Abstract

In order to understand utterance based human-robot interaction, and to develop such a system, this paper initially analyzes how loud humans speak in a noisy environment. Experiments were conducted to measure how loud humans speak with 1) different noise levels, 2) different number of sound sources, 3) different sound sources, and 4) different distances to a robot. Synchronized sound sources add noise to the auditory scene, and resultant utterances are recorded and compared to a previously recorded noiseless utterance. From experiments, we understand that humans generate basically the same level of sound pressure level at his/her location irrespective of distance and background noise. More precisely, there is a band according to a distance, and also according to sound sources that is including language pronounce.

According to this understanding, we developed an online spoken command recognition system for a mobile robot. System consists of two key components: 1) Low side-lobe microphone array that works as omnidirectional telescopic microphone, and 2) DSBF combined with FBS method for sound source localization and segmentation. Caller location and segmented sound stream are calculated, and then the segmented sound stream is sent to voice recognition system. The system works with at most five sound sources at the same time with about at most 18[dB] sound pressure differences. Experimental results with a mobile robot are also shown.

1. Introduction

Voice based human-robot interaction can be the most intuitive way to communicate with a robot. In order to achieve utterance based communication, there are technical problems to solve such as sound localization, sound segmentation, sound recognition and so on. Many research results related to these topics are proposed in the field of sound signal processing, natural language processing, microphone array design and so on. In general, systems developed by those results are designed with headset microphone or with a microphone close to the user’s mouth for human-robot interaction with utterance.

However, a natural and more useful way of using utterance for communication is in the case where the human is at some distance to a robot and he/she is not equipped with a microphone in a natural noisy environment. Difficulty to achieve such an utterance based communication for a mobile robot is that there is distance in between human and robot, so the arriving sound is changed and is mixed according to attenuation, environmental noise, echo & resonance effect, obstacles, interferences, and so on.

Even how loud human being speaks to a robot in noisy environment is not known. Therefore, in this paper, we developed a system to analyse the loudness of human utterances in different conditions. The system uses frequency domain power subtraction of human utterance and without human utterance records (which is previously recorded).

Then according to the results, the authors developed an integrated online sound source localization, segmentation and recognition system for a mobile robot. A 32ch low side-lobe high-gain microphone array that is optimized for DSBF(Delayed-Sum Beam Forming) sound source separation system is developed. The system works as an omni-directional telescopic microphone. Finally, experiments which demonstrate the robot reacting to human utterances are reported.

1.1. Related Works

Human being in noisy environment attempts to communicate more effectively by changing its pitch and so on in various way. This phenomena is called the Lombard effect. There is many researches in voice recognition software (ex. [7]). In this paper, we consideres the first human voice (order) to the robot, and how it can be effectively given.

There are state-of-the-art human-robot interaction systems using a single microphone arranged close ($\leq 30$cm) to the human mouth to achieve human-robot interaction in relatively quiet and single user environment(ex. [2]).

There are difficulties in extending the distance between humans and robot when interacting, and these are compounded when there are several users, or the interaction...
occurs in a noisy environment.

In order to overcome those conditions, a microphone array approach is typically used. Pioneer work to show beam forming performance by a 1024ch microphone array to locate a human voice within a room is shown[3]. Recently, tracking human voice using a distributed microphone array in a room by particle filtering techniques to improve tracking performance is reported[12, 11].

As for robot audition, robots with an embedded microphone array to interact human being are reported[13, 4]. Several methods are proposed to tackle issues in robot audition field. For example, GMM-based speech end-point detection method and outlier-robust generalized sidelobe canceler are proposed using a 12ch microphone array mounted on the torso[5]. Also, the missing feature algorithm is proposed in order to achieve a robust localizing and segmentation method for robot audition[15, 10].

As for mobile robot audition area, pioneer work was done by Jijo-2 robot[9]. Recently, particle filter based sound source localization technique for mobile robot audition is proposed[14]. Another aspect of mobile robot audition, sound field measurement and re-use for mobile robots is proposed [8].

Our aim in this research is to know how loud humans speak commands to a mobile robot in different condition in order to know what is the base requirement for utterance based communication. Then according to the result, we designed and developed a low-sidelobe microphone array, to embedded in our mobile robot. The final system enables users to command a mobile robot to go to different places using an utterance based interface.

2. Loudness of Human Utterances

In order to understand the required loudness of human utterance for giving an order to a robot, we conducted experiments with humans giving commands at various distances and volume of speech. Fig.1 (above) shows the experimental configuration. The robot is in the center of the room and five humans, arranged around the robot, command the robot by saying “¡Robot name¿ come here” from a distance of 1, 2.5, 5, or 10 [m]. Entire experimental setup is shown in Fig.1, although during the experiment only one human at a time is present in the experimental area.

