Optimizing of the Quality of Speech Output using Phase Difference in Signals

Obikwelu R Okonkwo, Nelson O Madu

Abstract:

The fluctuations in air pressure through which speech is transmitted, are regularly corrupted by a variety of sounds from other sources in both enclosed and open environments, including the humming of air conditioners, computer machines, reverberation of the fan, the bustling noises of a crowded street, car ground noise, the ambient rushing of wind in an open field or the speech babbles of other individuals at a social gathering. Signal processing algorithms struggle to process speech in the presence of even modest background noise in order to optimize the quality of such speech output. This work describes a study of techniques for optimizing speech quality via noise attenuation that entails the use of microphone arrays based on combination of beamforming and multi-filtering technique. This technique allowed distortionless signal components within a specified angle range of incidence from the desired direction and attenuated the interference signals outside this range. Consequently, the result indicated that the output speech quality improved as processed Speech Signal.

Keywords: *Optimization, phase difference, binaural beamforming, spatial filtering, technique, distortionless, algorithm*

1. INTRODUCTION

Ultimately, speech is purposely used for communication. Human speech communication takes place in complex acoustic backgrounds. In a natural environment, target speech is usually corrupted by such acoustic interference (ie sound sources), contending voices, and ambient noise, creating an output speech quality challenge [1]. Another scenario is where human displayed the ability to focus aural attention on a particular conversation while filtering out a range of other sound sources in a noisy room. Accordingly, this is described as a cocktail-party problem [2]

It is notable that human speech understanding remains robust in the presence of such interference. [3] speech-in-noise research showed that speech signals combine several acoustic properties that contribute to compensating for signal distortions and noisy interferences. Essentially, noise source can also be stationary (e.g., generator, fan humming or an air-conditioner) or nonstationary (e.g. moving vehicles, wind etc.). It represents one of the foremost disquiets that constituted the acoustic background of verbal communication over human history.

The advent and the evolution of the adaptive listening capabilities in humans have developed under the continuous influence and pressure of such acoustic worries [3]. In all practical situations, the received speech waveform contains some form of noise component. The noise may be a result of the finite precision involved in coding the transmitted waveform (quantization noise), or due to the addition of acoustically coupled environmental noise. Depending on the amount and type of noise, the quality of the received waveform can range from degraded to annoying to listen, and totally unintelligible. However, for machines this speech signal separation is still a daunting challenge. The challenge of eliminating the unwanted noise component from a received signal has been the subject of numerous studies.

2. REVIEW OF RELATED LITERATURE

In recent years, tremendous and remarkable progress has been made in the domain of optimizing the quality of speech output in Signal Processing using enhanced technique. [11] opined that signal processing techniques are imperative to extracting reliable acoustic features from the speech signal, and stochastic modeling algorithms are useful for representing speech utterances in the form of efficient models, such as hidden Markov models (HMMs), which simplify the speech recognition task.

[12] indicated that a time-honored objective in signal processing is an improvement technique capable of separating speech signals from noise (ie speech signals and noise picked up by the same microphone). They are largely based on statistical analysis of speech and noise, followed by estimation of clean speech from noisy speech. Model approaches to make this happen comprise of spectral subtraction, Wiener filtering, and mean-square error estimation.

[4] claimed that the mean-square error estimation, models speech and noise spectra as statistically independent Gaussian random variables and estimates clean speech accordingly.
[5] and [13] provided an optimal method for deriving a filter that tends to suppress the noise while leaving the desired signal relatively unchanged. The design of these filters requires that the signal and the noise be stationary and that the statistics of both signals be known a priori. In practice, these conditions are hardly achievable. The classical approach to noise cancellation is a passive acoustic approach. Passive silencing techniques such as sound absorption and isolation are inherently stable and effective over a broad range of frequencies. However, its observed that these tend to be expensive, bulky and generally ineffective for cancelling noise at the lower frequencies. The performance of these systems is also limited to a fixed structure and proves impractical in a number of situations where space is at a premium

and the added bulk can be a hindrance. The shortcomings of the passive noise reduction methods have given impulse to the research and applications of other methods of controlling ambient noise in the environment.

