A PORTABLE USB-BASED MIROPHONE ARRAY DEVICE FOR ROBUST SPEECH RECOGNITION

*Qi (Peter) Li, Manli Zhu, and Wei Li**

Li Creative Technologies, Inc. (LcT)
30 A Vreeland Road, Suite 130, Florham Park, NJ 07932, USA
E-mail: li,manlizhu@licreativetech.com; *iewil2000@yahoo.com

ABSTRACT

We present a USB-based, highly directional, and portable microphone array device that delivers a crisp, clear and noise-reduced speech signal. This device consists of four linearly distributed microphone sensors and a filter-and-sum beamformer designed using broadband beam-forming algorithm. The device has a narrow acoustic beam pattern and identical frequency responses for almost all speech bands. In addition to beamforming, an adaptive noise reduction algorithm is used to further reduce the background noise. By utilizing both the spatial and temporal information, the SNR of speech signals is improved and speech recognition performance in noisy environments is significantly improved as reported in our experiments.

***Index Terms* —** Microphone array, beamforming, noise reduction, robust speech recognition.

1. INTRODUCTION

A microphone array consists of multiple microphone sensors located at different positions. It can be used for both sound source location [1] and speech enhancement by processing the signals from each individual sensors [2][3]. While most of the current speech processing software can only use the temporal information, the designed device utilizes both the spatial and temporal information; thus, significantly improving speech recognition performance and robustness.

One of the major challenges in applying a microphone array in speech recognition is that speech is a wideband signal. The traditional narrowband beamforming techniques are not appropriate anymore [4]. The problem has been addressed by the spatial Fourier transform of a continuous aperture [5] and the joint optimization of the spatial and frequency response [6]. In these approaches, to keep the constant response over the wide frequency range, the array size is usually large; thus most of the prototypes or products using microphone arrays on the market are quite large and cannot be used as a portable device [7]. The large size of the array prevents the array products from broad applications, such as handheld devices, wireless handsets, and PDA. Another challenge of a microphone array compared to single close-talking microphones is the decreased performance in speech recognition because of the variation of their frequency responses [8]. These problems need to be addressed by improving the array design algorithms, in order to develop a portable and high performance device.

Figure 1. The 4-sensor microphone array named CrispMic^{TM} clipped on a laptop

In this paper, we present a portable microphone array device, named CrispMic™, where both the size and performance have been properly addressed. As shown in Figure 1, the size of the microphone array is only 9.6 x 2.6 x 1.3 cm which is less than the half of the size of a similar microphone array on the market. Also, it has a constant frequency response in a wide range of speech bands as shown in Figure 4. The experimental results in this paper show that the array can improve speech recognition performances significantly, especially in noisy environments.

2. SYSTEM DESCRIPTION

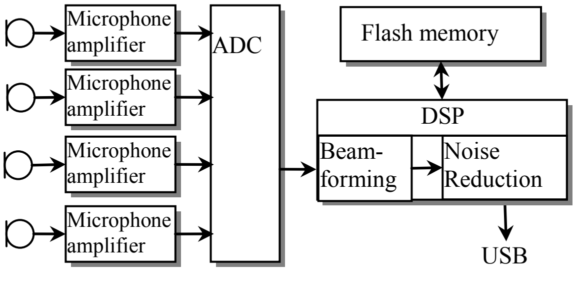


Figure 2. Illustration of the hardware structure

Figure 2 shows the structure of the device. It consists of four identical omni-directional microphones, an audio codec chip with an ADC and pre-amplifiers, a DSP (digital signal processor) chip, a flash memory and a USB interface. The acoustic signal is picked up by four microphone components arranged as a linear array with 20 mm intervals. The audio codec provides an adjustable gain, and converts the four channels of analog signals into digital signals for the DSP. The beamforming algorithm combines the 4 channels of speech into one channel and then a noise reduction algorithm is applied to further reduce the background noise. The processed clean speech signals are then transmitted to a laptop or PC through the USB interface.

