Machine Learning in Sound Source Localization for the Hearing Impaired

LITERATURE REVIEW

ALEXANDER SUN

Machine Learning in Sound Source Localization for the Hearing Impaired

Literature Review

Introduction

Purpose

People with hearing impairments are unable to sense the environment outside of their field of vision which often leaves them with the risk of missing crucial noises that could pose a danger to their lives in public and creates heavy communication barriers between other people and themselves. Over five percent of the world's population faces some aspect of impaired hearing, and the cost of hearing aids and cochlear implants are widely unaffordable. Currently, around 466 million people suffer from disabling hearing ("Deafness and Hearing Loss," 2019). The main impact of disabling hearing loss is the communication barrier. Often, they are left isolated from others because they are socially ignored, and because the majority of the population fails to learn sign language, which cuts off much of their communication. Feelings of isolation and loneliness surface as even when they are surrounded by others, they are ignored and treated as if they are invisible ("Deafness and Hearing Loss," 2019). In families with hearing-impaired children, less than a third of the parents sign regularly, which creates a permanent social divide between even those closest to one (Correll, 2019). This problem requires a solution that allows for wide accessibility at a low cost, even if its effectiveness cannot match that of modern expensive solutions.

Method of Attack

Sound Source Localization (SSL) describes the use of acoustic technology to determine the location of sound in robots. In humans, this capability is crucial as it allows for increased comprehension of speech, by allowing for separation between different sound sources. Furthermore, in the animal kingdom, this awareness allows for the rapid location of prey, making it a crucial capability for survival. Currently, researchers aim to recreate these systems for a variety of uses including sound source tracking, speech enhancement, virtual reality, and human-robot interaction (Risoud et al., 2018). Researchers are studying how cameras can use this technology to track different speakers during a meeting and recreate it virtually for a more immersive experience. They aim to create faster, more efficient, and smaller designs in order to reach their goal. SSL technology is evolving to meet modern needs and hopes of increased digital connection. Currently, societal standards and needs require improvements in processing speed, size, and computational cost, all while maintaining a similar level of accuracy and performance. The Internet of Things (IoT) is expanding quickly, and many want to have SSL capabilities. SSL improvements could better the safety of autonomous robots through safeguards (Zhao et al., 2012). Overall, SSL is quickly expanding fields intending to bridge the gap between human and robotic capabilities.

Challenges

SSL may seem simple at first, but many challenges stand in its ways. Modern microphones are overly sensitive to the reverberation of sound, as well as noises in frequencies that human ears cannot comprehend. Reverberation describes how sound reflects off objects which can create other sources of sound from the point of view of the microphone array (Zhao et al., 2012). This makes many SSL devices environmentally dependent and less robust. Furthermore, noise is difficult to filter in outdoor environments and often time computers cannot locate the important parts of the sound. Also, multiple sources of sound can easily confuse these systems. These issues need to be overcome for systems to improve and eventually meet the high standards of the industry (Takeda & Komatani, 2016). SSL research is expanding quickly to meet the requirements of the future, but many requirements have to be transferred from research and testing to robust devices capable of doing their jobs efficiently and accurately.

Current Approaches

SSL technology either aims to mimic the human capacity for human spatial awareness using similar methods, while others aim to gain more accuracy through the use of multiple arranged microphones. Microphone arrays represent the hardware used to gain the crucial information that needs to be processed. They are a device consisting of multiple precisely located microphones. In the 1900s, microphone arrays would be placed within a large room, and locate sources within its bound (Mandlik, Nemec & Dolecek, 2012). Now, they have evolved significantly to allow sizes that can fit in one's palm and can locate outside its bounds of up to 20 meters away (Zhao et al., 2012). Currently, most arrays consist of three or more microphones in a symmetrical pattern that all record simultaneous data. This data is then processed and used to locate the sound source. Some arrays aim to mimic human binaural localization by only utilizing two microphones, but only display accuracy similar to that of human awareness. More commonly, three or more microphones are used and now researchers are aiming to pinpoint locate the position of sound in any orientation (Mandlik, Nemec & Dolecek, 2012). The data recorded from the microphones also needs to be processed. As requirements become more strict in the areas of computation time, and accuracy, different algorithmic and non-algorithmic, methods are constantly being innovated upon. The most common methods involve microphone arrays consisting of multiple spaced microphones and applying a time difference of arrival algorithm (TDOA). Common examples of these algorithms include Power Phase Transforms (PHAT) or Multiple Signal Classification (MUSIC) paired with a generalized cross-correlation (GCC) algorithm to determine the correct position that the source of sound comes from. These algorithms vary and allow the processing of data from different microphone setups

