Matching Feature Distributions for Robust Speaker Verification

Marshalleno Skosan, Daniel Mashao
Department of Electrical Engineering, University of Cape Town
Rondebosch, Cape Town, South Africa
mskosan@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Abstract—**In this work we improve the performance of a speaker verification system by matching the feature vector distributions obtained when training and testing the system. In particular, we perform experiments using speech that has been degraded by telephone transmission. Speaker Verification experiments are performed on the NIST 2000 database. Significant improvements, above the baseline, are reported.**

***Index Terms*—Speaker verification, Histogram Equalization, Gaussian mixture models**

1. Introduction

SPEAKER verification (**SV**) is concerned with verifying that an individual is who he/she claims to be. In ideal conditions speaker verification systems perform extremely well. However, as soon as these systems are exposed to real-world conditions, their performances degrade considerably [4]. From a statistical point of view, these degradations in performance can be attributed to the mismatch between a particular speaker's training and testing data distributions caused by the exposure to real-world conditions. In this work, we improve SV performance by using a technique that has its origins in digital image processing. The technique is known as histogram equalization and is used here to optimally minimize the mismatch between training and testing distributions. Experiments are performed on the telephone degraded NIST 2000 speech database. Large improvements, above the baseline system, are reported. In addition, we show that histogram equalization outperforms two commonly used normalization techniques namely, cepstral mean normalization and mean and variance normalization.

2. An Overview Of Speaker Verification

There are many papers that provide extensive overviews of speaker recognition research (eg [1, 2, 3, 4]). This section summarizes some of the concepts discussed in these papers. Fundamentally, an SV system needs to make a 2-class decision. That is, to either accept or reject the current identity claim. Figure 1 depicts a typical SV system. Here the system must decide whether the input speech signal better matches a model of the claimed speaker or a background model of non-claimant speakers (imposters). Features extracted from the front-end processing unit are compared to the claimed speaker model and to the background model. Following this a likelihood ratio statistic $\Lambda(X)$ 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.

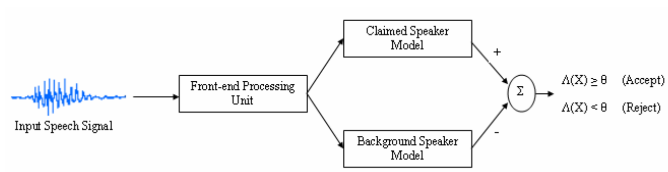

Figure 1: A typical speaker verification system

An SV system can make two types of errors, i.e. it can falsely accept imposters (**FA**) and falsely reject true identity claims (**FR**). In practice, a detection error tradeoff (**DET**) curve is used to illustrate the tradeoff between FA and FR errors as the decision threshold is adjusted. The equal error rate (**EER**) is the point on a DET curve where FA = FR and is used and as a single performance indicator for these two types of error. Another performance indicator that is often used in speaker verification research is the detection cost function (**DCF**) [2, 17]. The is DCF is the weighted arithmetic mean of the FA and FR rates and is defined as

$$DCF = C_{FR} \bullet P_{FR} \bullet P_{\text{true speaker}} + C_{FA} \bullet P_{FA} \bullet P_{\text{imposter}}$$

- Cost of a false reject - $C_{FR} = 10$
- Cost of a false accept - $C_{FA} = 1$
- A priori probability of a true speaker - $P_{\text{true speaker}} = 0.01$
- A priori probability of a false speaker - $P_{\text{imposter}} = 0.99$
- Probability of false accept - P_{FA}
- Probability of false reject - P_{FR}

The minimum value of the DCF is usually computed over all operating points (as the decision threshold is varied).

