Evaluation of the Quality of Microphone Array Enhanced Speech

Nicholas Zulu, Daniel Mashao
Department of Electrical Engineering, University of Cape Town
Rondebosh 7701, Cape Town, South Africa
pzulu@crg.ee.uct.ac.za daniel@eng.uct.ac.za

Abstract — Microphone arrays offer the possibility of hands free speech acquisition. This increases the convenience for those using speech technologies as they do not need to hold a microphone in order to interact with a speech system. In addition, a microphone array also has the advantage of potential gains in signal-to-noise ratio in noisy and reverberant environments. In this paper we evaluate the quality of a locally designed four element linear microphone array. The microphone array enhanced speech is evaluated on distortion, noise and speaker identification performance. The reported results show the noise canceling beamformer with post filter as having produced low distortion, high signal-to-noise ratio speech and the best speaker identification rate when compared to other general beamforming techniques.

I. INTRODUCTION

Speech technology 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 speech 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 will significantly degrade the performance of most speech technology systems. 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. This paper is aimed at contributing to research in the use of microphone arrays for speech acquisition for speech technology systems. We present an alternative beamforming technique with a post filter and compare its performance to other beamforming techniques. The speech signals obtained from these beamforming techniques are evaluated using objective quality tests that rely on mathematically based measures between original clean speech captured with a close-talking microphone and beamformed speech from a microphone array.

In exploring this topic, the principles of some basic beamforming techniques are discussed and evaluated. Thereafter, reviews of the objective quality assessment methods are given and results on the performance of the four element linear microphone array are presented and conclusions made.

II. BEAMFORMING TECHNIQUES

In this section the array processing techniques used in our experiments are examined. 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 [3]. 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 gives 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

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

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

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

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. 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) [4] 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 [5]. 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

$$y(f) = y_u(f) - y_a(f) \qquad (3)$$

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 1.

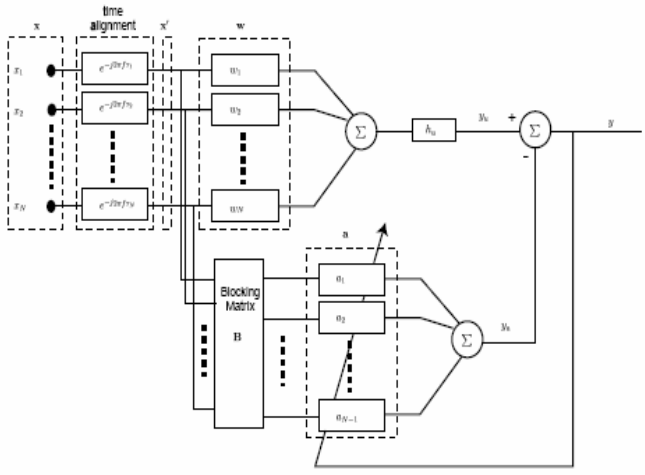

Figure 1: Generalized sidelobe canceller structure

C. Noise Canceling (NC)

This beamformer 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

$$\mathbf{x}''(f) = \mathbf{B}\mathbf{x}'(f) \qquad (4)$$

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 [5]. The standard Griffiths-Jim blocking matrix is [4]

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & 0 & \Lambda & 0 \\ 0 & 1 & -1 & 0 & \Lambda & 0 \\ \mathrm{M} & \Lambda & \mathrm{O} & \mathrm{O} & \Lambda & \Lambda \\ 0 & \Lambda & 0 & 1 & -1 & 0 \\ 0 & \Lambda & 0 & 0 & 1 & -1 \end{bmatrix} \qquad (5)$$

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

$$y_B(f) = \mathbf{a}(f)^T \mathbf{x}''(f) \qquad (6)$$

where y_B is a vector containing only noise samples. The positions of these samples are extracted in the noise canceling module and the corresponding positions in the upper path output are replaced with nulls. Thus effectively canceling noise in the overall system output, y. Figure 2 illustrates the proposed beamforming technique.

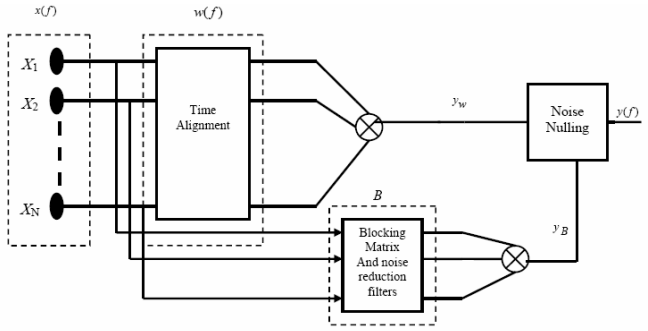

Figure 2: Active noise canceling beamforming structure

D. Post - Filtering

In the study of microphone arrays, the term post-filter refers to the post-processing of the array output by a single channel noise reducing filter. The Wiener filter provides a Minimum Mean Squared Error (MMSE) solution for a broadband input such as speech [6]. The squared error minimized by the post-filter is the sum of residual noise and signal distortion components at the output of the beamformer. The Wiener post-filter tries to find a compromise between signal distortion reduction and noise reduction. As a result signal distortion cannot be avoided entirely.

