

Histogram Equalization for Robust Speaker Verification in Telephone Environments

♦Marshalleno Skosan and Daniel Mashao

Department of Electrical Engineering, University of Cape Town
Rondebosch, Cape Town, South Africa

♦Tel.: +27-21-650-2813 Fax: +27-21-650-3465

♦mksosan@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Abstract—While it is common for speaker recognition systems to perform well in ideal conditions, performance degrades when these systems are exposed to adverse conditions. This degradation in performance becomes more evident when speaker recognition systems are trained and tested in different recording conditions. For the success of the technology, speaker recognition needs to perform reliably regardless of the conditions under which training and testing is done. This paper is aimed at mitigating the problem of mismatched training and test conditions by using a technique known as Histogram Equalization. Here, it is used to improve the robustness of a speaker verification system evaluated on the entire NIST 2000 database. This database contains the speech of more than 1000 speakers which has been degraded by telephone transmission. Histogram Equalization is applied directly to the features extracted from a particular speaker's training and test speech. In so doing, it modifies the underlying feature distributions such that they become less environment-dependent and more consistent across different recording conditions. The technique is shown to lead to a relative improvement in the equal error rate of a speaker verification system employing cepstral mean normalization of more than 11%.

Index terms—Histogram Equalization, speaker verification

I. INTRODUCTION

The potential for the application of speaker recognition technology exists anytime speakers are unknown and their identities are required. For example, it can be used to control access to restricted sites and confidential information on computer networks; authenticate financial transactions over the telephone such as telephone banking and remote credit card purchases; identify individuals who make threats over the telephone; monitor criminals place on parole; browse voicemail for messages from specific individuals or it can be used in meetings and conferences to segment and store the speech from different individuals. Many of these applications require speaker recognition to be performed over the telephone. Thus, the ability to reliably recognize individuals over the telephone has great commercial potential. However, speech collected over telephone networks is subject to distortions caused by various telephone handset microphones; unpredictable levels of noise in the background or on the line; as well as different transmission channels and network types (mobile versus fixed-line networks for example). The bandwidths of telephone channels as well as

the various compression techniques in use, also distort speech signals during telephone transmission. The growth in the use of mobile devices also means that speaker recognition systems can be expected to operate in several uncontrolled environments (e.g., in crowded shopping malls, in cars or in rooms with inconsistent or poor acoustics).

When the recording equipment and environment vary between recording sessions, speech is acquired in what is known as *mismatched conditions*. Mismatched conditions generally lead to an acoustic mismatch between the speech acquired during training and testing. Statistical speaker modeling strategies, like Gaussian mixture models [1] for example, are aimed at modeling the underlying distribution of feature vectors extracted from a particular speaker's speech during training. During testing, speakers are classified according to the statistical similarity between the features extracted from their speech and the speaker models generated during training. This approach is based on the implicit assumption that for the same speaker, the features extracted from his/her will speech have similar statistical properties. A mismatch between the statistical properties of the training and test speech however, to some extent violates this assumption and leads to a deterioration in classification performance. This work is aimed at improving the robustness of a speaker verification system operating in mismatched telephone environments. Section II provides a general overview of speaker verification technology after which Section III covers previous work involving Histogram Equalization. Section IV covers the mathematical formulation of Histogram Equalization and other compensation techniques. The experimental framework for evaluating these techniques is then described in Section V. An analysis of the obtained results is conducted in Section VI.

II. A BRIEF OVERVIEW OF SPEAKER VERIFICATION

Fundamentally, a speaker verification system makes a 2-class decision. That is, to either accept or reject the current identity claim.