3. Histogram Equalization

In many pattern recognition tasks, improvements in performance can be expected if one reduces the mismatch between training and testing conditions. In speaker recognition (**SR**) systems this mismatch can to a large extent be attributed to varying ambient conditions, speech acquisition equipment and transmission channels [3]. One way of reducing this mismatch is by defining transformations that normalize feature distributions obtained during the training and testing of an SR 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 [5]. CMN however, also has the dual effect of normalizing the mean of each speaker's training and test data distributions [5]. It does this by using the following transformation

$$x_{new} = x_{old} - \mu_{x_{old}} \,. \tag{1}$$

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 [6]

$$x_{new} = \frac{x_{old} - \mu_{x_{old}}}{\sigma_{x_{old}}} \,. \tag{2}$$

In equations (1) and (2), $\mu_{x_{old}}$ is the global mean of the variable x_{old} for a particular utterance, whereas $\sigma_{x_{old}}$ is the standard deviation. However, these techniques are linear and can thus not adequately compensate for the non-linear effects caused by telephone transmission.

To this end, a technique known as Histogram Equalization (**HEQ**), which is used extensively in digital image processing [7] and, which has recently been applied to speech recognition with great success [8, 9], is applied in this research. The aim of HEQ is to completely match the distributions of the training and test data, not just the mean and/or variance (like CMN and MVN) [10]. It does this by non-linearly transforming the probability distribution of a particular speaker's feature vectors, obtained during training and testing, into a reference distribution.

The formulation of HEQ is as follows [11, 12, 13, 14]: Let x_0 be a one-dimensional variable with a probability distribution $p_0(x_0)$. Let $x_1 = T(x_0)$ be a single-valued and monotonically increasing transformation that converts the probability distribution $p_0(x_0)$ into a reference probability distribution $p_{ref}(x_1)$. In other words, it is a transformation that makes the probability of finding x_0 in a differential range dx_0 equal to the probability of finding x_1 in the corresponding range dx_1 i.e.

$$p_{ref}(x_1)dx_1 = p_0(x_0)dx_0 \tag{3}$$

Thus the transformation $x_1 = T(x_0)$ modifies the original probability distribution $p_0(x_0)$ according to the expression

$$p_{ref}(x_1) = p_0(x_0)\frac{dx_0}{dx_1} = p_0(G(x_1))\frac{dG(x_1)}{dx_1} \tag{4}$$

where $G(x_1)$ is the inverse transformation of $T(x_0)$.

Using equation (4), the relationship between the cumulative probabilities associated with $p_0(x_0)$ and $p_{ref}(x_1)$ is given by

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' \\ &= \int_{-\infty}^{T(x_0)} p_0(G(x_1'))\frac{dG(x_1)}{dx_1}dx_1' \\ &= \int_{-\infty}^{x_1} p_{ref}(x_1')dx_1' \\ &= C_{ref}(x_1) \\ &= C_{ref}(T(x_0)) \end{aligned} \tag{5}$$

Thus the transformation $T(x_0)$ can be obtained as

$$T(x_0) = C_{ref}^{-1}(C_0(x_0)) \tag{6}$$

where C_{ref}^{-1} is the inverse of the cumulative distribution function of the reference probability density function (**PDF**).

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 in equation (6) cannot however be easily be applied to the multi-dimensional feature vectors obtained from the signal processing front-end of speaker recognition systems. As a result, it is assumed that the all the dimensions of the feature space are independent. Under this simplifying assumption, the transformation can be applied to each feature space dimension independently. A graphical illustration of the transformation is depicted in the figure 2. It shows how the cumulative histograms of the original variable and the transformed variable (the reference cumulative histogram) can be used to perform the transformation. Here each test/training set value x_0 is replaced the value x_1 that corresponds to the same point in the reference cumulative histogram. This illustration shows that HEQ is computationally attractive as it can be implemented by using a simple look-up table.