The flash memory stores the software code for the DSP chip. Once the system boots up, the DSP chip reads the code from the flash memory into the internal memory and starts to execute the code. The device is powered through the USB. It is a plug-and-play device and can be used without installing any software. Since both the analogue and digital circuits are implemented in a small PCB, the circuit board was especially designed to avoid the noise interferences.

2. BROADBAND BEAMFORMING

Due to the special requirements in size and performance, we developed a robust, far-field broadband beamforming algorithm and implemented it in the DSP chip. The beamformer has a constant response in the speech frequency bands between 300-4000Hz. In theory, it can significantly improve spatial SNR of the speech signals without distortions in different frequency bands.

The linear microphone array configuration is shown in Figure 3. It is comprised of four equally spaced microphone sensors, where d_i is the distance between the i^{th} microphone and the center of the array. The output y of the array is the filter-and-sum of the four microphone outputs, $y = \sum_{n=1}^{4} w_n^T x_n$.

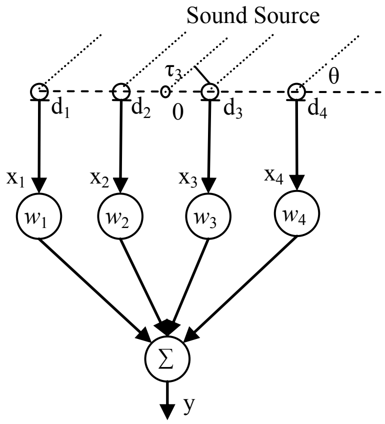


Figure 3. The configuration of linear microphone array.

The spatial directivity pattern $H(\omega,\theta)$ for the sound source form angle θ with normalized frequency ω is defined as [2]:

$$H(\omega,\theta)=\frac{Y(\omega,\theta)}{\overline{X}(\omega,\theta)}=\frac{\sum_{n=1}^{4}W_n(\omega)X_n(\omega,\theta)}{\overline{X}(\omega,\theta)} \tag{1}$$

where $\overline{X}$ is the signal received at the center of the array and W is the frequency response of the real-valued FIR filter w. If the sound source is far enough from the array, the difference between the signal received by the n^{th} microphone x_n and the center of the array is a pure delay. We use $\tau_n = f_s d_n \cos\theta / c$ to measure the delay by the number of samples, where f_s is the sampling frequency, c is sound speed, and $X_n(\omega,\tau)=\overline{X}(\omega,\theta)e^{-j\omega\tau_n}$ is the microphone signal. The spatial directivity pattern H can be re-written as:

$$H(\omega,\theta)=\sum_{n=1}^{4}W_n(\omega)e^{-j\omega\tau_n(\theta)}=\mathbf{w}^T\mathbf{g}(\omega,\theta) \tag{2}$$

where $\mathbf{w}^T = [\boldsymbol{w}_1^T,\boldsymbol{w}_2^T,\boldsymbol{w}_3^T,\boldsymbol{w}_4^T]$ and $\mathbf{g}(\omega,\theta)$ is the steering vector.

Let the desired spatial directivity pattern equal 1 in the pass band and 0 in the stop band. The cost function is then defined as:

$$J(w)=\int_{\Omega_p}\int_{\Theta_p}|H(\omega,\theta)-1|^2\,d\omega\,d\theta+\alpha\int_{\Omega_s}\int_{\Theta_s}|H(\omega,\theta)|^2\,d\omega\,d\theta \tag{3}$$

Let $\partial J/\partial w = 0$. We can then obtain the best parameter set w.

A Computer simulation was conducted to verify the performance of our designed beamformer with the following parameters: The distance between microphones is 0.02m, the sampling frequency f_s = 48k, and FIR filter taper length L=128. When the pass-band (Θ_p, Ω_p) = {300-4000Hz, 70°-110°}, the designed spatial directivity pattern is 1. When the stop-band (Θ_s, Ω_s) = {300~4000Hz, 0°~60°+120°~180°}, the designed spatial directivity pattern is 0.

