NOISE-CON
2022 June 13-15
Lexington, KY

A multi-stage filter for separating speech from background noise

Curtis L. Garner[1]
Brigham Young University
Department of Mechanical Engineering, 350 Eng. Bldg., Provo, UT 84602

Tyler Sanders[2]
Brigham Young University
Department of Mathematics, 275 TMCB, Provo, UT 84602

Scott D. Sommerfeldt[3]
Brigham Young University
Department of Physics and Astronomy, N283 ESC, Provo, UT 84602

Jonathan D. Blotter[4]
Brigham Young University
Department of Mechanical Engineering, 350 Eng. Bldg., Provo, UT 84602

ABSTRACT

This paper proposes a multi-stage method for filtering speech signals that contain large-amplitude background noise. Many commonly encountered noise sources such as engines, propellers, or turbines produce noise fields that are too spatially incoherent to be effectively filtered using traditional methods. The proposed filtering method uses two filter stages to separate speech from background noise: one to remove the coherent components of the noise, and the other to reduce the incoherent components of the noise. The first stage uses the LMS algorithm. In this stage, strategically placed reference microphones are used to eliminate coherent noise from an array of error microphones. The second stage uses a beamforming algorithm to reduce the remaining incoherent noise. Results are presented that demonstrate the effectiveness of the multi-stage filter.

1. INTRODUCTION

Digital filtering of acoustic signals to reduce noise has been employed in many applications over the past several decades. The filtering algorithms are extremely diverse, ranging from computationally expensive processes used to improve music recordings to the rapid filters used to enhance telecommunications. The method presented herein focuses on developing an acoustic filtering method specifically for detecting and enhancing speech signals in the presence of high-amplitude complex noise. Noise sources such as engines, propellers, and turbines typically have both coherent and incoherent components. The mixed nature of these noise signals makes them difficult to filter effectively using traditional methods. These difficulties can greatly hinder verbal communication when complex noise sources are present.

[1] curtislgarner@gmail.com
[2] pianomanty@gmail.com
[3] scott_sommerfeldt@byu.edu
[4] jblotter@byu.edu

2. BACKGROUND

Speech filtering has been applied in many situations across multiple fields. Some of the earliest efforts to filter speech are passive methods, such as absorption or isolation [1] [2]. These methods rely on physical structures to alter noise propagation. As digital signal processors (DSPs) have become less expensive and more powerful, various digital methods have emerged to address the shortcomings of passive methods. Digital speech filtering is utilized in many everyday applications such as telecommunications and hearing aids. [3] [4] [5]. In recent years, many different specialized algorithms have been shown to be well suited for removing certain types of noise from speech signals. [6] [7].

One common digital method for reducing background noise is the LMS algorithm and its variants, such as the filtered-X algorithm. The filtered-X algorithm, for example, is generally used to perform active noise cancellation and is based on the underlying LMS algorithm. The LMS algorithm essentially uses one microphone (called a reference microphone) to track the noise source, then passes this signal through an adaptive filter. The filtered signal is then added to the signal from an additional microphone called an error microphone. The controller tries to adapt the filter such that the filter output signal is the inverse of the unfiltered error signal. If multiple error microphones are included, the controller applies a separate filter to the reference signal for each error microphone. Because the reference microphone is placed near the noise source, the filter tends to remove noise, while leaving the desired signal largely unaltered. An overview of the elements of the LMS algorithm can be seen in Figure 1.

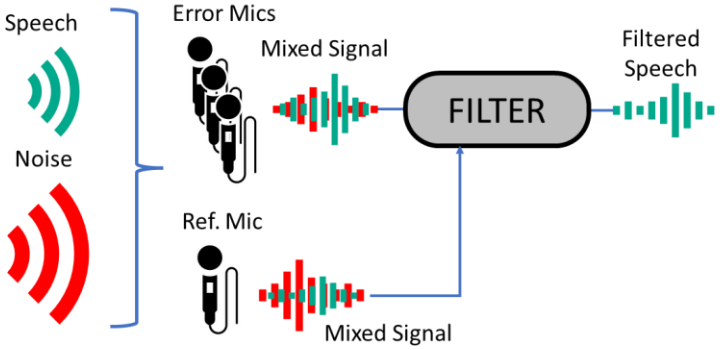

Figure 1: A simplified diagram of the LMS method. The error array and the reference microphone both pick up a combination of speech and noise, though the reference signal is dominated by noise. The filter controller uses the reference signal to remove noise from the error signal(s), resulting in a clearer speech signal.