and output requirements (Mandlik, Nemec & Dolecek, 2012). Neural Networks (NN) are the newest addition to the methods used in SSL. Now, researchers are training models to output the correct location from the input through machine learning. This approach provides promise as the models serve as much less computationally intensive compared to the algorithms with negligible accuracy difference and provide a new route into real-time SSL and wearable technology (Takeda & Komatani, 2016). Despite this, machine learning methods require tons of data that may be unacquirable in the real world, which often leads to the use of simulations to obtain training data, which fails to meet the unpredictability of a real-world environment, even an indoors or contained one (Takeda & Komatani, 2016). Furthermore, success has not been proven for high accuracy tasks involving many microphones, due to the difficulties training the data innately brings, especially something of high complexity. Many approaches have been taken to both the software and hardware sides of the problem, and success has been achieved in both, but newer research and innovations still need to be pursued to make SSL a reality.

Needs of the Hearing Impaired

American Sign Language

American Sign Language (ASL) is the main form of communication that the deaf can use. ASL has its strengths, but its viability to help all deaf people is lacking. There are an estimated between 100,000 and one million ASL interpreters in the US, which is far from the amount necessary for every deaf person to have someone available (Correll, 2019). Furthermore, with the presence of an interpreter all the time, a deaf person loses a sense of independence and freedom as all communication has to be relayed through another. In order to communicate with others at all, the interpreter would have to go everywhere with the hearing impaired (Correll, 2019). More often than not, interpreters are not available at all, especially in third world countries. Those without access to this help face not just social barriers, but economic barriers in the future when one needs employment ("How is Deafness Affecting your mental Health," 2016). Lip reading is another method that allows a deaf person to at least, understand what another is saying. Unfortunately, not all deaf people can master this, and even those who cannot always consistently get accurate meaning (Correll, 2019). Therefore, current communication methods are lacking in function and accessibility to all of the impaired.

Available Technology

Cochlear implants and hearing aids improve everywhere in terms of size, invasiveness, and effectiveness. On the other hand, their cost increases because of these improvements or stays the same at least. For most, these devices pose economic challenges, as hearing aids can range between $1000 and $4000, and cochlear implants costing around $400,000.

The adoption rate of hearing aids is approximately only one third in the US, with the main cause being cited as price (Valente and Amlani, 2017). The production of hearing aids also fails to meet the demand as global production of hearing aids meets less than 10% of global need. Often availability lacks because of an inability to set up fitting appointments, maintenance, and lack of batteries in third world countries or for those in poverty ("Deafness and Hearing Loss", 2019). These devices are not accessible to all either. Cochlear implants are far more expensive and require surgical installation in order to function. The cost and invasiveness of the procedure dissuades most deaf people from even considering the option. Furthermore, for older individuals that get the operation, they can still face trouble interpreting the sound and new capabilities that they gain. Corrective devices can also create issues and problems. For example, often they over amplify background noise which can blur the words of others ("Impact of hearing loss on daily life and workplace," 1970). The transition is very difficult for many, causing them to remain as a good solution for the very young only ("Pros and Cons of Cochlear Implants," n.d).

Microphone Arrays

Design Requirements

Microphone (Mic) Arrays represent the majority of the hardware used in a SSL system. All SSL systems require some use of microphones, and its specific configuration is the array. Certain requirements are accepted as necessary in order for them to accomplish localization of sound to different degrees. For example, humans have two ears which can be seen as a human's microphones. Most people cannot precisely locate the exact position or orientation sound comes from with their eyes closed, they still gain a sense of the general direction. Generally, to achieve a pinpoint location three different spaced microphones are necessary, which helps explain why humans cannot achieve such a feat. The reason multiple microphones are required is to sense minor differences between each and use algorithms to compare them to predict the source's location. Symmetricity shows importance as well because it allows the balancing of sound from all directions. The speed of sound is considerably quick, but still slow enough to be measured y even microphones close together. The more space or area between separate microphones allows increased precision and easier calculation because the differences between microphones are amplified. Furthermore, more microphones also mean more comparisons can be made also increasing accuracy. Currently, the problems do not involve the precision achievable with multiple microphones, but really by gaining more with less equipment and size. Common shapes of arrays are stars, lines, and squares. Squares function well for 360-degree recognition, while lines serve for specific directional location. Microphone arrays have become extremely precise tools, which can often lead to their detriment as well. Their increased sensitivity makes them more susceptible to false sources like reverberation off objects. If an array is sensitive enough to the differences between microphones, then it will surely pick up

on reverberation causing data to be disturbed. Researches have yet to combat issues like this (Mandlik, Nemec & Dolecek, 2012).

