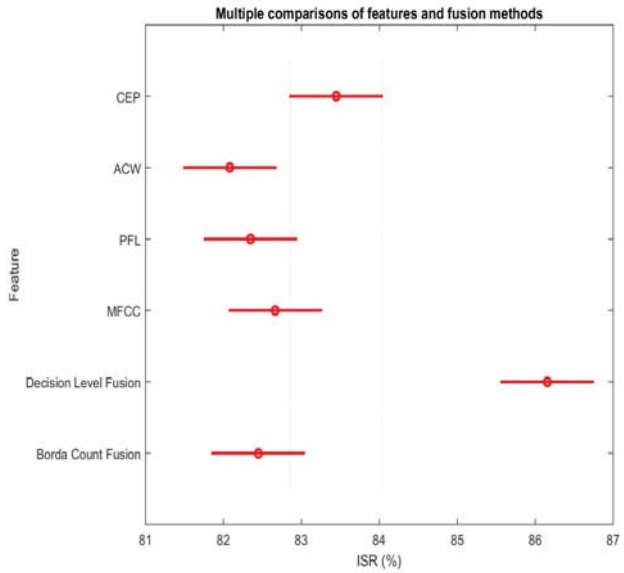

Fig. 6. Speaker identification: Comparison of the features (GSM-AMR).

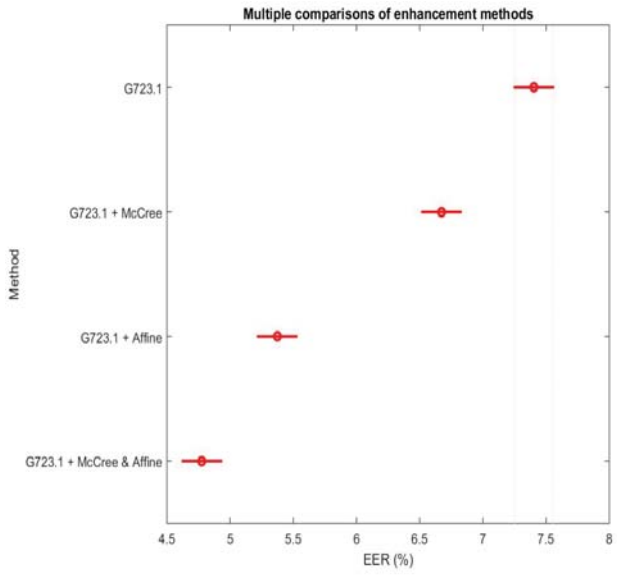

Fig. 7. Speaker verification: Comparison of the methods (G.723.1).

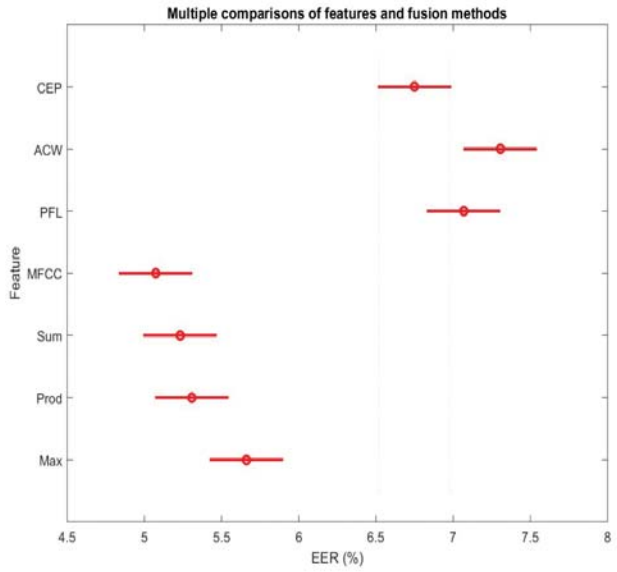

Fig. 8. Speaker verification: Comparison of the features (G.723.1).

