RUTALKING2ME? AN ASSISTIVE DEVICE COMBINING BEAMFORMING AND SPEECH RECOGNITION

Colin Miyata[1], Raymond Greiss[1], James R. Green[1], and Jim Ryan[1,2]
[1]Carleton University, Ottawa, Ontario; [2]On Semiconductor, Burlington, Ontario

INTRODUCTION

Individuals with hearing impairments can face considerable adversity managing the activities of everyday life. One such adversity is the ability to actively notice, as well as appropriately respond to, being addressed. While such a customary gesture may seem pedestrian, the inability to participate in such a social necessity draws attention to the impairment and may ostracize the individual. In response to such a need, the present work details the development of RUTalking2Me, an assistive device which recognizes user-specified speech cues, determines the relative direction of the speaker, and alerts the user to this location.

RUTalking2Me seeks to incorporate a portable microphone array with a combination of simple beam forming and speech recognition to determine the angle of arrival of the user's name. Through a tactile interface, the angle can be transmitted to the user so that they may respond appropriately.

METHODS

The system was operated by beam forming the data to distinguish between directions and speech recognition was used on the processed data to recognize and isolate the utterance. Angle of arrival was determined as the direction of greatest average power density. The steps used to develop the prototype are:

1. Determine the microphone array specifications (number of microphones, array geometry, and spatial orientation) to prevent spatial aliasing and to optimise for minimum error during beam forming.
2. Bandpass filter each isolated signal to isolate the human speech frequency range. Beamform the filtered data in eight different directions.
3. Using a speech recognition algorithm, isolate the utterance of interest in each recording. Determine the average power density for each beamsteered direction.
4. Determine the angle over which any beam is dominant.

Array Specifications

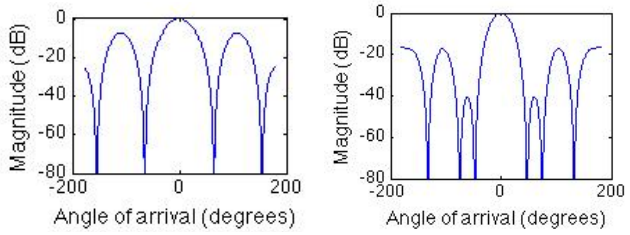
A square array of omni-directional microphones was chosen due to symmetry for beam forming and the low number of inputs for data processing. The microphone array was placed in free space to limit the effect of diffraction. A sampling rate of 44.1 kHz was used to increase the accuracy of beam forming. The frequency band of speech and the speed of sound were assumed to be 300-3000 Hz and 340.29 m/s respectively. To prevent spatial aliasing, the diagonal of the square was restricted to being less than the wavelength of sound at 3000 Hz (11.343 cm). The dimensions of the square were selected to produce sample delays as close to integers as possible when beam forming. Final diagonal length was selected as 7.7 cm (Figure 1).

Figure 1: Microphone array set-up

Bandpass Filtering and Beamforming

Prior to beamforming, the signals were passed through digital filters to remove sound outside the 300-3000 Hz frequency band of human speech. A fifth order, bandpass, Butterworth filter was chosen.

Delay and sum beamforming was used in the experiment [1]. Given the width of a typical human field of view, a user will be able to locate the speaker if they are directed to the 45° window that the speech originated from. Consequently, the signals were beam formed in 8 equally spaced directions. Due to the symmetry of the array only 2 sets of delays were required, creating two different beam forming patterns. Type I beams are oriented along the diagonals and have delays of 0, 5, 5, and 10 samples for the lead, two side and rear microphones, respectively. Type II beams are oriented perpendicular from the faces of the square and have delays of 0 samples for the lead pair of microphones and 7.07 samples for the rear pair (7.0 samples actually used). Theoretical beamforming patterns with ideal delays for angles from -180° to 180° and at 3000 Hz can be seen in figure 2.



(a) (b)

Figure 2: Theoretical beam forming patterns at 3000 Hz for (a) type I and (b) type II beams.