3. TECHNIQUE ENHANCEMENT AND THE OUTPUT OF SPEECH SIGNAL PROCSSING

There has been substantially more signal processing that applied direct technique than combination of techniques. [6] noted that one of the simplest and oldest ways to do this is with the Delay and Sum (DAS) beamformer. Given that, the direction of the target source is specified, the task of forming the beam and filtering the incoming signal remains. DAS basically reverses this process of beamforming by delaying the outputs of all microphones except the last one, which is the reference microphone. Considering that the delay between the microphones usually consists of a non-integer number of time-samples, the delay is accomplished by a phase change in the Fourier domain rather than in the time domain. [7] noted that the weighted signals are then summed and transformed back to the time domain to create the output signal.

[8] highlighted that focusing the response of the array to a certain direction, is the approach by which the MVDR beamformer takes a technique different from the DAS. The MVDR firstly tries to minimize the signals from interferers and noise. MVDR primarily aims at minimizing the total output power. Thus the minimum average output power of the beamformer per time frame i and per frequency-bin k is given as

$$\min\{P\} = \min\{w^H Rw\};\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.0)$$

where ^{H} denotes the Hermitian transpose, *w* is the weight vector and *R* is the autocorrelation matrix of the noise. Thus, it is required to make an estimate of the noise signal before performing the calculations. One way to obtain this signal is to use the fact that the first few time frames are likely to consist only of noise.

With this estimation the autocorrelation matrix can be calculated with:

$$R = \begin{matrix} |\hat{v}_1|^2 & \hat{v}_1\hat{v}_2 H \\ \hat{v}_1\hat{v}_2 & |\hat{v}_2|^2 \end{matrix} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.1)$$

where $\hat{v}_n$ denotes the noise received in the frequency domain at the nth microphone.

From equation (3.0) it can easily be seen that **P** is minimized if *w* is the zero vector **0**. Despite satisfying the primary aim, this trivial solution will not steer the response in the wanted direction

[8] stated that MVDR requires a constraint: the implication is that unity gain should be maintained in the target direction. This constraint can be defined by equation (3.2).

$$c^H w = 1:\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.2)$$

This is because the delay in the signal from the different sensors can be written as

$$c = [1; e^{j\zeta}]^T$$

When (equation 3.0) is combined with (equation 3.2) we end up with the final solution for

$$w = R^{-1}c(c^H R^{-1}c)^{-1}:\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.3)$$

It can be noted that if *R* is the identity matrix *I*, the weights *w* would simplify to the same weights as with a DAS beamformer.

$$w = I^{-1}c(c^H I^{-1}c)^{-1} = c(c^H c)^{-1} = c(||c||^2)^{-1} = 0.5c \;\ldots\ldots(3.4)$$

[9] stated that R simplified to I would correspond to the correlation between the 8 microphones being zero, though this does not imply that the 8 microphones are independent of each other. [8] also inferred that since we are not dealing with a deterministic signal this may not always be possible to calculate autocorrelation for every time frame and every frequency bin.

[10] has used a model-based methods to capture the feature statistics of isolated sources, and consequently source separation becomes the problem of using prior models to identify a set of source signals that combine to result in the observed mixture signal. These approaches, ultimately target either to enhance speech or attenuate noise in noisy speech. In circumstances where the interference is a competing speech, speech enhancement algorithms are not able to separate them.

3.1 METHODOLOGY

The specific methodology adopted in this study was an enhanced binaural beamforming which entails spatial filtering technique and independent component analysis combined to optimize the quality of speech output in signal processing. In [14], it was revealed that Paul Lueg was the first to realize the possibility of attenuating background noise by superimposing a phase flipped wave for the concept of active noise cancellation. Also this approach requires spatial selectivity. [6], has stated that the beamformer is required to distinguishes between the components from different directions of an incoming signal in order to suppress hugely and efficiently the parts of the signal that are not coming from the target source(s). Consequently, there is need to analyse the incoming signal and differentiate between different angles of incidence. In practical beamforming, when the microphone array picks up a signal coming from an angle other than 0 or 180, every consecutive microphone will experience an increased delay. This is because the signal entering from an angle needs to travel an additional distance, Δx, to the next microphone in the array. Δx is proportional to the distance between the microphones and the angle of incidence, as has been expressed in figure 3.1

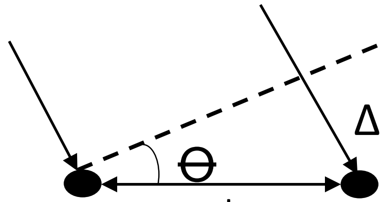

Figure 3.1 - The signal travels an additional distance Δx to the next microphone [6],

Considering that the microphones are fixed, the distance between them is fixed too. It is therefore possible to relate the additional distance to the angle of incidence of the signal. However, since the speed of sound is also fixed, [6] noted that the time delay can be deduced from the direction of the signal by

$$t = \frac{d}{C}\ \sin(\theta):\text{……………………..................... } (3.5)$$

Where t is the time delay in [s] and d, are the distance between the microphones in [m], c the speed of sound in [m/s] and θ the angle between the microphones in [rad].

