Evaluating Microphone Arrays for a Speaker Identification Task

Nicholas Zulu, Daniel Mashao
Department of Electrical Engineering, University of Cape Town
Rondebosh, Cape Town, South Africa
pzulu@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Abstract—Microphone array systems have been an area of active research for several years. The potential for high quality hands-free speech acquisition in noisy and reflecting environments makes microphone arrays an attractive alternative to conventional close-talking microphones. The signal-enhancement and source-location capabilities of microphone arrays make them applicable to a variety of tasks including teleconferencing, speaker tracking, speaker recognition and speech recognition. In this paper we evaluate techniques for setting up microphone arrays for speaker identification. We propose the use of an active noise canceling beamformer based on the generalized sidelobe canceller (GSC) beamformer. Significant improvements in identification rate are achieved using this method compared to other beamforming techniques investigated in this paper.

I. INTRODUCTION

Speaker identification systems are known to perform well when the speech signals are captured in a noise-free environment using a close-talking microphone worn near the mouth. However, many of the target applications of this technology do not take place in noise-free environments and it is often inconvenient for the user to wear a close-talking microphone. As the distance between the speaker and the microphone increases, the speech signal becomes increasingly susceptible to background noise and reverberation effects that significantly degrade speaker identification accuracy. This problem can be greatly alleviated by the use of multiple microphones to capture the speech signal.

Microphone arrays provide a means of localizing sound pickup and improving sound quality in noisy and reverberant conditions [1]. A microphone array uses multiple spatially distributed sensors to capture speech signals. The speech signals are captured simultaneously by each of the microphones and then processed jointly using one or more of a variety of methods to obtain a cleaner output signal [2]. The most important objective of a microphone array is to provide a high quality version of the desired speech signal for a specified application.

Microphone array speech enhancement techniques achieve this by beamforming, which reduces the level of localized and ambient noise signals, while minimizing distortion to speech from the desired direction. Beamforming has been applied to speaker identification as in [3], using speech signals generated by a computer model of room acoustics. This paper is aimed at contributing to research in the use of microphone arrays for speaker identification and proposes a beamforming technique based on the Generalized Sidelobe Canceller, (GSC) beamformer using real speech signals. This technique is aimed at reducing coherent and incoherent noise in speech signals acquired in an office environment, with minimal distortion to the desired speech.

Microphone array speaker identification has as one of its applications, automatic meeting transcription, where in conjunction with speech recognition, speakers in a conversation or conference can be identified. An example of such a deployment is being done at the Laboratory for Engineering Man/Machine Systems (LEMS) [4].

In exploring this topic, the principles of some basic beamforming techniques are discussed and evaluated. Thereafter, a review of current speaker recognition is given. A generalized sidelobe canceller is discussed and a slight modification to the GSC introduced. An overview of the system follows, with speaker identification results and conclusions.

II. BEAMFORMING TECHNIQUES

In this section three array processing techniques are reviewed. We present the theory behind these beamforming techniques, indicating their advantages, disadvantages and applicability to different noise conditions.

There are two classes of beamformers; data-independent (also known as fixed beamformers) or data-dependent (also known as adaptive beamformers). Data-independent beamformers are so named because their parameters are fixed during operation. Whereas, data-dependent beamformers continuously update their parameters based on the received signals.

A. Delay-and-sum Beamforming

The simple Delay-and-Sum beamformer is an example of a data independent beamformer [5]. The delay and sum beamforming algorithm adds the captured signals from the array sensors with corresponding delay in such a way that signal components originating from a desired location are combined coherently, while signals originating from other locations are combined in an incoherent fashion. This lends

the desired signal gain over undesired noise that increases as a function of the number of sensors [1]. By applying phase weights to the input channels, we can steer the main lobe of the directivity pattern to a desired direction. Phase shifts in the frequency domain can effectively be implemented by applying time delays to the sensor inputs. The delay for the n^{th} sensor is given by

$$\tau_n = \frac{(n-1)d\cos\phi'}{c} \tag{1}$$

which is the time the plane wave takes to travel between the reference sensor and the n^{th} sensor. Where ϕ' is the direction of arrival of the wave, c is the speed of propagation and d is the inter-element spacing.