In addition to the human sound sources, five speakers are connected to a commercial 5.1ch PCM 44.1[khz] surround system and placed 3[m] apart from the robot. Synchronized sounds are played through the speakers, starting with a 100[ms] at 2500[hz] sinusoidal wave at 60% of maximum power. In the beginning, the microphone array system records the sound scene with only the speakers playing. In the human utterance experiment, the system also records the human utterances along with the speakers playing. In this case, the white noise is detected by the system and the speaking part is segmented by subtracting the FFT results of both records. FFT is from 300[hz] to 5[khz] by 16[hz] window size in 1024 sampling (64[ms]).

Sampling rate of the system is 16[khz].

Five sound streams are prepared:

a) radio news voice
b) BGM music
c) radio news voice, vacuum cleaner sound
d) BGM music, vacuum cleaner sound
e) radio news voice, BGM music, radio weather forecast voice, vacuum cleaner sound, sound taken at a park where children is playing.

Here, stream a)&b) contain one sound, c)&d) contain two sounds, e) contains five sounds. Combined sounds are normalized by the maximum peak. We also prepared low and medium volume level versions from the above (a) - e) sound streams, by reducing each streams maximum power to -20[dB] and -10[dB] respectively.

2.1. Human utterance without noise

Fig.2 shows the averaged results of recording human utterances at different distances, without any explicit noise. Background noise power is about 40 [dB(A)] (a weighted sound pressure level defined at ISO R226). In the experiment the human subject (N=8) calls a robot at a distance of 1, 2.5, 5, 10[m] and the vertical axis of the figure shows the power [dB(A)] that is measured by the robot microphones.

From this experiment, we can understand that the human subjects basically generate louder sounds as the distance to the robot increases. However from the robot position, the distance effect is not well compensated (records power of sound source still decreases).
2.2. Human utterance with one sound source

Fig. 3 shows the averaged results of human subject (N=4) utterance with a sound stream a) from speaker No.3. There are four distances from robot to human subjects (1, 2.5, 5, 10[m]), and three degrees of loudness of the sound stream a). At each experiment, the human subject listens to the sound stream for 15[s], then calls the robot. The vertical axis of Fig.3(above) shows the sound power [dB(A)] that is measured by the robot microphones. The horizontal axis of Fig.3(above) shows noise power [dB(A)] at the human subject position. Fig.3(below) shows the relationship of the noise [dB(A)] at the human subject position, and the distance from the human subject to the robot, to the voice power[dB(A)] detected at robot microphone. In this figure, the weighted average of sampling points are used to generate 3D curved meshes.

From this experiment, not much difference in the power of human utterances is seen from the given noise levels.

2.3. Human utterance with two sound sources

Fig.5 shows the averaged results of human subject (N=4) utterance with sound stream c) from speakers No.3 and No.4 respectively. Other conditions and figures are the same as in the one sound source experiments.

In this experiment, the noise level is increased from the previous one sound source experiment, but not much difference is seen to the voice power level.

Fig.6 shows the averaged results of human subject (N=4) utterance with sound stream d) from speakers No.3 and No.4 respectively. Other conditions and figures are the same as in the one sound source experiments.

In this experiment, not much difference is seen from the given noise level.
2.4. Human utterance with five sound sources

Fig.7 shows the averaged results of human subject (N=8) utterance with sound stream e) from speaker No.1 to No.5. Correspondence are as follows: Speaker No.1 - radio weather information voice, No.2 - kids at park, No.3 - BGM music, No.4 - radio news voice, No.5 - vacuum cleaner. Other conditions and figures are the same as in the one sound source experiments.

In this experiment, the entire input noise level increased and also the output voice power as well.

2.5. Voice power range vs. distance

Fig.8 shows the relationship between different sound streams, for different loudness levels (10[dB] difference to each), and voice power level [dB(A)]. Fig.8 (above) shows the results for a distance of 1[m] and Fig.8 (below) shows the results for a distance of 10[m].

For each sound stream, the difference in between “small” and “loud” was 20[dB], but the difference in output voice power was at most about 4[dB]. Therefore, the utterances of the human subjects can be considered to have a small dynamic range in voice power performance. At 1[m], the human utterances result in an 83 to 93 [dB(A)] signal at the robot location, when competing with noise levels of 43 to 72[dB(A)], and at 10[m], the generated signal is from 74 to 81 [dB(A)] when competing with noise levels of 36[dB(A)] to 63[dB(A)].

Another thing that is noticeable is, as the distance from human to robot increases, the receiving sound power level at the robot decreases. Humans do not appear to be good at compensating for the distance effect.

The difference between Fig.3, 5 and Fig.4, 6 is that the noise of the former contains clearly pronounced language and the latter does not. Especially at closer range, these noise streams only produce about a 3[dB] difference in how loud the human subject speaks. Therefore, we can say that, clearly pronounced language in the noise streams causes an increase of voice power level about 3[dB].