Waves are usually described by variations in some parameter via space and time. The value of this parameter is called the amplitude of the wave, and the wave itself is a function specifying the amplitude at each point. In many cases such as classic wave equation, the equation describing the wave is linear. Once this is real, the superposition principle can be applied. As expected, the undesired sine has amplitude larger than the amplitude of the desired sine.

Based on the beamforming methodologies reviewed, several techniques were identified. The proposed system would achieve the main research objective which is to develop a new microphone array processing to attenuate noise interference (ambient noise) using an enhanced MV beamforming technique. One of the first major steps in beamforming is focusing. In signal processing, delays are applied to control the signals coming from all microphones to hit a given point. The given point is called the focal point. The signals received by the elements (raw signal data) are delayed such that they sum up signals coming from the same directions (correlated signals) constructively and destructively sum those signals coming from different directions (uncorrelated signals).

The Figure 3.2 below shows a vector superposition of a large amplitude N of simple harmonic vibration of equal amplitude A and equal successive phase difference δ.

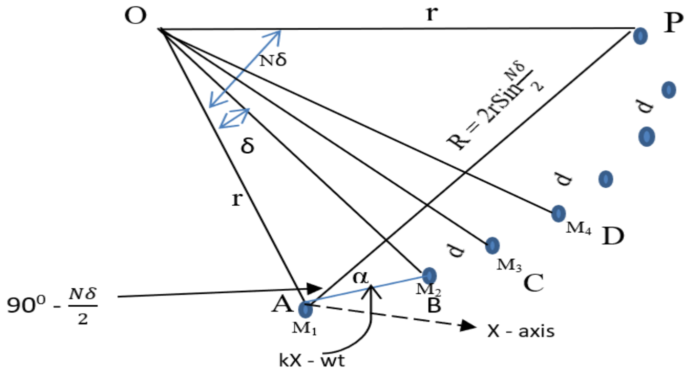
O
r
P
Nδ
$R = 2r\text{Sin}\frac{N\delta}{2}$
d
δ
d
r
d
M₄
D
M₃
C
d
α
M₂
B
$90^0 - \frac{N\delta}{2}$
A
M₁
X - axis
kX - wt

Figure 3.2 - Vector Superposition

The amplitude of all the resultant (ie Chord)

(3.5)

$$R \;=\; 2\,r\,\text{Sin}\,\frac{N\delta}{2} \;=\; A\,\frac{Sin\frac{N\delta}{2}}{Sin\,\frac{\delta}{2}}$$

……………………………..*(3.6)*

Consider the triangle AOP. Splitting the triangle AOP into two right angles, we have;

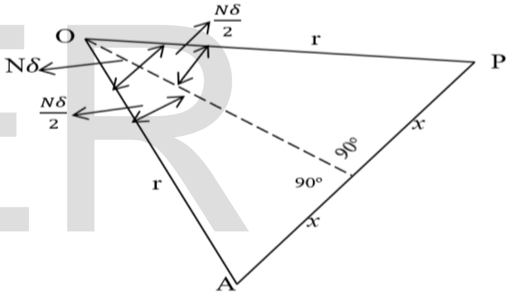

Figure 3.3 – Splitting the Triangles for Microphone positions

Therefore, for one of the Right-Angled triangles,

$$\text{Sin}\frac{N\delta}{2} = \frac{x}{r} \qquad \text{………………………..}(3.7)$$

Where *x* is Opposite, and r is the Hypotenuse.

$$X = r\ \text{Sin}\frac{N\delta}{2} \qquad \text{…………………………}(3.8)$$

Then, for both right angled triangles,

$$x + x = r\ \text{Sin}\frac{N\delta}{2} + r\ \text{Sin}\frac{N\delta}{2} = 2r$$
$$\text{Sin}\frac{N\delta}{2}\text{…………………………..}(3.9)$$

$$\text{therefore, } R = 2r\ \text{Sin}\frac{N\delta}{2} \qquad \text{………………………………}(3.10)$$