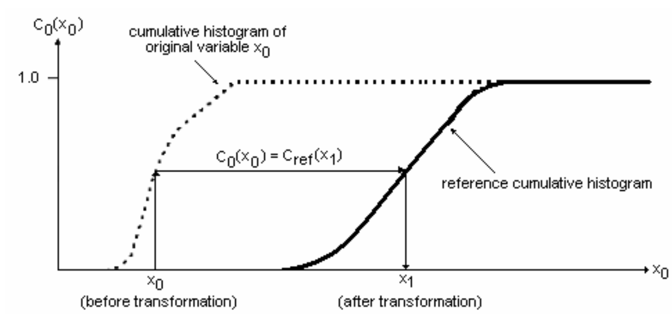

Figure 2: The histogram equalization transformation

4. THE SPEECH DATABASE

Moreno [15] states that both stationary and non-stationary noises can be encountered in a telephone network. Stationary noise appears in the form of low frequency tone-like signals, or white noise caused by thermal and other physical phenomena. He goes on to state that these single frequency noises can be produced by the harmonics power lines and by signaling tones that get transmitted by error through the telephone channel. Non-stationary noises on the other hand can be attributed to clicks and other transient phenomena caused by intermittent connections. As a result, evaluating histogram equalization on speech degraded by telephone transmission will give one a true idea of its ability to compensate for both linear and non-linear distortions.

In a previous contribution [20], we evaluated HEQ on a speaker identification task using the NTIMIT database. 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 [21]. Although HEQ was shown to outperform CMN and MVN, the effect of conversational-like speech, different telephone handsets and various periods of intersession could not be evaluated using this database.

In this work we evaluated the performance of CMN, MVN and HEQ on the NIST 2000 speaker recognition evaluation database [16, 17]. This database includes conversational telephone-quality speech taken from the Switchboard 2 corpus. The test segments are recorded from calls made from a telephone number that is different from the one used to enroll. Therefore, all test utterances may be considered to be collected using a different handset than the one used for training the speaker models. Each speaker model is trained using a single two minute session of speech, while testing utterances range between 15 and 45 seconds. This database allows one to evaluate speaker verification systems under very challenging real-world conditions as the speech, in addition to being degraded by telephone transmission, is also affected by the use of different handsets, different periods of intersession, conversational speech and different test segment lengths. We used this database to perform 1561 true speaker trials and 15501 imposter trials (all trials consisted of male speakers only).

5. EXPERIMENTAL RESULTS

5.1. The baseline system

In this work the front-end processing unit extracts mel-frequency cepstral coefficients (**MFCC**) from the input speech signal. These features are aimed at emulating the spectral compression applied by the human auditory system to an incoming speech signal [3]. MFCCs are spectrum-based features and are used here as a result of the speech spectrum having been shown to be very effective in speaker recognition (**SR**) research [2]. This is as a result of its ability to provide an adequate representation of an individual's vocal tract structure, which is one of the main speaker dependent characteristics that SR systems use to discriminate between speakers [1].

The MFCCs were generated as follows: the incoming speech signal was first multiplied by overlapping Hamming windows which divided it into a sequence of 20ms frames with an overlap of 10ms between frames. These speech frames were then Fourier transformed into the frequency domain where a sequence of log-magnitude spectra were computed. To obtain the mel-frequency cepstral coefficients, these log-magnitude spectra were filtered by a bank of mel-scaled triangular filters distributed over a bandwidth of 0Hz to 3800Hz. The outputs of the filterbank where then discrete cosine transformed into 30 dimensional feature vectors. In the subsequent experiments, CMN, MVN and HEQ were applied at this stage to modify the distributions of these feature vectors.

In order to model the distribution of feature vectors obtained for each speaker, we used Gaussian mixture models (**GMM**) [4, 18]. A GMM can be viewed as a non-parametric, multivariate PDF model that that is capable of modeling arbitrary distributions and is currently the most dominant method of modeling speakers in speaker recognition research. The GMM of the distribution of feature vectors for speaker S is a weighted linear combination of M unimodal Gaussian densities $b_i^s(\mathbf{x})$, each parameterized by a mean vector $\boldsymbol{\mu}_i^s$ and a covariance matrix Σ_i^s. These parameters are collectively represented by the notation

$$\lambda_s = \{p_i^s, \boldsymbol{\mu}_i^s, \Sigma_i^s\} \qquad \text{for} \quad i = 1, \ldots, M \tag{7}$$

where p_i^s are the mixture weights satisfying the constraint

$$\sum_{i=1}^{M} p_i^s = 1 \tag{8}$$

For a feature vector $\mathbf{x}$, the mixture density for speaker S is computed as

$$p(\mathbf{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(\mathbf{x}) \tag{9}$$

where

$$b_i^s(\mathbf{x}) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i^s \mid^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i^s)' \Sigma_i^s (\mathbf{x} - \boldsymbol{\mu}_i^s)\right) \tag{10}$$