Speech cue isolation and power density estimation

To enhance the accuracy of the power density measurement, the speech cue is isolated from the bandpass filtered data. This isolation is achieved through the use of Sphinx-4, an open-source speech recognition system, which makes use of Hidden Markov Models [2]. A “digits” level of vocabulary, containing up to 11 words [3] was incorporated into the software. The isolation algorithm halves the 15 second beamformed data, streams the halves into Sphinx-4 to perform speech recognition and eliminates the half without speech. The process is repeated on the remaining half until the word is separated into the two halves and speech recognition fails for both. Prior to angle of arrival calculations, the algorithm eliminates sections of the beamformed data that do not contain a recognized speech pattern. The accuracy of system was tested using pre-recorded sound files.

Configuration of the Sphinx-4 system is handled through an XML file, which specifies connectivity and configuration of the system’s resources. The default specifications are the canonical paths of Sphinx-4 acoustic models and grammar files. User-generated grammar files specify the search cue(s) of the software, and facilitate seamless alterations of search terms. Documentation for configuring a grammar file can be found in the Sphinx-4 programmer’s guide [3].

Average power density values are calculated for isolated speech portion of the beamformed signals. This computation is based on an integration of the power spectral density. Power density values are then passed to a separate function to determine the angle of arrival.

Experiment 1: Determination of dominant angles for each beam

Experiments were concerned with the effect of the angle of arrival on the power density of isolated mono-syllable speech terms in each beamformed channel. Accordingly, the array was affixed to a freely rotatable platform, which allowed the microphones to swivel concurrently relative to a point sound source (Figure 1). A recording of the search term was played from a stationary point, and microphone data was recorded for rotations 0-345° in increments of 15°. As the magnitude of the power densities varied between trials, power densities were standardized by dividing by the peak value for the trial. The effect of angle of arrival on the amplitude of the power density was then characterized for each type of beam forming. This was done by averaging the standardized power density for the 4 beams of the given type at the angle from their oriented direction.

Experiment 2: Determination of Speech Recognition and Angle of Arrival Accuracy

To determine the correct classification rate of speech recognition and the accuracy of the angle of arrival determination, 4 speakers were placed 50 cm away from each corner of the array in the direction of the array diagonal. Monosyllabic words ("red", "blue" or "green") were played from random speakers one hundred times at 15 second intervals. The experiment was performed with ambient noise and the hundred samples were recorded into 3 files. The collected files were then parsed into 15 second intervals and the parsed data was processed with speech recognition and beamforming. The determined word and angle were compared with the known values to determine accuracy.

RESULTS

Experiment 1: Figure 3 shows the average power density values of the 3 most powerful beams over 0° to 90°. The graph illustrates that the beam with greatest power for any given angle corresponds to the beam whose ideal orientation most closely matches the angle. The region over which each beam has maximal power appears to have a width of ± 22.5° from its ideal orientation. Given this result, the angle of arrival can be narrowed to a 45° window; this provides sufficient accuracy to direct a user to the speaker of the utterance.

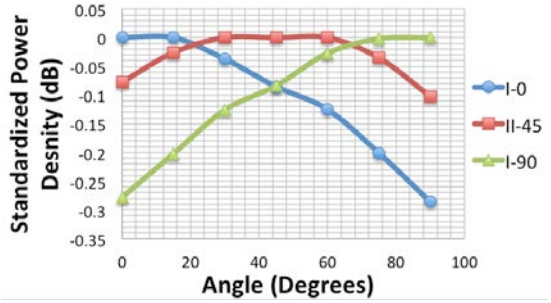


Figure 3: Dominant beam for given angular regions; legend uses form beam_type-oriented_angle

Figure 4 and 5 show the experimentally observed effect of angle of arrival on the standardized power densities of type I and II beams. Both beams displayed the expected symmetry due to the square shape of the array. The difference in standardized power density from the direction of orientation (i.e. 0 degrees) to the opposite direction was approximately 0.5 dB for both beam types. It should be noted, the attenuation changed based on the volume, geometry and position of the speaker.

