Stereo Voice Detection and Direction Estimation in Background Noise or Music for Robot Control

using Raspberry Pi and Python

Samyukta Ramnath

December 22, 2016

Outline

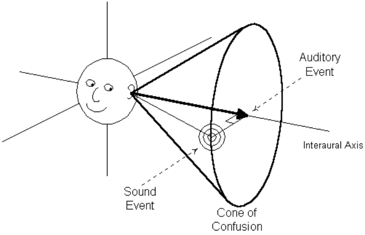
Problem to be Solved

- Stereo Voice Detection and Direction Estimation in Background Noise or Music for Robot Control
- Audio Source Localization in horizontal plane using two microphones
- Voice Detection
  a. Detect a sound event
  b. distinguish between tonal sounds and noise
- Source Distinguishing
  Distinguish between voice and music
- Source Separation
  Separate voice from music

- Find the location of a single acoustic source

- Number of dimensions - minimum number of microphones required
- Cone of confusion - ambiguity with 2 microphones

Figure: Cone of confusion



- 3 microphones - localization in a plane
- 4 microphones - localization in 3 dimensions, intersection of 3 cones

Previous Approaches

Acoustic Source Localization

- ITD: Inter-aural Time Difference
- ILD : Inter-aural Level Difference

Previous Approaches

Acoustic Source Localization

- Steered beamformer approach
- Microphone arrays of 4 microphones in the shape of a pyramid. [2]
- Direction estimation using a single microphone and an artificial 'pinna', using machine learning methods [1]
- 3 microphones kept in a triangle arrangement to find direction in a plane

The Approach

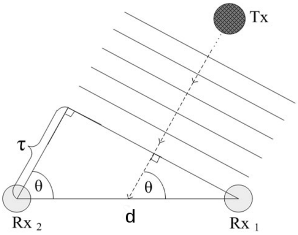
Acoustic Source Localization

- Two microphones can localize a sound in 180 degrees in a plane
- Time difference found using cross-correlation of the signals obtained at the left and right microphone
- Assumption of a planar wavefront and a large distance between source and sound compared to the distance between the microphones

The Approach

Acoustic Source Localization

Figure: Diagram to find angle of acoustic source
Image Source : Direction of Arrival Estimation and Localization Using Acoustic Sensor Arrays,Vitaliy Kunin, Marcos Turqueti, Jafar Saniie, Erdal Oruklu



The Approach

Acoustic Source Localization

- The path difference between the distance traveled by the sound waves at microphone i and j is τ c, where τ is the time difference calculated by cross correlation, and c is the speed of sound in air.
-
$$\tau c = d \cos \theta \tag{1}$$
so that
$$\cos \theta = \frac{\tau c}{d} \tag{2}$$
giving us the angle at which the sound source is.

The Approach

Acoustic Source Localization

Figure: Top view of Roomba with angle sign conventions indicated

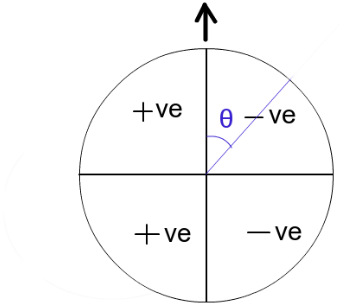
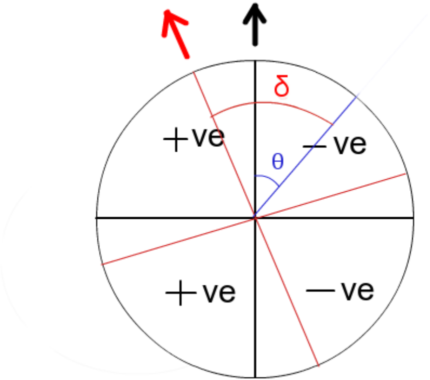


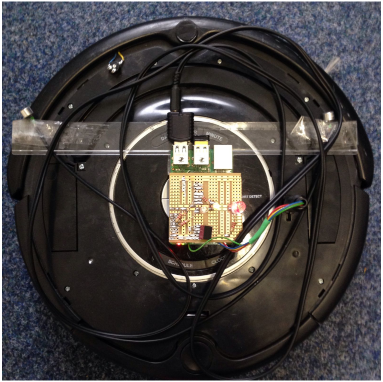
Figure: Top view of Rotated Roomba



The Approach

Acoustic Source Localization

Figure: Setup:Roomba Vacuum Cleaner Robot setup with two microphones and a Raspberry Pi with Serial Interface



The Approach

Acoustic Source Localization

- Roomba controlled using a Serial Interface with Raspberry Pi.
- 12V power supply output
- 5V serial Tx and Rx pins
- Voltage converter used for 12V - 5V conversion to power RPi
- Roomba can be programmed to move straight, or turn by a certain radius, or rotate in place

The Approach
Acoustic Source Localization

- Distance between microphones sets upper limit of frequency of sound that can be localized
- Upper limit 1kHz: $\lambda/2 \geq d$,
  where d : distance between microphones,
  λ : wavelength of the input signal.