Given a sequence of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, which are assumed to be independent, the log-likelihood of a speaker model λ_s is given by

$$L_s(X) = \log p(X \mid \lambda_s) = \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \lambda_s) \tag{11}$$

For speaker verification, equation (11) is computed for the claimed speaker model as well as for the background model of non-claimant speakers. The difference between these values is termed the likelihood ratio $\Lambda(X)$ and is subsequently compared to a threshold θ to determine whether to accept ($\Lambda(X) \geq \theta$) or reject ($\Lambda(X) < \theta$) the identity claim [4]. In this work we used GMMs with 64 mixtures to model each speaker. These GMMs were obtained from well-trained a background model with a form MAP adaptation according to the work done in [19].

5.2. The effect of CMN, MVN and HEQ

This section evaluates the performance of all the feature normalization techniques discussed in section 3. The various statistics for these techniques (such as the means, standard deviations, probability distributions and cumulative distributions) were estimated on an utterance by utterance basis.

Also, we chose a Gaussian PDF with zero mean and unity variance as the reference PDF for the HEQ technique. Table 1 displays the performance of CMN, MVN and HEQ on the male portion of NIST 2000 database.

Compensation Technique	Equal Error Rate	Relative Improvement	Minimum DCF
No compensation	31.35%	-	0.0843
CMN	24.57%	21.63%	0.0742
MVN	10.76%	65.68%	0.0403
HEQ	**10.16%**	**67.59%**	**0.0389**

Table 1: The effect of the feature normalization techniques

Table 1 clearly illustrates that HEQ performs better than both MVN and CMN but, that MVN outperforms CMN. This result is to be expected as HEQ can be viewed as an extension of MVN which, in turn, can be viewed as extension of CMN. This result emphasizes HEQ's ability to compensate for non-linear distortions of the probability distributions of the feature vectors (as discussed in section 3) which cannot be eliminated by linear methods such as MVN and CMN. However, from table 1 it can be seen that normalization of the variance of the training and testing distributions accounts for the largest improvement in performance and that normalization of other moments improves performance only slightly. The trend of the results obtained in this research corresponds to those reported in [9] and [10] which use CMN, MVN and HEQ to improve the performance of speech recognition systems in noisy environments. In figure 3 we show the significant improvements that can be obtained by minimizing the mismatch between training and testing distributions when speech is obtained in adverse environments.

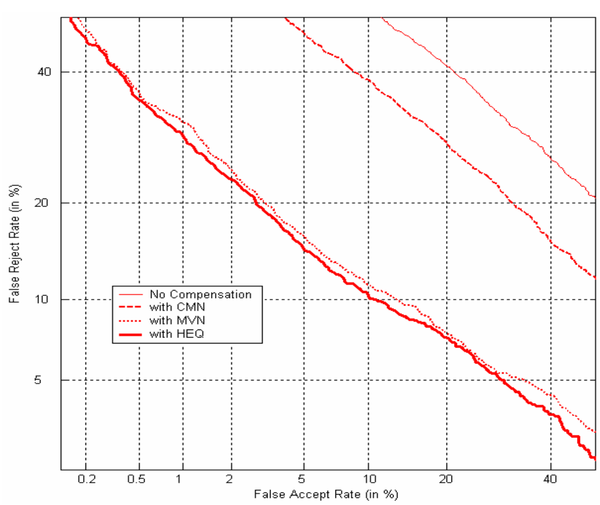

Figure 3: The improvements obtained when applying CMN, MVN and HEQ to minimize the mismatch between training and testing distributions

6. Conclusion

In this work we have shown that histogram equalization is very effective in compensating for both linear and non-linear effects caused by the various noise sources encountered in a telephone network. In particular, histogram equalization's ability to match training and testing distributions improved speaker verification performance above the baseline by over 67%.

7. REFERENCES

[1] D.A. Reynolds, "An overview of automatic speaker recognition technology", *Proceedings of IEEE ICASSP*, 4, pp. 4072-4075, 2002.