The method used for updating the filter involves a least-mean-squares (LMS) solution, so the filter is typically referred to as the LMS algorithm [8], [9]. The effectiveness of an LMS filter is closely related to the coherence between the reference microphone(s) and the error microphone(s) [10]. In real-world scenarios, this is often the limiting factor for filter performance. Over the past few decades, the LMS algorithm and its variants have been applied to many types of noise control problems [11], [12]. This has led to the development of many adaptations of the basic algorithm intended to address specific types of problems [13], [14]. While the simplest forms of the LMS filter are based in the time domain, several frequency domain methods have also been proposed [15], [16]. A multiple reference approach is needed when the noise is produced by multiple sources. The basic procedure is to assign a reference microphone to each source, then remove them sequentially from the error microphone signal until only the desired signal remains [17].

Another common method for separating a desired signal from background noise is beamforming. A beamformer utilizes a microphone array of known dimensions. In many cases, this is a uniform linear array, which means that all the microphones are placed in a line with equal spacing between each microphone, [18], though other array shapes have been studied as well [19]. The physical separation of the microphones results in each microphone receiving a slightly different version of the same signal. In the simplest case, the received signals will simply be time-shifted. The beamformer uses information about the direction of arrival (DOA) of the incoming signal to alter and sum the received signals to maximize the signal coming from the desired direction, also called the “look” direction.

The simplest beamformers use the delay-and-sum method, which simply time-shifts the received signals to align the sound coming from the look direction. Another consideration in beamformer design is bandwidth. Many beamforming algorithms are designed for narrow-band signals [20]. Others, such as a sub-array beamformer, analyze broadband signals by first dividing them into several, narrower bands, then recombining the output signals [21].

3. PROPOSED METHOD

This paper presents an acoustic filtering method for detecting and enhancing speech signals in the presence of high amplitude background noise. The filtered speech signal can then be played back to the listener in real time. This method can be divided into three major steps. First, sound is recorded by several microphones. Next, these recorded signals are then processed to remove unwanted noise. This filtering process includes an LMS stage and a beamforming stage. All of this filtering is accomplished by a digital signal processor (DSP) in real-time. Finally, the filtered signal is played back to the listener either through headphones or speakers. These steps are described in greater detail below.

3.1. Sound Acquisition

Sound is captured by two sets of microphones placed in the environment. The first set of microphones is arranged in an array of known dimensions. These microphones are referred to as error microphones, or collectively as the error array, and are placed away from dominant noise sources. The purpose of the error array is to capture the voice signal with minimal background noise, although the level of background noise is still expected to be substantial. One or more microphones are also placed in close proximity to major noise sources. These are referred to as reference microphones. Unlike the error microphones, these microphones are intended to capture as much noise as possible, with minimal voice content. Signals from both the error microphones and reference microphones are sent to the DSP for processing.

3.2. LMS Stage

The signals received by the error microphones are comprised of three major components: coherent noise, incoherent noise, and coherent speech. Coherence in this case specifically refers to the mean-squared coherence of the signal received at an error microphone and the signal received at a reference microphone. The LMS algorithm is applied in the first filter stage to remove the coherent component of the noise received by each microphone in the error array.

This algorithm performs the calculations and updates every time new data are available from the microphones. Each incoming reference signal is stored in a circular buffer of length L_1. The value of L_1 is set arbitrarily, and must be larger than the delay time, measured in number of samples, between the reference microphones and the error microphones. To simplify the calculations, these circular buffers are concatenated into a single vector R_{bf} with length $N_r \cdot L_1$, with N_r being the number of reference microphones.

$$R_{bf} = [R_1 \quad R_2 \quad \cdots \quad R_{Nr}] \tag{1}$$

where

$$R_i = [R_{i,t} \quad R_{i,t-1} \quad \cdots \quad R_{i,t-L_1+1}] \tag{2}$$

Similarly, the concatenation of the control filters from each reference microphone to each error signal can be written as

$$W_i = [W_{1i} \quad W_{2i} \quad W_{3i} \quad \dots \quad W_{Nri}]$$

At each time step, the filtered error signals are calculated by multiplying R_{bf} by W_i, which is also of length $N_r \cdot L_1$