Binaural Approaches

Humans possess enough capability of sound localization to sense general directions of sound. Researchers aimed to push dual microphone arrays to their limit and use in their experiments to test simple applications like basic tracking adjustment. Some researchers claim that with near-perfect setup and calculation, a binaural approach could achieve the same level of accuracy and precision as that of a system with many more microphones because many blind deaf humans display super-human sound location capabilities (Yang, Jongdae & Donggug, 2008). Researchers aimed to attempt to solve the problem with a system consisting of microphones in positions mirroring that of human ears. The system faced difficulties achieving under five degrees of inaccuracy while using the Generalized Cross Correlation algorithms but managed more success using a neural network that acted similarly to a human brain. The algorithm faced computational difficulty as the inputs to this device gave significantly less information to act on, but supervised learning overcame that because it only required the result to learn the minuscule differences and patterns. The second attempt with a NN achieved high accuracy and precision of less than one degree off the location when predicting (Yang, Jongdae & Donggug, 2008).

Three Dimensional Designs

3D Sound Source Localization has also been a new advancing topic as well. Currently, most systems are designed to locate a sound source by their x and y coordinates, which allows the microphone arrays to fit onto one plane. With the goal to identify the elevation of the object in addition to the other information, the microphone array also must expand to three dimensions while maintaining symmetricity. One approach involved spacing microphones equally around a cube allowing the different microphones to meet the above conditions. Furthermore, all the microphones provided unique information in all three dimensions because they could be offset. Other approaches have involved using a spherical arrangement, which also has success because of its perfect evenness in all dimensions, directions and surfaces. Many patents have also been created with unique solutions in arrays to the three-dimensional problem (Belloch et al., 2015).

Algorithmic Approaches

TDOA Approaches

Time Difference of Arrival (TDOA) denotes the name for a method of Sound Source Localization which involves the measurement differences between microphones. The algorithm applies to microphone arrays that can have multiple different microphones record the same sound event. Because they are spaced and the traveling speed of sound, each microphone records the sound eventmin but shifted on a time scale. This minuscule difference is what the following algorithms analyze to localize the sound. Different algorithms are used to analyze these differences, and TDOA can often be paired with other approaches as well to create stronger and more accurate methods of analyzing this. The most basic form is Generalized Cross correlation which forms the base for most SSL algorithms that are currently being used. Mathematical deterministic algorithms are quite accurate and reliable, but computational cost serves as the main barrier to their expansion and improvement (Cobos et al., 2017).

Generalized Cross Correlation

Generalized Cross Correlation (GCC) refers to the use of cross correlation to relate the information between the microphones in a mic array. Cross correlation is a measure of similarity between two series, which is very helpful in assigning costs in a TDOA approach. GCC relates each pair of microphones and assigns a cost representing the signal differences to each one (Cobos et al., 2017).

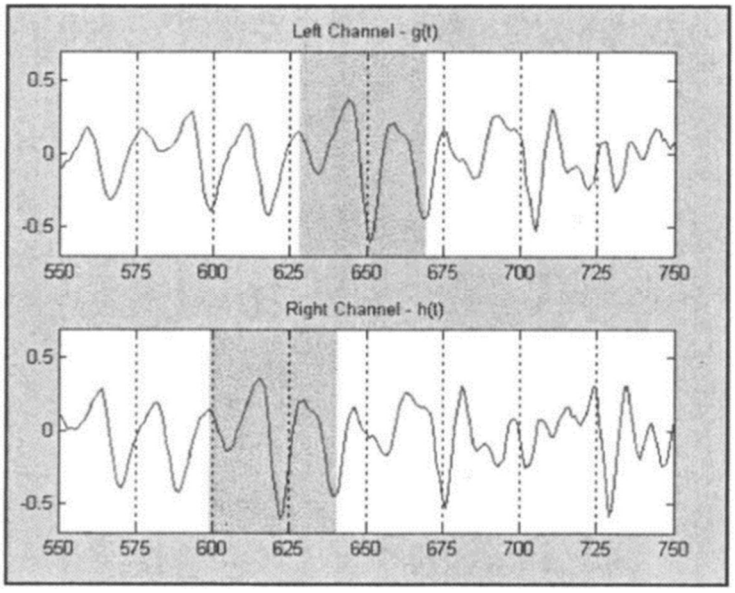

Figure 1. An image of cross correlation between two microphones. Retrieved from J. C. Murray, H. Erwin and S. Wermter, "A recurrent neural network for sound-source motion tracking and prediction," *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Montreal, Que., 2005, pp. 2232-2236 vol. 4.doi: 10.1109/IJCNN.2005.1556248.