[2] G.R. Doddington, M.A. Przybocki, A.F. Martin and D.A. Reynolds, "The NIST speaker recognition Evaluation – overview, methodology, systems, results, perspective," *Speech Communication* 31, pp. 225-254, 2000.

[3] J. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, Vol.85, No.9, pp. 1437-1462, September 1997.

[4] D.A. Reynolds, "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models," *MIT Lincoln Laboratory Journal*, Vol. 8, No. 2, pp. 173-192, 1995.

[5] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE transactions on Speech and Audio Processing*, vol.3, No.1, January 1995.

[6] R. Duncan, "A description and comparison of the feature sets used in speech processing", Mississippi State University, 2000.

[7] H.D. Cheng and X.J. Shi, "A simple and effective histogram equalization approach to image enhancement", Digital Signal Processing 14, pp.158–170, 2004.

[8] S. Molau, M. Pitz, and H. Ney, "Histogram Based Normalization in the Acoustic Feature Space," *Proc. of ASRU*, December 2001.

[9] A. de la Torre, J. C. Segura, M. C. Ben´itez, A. M. Peinado, and A. Rubio, "Non-linear transformations of the feature space for robust speech recognition", *Proc. ICASSP*, pp. 401–404, 2002.

[10] S. Molau, F. Hilger and H. Ney, "Feature Space Normalization in Adverse Acoustic Conditions", *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. I, pp. 656-659, Hong Kong, China, April 2003.

[11] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez, C. Benitez, A.J. Rubio: "Histogram equalization of the speech representation for robust speech recognition". IEEE Transactions on Speech and Audio Processing, Article In Press. 2003

[12] S. Molau "Normalization in the Acoustic Feature Space for Improved Speech Recognition". PhD Dissertation, Aachen, Germany, February 2003

[13] Cepstral domain segmental nonlinear feature transformations for robust speech recognition. J. C. Segura, C. Benítez, Á. de la Torre, A. J. Rubio, J.

Ramírez. IEEE Signal Processing Letters, Vol. 11, no. 5 may 2004, pp.517-520.

[14] S. Dharanipargda and M. Padmanabhan, "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition", in Proc. ICSLP 2000 Peking, china, October 2000, pp. 556-559

[15] P.J. Moreno, “Speech recognition in Telephone Environments”, MSc dissertation, Carnegie Mellon University, Pittsburg, Pennsylvania, December 1992.

[16] http://www.itl.nist.gov/iad/894.01/tests/spk/2000/doc/spk-2000-plan-v1.0.htm Accessed: 21/09/2004

[17] M.A. Przybocki and A.F. Martin, “Odyssey Text Independent Evaluation Data,” in Proceedings of *2001: A Speaker Odyssey, A Speaker Recognition Workshop,* pp 21- 24, June 2001.

[18] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE transactions on Speech and Audio Processing*, vol.3, No.1, January pp. 72-83, 1995.

[19] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker verification using adapted Gaussian speaker mixture models,” *Digital Signal Processing*, pp. 19-41, 2000.

[20] M. Skosan and D.J. Mashao, “Improving Speaker Identification Performance for Telephone-based Applications”, *Proceedings of the South African Telecommunication Networks and Applications Conference*, September 2004

[21] J. Campbell and D.A. Reynolds, "Corpora for the Evaluation of Speaker Recognition Systems", *Proceedings of IEEE ICASSP*, pp. 2247-2250, 1999.

.

technique.

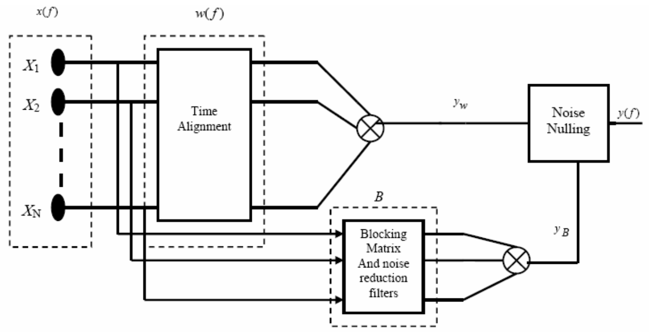

Figure 4: Active noise canceling beamforming structure

B. System description

The microphone array used in the evaluation is a 4 element (N) array placed on a table. The array is 9cm long with an equal inter-element spacing d, of 3cm giving it an effective length, L = N*d, of 12cm. It accommodates the frequency band; 2 kHz < f < 6 kHz. All signal sources are considered far-field to simplify calculations and Figure 5 shows the directivity pattern for a linear, equally spaced array of 4 microphones.