95% confidence interval has considerable overlap with the product fusion and partial overlap with the MFCC. Due to the interaction of the feature and method, the best performance (average EER of 3.13%) is obtained using the McCree technique and the affine transform in conjunction with sum fusion.

3) GSM-AMR: Figure 11 and 12 show the results. The best method is using only the affine transform. The best features are the MFCC, sum fusion and product fusion. Due to interaction, the three best approaches are MFCC with McCree (3.29%), sum fusion with affine (3.26%) and product fusion with affine (3.24%). All three are statistically indistinguishable.

C. Comparison with Testing on Clean Speech

In the case of testing on clean speech, neither signal nor feature enhancement is necessary. Also, there is no statistical difference among the features and fusion methods for both SI and SV systems. The purpose is to compare the performance of the best approaches for each speech coder with the performance on clean speech.

Table III gives the average ISR comparisons for the SI case. There are two approaches that achieve the best average ISR for the G.729 codec. The MFCC feature is selected as the benchmark for clean speech as it achieves the highest average ISR. The best approach for each codec is individually compared to the test case of clean speech only. Therefore, a two sample statistical t-test with a 5% significance level

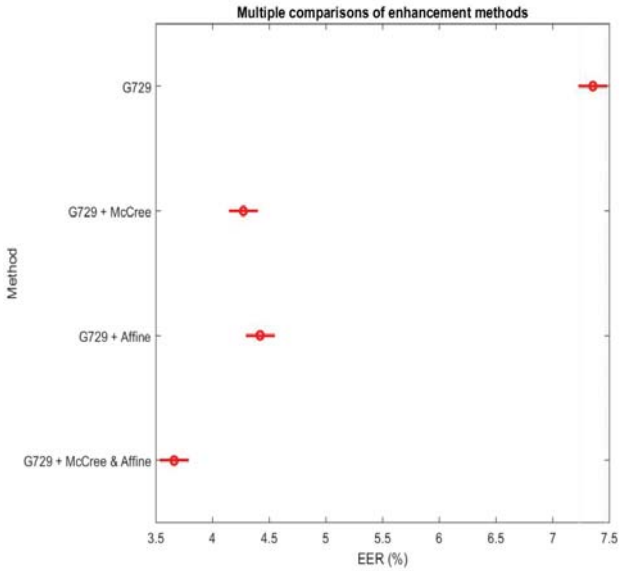

Fig. 9. Speaker verification: Comparison of the methods (G.729).

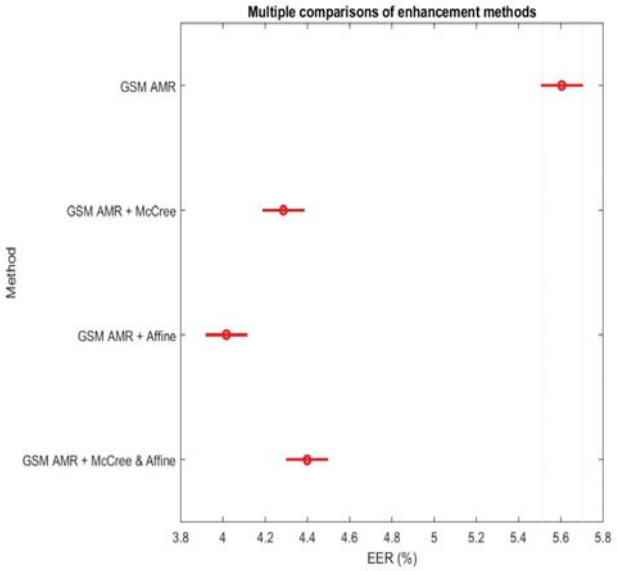

Fig. 11. Speaker verification: Comparison of the methods (GSM-AMR).