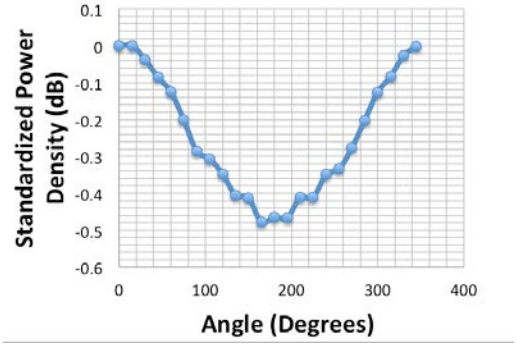


Figure 4: Experimental beam forming pattern for type I beam

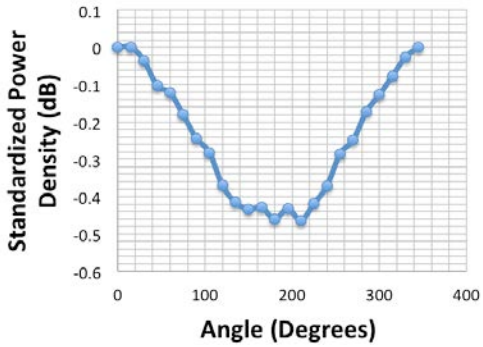


Figure 5: Experimental beam forming pattern for type II beam

Experiment 2: The average word correct classification rate for "red", "blue", and "green" was 68% when evaluated in the presence of white ambient noise. The signal to noise ratio was estimated to be 5.60±3.66 dB by comparing the signal power of speech and non-speech portions of 15 randomly chosen audio samples. The error rate rose when more complex noise was present (e.g. unrelated speech).

Implementation of the speech recognition isolation algorithm increased the ability to determine the angle of arrival of the sound by increasing the observed difference in signal power between beam directions. Even when the incorrect word was detected, simply identifying

the portion of the sample that contained speech led to excellent angle of arrival determination Angle of arrival, as determined by average power density, was correctly determined in 95 of the 100 test trials.

DISCUSSION

While the experimental beam forming patterns display the correct directionality, for both beam types the change in standardized power density is relatively small. Slight attenuations result in Sphynx-4 still being able to recognize the utterance from the attenuated region. For this reason, power density of the isolated speech portions was used to determine the angle of arrival. However, it's worth noting that the present system is unable to differentiate between multiple words spoken in a single trial. When multiple words were present within a single sample, different processed beams isolated different terms, consequently reporting inaccurate power density values.

Future Work

If another sound is played at the same time within the speech frequency range, the calculated angle of arrival will correspond to the direction of the louder noise. This must be resolved with more effective beam forming and improved speech recognition.

Processing of the recorded microphone data in MATLAB is resource intensive. Optimizing the code, or porting the final algorithms to DSP or GPGPU technologies may be necessary to achieve real-time processing. Beam forming and speech recognition algorithm efficiency should be revisited for possible improvements. Alternatively, mobile development can be facilitated through the use of PocketSphinx, a speech recognition system for use in embedded devices [4].

Experimentation revealed considerable limitations with the present configuration of the speech recognition system. In practice, the device should be able to differentiate between distinct utterances. Improvements to the speech recognition configuration are necessary to enhance accuracy. Acoustic model adaptation techniques should be considered to reduce the frequency of misinterpretation.

CONCLUSIONS

A prototype of the RUTalking2Me device has been developed, which has been shown to accurately compute the angle of arrival of a spoken keyword using beam forming and speech recognition techniques. This device is hoped to provide the cue required to alert an individual with a hearing impairment when someone addresses them by name. The present prototype is limited to relatively silent environments with nominally demanding speech interpretation requirements. Further development of this prototype should focus on miniaturizing the microphone array equipment, improving speech recognition accuracy, optimizing processing algorithms to permit real-time analysis, and implementing a physical actuating element to alert the user to a speaker's location. Ultimately, one can envision implementing this system on a binaural hearing aid system with multiple microphones.

AKNOWLEDGEMENTS

The authors would like to extend thanks to Dr. Rafik Goubran for providing the microphone array equipment and Mr. Payam Moradshahi for technical guidance and assistance.

REFERENCES

[1] A. Greensted. "Microphone Array Beamforming." Internet: http://www.labbookpages.co.uk/audio/beamforming.html, Nov 29, 2010 [Feb. 14, 2013].

[2] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea et al. "Sphinx-4: A Flexible Open Source Framework for Speech Recognition," Sun Microsystems, Inc., Mountain View, CA, USA, 2004.

[3] Carnegie Mellon University. "Sphinx-4 Application Programmer's Guide." Internet: http://cmusphinx.sourceforge.net/sphinx4/doc/ProgrammersGuide.html, [Feb. 16, 2013].

[4] D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, A.I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* , vol.1, no., pp.I, 14-19, May 2006.