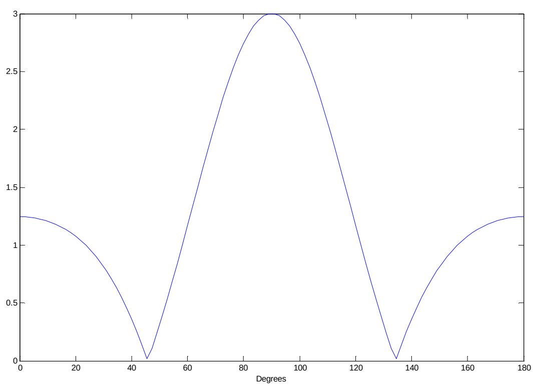

Figure 5: Directivity pattern for 4 element microphone array

The complete microphone array system comprises three main components; *the linear array, data acquisition module and processing module*. Figure 6 illustrates these three components and includes the speaker identification system.

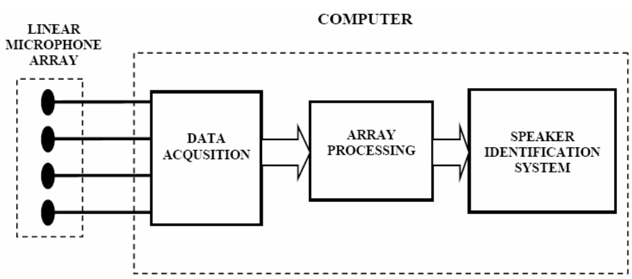

Figure 6: Microphone array system

The three components perform the following tasks:

1) *Linear Microphone Array*

The microphones act as transducers that convert sound pressure waves into electrical signals. Let us assume that a talker produces a speech message *x(t)* that is acquired by microphones 1, …., N as signals $x_1(n),\dots, x_N(n)$. Signals sampled by microphones *i* and *k* are characterized by a relative time delay τ_{ik} of the direct wavefront arrival [12].

2) *Data Acquisition Module*

Signals from the microphone array are acquired for computer processing using a PCI703 series 16 analog input channel data acquisition board from Eagle Technology. The board has a maximum analog sample rate of 400 kHz with 14-bit accuracy. For 4 channels the sample rate used is 64 kHz (16 kHz per channel). After acquisition the data is converted to a suitable file format for processing.

3) *Array Processing Module*

Generally, array processing with regard to microphone arrays refers to beamforming. A beamformer performs spatial filtering. The beamforming capabilities of microphone array systems allow highly directional sound capture, providing superior signal-to-noise ratio (SNR) when compared to single microphone performance [1].

A total of 40 speech samples, comprising 20 training and 20 testing speech utterances, from 20 speakers were acquired using the microphone array. Each speaker was seated 50cm directly in front of the array. The speech was recorded in an office environment with interfering noise mainly from an air conditioner and other randomly distributed speakers. No additional noise was artificially introduced to the data.

C. Results

It has been shown that for clean speech recorded using a close-talking microphone, a GMM based speaker identification system similar to the one used in this research obtained a 100% identification rate [13]. It should be noted that the experimental setup and data used in [13] were different to that used in our evaluation. The baseline for the experiments to which further improvements will be compared, is the identification rate obtained using a single microphone under the same conditions as the microphone array. We obtained an identification rate of 60% for a 20 speaker database as a baseline. The performances of the delay-and-sum beamformer, filter-and-sum beamformer and the active noise canceling beamformer were evaluated and compared. All the systems compared fairly well to the baseline, with the active noise canceling beamformer attaining the highest improvement in identification rate of 85%. Table 1 displays the performance of the beamforming techniques on a 20 speaker database.