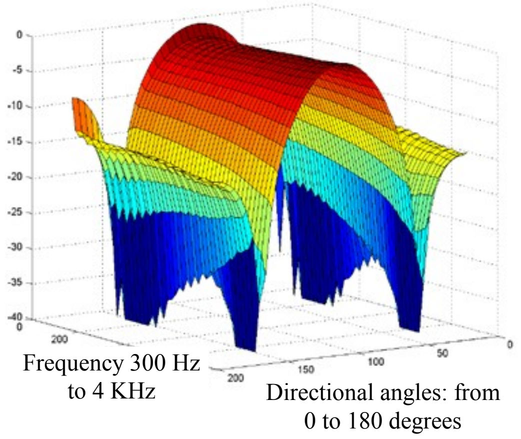


Figure 4. Directivity pattern of the designed microphone array: the frequency bands from 300 Hz to 4 kHz of the sound from the front of the microphone array are enhanced and the sound from other directions are reduced by about 15 dB.

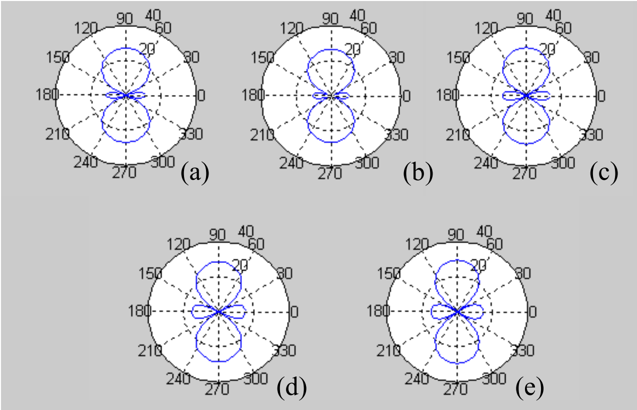


Figure 4. Computer simulation results for the designed 4-sensor microphone array: The directional response for frequencies of (a) 0.5, (b) 1.0, (c) 2.0, (d) 3.0, and (e) 4.0 KHz.

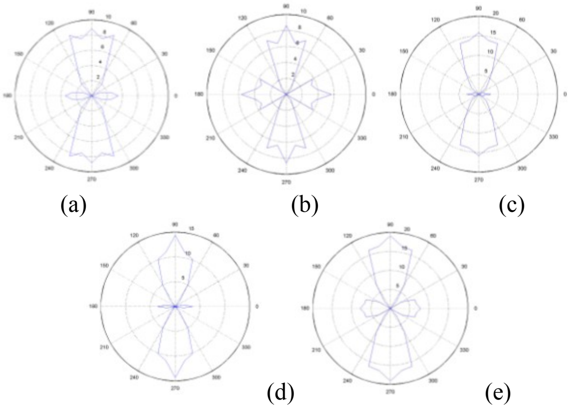


Figure 5. The directivity pattern of the microphone array based on our measurement in an anechoic chamber: The direction response for frequencies of (a) 0.5, (b) 1.0, (c) 2.0, (d) 3.0 and (e) 4.0 KHz. (The center is 0 dB.) The directional gains are significant and match the theoretical design and simulation.

Figure 4 shows the directivity pattern of designed microphone array. In all frequency bands, the main lobe has the same level, which means the speech signal has little distortion in frequency. The main lobe is about 15dB higher than the side lobe; therefore the background sound from other directions will be highly suppressed compared to the sound in the desired pass direction.

To verify the design simulation, we conducted an experiment to measure the spatial response of the microphone array in an anechoic chamber using a white noise stimulus. The microphone array was placed in a fix positioned and rotated in 20 degree increments. For each rotation, we applied the designed filter $\boldsymbol{w}_i$, i = 1,2,3 and 4, on the signal picked by each microphone and calculated the energy gain. The spatial directivity pattern is shown in Figure 5. The main lobe has a similar width among all the bands. The result shows that the directional gains are significant and are similar to our theoretical simulation.