Delay-and-sum beamforming is so-named because the time domain sensor inputs are first delayed by τ_n seconds, and then summed to give a single array output. Expressing the array output as the sum of the weighted channels, we obtain in the time domain

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t-\tau_n) \tag{2}$$

There exists a variation of delay-and-sum beamformers that combine the conventional delay-and-sum beamformer with channel filters to implement a desired shaping and steering of the beam pattern.

B. Filter-and-sum Beamforming

While the delay-and-sum beamformer is easy to understand, it offers minimal noise reduction and requires a large number of microphones to improve SNR [5]. It belongs to a more general class of beamformers known as *filter-and-sum beamformers,* where both the amplitude and phase weights are frequency dependent. In practice, most beamformers are a class of filter-and-sum beamformer.

The filter implemented in this research was a *multi-dimensional wiener filter.* The filter has as its inputs two correlation matrices: the correlation matrix of the *background noise* affecting the signal of interest and the correlation matrix of the *signal* affected by the noise. It is assumed that speech, ***s***, and affecting noise, ***n***, are statistically uncorrelated, and that noise is linearly added to speech: $\mathbf{x} = \mathbf{s} + \mathbf{n}$, where, for example, $\mathbf{x}$ is the output from the *N* channels of the microphone array for a given frame of analysis where each channel has a block of L_S samples being considered:

$$\mathbf{x} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(L_S) \\ x_2(1) & x_2(2) & \cdots & x_2(L_S) \\ \vdots & \vdots & \ddots & \vdots \\ x_N(1) & x_N(2) & \cdots & x_N(L_S) \end{bmatrix} \tag{3}$$

The objective is to estimate ***s*** given ***x*** and ***n*** for a defined filter order *L*. The algorithm has two correlation matrices as input, the background noise correlation matrix $\mathbf{R}_N$ and the signal correlation matrix $\mathbf{R}_X$. The optimal multi-dimensional wiener filter, $\mathbf{W}_{WF}$, is calculated as

$$\mathbf{W}_{WF} = \mathbf{R}_X^{-1}(\mathbf{R}_X - \mathbf{R}_N). \tag{4}$$

As presented in [6], matrix $\mathbf{R}_X^{-1}$ above can be replaced by $(\mathbf{R}_X + \rho\mathbf{R}_N)^{-1}$, where $\rho \geq 0$. Increasing ρ improves the intelligibility at a cost of increasing signal distortion. The filtered signal matrix can then be computed from,

$$\mathbf{Y} = \mathbf{W}_{WF}.\mathbf{X}^T. \tag{5}$$

The matrix $\mathbf{Y}$ comprises *N* filtered channel outputs which are separated and summed to give the beamformed output, y_W [7]. A block diagram showing the structure of a general filter-and-sum beamformer is given in Figure 1.

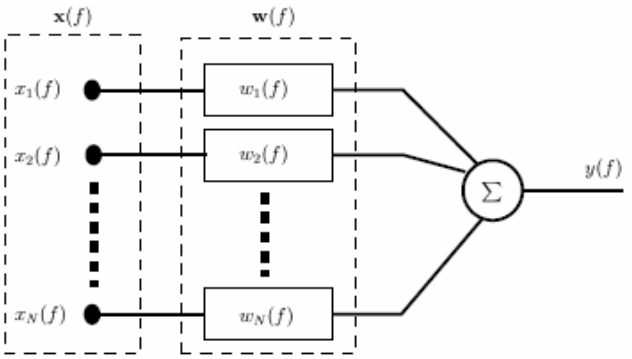

Figure 1: Filter-and-sum beamformer structure

C. Generalized Sidelobe Canceller (GSC)

A limitation of data independent beamforming techniques, such as the delay-and-sum and the filter-and-sum is their inability to adapt to changing noise conditions. Data-dependent beamforming techniques, such as the Generalized Sidelobe Canceller (GSC) [8] aim to solve this problem. The GSC separates the adaptive beamformer into two main processing paths. The first path implements a standard fixed beamformer with constraints on the desired signal. The second path is the adaptive part, which provides a set of filters that adaptively minimize the noise power in the output. The desired signal is blocked from the second path by a blocking matrix, ensuring that the noise power is minimized. Such an adaptive beamforming technique succeeds in significantly reducing the noise level for coherent noise signals emanating from localized sources [9]. Due to the blocking matrix, the lower path output only contains noise signals. The overall system output is calculated as the difference of the upper and lower path outputs