In this section we have reviewed three beamforming algorithms and a post-filtering technique. In the next section we discuss the quality measures used to evaluate thee microphone array beamforming algorithms.

III. QUALITY MEASURES

A. Itakura-Saito Distortion Measure (IS)

The Itakura-Saito distortion measure, also known as the maximum likelihood distortion, was first used for short-time spectral estimation of speech signals [7]. For an original clean frame of speech with linear prediction (LP) coefficient vector, $_{\phi}$, and processed speech coefficient vector, $_{d}$, the Itakura-Saito distortion measure, denoted as d_{IS} is given by,

$$d_{IS}(a_d, a_\phi) = \left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{{}_dR_\phi \quad {}^T_d}{{}_\phi R_\phi \quad {}^T_\phi}\right] + \log\left(\frac{\sigma_d^2}{\sigma_\phi^2}\right) - 1 \qquad (7)$$

where σ_d^2 and σ_ϕ^2 represent the all-pole gains for the processed and clean speech frame respectively [8].

B. Segmental SNR Measure (SegSNR)

Segmental SNR is formed by averaging frame level SNR estimates as follows,

$$d_{SEGSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1} s_\phi^2(n)}{\sum_{n=Nm}^{Nm+N-1}\left[s_d(n) - s_\phi(n)\right]^2} \qquad (8)$$

Frames with SNRs above 35dB do not reflect large perceptual differences, and can generally be replaced with 35dB in Eq. 8. Similarly, during periods of silence, SNR values can become very negative since signal energies are small. Therefore a lower threshold bound of -10dB is set for the segmental SNR [8].

C. Speaker Identification Performance

As speech acquisition is an integral part of most speech technology systems, it is important that a speech system be used as a measure of performance of the processed speech. Speaker identification (SID) 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 [9]. The system used here is Gaussian Mixture Model (GMM) [10] based speaker identification system. Figure 3 illustrates a typical speaker identification system.

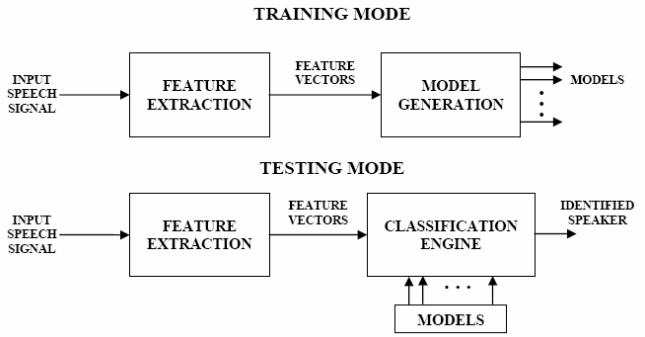

Figure 3: A typical speaker identification system

The section that follows describes the experimental configuration and results obtained from the quality measures described above.

IV. SYSTEM SETUP AND RESULTS

A. System Setup

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.

The whole microphone array system comprises three main components; *the linear array, data acquisition module and processing module*. Figure 4 illustrates these three components with the output being either an audio output or an input to another system.

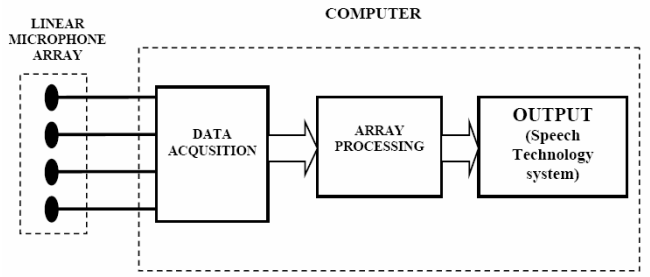

Figure 4: Microphone array system

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 [11].

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 (www.eagle.co.za). 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 200 speech files, comprising 100 training and 100 testing speech utterances, from the first 100 speakers from the TIMIT database were used. Each utterance was acquired from 50cm and perpendicular to 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.

B. Results

The objective quality measure results are presented in three areas; distortion measure, segmental SNR and speaker identification performance. There are several ways of obtaining overall quality scores. For most measures, finding a mean across a large test set is reasonable [8]. Table 1 below summarizes the results for four beamforming algorithms.

Table 1
Speech quality and speaker identification scores

	IS	SegSNR	SID
Single Mic.	7.79	-7.92	53%
Delay & Sum	6.99	-7.92	62%
GSC	7.69	-8.00	54%
NC	5.36	-7.49	63%
NC & Wiener	5.24	-6.83	65%

For IS and SegSNR measure, values closer to 0.0 reflect higher quality, whereas for SID, the goal is to achieve a performance as close as possible to 100%. We see that all the beamforming algorithms provide some quality improvement compared to single microphone speech. Since our main interest is in speech technology we regard the noise cancellation with Wiener filtering as having out performed the other three techniques.