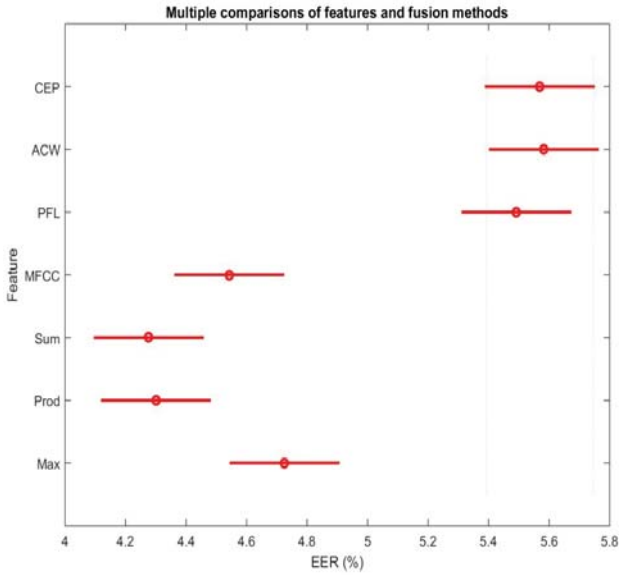

Fig. 10. Speaker verification: Comparison of the features (G.729).

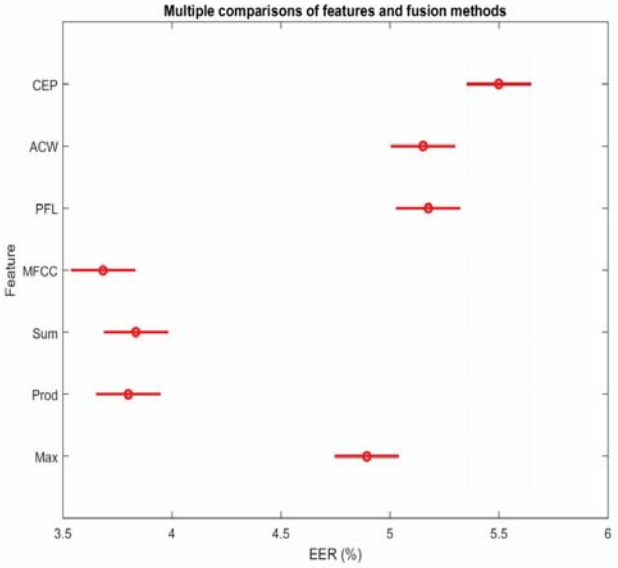

Fig. 12. Speaker verification: Comparison of the features (GSM-AMR).

and unequal variances is performed to determine if the performance on clean speech is significantly better than the technique used for each codec. The test is based on the 10 trials that are performed for each experiment. Table III gives the obtained p-values. Although the methods have mitigated the train/test mismatch and led to a substantial performance improvement, the low p-values indicate that the ISR values are not statistically comparable to that of clean speech.

Table IV gives the average EER comparisons for the SV case. Product fusion is selected as the benchmark for clean speech as it achieves the lowest average EER. Again, the best approach for each codec is individually compared to the test case of clean speech only using a two sample statistical t-test with a 5% significance level and unequal variances. Again, the methods mitigate the train/test mismatch but are not statistically comparable to that of clean speech.

V. Summary and Conclusions

Both the signal (McCree method) and feature (affine transform) enhancement strategies are highly useful in improving the performance of SI and SV systems that are trained on clean speech and tested on the decoded speech. For the G.723.1 and G.729 codec at the relatively lower bit rates, the combination of the McCree technique and the affine transform in conjunction with a fusion strategy is generally the best approach. For the higher bit rate 12.2 kb/s GSM-AMR coder, using only the affine transform with a fusion strategy is the best approach.