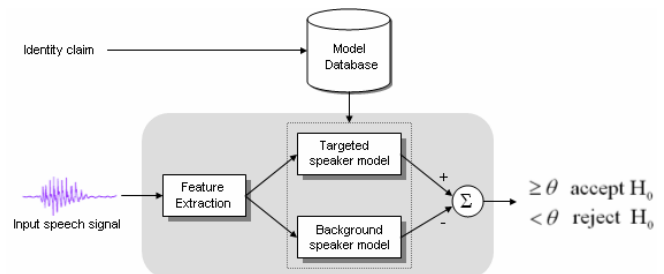


Figure 1: A typical speaker verification system

In contemporary speaker verification systems, the computer must decide whether the input speech signal better matches a model of the targeted speaker or a background model of non-claimant speakers (imposters) [1]. This process is illustrated in Fig. 1. Features generated by the feature extraction unit are compared to the targeted speaker model and to the background model. Following this a likelihood ratio statistic $L(X)$ is computed as the ratio (or difference in the log domain) of these scores. This value is then compared to a decision threshold θ to determine whether to accept or reject the current identity claim. A speaker verification system can make two types of errors, i.e. it can falsely accept imposters or falsely reject legitimate speakers. In practice, a detection error tradeoff (DET) curve [2] is used to illustrate the tradeoff between the false accept rate and false reject rate as the decision threshold is adjusted. The equal error rate (**EER**) is the point on a DET curve where the false accept rate equals the false reject rate and is often used and as a single performance indicator for these two types of error.

III. PREVIOUS WORK

In previous contributions we showed that a feature-based compensation technique, known as *Histogram Equalization* (HEQ), could be used to minimize the mismatch between feature distributions caused by mismatched telephone environments [3, 4]. It does this by non-linearly transforming the characteristics (i.e., the scale, shape and location) of one probability distribution to that of another such that their statistical properties (i.e., the mean, variance, skew) match. The technique was originally used to alleviate brightness and contrast alterations in digital images [5], but has in recent years also successfully been applied to improve the robustness of speech recognition systems in adverse environments [6, 7]. In [3] and [4] HEQ was used to warp the features extracted from an utterance such that their statistics conform to that of a Gaussian distribution with zero mean and unity variance across all recording conditions. In so doing, the statistical mismatch between the training and test feature distributions was reduced, which led to large improvements in performance. In [3] and [4] HEQ was also shown to outperform other compensation techniques namely, cepstral mean normalization (CMN) [8] and mean and variance normalization (MVN) [9]. Its use was motivated by the fact that feature-based compensation techniques that normalize the first and second moments of feature distributions, like CMN and MVN for example, have been shown to be effective in improving speaker verification performance in adverse environments [8, 9]. However, CMN and MVN are linear techniques which limits their ability to compensate for non-linear distortions of the feature space (such as those caused by additive noise for example). The non-linear compensation provided by HEQ however, can be used to not only normalize the first two moments of feature distributions, but all the other moments as well.

The limitations of the work done in [3] and [4] are as follows. In [3] we evaluated HEQ on a speaker identification task using the NTIMIT database [10]. This database contains phonetically rich speech that was captured in a sound booth during a single session. The speech was then transmitted through a carbon-button telephone handset and recorded over local and long distance telephone loops. Although HEQ was shown to

outperform CMN and MVN, the effect of conversational speech, different telephone handsets and various periods of intersession could not be evaluated using the NTIMIT database. In [4] we evaluated the performance of CMN, MVN and HEQ on a subset of the NIST 2000 speaker recognition evaluation database [11] which only included male speakers using telephone handsets with electret microphones. In this work we evaluate CMN, MVN and HEQ on the entire NIST 2000 database. This evaluation is more realistic as it includes the speech from both male and female speakers as well as the use of telephone handsets employing both electret and carbon-button microphones.

IV. HISTOGRAM EQUALIZATION & OTHER TECHNIQUES

In many pattern recognition tasks, improvements in performance can be expected if one reduces the mismatch between training and testing conditions. As far as speaker recognition systems are concerned this mismatch can to a large extent be attributed to varying ambient conditions, speech acquisition equipment and transmission channels. One way of reducing this mismatch is by defining transformations that normalize feature distributions obtained during the training and testing of a speaker recognition system. Two such transformations are cepstral mean normalization (CMN) and mean and variance normalization (MVN). CMN is a channel compensation technique that has successfully been used to reduce the convolutional effects of telephone channels on input speech signals [8]. CMN, however, also has the dual effect of normalizing the mean of each speaker's training and test data distributions. It does this by using the following transformation

$$T(x) = x - \mu_x. \quad (1)$$

MVN, on the other hand, uses the transformation given in Equation (2) to normalize not only the means but, the variances of these distributions as well [9]

$$T(x) = \frac{x - \mu_x}{\sigma_x}. \quad (2)$$

In Equations (1) and (2), μ_x is the global mean of the variable x for a particular utterance, whereas σ_x is the standard deviation and y is the compensated version of x .