$$E_{i,t} = d_{i,t} - R_{bf} \cdot W_i^T \tag{3}$$

where $E_{i,t}$ is the current value in the i_{th} filtered error signal, $d_{i,t}$ is the current data value from the i_{th} error microphone, and W_i^T is the transpose of W_i. After each value in each filtered error signal is calculated, the respective filter vector W_i is updated:

$$W_i = W_i + \mu R_{bf} E_{i,t} \tag{4}$$

μ is an arbitrary value that controls how fast the filter converges. This value can either be fixed, or be dynamically allocated based on the values in R_{bf}. In the cases presented in the validation section, μ is calculated as follows:

$$\mu = 0.1/(R_{bf} \cdot R_{bf}^T) \tag{5}$$

As the filtered error signals are calculated, they are stored in circular buffers similar to the reference microphone signals, albeit with length L_2 instead of L_1.

$$E_i = [E_{i,t} \quad E_{i,t-1} \quad \cdots \quad E_{i,t-L_2+1}] \tag{6}$$

The amount of noise removed by the primary filter stage is heavily dependent on the coherence between the error microphones and the reference microphones [10]. Regardless of the overall noise reduction, this stage will reduce the coherence between the error signals and the reference signals, such that $C_{xy}(E_i, R_j) < C_{xy}(d_i, R_j)$ for all i and j.

3.3. Beamforming Stage

Unlike the LMS algorithm, which can only filter coherent noise, beamforming reduces noise based solely on its direction of arrival. Thus, the beamforming filter stage can reduce the incoherent noise that remains in the filtered error signals, provided the direction of arrival of the speech signal does not align with that of the noise signal. In this work, it is assumed that the DOA of the speech signal is known.

To minimize computational expense, the beamforming filter stage uses a standard delay-and-sum algorithm. The beamforming algorithm requires one of the error microphones to be selected to be the “center microphone”, to which all other microphone signals will be aligned. The filtered error signal associated with the center microphone is left unaltered. Each of the other filtered error signals is time shifted such that the speech signals are all aligned. The time shift for each filtered error signal is found using the relative position of the microphone, as well as the direction of arrival of the speech signal.

$$\Delta s_i = \frac{Fs}{c} d\, cos(\angle DOA + \pi - \alpha) \tag{7}$$

As before, d is the distance between the i_{th} microphone and the center microphone, and α is the physical angle between the same. Δs_i is the delay (in samples) applied to the i_{th} filtered error signal, and is rounded to the nearest integer. Fs is the sampling frequency. In many cases, the delay time Δs_i is negative. In order to accommodate this, all of the filtered signals are delayed by a static number of samples. This delay can be found as follows:

$$\Delta s_{min} = \frac{d_{max} Fs}{c} \tag{8}$$

where d_{max} is the largest distance between the center microphone and any other error microphone and Δs_{min} is the minimum static delay that will work for all possible directions of arrival.
After the delay times have been determined, the appropriate data points from the filtered error signal buffers are added together, which gives the final output at the current time step.

$$S_t = \sum_{i=1}^{N_e} \frac{E_{i,\Delta s_{min}+\Delta s_i}}{N_e} \tag{9}$$

Here S is the output of the beamforming filter stage, and N_e is the number of error microphones in the array. This is the signal sent to the headphones or speakers to be played back to the listener. An overall schematic of the proposed filter can be seen in Figure 2.

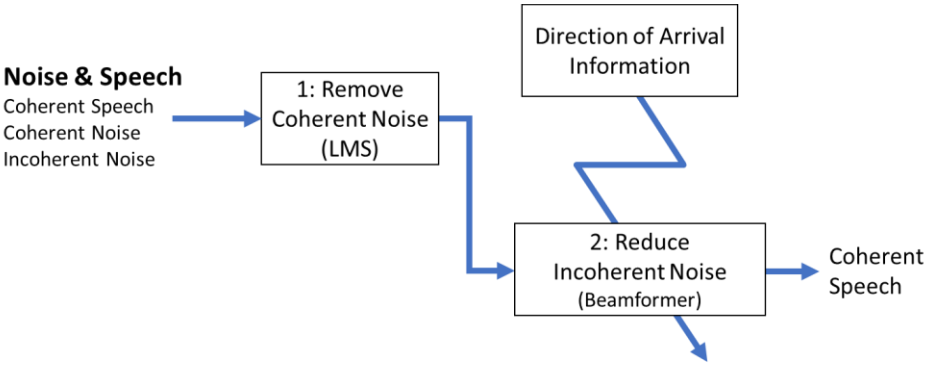
Noise & Speech
Coherent Speech
Coherent Noise
Incoherent Noise
1: Remove Coherent Noise (LMS)
Direction of Arrival Information
2: Reduce Incoherent Noise (Beamformer)
Coherent Speech

Figure 2: An overview of the proposed filter structure. The incoming signal contains coherent speech, coherent noise, and incoherent noise. Coherent noise is removed by the LMS filter. Incoherent noise is reduced by the beamforming filter using the provided DOA. This results in a much more comprehensible speech signal.

