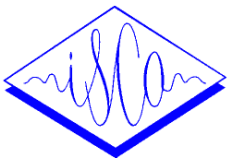

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18–22, 2001

Robust Speaker Recognition using Microphone Arrays

Iain A. McCowan *Jason Pelecanos* *Sridha Sridharan*

Speech Research Laboratory, RCSAVT, School of EESE
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
[i.mccowan, j.pelecanos, s.sridharan]@qut.edu.au

Abstract

This paper investigates the use of microphone arrays in hands-free speaker recognition systems. Hands-free operation is preferable in many potential speaker recognition applications, however obtaining acceptable performance with a single distant microphone is problematic in real noise conditions. A possible solution to this problem is the use of microphone arrays, which have the capacity to enhance a signal based purely on knowledge of its direction of arrival. The use of microphone arrays for improving the robustness of speech recognition systems has been studied in recent times, however little research has been conducted in the area of speaker recognition. This paper discusses the application of microphone arrays to speaker recognition applications, and presents an experimental evaluation of a hands-free speaker verification application in noisy conditions.

1. Introduction

Currently, research is being undertaken to improve the robustness of speech and speaker recognition systems to real noise environments. In an effort to improve robustness and ease-of-use, microphone arrays have been investigated for their ability to reduce input noise, and also because they remove the burden of a close-talking microphone from the user. While the use of microphone arrays for speech recognition applications has been investigated for some time, to date, speaker recognition has not received the same attention.

Speaker recognition technology has a wide range of potential applications. Accurate speaker recognition can be an integral part of many security applications, controlling access to information, property and finances. In particular, with the increased use of automated services for applications such as banking, speaker recognition has the potential to become an important means of authentication over telephone networks. Access to automatic teller machines could also be improved by including voice authentication with PIN verification. In addition to security applications, the ability to correctly identify a person from their voice can be used in conjunction with speech recognition to produce automatic transcripts of conversations and conferences. Speaker recognition may also be used in forensic applications, such as helping determine the identity of speakers in recorded telephone calls.

The above list of applications is by no means exhaustive, yet it serves to illustrate the point that speaker recognition systems must be capable of performing well in a variety of environments and configurations. In addition, it is apparent that many potential applications require hands-free sound capture, such as automatic teller machine authentication, the production of video conference transcripts, and security access to buildings or vehicles. In such applications, a microphone array capable of enhancing the desired speech from a known location offers a means of meeting the requirements for hands-free operation and robustness to noise conditions.

This paper commences by explaining the principles of microphone arrays and beamforming algorithms. Following this, a review of the current state of microphone array speaker recognition research is given, and issues requiring further investigation are identified. A microphone array speaker recognition system addressing these issues is then assessed in an experimental evaluation.

2. Microphone arrays and beamforming

An array of sensors is essentially a discretely sampled continuous aperture, and the response of the array approximates that of the continuous aperture which it samples. The array response as a function of direction is known as the *directivity pattern*. A linear array of N sensors with uniform inter-element spacing, d, has a far-field horizontal directivity pattern given by

$$D(f,\phi) = \sum_{n=1}^{N} w_n(f) e^{j2\pi\alpha(n-1)d} \qquad (1)$$

where w_n is the complex weight associated with the n^{th} sensor, $\alpha = \frac{\cos\phi}{\lambda}$, ϕ is the angle measured from the array axis in the horizontal plane, and λ is the wavelength. A sample horizontal directivity pattern for equally weighted sensors ($w_n(f) = \frac{1}{N}$) is shown by the bold line in Figure 1, illustrating the directional nature of the array response. From the directivity pattern, we see that a sensor array is capable of enhancing a signal arriving from a certain direction with respect to signals arriving from all other directions. This enhancement is based purely on the direction of arrival, and is independent of the characteristics of the desired and undesired signals.