As mentioned previously, HEQ provides a transformation that allows for the conversion of one probability distribution to another. It does this by matching the cumulative distributions of a reference distribution and that of the variable to be transformed. This is accomplished as follows [12]: Let x be a random variable with a probability distribution $p_x(x)$, and let $y = T(x)$ be a single-valued and monotonically increasing transformation that converts the probability distribution $p_x(x)$ into a reference probability distribution, $p_{ref}(y)$. In so doing, $T(x)$ makes the probability of finding x in the differential range dx equal to the probability of finding y in the differential range dy , i.e.

$$p_{ref}(y)dy = p_x(x)dx. \quad (3)$$

Thus, the transformation $y = T(x)$ modifies the original probability distribution $p_x(x)$ according to the expression:

$$p_{ref}(y) = p_x(x) \frac{dx}{dy} = p(G(y)) \frac{dG(y)}{dy}, \quad (4)$$

where $G(y) = x$ is the inverse of $T(x)$. Using Equation (4), the relationship between the cumulative distribution functions (CDFs) associated with $p_x(x)$ and $p_{ref}(y)$ is as follows:

$$\begin{aligned} C_x(x) &= \int_{-\infty}^x p_x(x') dx' = \int_{-\infty}^{T(x)} p_x(G(y')) \frac{dG(y')}{dy'} dy' \\ &= \int_{-\infty}^y p_{ref}(y') dy' = C_{ref}(y) = C_{ref}(T(x)). \end{aligned} \quad (5)$$

Thus, the transformation $T(x)$, that converts $p_x(x)$ into $p_{ref}(y)$, is given by:

$$T(x) = C_{ref}^{-1}(C_x(x)), \quad (6)$$

where C_{ref}^{-1} is the inverse of the CDF of the reference probability distribution.

For practical implementations only a finite number of observations are available. As a result, cumulative histograms instead of cumulative probabilities are used. This is the reason that the transformation is called Histogram Equalization and not probability distribution equalization. The transformation given by Equation (6) cannot however easily be applied to the multi-dimensional feature vectors obtained from the feature extraction module of a speaker recognition system. For this reason, it is assumed that all the dimensions of the feature vectors are independent. Under this simplifying assumption, the transformation can be applied to each feature vector component independently. A graphical illustration of the cumulative distribution matching performed by HEQ is depicted in Fig. 2. It shows how the cumulative histogram of the original variable x and the reference cumulative histogram can be used to perform the transformation. Here, each value of x is replaced the value of y that corresponds to the same point in the reference cumulative histogram. This illustration shows that HEQ is computationally attractive as it can be implemented using a simple look-up table.

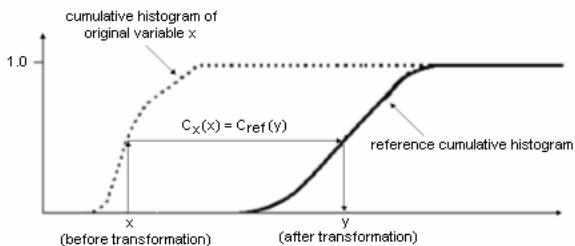


Figure 2: The cumulative distribution matching performed by HEQ

V. EXPERIMENTAL SETUP

For the experimental work done in this study, the database used in the National Institute of Standards and Technology (NIST) 2000 speaker recognition evaluation [11] was used. The database includes data from 1003 (546 female and 457 male) speakers and requires the evaluation of 6096 legitimate speaker trials and 60476 impostor trials (a total of 66572 verification

trials). The database includes conversational telephone-quality speech taken from the Switchboard 2 corpus. All test segments are recorded from calls made from a telephone number that is different from the ones used to collect the training speech. Therefore, all test utterances may be considered to be collected using a handset different than the one used for training the speaker model. Each speaker is trained using a single two minute session, while testing utterances range between few seconds and a minute (with a primary focus on utterances with a varying between 15 and 45 seconds). Since the data is collected in what is referred to as telephone environments, “the challenges presented by this data include limited bandwidth, channel noise from various sources, the use of different microphones, recordings from different locations, and recordings collected over a period of time” [13]. All these factors contribute to mismatched training and test conditions.