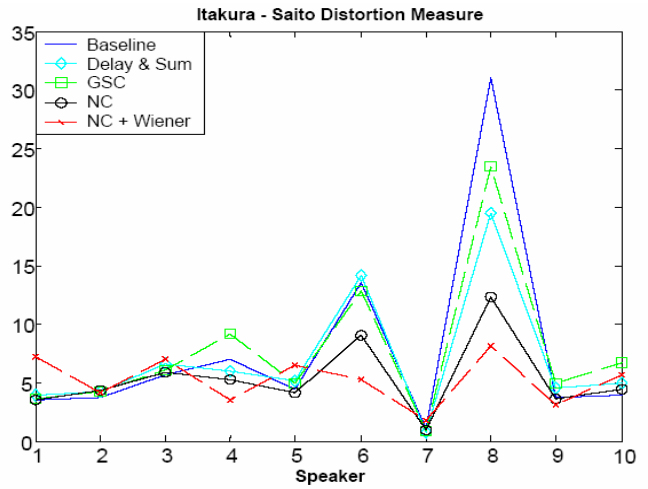

Figure 5: Overall mean IS quality measure for 10 speakers

Figure 5 shows a plot of average Itakura-Saito distortion measure for utterances from 10 speakers. The distortion is a comparison of beamformed speech and clean speech. Values close to zero show low distortion. As seen from the graph the NC + Wiener beamformer is more stable and generally outperforms the other techniques.

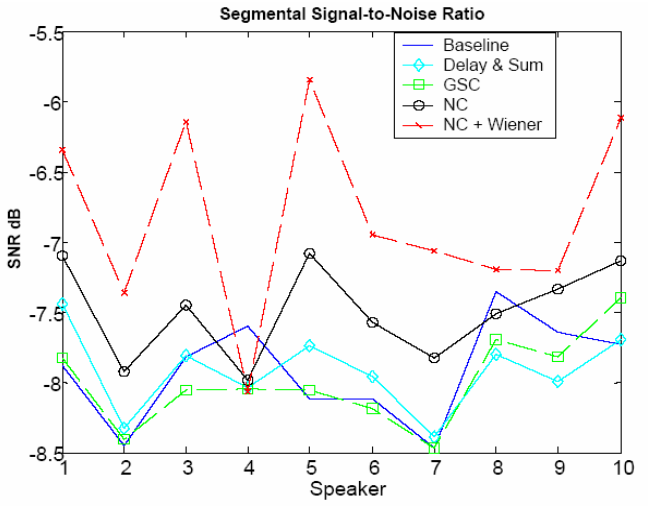

Figure 6: Overall mean Segmental SNR for 10 speakers

It has been shown in [12] that for clean speech recorded using a close-talking microphone, a GMM based speaker identification system similar to the one used here obtained a 100% identification rate. It should be noted that the experimental setup and data used in [12] 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.

V. CONCLUSIONS

The work presented here has demonstrated that using a microphone array for speech acquisition offers a performance advantage for speech technology applications in distant-talking environments. We found that beamforming algorithms reduced distortions to the desired speech and improved signal-to-noise ratios. The active noise cancellation with Wiener filter beamformer performed well in reducing distortion, noise and providing a superior signal-to-noise ratio compared to other beamformers.

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 other speech technology systems.
3. Investigate the effect of real people speaking towards and away from the microphone array.

REFERENCES

[1] D.V. Rabinkin, R.J. Renomeron, J. C. French, and J. L. Flanagan, "Optimum microphone placement for array sound capture," presented at SPIE, 1997.

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

[3] V. C. Raykar, "A study of various beamforming techniques and implementation of the constrained least mean squares (LMS) algorithm for beamforming," presented at ICASSP, Salt Lake City, 2001.

[4] 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, 1982.

[5] I. A. McCowan, "Robust speech recognition using microphone arrays," in *Electrical Engineering*: Qeensland University of Technology, Australia, 2001.

[6] M. Brandstein and D. Ward, *Microphone Arrays - Signal Processing Techniques and Applications*, 1 ed: Springer, 2001.

[7] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt, "Comparative Study of Several Distortion Measures for Speech Recognition," *Speech Communication*, vol. 4, pp. 317 - 331, 1985.

[8] J. H. L. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol for Speech Pprocessing Algorithms," presented at ICSLP, Sydney, Australia, 1998.

[9] H. Gish and M. Schmit, "Text-Independent Speaker Identification," in *IEEE Signal Processing Magazine*, 1994, pp. 18 -32.

[10] 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, 1995.

[11] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," presented at ICASSP, Seattle, Washington, 1998.

[12] D. A. Reynolds, "Large Population Speaker Identification Using Clean and Telephone Speech," *IEEE Signal Processing Letters*, vol. 2, 1995.

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