$$y(f) = y_u(f) - y_a(f) \tag{6}$$

The GSC is a flexible structure due to the separation of the beamformer into a fixed and adaptive portion. In practice, the GSC can cause a degree of distortion to the desired signal due to what is termed signal leakage. This occurs

when the blocking matrix fails to remove all of the desired signal from the lower noise canceling path. The block structure of the generalized sidelobe canceller is shown in Figure 2.

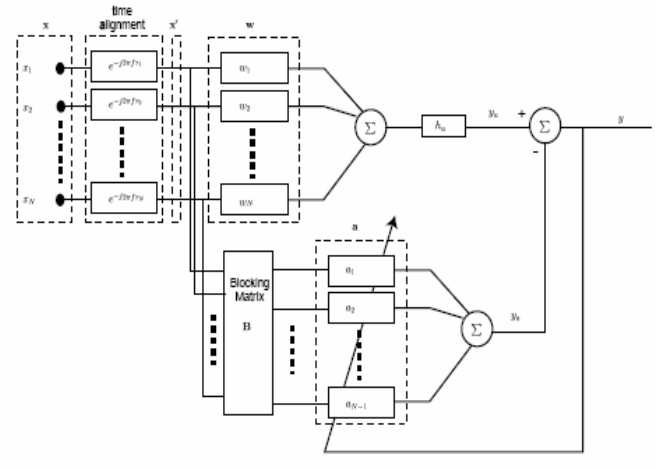

Figure 2: Generalized sidelobe canceller structure

In this section we have reviewed three common beamforming techniques. The delay-and-sum, filter-and-sum and the generalized sidelobe canceller. In the next section we discuss the speaker identification system we used to evaluate our microphone array.

III. SPEAKER IDENTIFICATION SYSTEM

Speaker recognition applications can be classified as either verification or identification tasks. Speaker verification tasks decide whether or not a speech segment was uttered by a specific speaker. On the other hand, speaker identification is concerned with recognizing an individual from a group of speakers based on a sample of his/her speech. The speaker identification system used in this research is text-independent. This type of speaker identification is concerned with determining who, from a group of known speakers, is speaking, regardless of what is being spoken. The speaker identification process can be summarized as follows: first the system needs to be trained with samples of speech collected from the speakers to be identified. Once this is complete, the system is tested (a speaker is identified) by comparing a speech sample from an unidentified speaker to the speech samples stored by the system and determining who the most likely speaker is [10].

Figure 3 illustrates a typical speaker identification system.

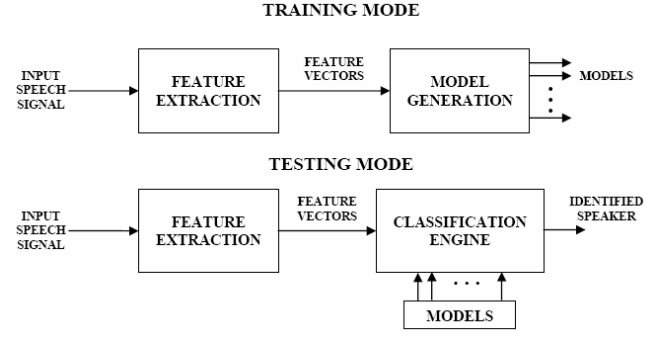

Figure 3: A typical speaker identification system

The system produces Mel-frequency Cepstral Coefficients (MFCC) in the feature extraction component. These features are aimed at emulating the spectral compression applied by the human auditory system to an incoming speech signal [10] and, are the most commonly used features used in speech-related research.