According to Przybocki and Martin [13], the performance range for ten speaker verification systems evaluated on the NIST 2000 database was as follows. Under the conditions that (1) only male verification trials where the test segment duration is in the 15 to 45 second range are used; and (2) the test segment and targeted speaker training data come from conversation sides that employ electret-type microphones in the telephone handset, the equal error rate (EER) varies between 7% and 18%. According to reference [13], verification trials involving females only resulted in slightly poorer results. The results reported on in [13] will be used as benchmark for performance on the NIST database. The set of tests conducted in this work however, considers the entire NIST 2000 corpus (both male and female data including both carbon-button and electret handsets) and as such is a considerably more difficult task. In the following sections we discuss the baseline speaker verification system used for evaluating the different compensation techniques.

A. Feature Extraction

The signal processing involved in the extraction of features was as follows: The speech signal was first filtered with a pre-emphasis filter of the form $H(z) = 1 - 0.97z^{-1}$ and partitioned into 25 millisecond frames at a frame rate of 80 Hz. These frames were then multiplied with a Hamming window to minimize signal discontinuities at the start and end of each frame. The frames were then passed through a voice activity detector (VAD) to eliminate all frames primarily containing silence, noise or unvoiced speech. The VAD was implemented as a simple energy-based detector that discarded all frames below a specified energy level. About 30% of all frames were discarded. From the remaining frames the extraction of mel-frequency cepstral coefficients (MFCCs) took place as follows: Each frame was first Fourier transformed into the frequency domain. The squared magnitude spectrum of each frame was then filtered by a bank of 26 mel-scaled triangular filters distributed over a bandwidth of 240-3480 Hz (which is approximately the bandwidth of the telephone channel). The logarithm of the filterbank outputs was then cosine transformed into 18-dimensional MFCC feature vectors. The procedure for extracting MFCCs was kept identical for all the tested compensation techniques.

B. Speaker Modeling

As for the feature extraction portion of the speaker verification system, the following model settings were kept identical for all the tested compensation techniques. As in [1], speaker modeling involved a two step process; a general Universal Background Model (UBM) was trained on a large quantity of exclusive speech, and a target speaker

model was then formed by adapting the parameters of the UBM. The UBM was comprised of a Gaussian Mixture Model (GMM) [1] which models the Probability Density Function of a collection of multi-dimensional feature vectors (like the MFCCs used in work). It does this using a composition of multi-dimensional Gaussians mixtures.

Four different UBMs were trained for the system - the UBMs were both gender- and handset specific (carbon-button or electret). The UBMs consisted of 128 mixtures and were trained using the entire test portion of the telephony database used in the NIST 1999 speaker recognition evaluation (about two hours of data for each model). The UBMs were trained using the Distance-based GMM procedure [14]. That is, the data was clustered using the K-means algorithm, and the GMM parameters were calculated as follows: the weights were given by the relative number of vectors in each cluster, and the means and variances were the sample mean and variance of each cluster. No iterations of the Expectation-maximization algorithm were performed. Once the UBMs were obtained, speaker models were formed by adjusting the UBM parameters by Bayesian Adaptation [1]. The choice of UBM to use for adaptation depended on the type of handset employed in the training session and the claimant speaker's gender. Only the means of the UBM mixtures were adapted.

C. Decision-making

As mentioned previously, a speaker verification system needs to make a 2-class decision. That is, to either accept or reject the current identity claim. The system must decide whether the input speech signal better matches a model of the targeted speaker or a background model of non-claimant speakers, i.e. the UBM. For each verification trial the features extracted from the test segment are compared to a targeted speaker model and to a UBM. Following this a likelihood ratio statistic is computed as the ratio (or difference in the log domain) of the scores obtained. This value is then compared to a decision threshold to determine whether to accept or reject the current identity claim. The likelihood ratio of a speaker model and UBM pair was approximated using the five highest scoring Gaussian components as described in [1].

D. Score Normalization

Before reporting the final performance of the speaker verification system, the scores obtained for the verification trials were pooled and the decision threshold was varied so as to obtain the full range of operating points for the system. However, to eliminate handset- and environment-dependent biases and scales in the scores, Test normalization (T-norm) [15] was used. T-norm uses the mean and standard deviation of the scores derived from testing a particular test segment against a set of standard models to adjust the score obtained when a targeted speaker model is tested with the same utterance. In so doing, T-norm normalizes score distributions.