3. NOISE REDUCTION

Our adaptive noise reduction algorithm consists of three key components: frequency analysis, adaptive Wiener filtering, and frequency synthesis. The frequency-analysis component is used for transforming the wideband noisy speech sequence into the frequency domain so that the subsequent analysis can be performed on a sub-band basis. This is achieved by the short-time discrete Fourier transform (DFT). The output from each frequency bin of the DFT represents one new complex valued time-series sample for the sub-band frequency range corresponding to that bin. The band width of each sub-band is given by the ratio of the sampling frequency to the transformed length.

The most crucial step of the temporal filtering technique is an adaptive Wiener filter, which estimates the clean-speech spectrum from the noisy-speech spectrum. The system explores the short-term and long-term statistics of noise and speech, as well as the segmental SNR, to support a Wiener gain filtering. The noisy-speech spectrum passes through the Wiener filter, which then generates an estimate of the clean-speech spectrum. In the last step, the frequency synthesis, as an inverse process of the frequency analysis, reconstructs the clean-speech signal given the estimated clean-speech spectrum.

5. ASR EXPERIMENTAL RESULTS

In order to verify the performance of the CrispMic™, we have conducted a sequence of experiments. The experimental configuration is shown in Figure 6. The recorded speech was played by an artificial mouth while the background noise was played by four loudspeakers.

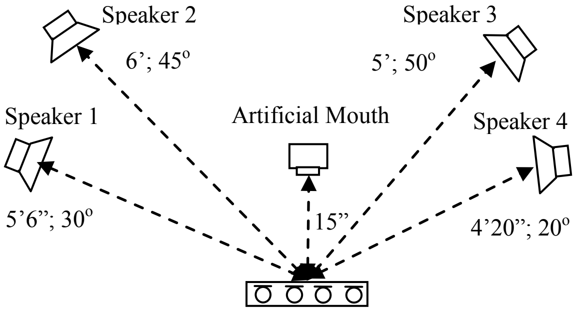
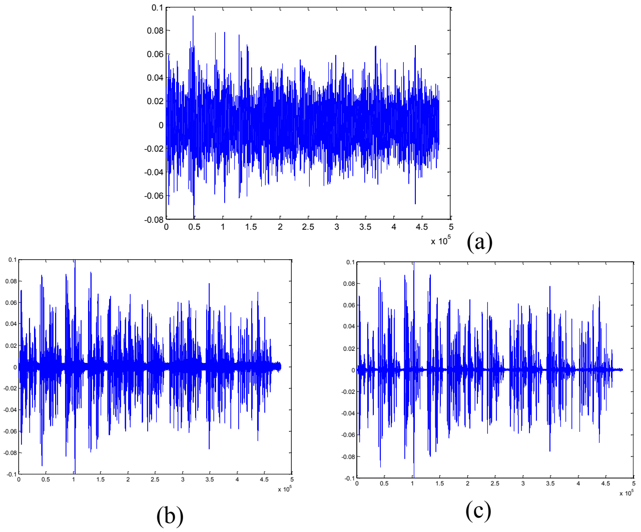


Figure 6. Experimental configuration

Experiment 1: Figure 7 shows the SNR improvement by using the CrispMic™. In this specific experiment, speech was played by an artificial mouth while white noise was played from Speaker 4 only as shown in Figure 6. Figure 7 (a) is the signal recorded by traditional omni-directional microphone, where the SNR is 3.3dB. Figure 7 (b) is the signal processed by beamforming. The SNR is increased about 31dB by the spatial approach. Figure 7 (c) is the signal after noise reduction, which is also the final output of the CrispMic™. SNR is further improved another 10dB by the temporal approach.



(a)

(b)

(c)

Figure 7. The improvements on SNR: (a) The original signal captured by a traditional mic, SNR = 3.3; (b) The signal after beamforming, SNR=36.9dB; and (d) The signal after noise reduction, SNR = 49.4.