The system overview that follows describes the experimental configuration and results obtained from three beamforming techniques evaluated on a Gaussian Mixture Model (GMM) [11] based speaker identification system.

IV. SYSTEM OVERVIEW

A. *Beamforming technique*

In section II, three beamforming techniques outlining the important characteristics of each technique were discussed. The proposed beamforming technique for the speaker identification task is a variation of the generalized sidelobe canceller, comprising only the path with the blocking matrix.

The blocking matrix eliminates the desired signal from the lower path, allowing only the noise power to be minimized. As the desired signal is common to all the time-aligned channels, blocking will occur if the rows of the blocking matrix sum to zero. If $\mathbf{x}''$ denotes the signals at the output of the blocking matrix, then

$$\mathbf{x}''(f) = \mathbf{B}\mathbf{x}'(f) \tag{7}$$

where each row of the blocking matrix sums to zero, and the rows are linearly independent. As $\mathbf{x}'$ can have at most $N-1$ linearly independent components, the number of rows in $\mathbf{B}$ must be $N-1$ or less [9]. The standard Griffiths-Jim blocking matrix is [8]

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \cdots & \ddots & \ddots & \cdots & \cdots \\ 0 & \cdots & 0 & 1 & -1 & 0 \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{bmatrix} \tag{8}$$

Following application of the blocking matrix, $\mathbf{x}''$ is filtered and summed to give the lower path output y_B. If we denote the lower path filters as $\mathbf{a}$, then we have

$$y_B(f) = \mathbf{a}(f)^T \mathbf{x}''(f) \tag{9}$$

where y_B is a vector containing only noise samples. The positions of these samples are extracted in the noise canceling module (Figure 3), and the corresponding positions in the upper path output are replaced with nulls. Thus effectively canceling noise in the overall system output, y. Figure 4 illustrates the proposed beamforming

technique.

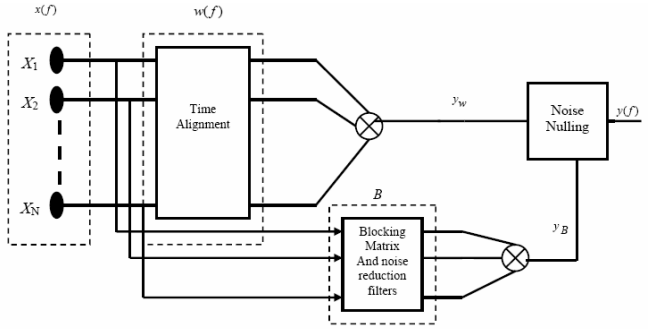

Figure 4: Active noise canceling beamforming structure

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.

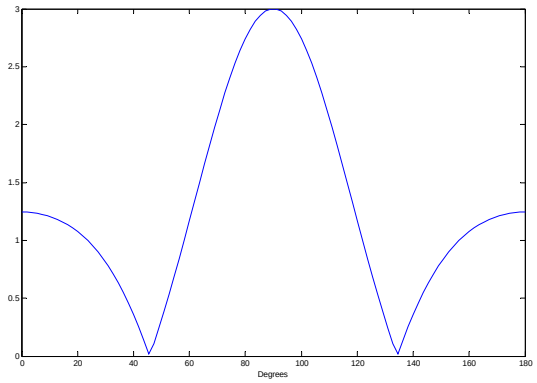

Figure 5: Directivity pattern for 4 element microphone array

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.

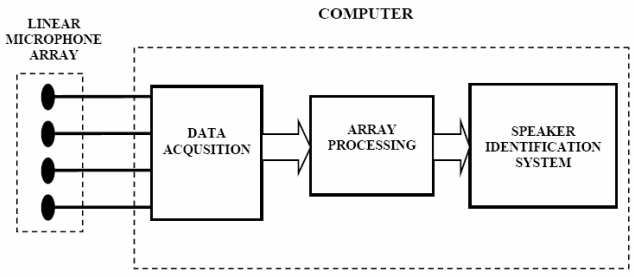

Figure 6: Microphone array system

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 $x_1(n),\ldots, x_N(n)$. 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	Identification 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.