4. VALIDATION

Validation of the proposed method has been performed both virtually and experimentally. Virtual validation was accomplished entirely in a MATLAB simulation, using pre-recorded speech and noise data and virtual microphones. This type of validation was chosen for its simplicity, and also because virtually constructed microphone signals allowed for greater control over experimental variables, such as signal to noise ratio, coherent to incoherent noise ratio, etc.). Another advantage of the virtual simulation was that it eliminated several possible confounding factors such as ground/wall reflections and additional noise sources in the environment.

Experimental validation was performed in an anechoic chamber. A real voice signal was used. Machine noise was simulated by playing pre-recorded machine noise over a pair of loudspeakers, and sound was recorded with several microphones. This validation method allowed the proposed method to be tested in an environment much closer to its intended application.

4.1. Evaluation Metrics

Because the proposed method is intended to remove background noise from speech signals, the most informative performance metric is the change in noise level from the signal originally received to the final filter output. The mixed nature of these signals (containing both speech and noise) makes this measurement somewhat difficult to make. The average sound power of a signal is easy to compute, but there isn’t a simple method to determine how much of the sound power comes from speech, and how much comes from noise. To address this issue, the equations above have been set up to leave the amplitude of the speech signal as unaltered as possible. The input and output signals are then trimmed so that only “silent” sections remain (silent meaning no speech content, only background noise). The average sound power of the trimmed signals can then be calculated directly, giving a close approximation of the overall noise reduction.

In addition to noise reduction, successful filtering also requires that the speech signal remains clear and distortion-free. The overall comprehensibility of the filtered speech signal is subjective, so qualitative observations were made in place of direct calculations.

4.2. Virtual Validation

Virtual validation of the proposed method was carried out by simulating the noise environment and all of the microphones in MATLAB, then processing the signals as detailed above. Seven error microphones were distributed within 1 meter of the origin. The virtual noise source and virtual speech signal were placed 4-6 maters from the origin. See Figure 3 for an approximate diagram of this setup. The signal received at each microphone was calculated by scaling the voice and noise signals proportional to the inverse of distance to the source, and applying the appropriate time delay to each based on the location of the virtual microphone. Incoherent noise was included separately in order to make the virtual microphone signals more closely match what a real microphone would receive. This was accomplished by adding white noise to each microphone, proportional to the noise signal. Thus each virtual microphone signal contained voice, coherent noise, and incoherent noise.

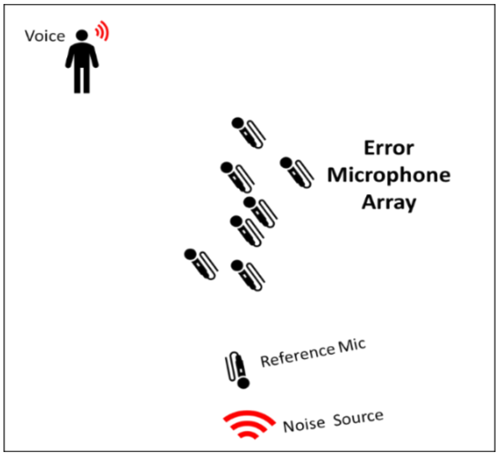

Figure 3: A rough visual of the layout used in the MATLAB simulation. Dimensions not to scale.

The results obtained from this simulation can be seen in Figure 4. The original signal (taken from one of the error microphones) is clearly dominated by noise. The filtering process reduced the noise level by about 15 dB, with 6.5 dB being removed by the LMS algorithm and the other 8.5 dB being removed by the beamforming algorithm. Listening to the filtered signal confirms that the voice is much more understandable after filtering.