In general, the complex weighting w_n can be expressed in terms of its magnitude and phase components as

$$w_n(f) = a_n(f) e^{j\varphi_n(f)} \qquad (2)$$

where $a_n(f)$ and $\varphi_n(f)$ are real, frequency dependent amplitude and phase weights respectively. By modifying the amplitude weights $a_n(f)$, we can modify the shape of the directivity pattern. Similarly, by modifying the phase weights, $\varphi_n(f)$, we can control the angular location of the response's main lobe. *Beamforming techniques* are algorithms for determining the complex sensor weights $w_n(f)$ in order to implement a desired *shaping* and *steering* of the array directivity pattern. In this way, the response of the array can be controlled in order to enhance

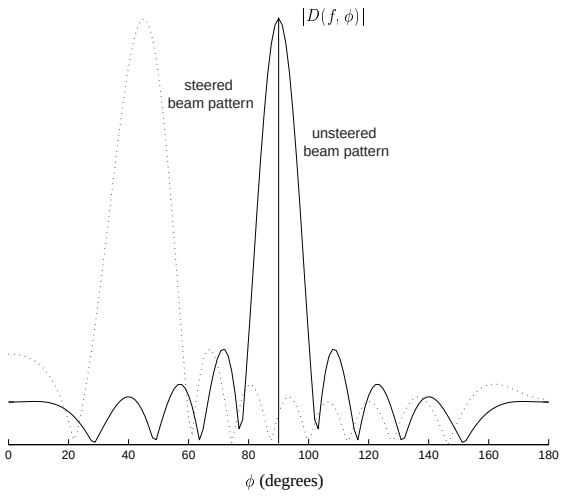

Figure 1: Unsteered and steered directivity patterns (ϕ'=45 degrees, f=1 kHz, N=10, d=0.15 m)

a specific signal, provided the direction of the signal source is known with some accuracy - a condition which is often met in many speech and speaker recognition applications.

2.1. Delay-sum beamforming

To illustrate the concept of beam steering, consider the case where the sensor amplitude weights, $a_n(f)$, are set to unity. If we use the phase weights

$$\varphi_n(f) = -2\pi\alpha'(n-1)d \tag{3}$$

where $\alpha' = \frac{\cos\phi'}{\lambda}$, then the directivity pattern becomes

$$D'(f,\alpha) = \sum_{n=1}^{N} e^{j2\pi(n-1)d\left(\alpha-\alpha'\right)} \tag{4}$$

or

$$D'(f,\alpha) = D(f,\alpha-\alpha') \tag{5}$$

The effect of such a phase weight on the beam pattern is thus to steer the main lobe of the beam pattern to the *direction cosine*, $\alpha = \alpha'$, and thus to the direction $\phi = \phi'$. The dotted line in Figure 1 shows the horizontal directivity pattern for $\phi' = 45^o$.

A negative phase shift in the frequency domain corresponds to a time delay in the time domain, and so beam steering can effectively be implemented by applying time delays to the sensor inputs. We see that the delay for the n^{th} sensor is given by

$$\begin{aligned}\tau_n &= \frac{\varphi_n(f)}{2\pi f} \\ &= \frac{(1-n)d\cos\phi'}{c}\end{aligned} \tag{6}$$

which is equivalent to the time a plane wave takes to travel between the reference sensor and the n^{th} sensor (with c representing the speed of sound propagation). This is the principle of the simplest of all beamforming techniques, known as *delay-sum beamforming*, where the time domain sensor inputs are first delayed by τ_n seconds, and then summed to give a single array output. Many more complex beamforming techniques exist, most of which calculate the channel filters w_n according to some optimisation criterion, or to implement a desired shaping and steering of the beam pattern.

2.2. Superdirective beamforming

One class of beamforming techniques is that of *superdirective beamforming* [1]. A key measure for sensor arrays is the *array gain*, which is defined as the improvement in signal-to-noise ratio between the reference sensor and the array output, and is dependent on the array geometry as well as the noise field characteristics. In the case of a diffuse noise field, the array gain is also known as the *factor of directivity*. A *diffuse* noise field is one in which noise of equal energy propagates in all directions simultaneously. Superdirective beamformers calculate the channel filters that maximise the array factor of directivity, and are thus optimal for diffuse noise conditions.

A near-field modification to the standard superdirective technique, termed *near-field superdirectivity*, was proposed by Täger [2] for the case where the desired speech source is located close to the array. A source is said to be located in the array's near-field if

$$|r| > \frac{2L^2}{\lambda} \tag{7}$$

where r is the distance between the source and the closest microphone, and L is the total array length. Within this range, the assumption of a planar wavefront no longer holds, and a spherical propagation model must be used. Previous work has demonstrated the suitability of near-field superdirectivity for speech recognition in the context of a computer workstation in a noisy office [3].