Its phase with respect to the first contribution is given by α;

To get the α which is angle between the microphones (M_1); we subtract the angle OAP from angle OAB. However, first let's calculate the two angles. From the previous, triangle OÂP is gotten, by subtracting $\frac{N\delta}{2}$ from 90°;

OÂP = $90^{o} - \frac{N\delta}{2}$ ………………………………..*(3.11)*

For angle OÂB ;

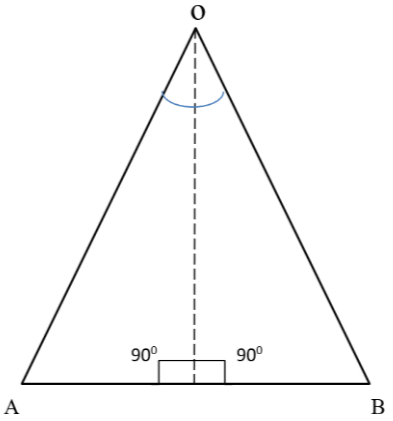

Figure 3.4 - Delayed distance

OÂB = 90° -…………………………..*(3.12)*

Therefore, θ = OÂB - OÂP

$= 90^{o} - \frac{\delta}{2} - (90^{o} - \frac{N\delta}{2})$……. *(3.13)*

$= 90^{o} - \frac{\delta}{2} - 90^{o} + \frac{N\delta}{2} = \frac{N\delta}{2} - \frac{\delta}{2} = (N - 1)\frac{\delta}{2}$…………..*(3.14)*

Therefore α = $(N - 1)\frac{\delta}{2}$ …………..*(3.15)*

From triangle OAP, following arrangement of microphones symmetry on y-axis we use the Cosine rule

$R = r^2 + r^2 - 2r^2 \text{ Cos } N\delta$ ……………….*(3.16)*

$= 2r^2(1 - \text{Cos}\delta)$ ……….…….*(3.17)*

$R^2 = \frac{4r^2(1 - \text{Cos}N\delta)}{2}$ ………………….*(3.18)*

$R = \frac{2r\sqrt{1 - CosN\delta}}{2}$ …………………..*(3.19)*

$R = 2r\text{Sin}\frac{N\delta}{2}$ ……………………..…*(3.20)*

From triangle OAB, it can be seen that delay distance from the speech source to the microphone is *dSin* θ where θ is the angle substituted by the microphones within an area

Speed of Sound (c) $= \frac{distance(\text{dSin}\frac{(N-1)\delta}{2})}{time(t)}$ …………………………..*(3.21)*

Therefore, $t = \frac{\text{dSin}\frac{(N-1)\delta}{2}}{c}$ …………………………..………...*(3.22)*

Where *c* is the Speed of Speech = 343m/s

Also, from triangle OAB using same Cosine rule;

$A = 2r\text{Sin}\frac{\delta}{2}$ ……………………………………………………..….*(3.23)*

Dividing $\frac{R}{A}$, we obtain:

$= \frac{2r\text{Sin}\left(\frac{N\delta}{2}\right)}{2r\text{Sin}\left(\frac{\delta}{2}\right)}$ ……………………..*(3.24)*

then;

$R = A\frac{Sin\frac{N\delta}{2}}{Sin\frac{\delta}{2}}$ ……………………..*(3.25)*

Note that δ (Phase difference) is the wave number multiplied by the extra distance travelled by the second wave.

$\delta = k$ x Sinθ (Let's see the extra distance travelled)…………………………………………….…*(3.26)*

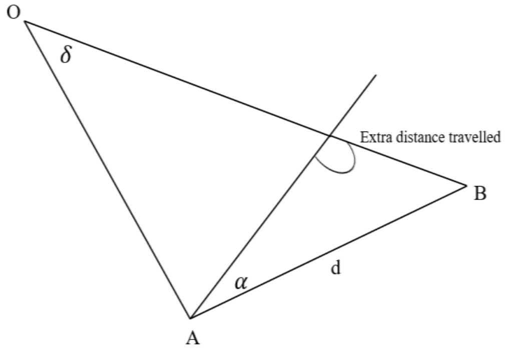

Figure 3.5 – Estimation of Extra Distance travelled

Extra distance travelled = dSinα

Recall also equation (3.15) *which is* $\alpha = (N - 1)\frac{\delta}{2}$

δ = KdSinα …………………………..*(3.27)*

Where K = wave number = $\frac{2\pi}{\lambda}$ ………..*(3.28)*

and λ = wavelength

$$R = A\frac{Sin\frac{N2\pi\, dsin\theta}{2\lambda}}{Sin\frac{2\pi\, dsin\theta}{2}} \quad \ldots\ldots\ldots\ldots(3.29)$$

$$R \;=\; A\frac{Sin\frac{Ndsin\theta}{\lambda}}{Sin\frac{\pi\, dsin\theta}{\lambda}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3.30)$$