E. HEQ Implementation

This section provides a simple algorithm for directly implementing the HEQ technique described in Section IV. Similar algorithms can be found in [6] and [16]. As mentioned previously, HEQ is applied separately to the distribution of each

feature vector component extracted from the utterance under consideration. As such, the following algorithm is applied to each feature vector component independently (subscripts are dropped for ease of notation).

The goal of HEQ is to modify the feature distributions obtained during training and testing such that their characteristics are similar to that of a reference distribution. Thus, the first step in performing HEQ involves selecting a reference distribution. Once a suitable reference distribution, $p_{ref}(y)$, has been selected and its cumulative histogram, $C_{ref}(y)$, has been computed, HEQ can be applied to the training and test feature distributions of each speaker as follows:

1. Determine the maximum and minimum values, x_{max} and x_{min} , across the entire set of observations (i.e. across all the observations of a particular feature vector component).
2. Divide the range $[x_{max}, x_{min}]$ into M equally-spaced non-overlapping bins (or intervals) B_i , where $x_{min} = b_1 < b_2 < \dots < b_{M+1} = x_{max}$ and bin $B_i = [b_i, b_{i+1})$.
3. Using these bins, construct a histogram of the observations in the set. This is done by scanning the set and counting the number of observations that fall into each bin.
4. Compute the normalized version of the histogram obtained after step (3) by using the following equation:

$$p_x(x \in B_i) = \frac{n_i}{N_x}, \quad (7)$$

where n_i is the number of observations in bin B_i and N_x is the total number of observations in the set. Equation (7) in effect approximates the probability of x being in bin B_i .

5. Compute the cumulative histogram of the set using the normalized histogram constructed in step (4) such that:

$$C_x(x : x \in B_i) = \sum_{j=1}^i \frac{n_j}{N_x}. \quad (8)$$

Equation (8) is a piecewise constant function approximation of the true cumulative distribution function.

6. Replace each value of x by the value of y that corresponds to the same point in the reference and computed cumulative histograms such that $C_x(x) = C_{ref}(y)$. This is in direct correspondence to Equation (6).

To efficiently implement step (6), one could construct two look-up tables $\{x, C\}$ and $\{y, C_{ref}\}$ from $C_x(x)$ and $C_{ref}(y)$ respectively, such that they take on values in the range $[0,1]$ in equal increments. This allows one to combine the two tables such that a new table $\{x, y\}$, which is a piecewise constant approximation of the true transformation, is formed [16]. Then, for every value of x , the closest value of y can be found by using a binary search. This value can then used as the histogram equalized version of x .

F. HEQ Algorithm Verification

In order to confirm that the HEQ algorithm was indeed implemented correctly we applied the technique to MFCCs extracted from replicas of the test utterance "she had your dark suit in greasy wash water all year", taken from the TIMIT and NTIMIT databases¹. These databases contain the same utterances but, recorded under different conditions. For the TIMIT database, all utterances are obtained in noise-free recording conditions, whereas for the NTIMIT database, all ut-

¹ The utterance was spoken by the same speaker.

terances are transmitted through a carbon-button telephone handset and recorded over local and long-distance telephone loops [10]. Figs. 3(a) and (b) shows the histograms of the first component of the MFCC feature vectors (hereafter referred to as MFCC₁) extracted from the test utterance for both the TIMIT and NTIMIT databases².