2.3. Adaptive beamforming

A limitation of fixed beamforming techniques, is their inability to adapt to changing noise conditions. Adaptive array processing techniques, such as the generalised sidelobe canceler (GSC) [4] aim to solve this problem. The GSC separates the adaptive beamformer into two main processing paths - a standard fixed beamformer with constraints on the desired signal response, and an adaptive path, consisting of a blocking matrix and a set of adaptive filters that minimise output noise power. The purpose of the blocking matrix is to exclude the desired signal from the adaptive path, ensuring that the output power minimisation does not degrade the desired signal.

Such an adaptive beamforming technique succeeds in significantly reducing the noise level for coherent noise signals emanating from localised sources. In addition to the noise reduction provided by the focused fixed beamforming portion, the adaptive noise canceling path is able to effectively construct a directivity pattern null in the direction of the principal undesired coherent sources.

2.4. Near-field adaptive beamforming

The beamforming technique chosen for the experiments in this paper is termed *near-field adaptive beamforming* (NFAB). The NFAB system is essentially a hybrid superdirective/adaptive beamformer, as seen from the block diagram in Figure 2. The upper path consists of a fixed near-field superdirective beamformer, while the lower path contains a near-field compensation unit, a blocking matrix and an adaptive noise cancelling filter,

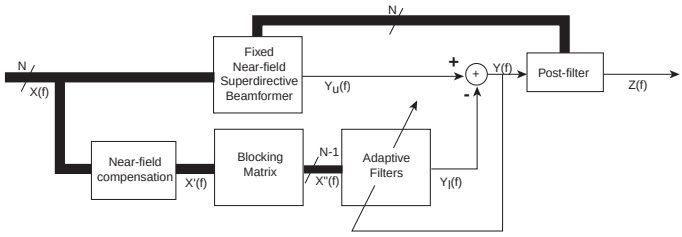

Figure 2: Near-field Adaptive Beamformer

similar to the GSC adaptive beamformer. The two paths combine before passing through a post-filter. The technique is described and analysed for speech enhancement in [5].

The motivation for a hybrid beamformer is the desire for good performance in a variety of noise conditions. While near-field superdirectivity performs well in a diffuse noise environment when localised noise sources exist, further noise reduction can be attained using an adaptive technique. By adding a GSC-style adaptive noise cancelling path to the superdirective beamformer, the resulting system demonstrates good noise reduction in both diffuse and coherent noise fields. Addition of a post-filter further reduces the output noise when used in conjunction with an effective beamformer [6]. In speech recognition experiments, the NFAB technique has shown to out-perform both standard near-field superdirective and GSC beamforming techniques.

3. Speaker recognition with microphone arrays

Although much research has been conducted into the use of microphone arrays with speech recognition systems, very little has been done for speaker recognition tasks. Lin *et al* [7] investigated the use of microphone arrays with speaker recognition, using a matched-filter array with a vector-quantization based speaker identification system. While their results showed significant performance improvements with the array in noisy conditions, the research is at least partially out-dated by the recent shift to Gaussian mixture model (GMM) speaker recognition systems.

More recently, Ortega-Garcia and Gonzalez-Rodriguez have produced a number of research papers investigating the use of low-complexity microphone arrays in GMM based speaker recognition systems in noisy conditions (eg. [8]). Their research has shown the benefits of using a microphone array over a single microphone in hands-free speaker identification experiments. In the experiments, the multi-channel input data is synthesised using impulse responses of the propagation paths between the source and each microphone, estimated using the image method [9]. While use of impulse responses is common for the purpose of microphone array recognition experiments, their estimation using the image method is based on a number of theoretical assumptions which are rarely met in practice.

Another limitation of current research is the lack of results for speaker verification. Speaker recognition applications can be categorised as either identification or verification tasks. *Speaker identification* tasks classify a speech segment as belonging to either the most likely speaker from a closed set of known speakers, or potentially as an unknown speaker. In contrast, *speaker verification* tasks decide whether or not a speech segment was uttered by a specific speaker. Speaker verification is thus the more likely task in most security and forensic applications. To date, all the research in microphone array speaker recognition has been confined to the task of speaker identification.