Problem Statement

Voice Activity Detection

- Sound event : Signal power threshold
- Threshold relative to background level of noise
- Tonal Sounds have one or more spectral peaks
- Distinguish between tonal sounds such as voice or music, and noisy sounds such as a door knock or background noise, which are flatter in spectrum

Previous Approaches

Voice Activity Detection

- Energy Based Methods : Energy of each frame is computed, and an energy threshold decided, above which a frame is considered to be voice and below which it is considered to be silence
- Machine Learning Methods by feature extraction. More complex to implement and run, but better accuracy
- Apply a filter in the range of human speech to reduce the likelihood of other sounds interfering

Approach

Voice Activity Detection

- A Measure of Spectral Flatness : Spectral Flatness Coefficient [4]
- $$SFM = log_{10}\left[AM(x(m))/GM(x(m))\right] \tag{3}$$
- For a collection of numbers x, $AM(x) \geq GM(x)$
- Equality when all numbers in x are equal. Thus, if x is the FFT of the frame, $AM(FFT(x_m)) = GM(FFT(x_m))$ for white noise, which has the same energy at all frequencies
- Low SFM indicates tonal sounds, High SFM indicates noisy sounds
- Tonal sounds with fundamental frequency less than 300 Hz assumed to be voice, higher than 300 Hz assumed to be music

Problem Statement

Speech-Music Discrimination

- Basis for Speech-Music separation
- Final goal : differentiate between instrumental music and speech/singing

Previous Approaches

Speech-Music Discrimination

- Feature Extraction and classification
- Frequency cutoff, which is only valid when the music used does not significantly overlap voice spectrum

The Approach

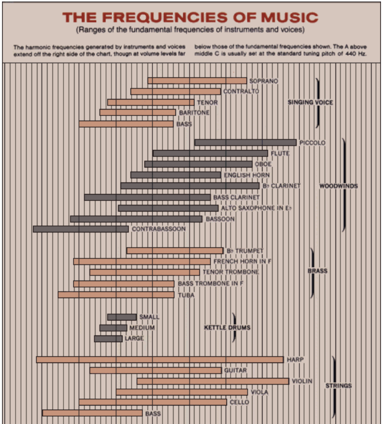
Speech-Music Discrimination

- Approaches mentioned like feature extraction and classification were too expensive for the Raspberry Pi to run real-time
- Frequency cutoff, assume that the music doesn't have low frequency components
- Piccolo (high frequency) would be suitable for this purpose

The Approach

Speech-Music Discrimination

Figure: Frequency Chart of various instruments, *Copyright 1980 by Hachette Filipacchi Magazines, Inc.*



Previous Approaches

Speech-Music Separation

- Non-Negative Matrix Factorization
- Neural Networks
- Frequency Filter

The Approach
Speech-Music Separation

- Music of high, non-overlapping frequency played
- Low-pass filter applied on right and left channels

Final System

- When only music plays : The Roomba hears the part of the music in the low-frequency range, moves towards it
- When music and voice play together : The voice becomes the dominant signal in low frequency, Roomba moves towards the voice
- Assumption : music used doesn't overlap much with voice

Summary

- The Audio Source Localization was done using two microphones in 360^o using a vacuum cleaner robot Roomba and a Raspberry Pi in Python.
- The Voice Activity Detection was done using Spectral Flatness Measure, and a frequency threshold to distinguish between voice and music.
- The Speech-Music Discrimination was attempted using Feature extraction and classification using a Support Vector Machine, but finally a frequency filter was used
- The Speech-Music separation was done using a filter applied to the right and left channels of the system

- Outlook
  - Improve precision in acoustic source localization by adding more microphones in the linear array, using sound suppression algorithms, performing in a room with absorbent material
  - Computations occur on a stronger processor, results sent to Pi

References I

- Saxena, Ashutosh, and Andrew Y. Ng. "Learning sound location from a single microphone." *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on.* IEEE, 2009.
- Valin, J-M., et al. "Robust sound source localization using a microphone array on a mobile robot." Intelligent Robots and Systems, 2003.(IROS 2003). *Proceedings. 2003 IEEE/RSJ International Conference on.* Vol. 2. IEEE, 2003.
- DL5BBN. Gerald Schuller. "Raspberry Pi Serial Interface." *Raspberry Pi Serial Interface.* N.p., n.d. Web. 01 Dec. 2016.

References II

Ma, Yanna and Akinori Nishihara. "Efficient Voice Activity Detection Algorithm Using Long-Term Spectral Flatness Measure". *EURASIP Journal on Audio, Speech, and Music Processing* 2013.1 (2013): 21. Web.