These are some background noise characteristics that a human-robot command interface relying on human utterances should consider.
3. Mobile Robot Equipped with Microphone Array for Human-Robot Interaction

According to the previous human utterance experiments, a voice commanded mobile robot system requires about a 15-20[dB] gain for focusing direction. Therefore, we designed and implemented such a microphone array.

3.1. Design and Implementation of 32ch Low-sidelobe Microphone Array

In order to detect sound source direction for audio input with an unknown frequency, we developed a microphone array and firewire microphone AD board.

The diameter of the microphone array is limited to 33cm due to the robot size. Through simulation of sound pressure distribution, we empirically decided the microphone arrangement to minimize sidelobes. Fig.9(left) shows the resulting microphone arrangement which consists of the octagonal arrangement of eight 4ch microphone boards that have an isosceles trapezoid shape.

Fig.9(right) shows the microphone direction pattern of the Delayed Sum Beam Forming method. At each frequency, the focus direction gain compared to sidelobe is 12[dB] at minimum and 16[dB] in average (from 700-2500[hz]). Fig.10 shows the beam forming simulation results at 1000, 1400, 2000 [hz]. Fig.11 is our 32ch microphone array and microphone AD board. The AD board has 16 bits resolution at maximum of 16[khz] simultaneous inputs.

3.2. Delay and Sum Beam Forming Method

Aligning the phase of each signal amplifies the desired sounds and attenuates ambient noise. If we let $L_i$ be the distance of the focus to the i-th ($i = 1, 2, \cdots , M$) microphone and let $L_{min}$ be the minimum of $L_i$, Equation(1) shows the delay $D_i$ for the i-th microphone and the synthetic sound wave $s(t)$:

$$D_i = \frac{L_i - L_{min}}{V_s}$$

$$s(t) = \sum_{i=1}^{M} x_i(t + D_i) \quad (1)$$

where $t$ is time and $x_i$ is the sound signal of the i-th microphone. In this paper, the focus is set at 2[m] distant from the array center and we obtain 180 data by 2 degrees resolution.
3.3. Frequency Band Selection Method

The DSBF is a simple and strong algorithm that enhances sound from a given direction, however the method does not remove other signals perfectly (just reduces). Thus, we apply the FBS method after DSBF for the detection of multiple sound sources. FBS is a kind of binary mask and segregates objective sound sources from mixed sound by selecting the frequency components judged to be from a common objective sound source.

The process is as follows. Let $X_a(\omega_j)$ and $X_b(\omega_j)$ be the frequency components of DSBF-enhanced signals for objective and noise sources, respectively. The selected frequency component $X_{as}(\omega_j)$ is expressed as Equation (2):

$$X_{as}(\omega_j) = \begin{cases} X_a(\omega_j) & \text{if } X_a(\omega_j) \geq X_b(\omega_j) \\ 0 & \text{else} \end{cases}$$

(2)

This process rejects the attenuated noise signal from the DSBF-enhanced signal. The segregated waveform is obtained by the inverse Fourier transform of $X_{as}(\omega)$.

When the frequency components of each signal are independent, FBS can perfectly separate the desired sound source. This assumption is usually effective for human’s voice or every day sound within a short time period.

3.4. FBS-based sound directional localization

We use the FBS method for direction-of-arrival estimation, as shown in Figure 12.

![Fig. 12 FBS Sound Localization Process](image)

The first step filters out the average signal of each microphone (no delayed signal) input by FBS and finds the loudest sound from the spatial spectrum. When the frequency component of the average signal is higher than any DSBF-enhanced signal from each direction, the system filters out the spectrum of that frequency. This process rejects omni-directional noise sounds.

The second step filters out the 1st sound signal by FBS, and finds the second strongest sound from the spectrum.

When the frequency component of the DSBF-enhanced signal of the 1st sound’s direction is higher than that of any other direction, the system filters out the spectrum at each frequency.

If there are more than two sounds, the system finds the third strongest sound, and so on, after filtering out the second strongest sound signal. The method localizes multiple sounds from the highest power intensity to the lowest at each time step. Then the system can continuously localize multiple sound sources and separate each sound source during movement.

3.5. Mobile Robot “Pen2”

A mobile robot platform is developed. The size is 40cm in length, 45cm in width and 42cm in height (without wireless LAN antenna)(Fig.13). The robot is designed to rotate on the spot in a 50cm diameter circle. There are two passive caster wheels attached to the rear of the body, and the robot can move over terrain with about 2cm height gaps and 7 degrees of slope. Maximum speed on a flat surface is about 2m/s. It uses 60W Maxon DC motor at each wheel and has spring and dumper suspension to support all wheels in order to augment moving stability. The caster wheels use belts in a semi-circle shape in order to augment effective wheel size. It also has low center of gravity (16cm) in order to augment stability in high speed movement.