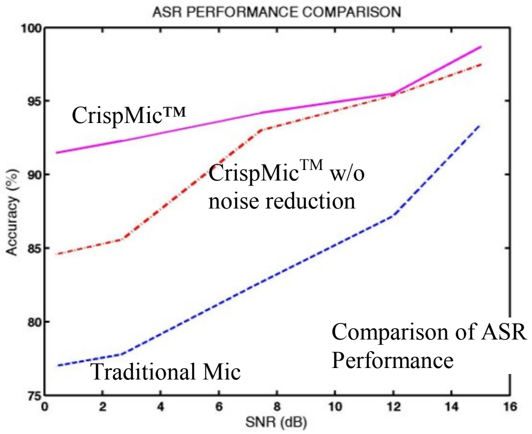


Figure 8. Comparison of ASR performance

Experiment 2: The CrispMic™ has also undergone a series of tests in the automatic speech recognition system and the performance is measured in terms of the recognition accuracy. In our experiments, background helicopter noise was played by four loudspeakers, Speakers 1-4 as in Figure 6. The CrispMic™ was connected to a laptop where Nuance's software was used for recognition. The experiments used 244 pre-recorded English phrases. During the test, if any word in the recognized phrase was different from the spoken phrase, we counted it as an error and the entire phrase was rejected. The experiment allowed for two attempts; if the first attempt failed, we allowed a second attempt using the same phrases to simulate the real applications. Our experimental results are plotted in Figure 8 and Table 1. The results showed that the contributions of our microphone array and noise reduction to speech recognition are very noteworthy.

Table 1. Comparison of Speech Recognition Performances on the CrispMic™: The number of loudspeakers is 4, the number of attempts is 2, and the number of test phrases is 244.

SNR (dB)	**0.40**	**2.66**	**7.44**	**11.99**	**15.01**
Traditional microphone (%)	77.0	77.9	82.7	87.2	93.4
LcT: 4-microphone array only (%)	84.6	85.6	93.0	95.4	97.5
CrispMic™: 4-mic array plus noise reduction (%)	91.5	92.3	94.2	95.5	98.7

4. CONCLUSIONS

In this paper, we presented a novel microphone array device. Compared to traditional microphones, the device utilizes both the spatial and temporal information; therefore it significantly improved SNR and speech recognition performance. Compared to the similar product or prototype on the market, this device is much smaller in size, which facilitates its application to portable devices, such as wireless phones, PDA, and Laptops.

5. ACKNOWLEDGEMENT

We would like to thank Jingdong Chen for useful discussions and Ashrith Deshpanda, Romnel Crasta, and Xiaoqian Xiao for conducting some of the experiments.

6. REFERENCES

[1] C.H. Knapp and G.C. Carter, The Generalized Correlation Method for Estimation of Time Delay, IEEE Trans. on Acoustic, Speech and Signal Processing, Vol.ASSP-24, No. 4, pp.320-327, Aug.1976.

[2] S. Doclo, M. Moonen, Design of Far-Field and Near-Field Broadband Beamformers Using Eigenfilters, *Signal Processing*, Vol. 83, pp.2641-2673.

[3] L.C. Parra, Steerable Frequency-invariant Beamforming for Aribitrary arrays, *J. Acoust. Soc. Am*, 199(6), pp.3839-3847, June 2006.

[4] M. Brandstein and D. Ward, *Microphone Arrays*. Springer, 2001.

[5] D.B. Ward, R.A. Kennedy and R.C. Williamson, Constant Directivity Beamforming, *Microphone Arrays*, pp3-17.

[6] S. Yan and Y. Ma, Design of FIR Beamformer with Frequency Invariant Pattern via Jointly Optimizing Spatial and Frequency Response, *ICASSP* 2005 pp.789-792.

[7] I. Tashev and H.S. Malvar, A new beamformer design algorithm for microphone arrays. *Proceedings of International Conference of Acoustic, Speech and Signal Processing ICASSP* 2005, Philadelphia, PA, USA, March 2005.

[8] M. Omologo, M. Matassoni and P. Svaizer, Speech Recognition with Microphone Array, *Microphone Arrays*, pp.331-34.