Now for the resultant wave, the amplitude is given by R above the angle with respect to the X-axis as:

$$(Kx - wt + \alpha) \ldots\ldots\ldots\ldots\ldots(3.31)$$

where α is :

$$\alpha \;=\; (N\text{-}1)\,\frac{\delta}{2} \qquad \text{in equation } (3.15);$$

Thus, the sum of all waves of a particular target point P, (or angle θ) is

$$Y1 + Y2 + Y3 + \ldots\ldots Yn. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.32)$$

This can be written as :

$$A\frac{Sin\frac{N\delta}{2}}{Sin\,\frac{\delta}{2}}\, \text{Sin}\left(kx - wt + \frac{N-1}{2}\,\delta\right)\ldots\ldots\ldots\ldots(3.33)$$

(ie Amplitude of the sum of the wave x Angle of sum wave with respect to the *X* –axis)

Thus, signal from the *X* axis doubles while the signal from *Y* axis attenuates.

3. CONCLUSION

From the study we were able to achieve enhanced signal output by the signals from the X axis additively summing up while the Y-axis signals attenuates. This was because, the accumulated normalised signal on the X-axis signals are correlated signals. However, the signals from the Y-axis are uncorrelated with increased angle difference which in turn makes the successive signals travel extra distance from the source of interest, introducing more noise. The signal power becomes uncorrelated, and they destructively add each other.

REFERENCES

[1] Wang W (2015); Frequency Domain Source Separation - UDRC Summer School, Surrey, 20-23 July.

[2] Cherry E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am., 25(5):975–979.

[3] Meyer J, Dentel L ; Meunier F (2013), Speech Recognition in Natural Background Noise Published: November 19, {https://doi.org/10.1371/journal.pone.0079279)

[4] Healy EW, Yoho SE, Wang Y, Wang D (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. J Acoust Soc Am. 2013 Oct;134(4):3029-38

[5] Kalman R. (1960).,"On the general heory of control," n Proc. 1st IFAC Congress. London: Butterworths,

[6] Ganesh T., Sekhar Rao R.V.Ch., Prasad P.H.V; Ashok kumar N, Vasundhra P. (2016) *; 'Noise Reduction Using Beamforming Algorithms'* - International Journal of Engineering Science and Computing, April.

[7] Mark A, Hendrik P, Arjan D (2012), 'Two Sensor Array Beamforming Algorithm'. July 4.

[8] Van de Sande, J.(2012.), "Real-time beamforming and sound classification parameter generation in public environments," Master thesis, Delft University of Technology, Feb.

[9] Annis, C (2017) "Correlation." www.statisticalengineering.com/ correlation.htm

[10] Ellis, D. P. W. (2006). "Model-based scene analysis," in Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, edited by D. L. Wang and G. J. Brown (IEEE Press/Wiley, New York) (in press).

[11] Li J, and Lee CH, (2007). Soft margin feature extraction for automatic speech recognition. Proc INTERSPEECH; 30–3.

[12] Loizou P. C. (2007). Speech Enhancement: Theory and Practice (CRC Press, Boca Raton, FL:), Chap. 5–8.

[13] Kalman R. and Bucy R. (1961), "New results in linear filtering and pre- diction theory," ***Trans. ASME, ser. D, J. Basic Eng.,*** vol. 83, pp. 95-107, Dec.

[14] Harris, Bill (2007). "How Noise-canceling Headphones Work." How stuff works. 16 July http://electronics.howstuffworks.com/noise-canceling-headphone.htm

Authors:

Obikwelu R Okonkwo
Department of Computer Science; Nnamdi Azikiwe University, Awka Nigeria.
ro.okonkwo@unizik.edu.ng

Nelson O Madu
Department of Computer Science; Nnamdi Azikiwe University, Awka, Nigeria.
noge_m2@yahoo.com