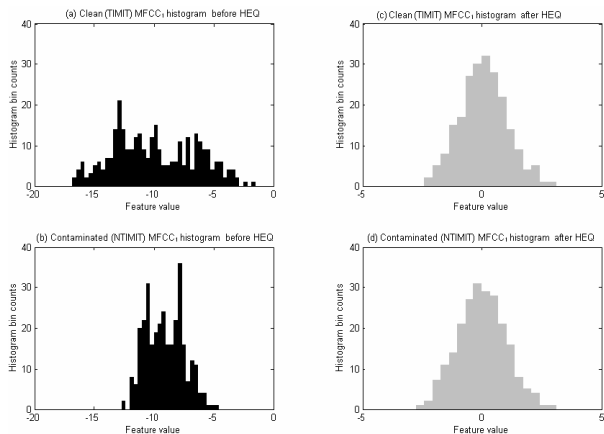


Figure 3: MFCC₁ histograms before and after the application of HEQ

As illustrated, the two histograms differ in their shape, scale, spread and location. The MFCC₁ histogram extracted from the NTIMIT version of the test utterance exhibits a shift in its mean and a reduction in variance when compared to the MFCC₁ histogram extracted from the TIMIT version of the test utterance. These distortions can be attributed to the linear filtering property of telephone channels and to additive noises encountered in the telephone network [17]. In Figs. 3(c) and (d) HEQ was used to equalize the two histograms so as to compensate for the degradations caused by telephone transmission. The reference distribution was chosen to be Gaussian with zero mean and unity variance. The cumulative histograms of the clean and contaminated MFCC₁ distributions were estimated according to steps (1) to (5) of the HEQ algorithm described in part E of Section IV. All histograms were estimated using 100 uniformly spaced intervals between the minimum and maximum values of the respective MFCC₁ values. From Figs. 3(c) and (d) it is clear that the clean and contaminated MFCC₁ histograms appear to be more alike in terms of their overall shape, scale, spread and location and that the two histograms are also very similar to a Gaussian distribution with zero mean and unity variance. Thus, Fig. 3 verifies that HEQ was indeed implemented correctly.

VI. RESULTS AND DISCUSSION

This section compares the performance of HEQ to CMN and MVN. In the experiments performed in [3] and [4] HEQ was shown to outperform CMN and MVN and, MVN was shown to outperform CMN. The purpose of this section is to determine whether these observations apply to under more challenging conditions of mismatch. CMN, MVN and HEQ were each ap-

² Similar plots were obtained for the other components of the extracted MFCC feature vectors.

plied separately to the baseline system. These feature-based compensation techniques were applied utterance-wise with the distribution of each MFCC feature vector component being processed separately. HEQ was implemented according to the algorithm described in part E of Section IV. A Gaussian distribution with zero mean and unity variance was used as the reference distribution for the technique and, all histograms were estimated using 250 equally spaced bins. The following results were obtained when CMN, MVN and HEQ were each applied to the MFCCs produced during the feature extraction phase. All the trials specified for the NIST 2000 database were performed and the results obtained are shown in Table 1. All the results were averaged over three runs. For comparative purposes, the performance of the baseline speaker verification system without any compensation technique applied is also shown.

Table 1: Performance of the baseline system with different feature-based compensation techniques (average \pm standard deviation)

| Feature-based compensation technique | EER |
|--------------------------------------|-------------------|
| No compensation | 26.84 \pm 0.01% |
| CMN | 14.87 \pm 0.11% |
| MVN | 13.56 \pm 0.06% |
| HEQ | 13.14 \pm 0.04% |

From the results tabulated in Table 1, it is clear that HEQ reduces the EER by 11.63% relative to the EER obtained when CMN was used and by 3.10% when MVN was used. Fig. 4 shows the performance of CMN, MVN and HEQ using DET curves. From this diagram it is clear that HEQ outperforms the other two techniques across all operating points. Thus, the progressive compensation of higher order moments of the feature distributions results in better speaker recognition performance.

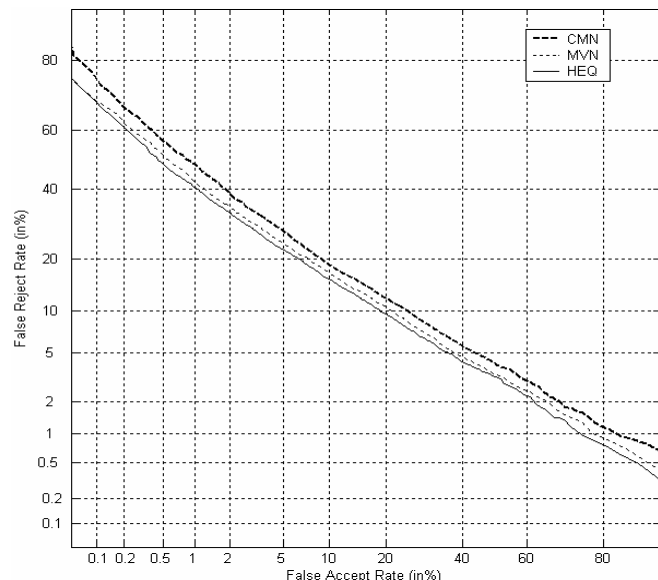


Figure 4: DET curves for the baseline system with different feature-based compensation techniques

As mentioned previously, the improvement in performance is primarily due to HEQ's ability to compensate for non-linear as well as linear distortions of the feature space. In so doing, it normalizes the shape, scale, spread and location of feature distributions. This consolidates the knowledge that HEQ outperforms linear techniques such as CMN and MVN. Using McNemar's test [18], all the improvements were found to be statistically significant as there was a greater than 95% chance that HEQ was better than MVN and that MVN was better than CMN.