The robot has a Pentium-M 2.0GHz (FSB400MHz, 1GB memory) processor that is operated by RT-Linux for motor control (1[ms] cycle PD control). The system is connected to the network by using 802.11a wireless communication. It has two Li-ion batteries (24V 7.5Ah for each) and can perform about four hours operation continuously. Stereo camera (Videre STH-DCSG-STOC), laser sensor (Sick LMS200), and 32ch microphone array are equipped on the robot.

Laser input is used for localization by adopting particle filter based method. The stereo camera is observing the floor about three meters ahead, and can detect about 2cm height obstacles. An optimized A* algorithm[6] is used to plan a trajectory of the robot. The entire cycle of these processes takes about 150ms for 15x12m size areas with 5cm grid cell.

![Fig. 13 Mobile Robot “Pen2”](image)
3.6. Sound Source Localization and Segmentation in Motion

We conducted an experiment by placing three loud speakers in a room and making the robot move in between them. Sound streams are male speech voice, female speech voice, and classical music.

Fig.14 shows the localization results in robot local frame. At each localization (shown as + mark), sound streams are extracted.

Data sampling points for one localization process is 2048 measurements (sampling rate is 16[khz], and localization process cycle is about 12[hz]. Even though there are false positives, three sound sources are correctly found and tracked.

Fig.15 shows the segmentation results. Fig.15(above) shows an input wave for a single microphone, Fig.15(middle) shows the original signal for the male speech sound, and Fig.15(below) shows the segmented male speech waves.

After segmenting a sound stream, the result is sent to a speech recognition system.

3.7. Combining Laser Range Finder to Reach to a Caller

The microphone array only detects speaker location in the yaw angle, and it cannot detect the distance from the speaker to the robot. Also, when a human stops speaking, the microphone array cannot detect the user’s location. Therefore, we adopted a laser range scanner to locate human existence. Fig.16 shows the result. The robot position is shown by a red circle, the laser range sensor output is shown with the green region. Given a microphone array input, the robot looks in that direction, and finds the minimum distance object that is not registered as an immobile object in the map (shown in yellow circle). This is assumed be the user location, and the system sets a position 50[cm] in front of the object as the goal location (shown in blue circle) and navigates toward it.

Fig.17 shows the experimental results. User’s call the robot from positions around the robot and the on-body microphone array system detects the sound locations. Then using laser range scanner, system locates users in 2D by turning to the detected sound direction (if needed).

In this experiment, reverberation time $T_{60}$ is about 500 [ms] and background noise is about 52[dB(A)].

3.8. Go to Indicated Place Task

A user placed a coffee cup on the robot and ordered to the robot to go to the kitchen, by using onbody microphone array (Fig.18 h). Julian[1] is used to recognize the
speech sound stream. The vocabulary for speech recognition is a few variations of the following three commands: 1) go to place, 2) come, 3) greetings. Each command includes about 20 words, as well as about 10 place names.

The experiment was conducted in our experimental house “Holone”. Reverberation time $T_{60}$ is about 750 [ms].

When robot detects that the given command is “go to place” the robot starts to navigate (Fig.18 a), and traverses the entrance (Fig.18 b), the living room (Fig.18 c, d) and reaches the destination point (kitchen) (Fig.18 e). Finally a different user obtained a cup and released robot by saying “thank you” to onboard microphone array (Fig.18 f).

Fig. 18 Snapshots on Go to Indicated Place Experiment and GUI display

4. Conclusion

This paper describes our experimental study on loudness measurement of human utterances to a robot in noisy environments. This paper initially analyzes how loud humans speak in a noisy environment. Experiments were conducted by measuring loudness of human commands with 1) different noise levels, 2) different number of sound sources, 3) different sound sources, and 4) different distances to a robot. Synchronized sound sources add noise to the auditory scene, and the resultant utterance is recorded and compared to a previously recorded without human speaking one.

From experiments, we understand that humans generate basically the same level of sound pressure level at his/her location irrespective of the environment conditions. More precisely, there is a band according to a distance. Also, as distance increases, received sound at the robot decreases. Humans do not seem to be good at compensating for this distance effect. Finally, clearly pronounced language sound causes increase of human command voice power level of about [3][db]. Those are the requirements that the system should detect among background noise.

According to this result, we developed an online spoken command recognition system for a mobile robot. The system consists of two key components: 1) a low side-lobe microphone array that works as an omni-directional telescopic microphone, and 2) DSBF combined with FBS method for sound source localization and segmentation. Caller location and segmented sound stream are calculated, and then segmented sound stream is sent to a voice recognition system.

The system works with at most five sound sources at the same time with at most approximately 8[db] sound pressure differences. Experimental results with the mobile robot are also shown.

References