Thus, while some research has been done on speaker recognition using microphone arrays, this has been minimal, and to further research in the field a number of issues should be addressed :

1. The use of more sophisticated beamforming techniques should be investigated.
2. More realistic methods of generating multi-channel speech databases for experiments should be used.
3. More research into the use of microphone array enhanced speech with state-of-the-art GMM based speaker recognition systems is required.
4. Experiments into the effect of microphone arrays on speaker verification performance should be performed.

The experimental evaluation that follows aims to address each of these issues.

4. Experimental evaluation

4.1. Beamforming technique

In order to investigate the use of more sophisticated beamforming techniques, the near-field adaptive beamforming (NFAB) technique discussed in Section 2.4 was used in the experiments. In previous work, the technique has been shown to be well suited for the task of speech enhancement for a near-field source in a high noise environment. In particular, the technique was shown to provide an additional 5-8 dB improvement in the signal to noise ratio as compared to standard delay-sum beamforming, while introducing negligible distortion to the desired speech signal [5]. For these reasons, it is expected that the technique will be well suited to the task of hands-free speaker recognition in noisy conditions.

4.2. Experimental configuration

The microphone array used in the evaluation is the 9 element array shown in Figure 3, consisting of a 7 element broadside array, with an additional 2 microphones situated directly behind the end microphones. The array is designed to sit on the top of a computer monitor, and is 40 cm wide and 15 cm deep in the horizontal plane. The broadside microphones are arranged according to a standard broadband sub-array design, where different sub-arrays are used for different frequency ranges. The two endfire microphones are included for use in the low frequency range where the amplitude difference between sensors is greater and can be exploited by the NFSD algorithm (for further explanation, see [2]). The three sub-arrays accommodating the different frequency bands are thus

- ($f < 1\ kHz$) : microphones 1-9;
- ($1\ kHz < f < 2\ kHz$) : microphones 1, 2, 4, 6 and 7; and
- ($2\ kHz < f < 4\ kHz$) : microphones 2, 3, 4, 5 and 6.

The experimental context is the computer room shown in Figure 4. Two different sound source locations were used, these being

1. the desired speaker situated 70 cm from the centre microphone, directly in front of the array; and

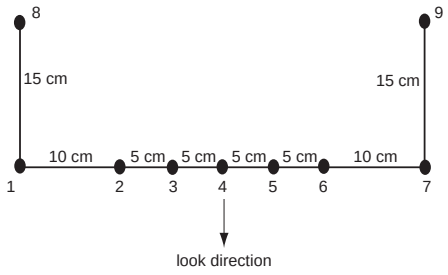

Figure 3: Array Geometry

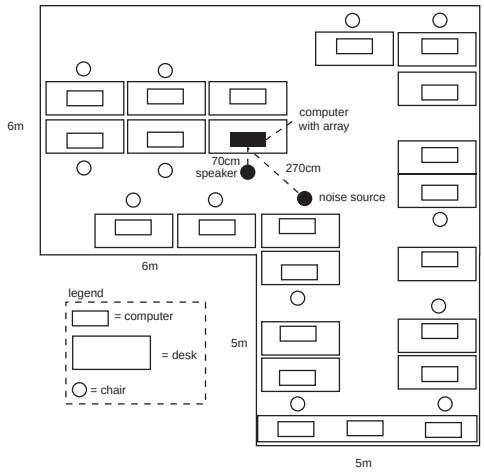

Figure 4: Experimental Setup

2. a localised noise source at an angle of 56 degrees and a distance of 2.7 metres from the array.

In order to generate different noise conditions, test signals were generated using impulse responses of the acoustic transfer function between the sources and each microphone in the array. As discussed earlier, the image method for estimating impulse responses makes a number of assumptions that are rarely satisfied in practice. In order to generate more realistic multi-channel test signals it is desirable to have more accurate impulse response measurements than available using the image method. For this reason, a *maximum length sequence* (MLS) technique as described by Rife and Vanderkooy [10] was used to measure the real acoustic impulse responses from actual recordings made in the room.