Beamforming Technique	Identification Rate
Single Mic. (Baseline)	60%
Filter-and-sum	65%
Delay-and-sum	70%

Noise Canceling	**85%**

Table 1: The effect of the beamforming techniques

It is clear from table 1 that all the beamforming techniques investigated improved the identification rate. These results are compared to the baseline, which is the identification rate achieved using a single microphone with speakers 50 cm from the microphone. The delay-and-sum beamformer outperformed the filter-and-sum beamformer due to signal distortions introduced by the multi-dimensional wiener filter used in these experiments [7]. The active noise cancellation technique produced the best results with a 25% increase in identification rate from the baseline.

Beamforming Technique	Identificatio n Rate
Close-Talking Mic.	100%
Single Mic. (Baseline)	60%
Noise Canceling	**85%**

Table 2: Baseline compared to Active Noise Cancellation

We suspect the active noise cancellation beamformer performs better because of the small population used for these experiments and the cleaner signal that it produces.

V. CONCLUSIONS

The work presented here has demonstrated that using a microphone array for speech acquisition offers a performance advantage for a speaker identification application in a distant-talking environment. We reviewed an active noise canceling beamformer, a delay-and-sum beamformer and a filter-and-sum beamformer, and found that the active noise canceling beamformer proved superior when evaluated on a speaker identification task.

We aim to further the research in the field by addressing the following:

1. Investigating the use of more sophisticated beamforming techniques used with speaker tracking.
2. More experiments into the effect of microphone arrays on speaker identification performance with respect to distance.
3. Increasing the speaker database.

REFERENCES

[1] D.V. Rabinkin, R.J. Renomeron, J.C. French and J.L. Flanagan, “Optimum microphone placement for array sound capture”, *Proc. SPIE, Vol. 3162, pp. 227-239, 1997.*

[2] M.L. Seltzer, B. Raj and R.M. Stern, “Speech recognizer-based microphone array processing for robust hands-free speech recognition”, *Proc. IEEE Conf. on Acoustics, Speech and Sig. Proc., May, 2002,* Orlando, Florida.

[3] Q.Lin, E. Jan and J. Flanagan, “Microphone arrays and speaker identification”, *IEEE Transactions on Speech and Audio Processing, Vol. 2, pp. 622-629, October 1994*

[4] LEMS Microphone-Array Papers [Online] : http://www.lems.brown.edu/ *Accessed: October,11th 2004.*

[5] V.C. Raykar, “A study of various beamforming techniques and implementation of the constrained least mean squares (LMS) algorithm for beamforming”, *Course project report ENEE 624, Fall 2001.*

[6] D.A. Florêncio and H.S. Malavar, “Multichannel filtering for optimum noise reduction in microphone arrays”, *ICASSP 2001, Salt Lake City, May 2001.*

[7] I. Sanches, A. Girardi, "Multi-dimensional Filtering for Speech Enhancement via Microphone Array", *2nd International Symposium of NAIST-IS 21st century COE program, Japan October 2003.*

[8] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming”, *IEEE Trans. on Antennas and Propagation, Vol. 30(1), pp. 27-34, January 1982.*

[9] I.A. McCowan, “Robust speech recognition using microphone arrays”, *PhD Thesis, Qeensland University of Technology, Australia, 2001.*

[10] H. Gish and M. Schmit, “Text-Independent Speaker Identification”, *IEEE Signal Processing* Magazine, *pp. 18-32, Oct.1994.*

[11] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE transactions on Speech and Audio Processing, Vol. 3, No. 1, January 1995.*

[12] M. Omologo, M. Matassoni, P. Svaizer and D. Giuliani, “Microphone array based speech recognition with different talker-array positions”, *Proc. ICASSP ’98,* Seattle Washington

[13] D.A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech", *IEEE Signal Processing Letters, Vol.2, No.3, March 1995.*

N. Zulu is currently pursuing an MSc in Electrical Engineering at the University of Cape Town and is in his first year of study.

Dr. D. Mashao is a senior lecturer at the University of Cape Town and head of the Speech research and Technology Group. He is also the supervisor of the above-mentioned author.