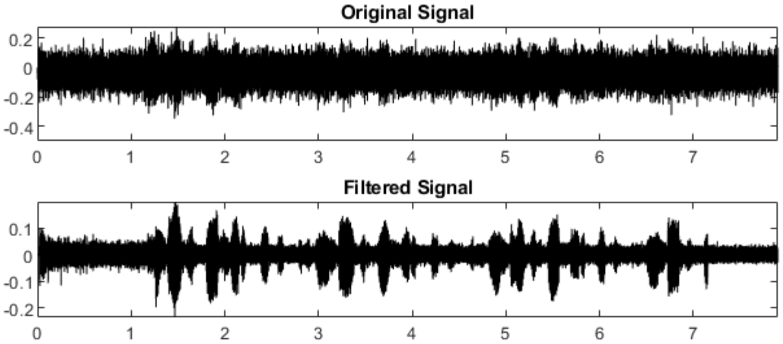

Figure 4: Results of the virtual validation of the proposed method. In the original signal, the voice content is entirely obscured by noise. In the filtered signal, noise levels are well below the voice level.

4.3. Experimental Validation

Experimental validation of the proposed method was carried out by conducting several tests in an anechoic chamber. Between six and fourteen error microphones were arranged in a cross-shaped array in the center of the chamber. A person stood a few meters away and read one or more of the Harvard sentences [22]. Pre-recorded noise was played over one or two loudspeakers also located a few meters away from the error array. A reference microphone was placed about thirty centimeters in front of each loudspeaker.

Below are some of the results obtained from a representative experimental setup. Figure 5 shows the approximate locations of the array, loudspeakers, and talking person. In this experiment, the two loudspeakers played separate, but similar, noise recordings.

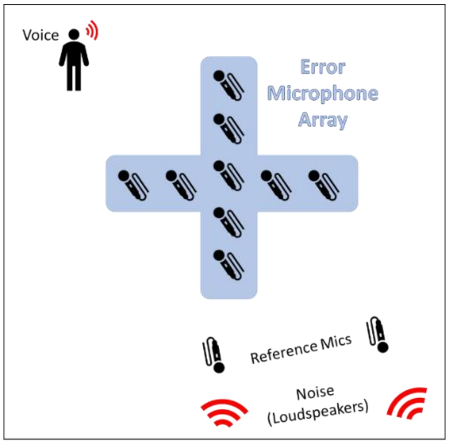

Figure 5: An experimental setup used in validating the proposed method. Dimensions are roughly to scale, with each side of the figure representing about 4 meters. The error microphone array contained 14 microphones (only 9 pictured).

Using the methods described in Section 3, the noise reduction obtained in this experiment was 27.5 dB, with the majority of the reduction coming from the first filter stage (LMS algorithm). These results can be seen in Figure 6.

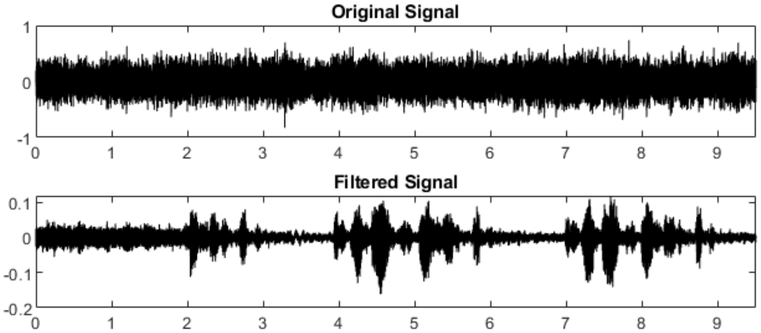

Figure 6: Results from experimental validation. In the original signal, the voice content is entirely obscured by noise. In the filtered signal, noise levels are well below the voice level.

As was the case with the simulated validation data, listening to the recordings before and after filtering confirms that the filtering process greatly reduces the background noise levels with minimal distortion of the speech signal.

5. CONCLUSION / FURTHER RESEARCH

The results presented above demonstrate that a hybrid filtering approach can effectively extract a speech signal from high background noise levels. Existing techniques in sound separation are not able to produce real-time results with enough clarity to be useful to a listener. The multi-stage filtering method presented above is still being developed, and notably has not yet been implemented in a true real time setting. Even so, the results obtained so far demonstrate the viability of the dual-stage approach and its ability to filter high-amplitude background noise with a significant incoherent component.