The multi-channel desired speech and localised noise inputs were generated by convolving the speech and noise signals with the measured impulse responses. In addition, a real multi-channel background noise recording of normal operating conditions was made in the room. This recording is referred to in the experiments as the ambient noise signal, and is approximately diffuse in nature. It consists mainly of computer noise, a variable level of background speech, and noise from an air-conditioning unit. The experiments were conducted for varying levels of signal to noise ratio (SNR), measured as an average segmental SNR. For the localised noise we used the speech-like noise from the NOISEX database. In this way, realistic multi-channel input signals were generated for varying levels of ambient and localised noise, testing diffuse and coherent noise conditions respectively.

4.3. Speaker recognition system

A GMM-based, text-independent, speaker verification system was used in the experiments. The core system consists of a large-mixture Gaussian mixture model to estimate the probability density of features for generalised speech. Individual speaker models are established by adapting the parameters of the generalised *universal background model* (UBM) to the statistics of each target speaker. The testing phase combines information from the adapted and background models in a likelihood ratio hypothesis test to examine the likelihood of the test speech segment being spoken by the target speaker. The core mechanism behind this speaker recognition approach is the *Gaussian mixture model* or GMM.

Gaussian mixture modeling is used for modeling the probability density function (PDF) of a multi-dimensional feature vector. A GMM forms a continuous density estimate of the PDF by the linear combination of multi-dimensional Gaussians. Given a single speech feature vector $\vec{x}$ of dimension D, the probability density of $\vec{x}$ given an N Gaussian mixture speaker model λ, with mixture weights w_i, means $\vec{\mu}_i$ and diagonal covariances Σ_i is given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{N} w_i g(\vec{x}, \vec{\mu}_i, \Sigma_i) \tag{8}$$

with a single Gaussian component density given as

$$g(\vec{x},\vec{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\vec{x} - \vec{\mu_i})^{'} (\Sigma_i)^{-1} (\vec{x} - \vec{\mu_i})) \tag{9}$$

where $(\cdot)'$ represents the matrix transpose operator. Note that the symbols D, w_i and λ are defined differently for the microphone array and speaker recognition theory.

In order to model the distribution of a set of training vectors, an iterative method is used to progressively refine the estimates using a form of the expectation-maximization (E-M) algorithm. The UBM was trained using a fast vector quantization Gaussian (VQG) [11] initialization before applying the E-M algorithm. In training, the speaker specific model is created by adapting the universal model towards the training speech [12].

For test trials, the set of speech feature vectors, X, comprising of T observations $\{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_T}\}$ was tested against both the adapted target, λ_{tar}, and the UBM, λ_{ubm}, models to determine a frame-averaged log-likelihood ratio score.

$$\Lambda = \frac{1}{T} \sum_{t=1}^{T} (\log p(\vec{x}_t|\lambda_{tar}) - \log p(\vec{x}_t|\lambda_{ubm})) \tag{10}$$

These results are compared across the global board of speakers to determine the error statistics in the form of an Equal Error Rate (EER), Detection Cost Function (DCF) or Detection Error Trade-off (DET) curve [13].

The speech was parameterised into vectors of 12 mel-frequency cepstral coefficients (MFCC's) with their corresponding delta coefficients. The MFCC's were determined by