Test Speech	Approach	ISR	p-Value
Clean	MFCC	95.4	
G.723.1	McCree and Affine, Borda Count	87.9	1.6e-07
G.729	McCree and Affine, Borda Count	91.1	6.4e-05
G.729	McCree with MFCC	91.1	1.06e-04
GSM-AMR	Affine transform, Decision Level	89.8	1.52e-07

TABLE III
IDENTIFICATION SUCCESS RATE (%) FOR COMPARISON WITH CLEAN SPEECH

Test Speech	Approach	EER	p-Value
Clean	Product Fusion	2.77	
G.723.1	McCree and Affine, Sum Fusion	4.10	2.3e-05
G.729	McCree and Affine, Sum Fusion	3.13	0.02
GSM-AMR	Affine transform, Product Fusion	3.24	9.9e-04

TABLE IV
EQUAL ERROR RATE (%) FOR COMPARISON WITH CLEAN SPEECH

VI. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation through TUES Type 2 Collaborative Research Grants DUE-1122296 and DUE-1122299 awarded to Rowan University and Tennessee State University respectively.

VII. REFERENCES

1) A. K. Vuppala, K. S. Rao and S. Chakrabarti, "Effect of Speech Coding on Speaker Identification", *Annual IEEE India Conference*, Kolkata, India, December 2010.
2) H. Beigi, *Fundamentals of Speaker Recognition*, Springer, 2011.
3) R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues", *IEEE Circuits and Systems Magazine*, pp. 23–61, June 2011.
4) A. Fazel and S. Chakrabartty, "An overview of statistical pattern recognition techniques for speaker verification" *IEEE Circuits and Systems Magazine*, pp. 62–81, June 2011.
5) T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Communication*, vol. 52, pp. 12–40, 2010.
6) K. Raval, R. P. Ramachandran, S. S. Shetty and B. Y. Smolenski, "Feature and Signal Enhancement for Robust Speaker Identification of G.729 Decoded Speech", *International Conference On Neural Information Processing*, Doha, Qatar, pp. 345–352, November 2012.
7) A. Moreno-Daniel, B.-H. Juang, J. A. Nolazco-Flores, "Robustness of bit-stream based features for speaker verification", *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. I-749–I-752, 2005.
8) A. McCree, "Reducing speech coding distortion for speaker identification", *IEEE Int. Conf. on Spoken Language Processing*, 2006.
9) J. Vicente-Pena, A. Gallardo-Antoln, C. Pelaez-Moreno and F. Daz-de-Mara, "Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition", *Speech Communication*, Vol. 48, pp. 1379–1398, 2006.
10) A. M. Gomez, A. M. Peinado, V. Sanchez and A. J. Rubio, "Recognition of coded speech transmitted over wireless channels", *IEEE Transactions on Wireless Communication*, Vol. 5, No. 9, pp. 2555–2562, September 2006.
11) T. Ogunfunmi and M. J. Narasimha, "Speech Over VoIP Networks: Advanced Signal Processing and System Implementation", *IEEE Circuits and Systems Magazine*, pp. 35–55, 2012.
12) "ITU-T: Recommendation G.723.1 - Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s", 1996.
13) P. Kabal, "ITU-T G.723.1 Speech Coder: A Matlab Implementation", Telecommunications and Signal Processing Laboratory, McGill University, 2004.
14) "ITU-T: Recommendation G.729 - coding of speech at 8 kbit/s using conjugate-structure algebraic-code-exited linear prediction (CS-ACELP)", 2007.
15) "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory speech CODEC speech processing functions; AMR speech CODEC; General description", 2012.
16) T. Hasan and J. H. L. Hansen, "A study on universal background model training in speaker verification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19, No. 7, pp. 1890–1899, September 2011.
17) 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.
18) C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
19) I. T. Nabney, *NETLAB: Algorithms for Pattern Recognition*, Springer, 2002.
20) M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 3, pp. 260–267, May 1998.
21) R. Polikar, "Ensemble based systems in decision making", *IEEE Circuits and Systems Magazine*, Vol. 6, No. 3, pp. 21–45, 2006.
22) F. Rastoceanu and M. Lazar, "Score fusion methods for text-independent speaker verification applications", *6th Conference on Speech Technology and Human Computer Dialogue*, 2011.
23) J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Brooks/Cole Cengage Learning, 2012.