The results show that the HEQ technique generalizes well to more challenging conditions of mismatch encountered in telephone networks. Table 1 also shows that with no compensation, the baseline system obtained an EER of 26.84%. When these results are compared to those obtained for CMN, MVN and HEQ it is clear that:

- a. Feature-based compensation is a crucial step in obtaining good performance in adverse environments. This is primarily due to the vulnerability of MFCCs when exposed to additive noise and linear filtering effects encountered in telephone networks.
- b. For the NIST 2000 database, the largest improvement in performance, when using feature-based compensation, is due normalization of the mean of the feature distributions. Normalization of other moments of the feature distributions leads to marginal, albeit statistically significant, improvements in performance. This result makes sense, as for the NIST 2000 database, the speech data is primarily degraded by linear filtering effects due to transmission by telephone.

VII. CONCLUSION

This study has shown that HEQ generalizes well to challenging conditions of mismatch encountered in telephone networks and as such can be used to make robust speaker recognition over telephone networks a reality in the near future. The results have also confirmed the knowledge that the progressive compensation of higher order moments of the feature distributions improves speaker verification performance.

REFERENCES

- [1] 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.
- [2] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of EuroSpeech 1997*, pp. 1895-1898, 1997.
- [3] M. Skosan and D. J. Mashao, "Improving speaker identification performance for telephone-based applications," in *Proceedings of SATNAC 2004*, 2004.
- [4] M. Skosan and D. J. Mashao, "Matching feature distributions for robust speaker verification," in *Proceedings of PRASA 2004*, pp. 93-97, 2004.
- [5] J. Matthews, "Histogram Equalization," 2004, [online] Available At: <http://www.generation5.org/content/2004/histogramEqualization.asp> [Last Accessed: 31/01/2005]
- [6] S. Molau, D. Keysers, and H. Ney, "Matching training and test data distributions for robust speech recognition," *Speech Communication*, vol. 41, pp. 579-601, 2003.
- [7] A. de la Torre, J. C. Segura, M. C. Benítez, A. M. Peinado, and A. J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proceedings of IEEE ICASSP '02*, pp. 401-404, 2002.
- [8] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 554-568, 1999.
- [9] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Digital Signal Processing*, vol. 10, pp. 113-132, 2000.
- [10] J. Campbell and D.A. Reynolds, "Corpora for the Evaluation of Speaker Recognition Systems", *Proceedings of IEEE ICASSP*, pp. 2247-2250, 1999.
- [11] "The NIST 2000 speaker recognition evaluation," [online] Available At: <http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm> [Last Accessed: 31/01/2005]
- [12] A. de la Torre, A. M. Peinado, J. C. Segura, M. C. Benítez, and A. J. Rubio, "Histogram equalization of the speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio processing*, (Accepted for publication)
- [13] M. Przybocki and A. Martin, "Odyssey text-independent evaluation data," in *Proceedings of ODYSSEY 2001*, pp. 21-24, 2001.
- [14] R. D. Zilca and Y. Bistriz, "Distance-based Gaussian mixture model for speaker recognition over the telephone," in *Proceedings of ICSLP '00*, pp. 1001-1003, 2000
- [15] R. Aukenthaler, M. Carey and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification system," *Digital Signal Processing*, vol. 10, pp. 42-54, 2000.
- [16] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proceedings of ICSLP '00*, pp. 556-559, 2000
- [17] P. J. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proceedings of IEEE ICASSP '94*, vol. 1, pp. 109-112, 1994.
- [18] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, pp. 1895-1923, 1998

M. Skosan completed his BSc (Elec. Eng.) in December 2003 and his MSc (Elec. Eng.) in June 2005. He is currently a researcher in the Speech Technology and Research Group at the University of Cape Town.
Dr. D. Mashao is a senior lecturer at the University of Cape Town and head of the Speech Technology and Research Group.