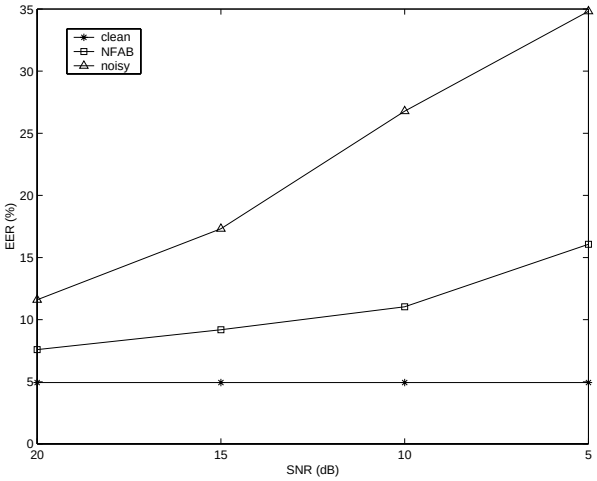

Figure 5: Equal Error Rate (EER) Comparison : Ambient Noise Only

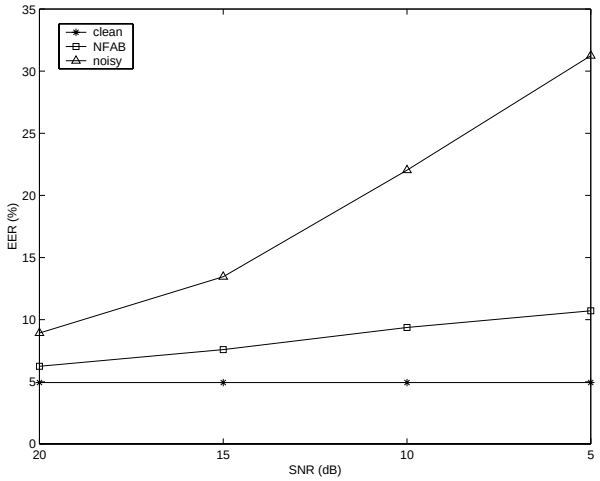

Figure 6: Equal Error Rate (EER) Comparison : Localised Noise Only

the application of a cosine transform to a set of 20 mel-spaced filter-bank energies across the 0-4000Hz spectrum. The filter-bank energies were derived using 32ms speech frames with 10ms frame advance. An energy based silence removal technique was used to discard silence frames in both training and testing.

4.4. Recognition task

An evaluation was performed on the TIMIT Acoustic-Phonetic Continuous Speech Corpus to examine the effect of microphone arrays on a speaker verification system. The TIMIT database is divided into two portions consisting of training and testing speech from exclusively different speakers. The male speakers in the training set were used to form the general background speaker model, while the 112 male speakers within the testing data set was used to perform the verification evaluation. For each speaker, there were 10 speech segments; the first 8 segments (totaling about 24 seconds) were extracted for speaker model training, and the remaining 2 segments (each of approximately 3 seconds) were used for testing against all other male speakers, producing a total of 25088 verification tests.

4.5. Results

In the first set of experiments, the level of ambient noise was varied over the SNR range 20-5 dB, with no localised noise present. This represents a diffuse noise condition, and thus tests the microphone array's ability to focus on the desired signal direction. The equal error rate (EER) is plotted in Figure 5 for three different signals :

- the clean input to the centre microphone (clean),
- the noisy input to the centre microphone (noisy), and
- the enhanced output of the NFAB microphone array (NFAB).

The speaker verification task is evidently highly sensitive to additive noise, as seen by the drastic degradation in results for the noisy input, which performs little better than a guess (50% EER) at the higher noise levels. It can be seen that the NFAB microphone array is successful in reducing the level of noise in the input signal, with the EER approaching that of the clean input at the 20 dB noise level. As the input noise level increases, the performance of the microphone array system degrades more gracefully than that of the single microphone.

Figure 6 plots the same results for the case of varying levels of localised noise. This represents a coherent noise condition, and as such, tests the microphone array's ability to attenuate undesired signals emanating from different directions to the desired speech. The plot demonstrates the same general trends as the preceding ambient noise case, however we note that the EER of the NFAB output is lower for localised noise than for the equivalent level of ambient noise. As expected, due to the adaptive noise canceling path, the microphone array system is better able to handle the situation of a single localised source than that of a diffuse noise field. This is due to the fact that a null can be placed in the direction of a single noise source, while this is not possible for a diffuse noise source which effectively contains an infinite number of noise sources, leading to an average gain over all undesirable directions that is non-zero. The degradation of verification performance with increasing noise is very slow in this case, with the EER increasing by only 4% over the 20-5 dB input noise range for the NFAB system.

The detection error trade-off (DET) curve is a common means of assessing the performance of speaker verification tasks [13]. Figure 7 plots the DET curves for the three signals for the case of equal levels both of ambient and localised noise (combined SNR level of 7 dB), representing a highly adverse noise condition. Noting the logarithmic axes, we see that the DET curve of the microphone array system is significantly closer to that of the clean input than the DET curve of the single microphone. From the figure, we see that the microphone array system has significantly reduced the EER from 29.5% to 12.7%.