6. REFERENCES

1. Lakshmikanth.S, Natraj.K.R and Rekha.K.R, "Noise Cancellation in Speech Signal Processing-A Review," *Journal of Advanced Research in Computer and Communication Engineering,* **3(1)**, 5175-5186 (2014).
2. C. M. Harris, *Handbook of Acoustical Measurements and Noise Control*, 1979.
3. R. Bentler and L.-K. Chiou, "Digital Noise Reduction: An Overview," *Trends in Amplification,* **10(2)**, 67-82 (2006).
4. R. Bentler, "Effectiveness of Directional Microphones and Noise Reduction Schemes in Hearing Aids: A Systematic Review of the Evidence," *Journal of the American Academy of Audiology,* **16(07)**, 473-484 (2005).
5. M. Karam, H. F. Khazaal, H. Aglan and C. Cole, "Noise Removal in Speech Processing Using Spectral Subtraction," *Journal of Signal and Information Processing,* **5(2)**, 32-41 (2014).
6. D. Wu, W.-P. Zhu and M. Swamy, "A Compressive Sensing Method for Noise Reduction of Speech and Audio Signals," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS),* pp. 1-4, 2011.
7. R. Aggarwal, J. K. Singh, V. K. Gupta, S. Rathore, M. Tiwari and A. Khare, "Noise Reduction of Speech Signal using Wavelet Transform with Modified Universal Threshold," *International Journal of Computer Applications (0975 – 8887),* **20(5)** (2011).
8. O. J. Tobias, J. C. M. Bermudez and N. J. Bershad, "Mean Weight Behavior of the Filtered-X LMS," *IEEE Transactions on Signal Processing,* **48(4)**, 1061-1075 (2000).
9. B. Widrow, "Adaptive Filters," *Aspects of Network and System Theory,* pp. 563-587, 1970.

10. Z. Jia, X. Zheng, Q. Zhou, Z. Hao and Y. Qiu, "A Hybrid Active Noise Control System for the Attenuation of Road Noise Inside a Vehicle Cabin," *Sensors,* **20(24)**, 7190 (2020).
11. P. Clarkson, *Optimal and Adaptive Signal Processing*, Boca Raton: Routledge, 1993.
12. S. Sarunta et. al., "Refrigerator Equipped with Active Noise Control System," *Proceedings of International Symposium on Active Control of Sound and Vibration,* pp. 9-11, 1991.
13. R. D. Gitlin, H. C. Meadors and S. B. Weinstein, "The Tap-Leakage Algorithm: An Algorithm for the Stable Operation of a Digitally Implemented, Fractionally Spaced Adaptive Equalizer," *The Bell System Technical Journal,* **61(8)**, 1817-1839 (1982).
14. T. J. Shan and T. Kailath, "Adaptive Algorithms with an Automatic Gain Control Feature," *IEEE Transactions on Circuits and Systems,* **35(1)**, 122-127 (1988).
15. M. Dentino, J. McCool and B. Widrow, "Adaptive Filtering in the Frequency Domain," *Proceedings of the IEEE,* **66(12)**, 1658-1659 (1978).
16. J. J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," *IEEE Signal processing magazine,* **9(1)**, 14-37 (1992).
17. W. B. Mikhael and P. D. Hill, "Acoustic Noise Cancellation in a Multiple Noise Source Environment," *IEEE International Symposium on Circuits and Systems,* **3**, 2399-2402 (1988).
18. J. E. Piper, "Beamforming Narrowband and Broadband Signals," in *Sonar Systems,* N. Kolev, Ed., Rijeka, IntechOpen, pp. 79-92, 2011.
19. P. Ioannides and C. Balanis, "Uniform Circular and Rectangular Arrays for Adaptive Beamforming Applications," *IEEE Antennas and Wireless Propagation Letters,* **4**, 351-354 (2005).
20. M. I. Kadir, S. Hoque and S. Islam, "Direction of Arrival Algorithms for Adaptive Beamforming in Next Generation Wireless Systems," in *Proceedings of 11th International Conference on Computer and Information Technology (lCCIT2008),* Khulna, Bangladesh, pp. 571-575, 2008.
21. I. A. McCowan, *Robust Speech Recognition using Microphone Arrays*, PhD Thesis, Queensland University of Technology, 2001.
22. S. Zhang, "The ‘Harvard Sentences’ Secretly Shaped The Development of Audio Tech," *Gizmodo Australia,* 10 March 2015.