While the results clearly demonstrate the ability of microphone arrays to provide considerable performance improvement in a speaker verification task, it is apparent that more research is required to attain performance levels acceptable for real applications. Greater performance may be achieved by combining the microphone array system with additional speaker verfication robustness techniques.

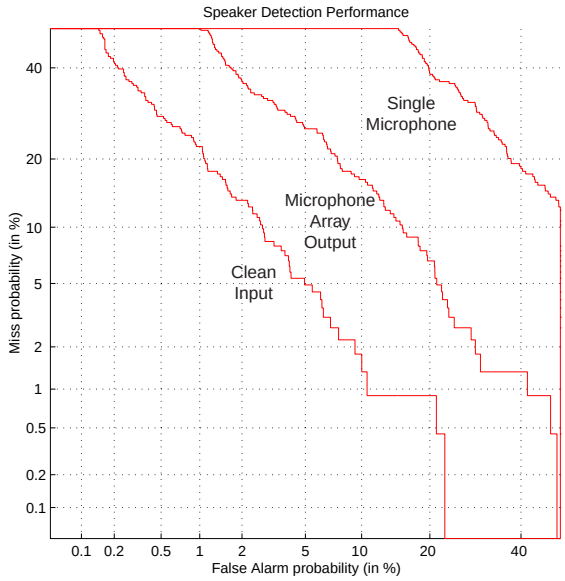

Figure 7: Detection Error Trade-off Curve : Ambient and Localised Noise (SNR = 7 dB)

5. Conclusions

This paper has investigated the use of microphone arrays for improving the robustness of hands-free speaker recognition applications in noisy environments. Microphone arrays have the benefit of providing a high level of enhancement based purely on knowledge of the speaker's location, without explicit use of the characteristics of the speech or the noise.

A review of the current state of research in the field was given, and a number of issues requiring further attention were identified, including :

1. the use of sophisticated beamforming techniques,
2. the use of realistic methods of generating multi-channel speech databases,
3. the need for speaker verification experiments, and
4. the use of state-of-the-art GMM based speaker recognition systems.

These issues were then addressed in an experimental evaluation of a hands-free speaker verification task in high noise conditions. The results indicate that the noise reduction provided by the microphone array succeeds in significantly improving the verification performance, as measured by the equal error rate, and as shown in the detection error trade-off curve.

With further research, and used in conjunction with other techniques, the results presented in this paper indicate that microphone arrays have the potential to achieve significantly higher performance levels in practical hands-free, high noise, speaker recognition applications.

6. References

[1] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, October 1987.

[2] W. Täger. Near field superdirectivity (NFSD). In *Proceedings of ICASSP '98*, pages 2045–2048, 1998.

[3] I. McCowan, C. Marro, and L. Mauuary. Robust speech recognition using near-field superdirective beamforming with post-filtering. In *Proceedings of ICASSP 2000*, volume 3, pages 1723–1726, 2000.

[4] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, 30(1):27–34, January 1982.

[5] I. McCowan, D. Moore, and S. Sridharan. Speech enhancement using near-field superdirectivity with an adaptive sidelobe canceler and post-filter. In *Proceedings of the 2000 Australian International Conference on Speech Science and Technology*, pages 268–273, 2000.

[6] Claude Marro, Yannick Mahieux, and K. Uwe Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259, May 1998.

[7] Q. Lin, E. Jan, and J. Flanagan. Microphone arrays and speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):622–629, October 1994.

[8] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez. Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays. In *Proceedings of ICSLP '96*, volume 3, pages 1333–1336, 1996.

[9] J. Allen and D. Berkley. Image method for efficiently simulating small room acoustics. *Journal of the Acoustical Society of America*, 65:943–950, April 1979.

[10] D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *Journal of the Audio Engineering Society*, 37:419–444, June 1989.

[11] J. Pelecanos, S. Myers, S. Sridharan, and V. Chandran. Vector quantization based Gaussian modelling for speaker verification. In *Proceedings of International Conference on Pattern Recognition*, 2000. Paper number 1219.

[12] D. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings of Eurospeech 97*, volume 2, pages 963–966, 1997.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *Proceedings of Eurospeech 97*, volume 4, pages 1895–1898, 1997.