GCC is computed by separating the input into different windows of time. Then the windows are compared to each other to calculate the cost. The algorithm is meant to locate the time differences of the desired signal as it appears in each channel separately. Examples of the different windows are above, with the dark grey windows representing the desired signal that is recorded and then GCC is applied to. Because all the microphones are identical, the exact sound waves are recorded, but with a time shift due to the spacing. GCC is meant to develop a cost based on the difference in the time window that the sound appears. While primitive, GCC lends itself in every TDOA SSL algorithm and creates a base for outputting the necessary raw differences between microphone pairs (Li, Zhang & Liu, 2015).

SRP-PHAT

SRP-PHAT stands for a Steered-Response Power Phase Transform, which is one of the most common algorithms dealing with the sound source localization problem because it performs robustly in reverberant conditions. The SRP portion of the algorithm aims to maximize and filter the incoming sound using a filter and a beamformer. Essentially, the process operates on the recorded audio waves from the microphone array and enhances or amplifies the important sound events from the ambient noise. This process creates the reliability of the SSL algorithm and makes it popular, but more computationally heavy. Specifically, the algorithm creates a spatial grid, where candidate locations, or locations where the sound is likely to come from based on the raw microphone inputs, are approached, and then run through Generalized Cross Correlation (GCC) which creates a function relating the TDOA information to the spatial location and determines the most likely point of positioning. This method combines noise filtering and TDOA techniques to create a robust and accurate technique for the SSL problem (Cobos et al., 2017).

Effectiveness

TDOA methods of SSL have proven effective for azimuth angle calculations, or where the direction where sound comes from. One group of researchers achieved high precision and accuracy using a four-microphone array running a simple time delay TDOA algorithm. The experiment displayed a standard deviation of less than two degrees for each tested measurement which shows that their implementation effectively and consistently located the correct sound source direction with very little inaccuracy (Mandlik, Nemec & Dolecek, 2012). On the other hand, their study further displayed the difficulties of estimating the distance away that the sound source lays. During their trials, the standard deviation of the measured distances from the true distance ranged between 0.07 meters and up to 14.34 meters. As the true distance increased the algorithm's prediction for the distance increased in variation (Mandlik, Nemec & Dolecek, 2012). Algorithmic SSL

shows strong promise in tasks that require accuracy and precision and should continue to be pursued and improved upon by researchers for an eventual implementation for consumer use.

Machine Learning Approaches

Neural Networks

Neural Networks (NN) are a set of algorithms based on the human brain. The structure is modeled after the neuron connections in the brain in order to calculate and solve more complex tasks that may require similar thinking to that of a human. NN are very effective at finding patterns and interpreting sensory data to develop a relationship between each input and output, which can then be used to predict other cases. NN need to be trained to gain the desired capabilities. This process is known as machine learning, and the process involves making a dataset containing a set of known output for a set of inputs. Then the modeled is repeatedly trained on this set, with each time generating a loss function, or what the model's error was. It then tries to self-adjust in a process called back-propagation to keep iteratively improving. The entire process is called gradient based learning. Once the model trains to the desired accuracy, predictions can then be made using the resulting model ("A Beginner's Guide to Neural Networks and deep Learning", n.d).

SSL Application

The most common approach to applying machine learning to SSL is to make a training set consisting of many recordings of the microphones with the sound source at a precise measured location. The data is then fed into the neural network and trained over many iterations to create a proper estimate of the sound sources true location. Certain implementations have trained models to higher accuracy then that of TDOA approaches. Their implementation at times achieved up to nine percent increase in accuracy of prediction over the MUSIC algorithm (Takeda & Komatani, 2016).

Convolutional Neural Network

Convolutional Neural Networks (CNN) are neural networks that effectively recognize images of two-dimensional matrices. Their structure differs from a typical DNN as they emphasize recursing over sections of an image to create a representation of the important spots in the image which is then used to create a prediction. A new method of sound classification makes use of such strong networks as sound can be transformed into spectrograms which map the frequencies present in the recording over time. They create a two-dimensional graph which can be analyzed like an image in a CNN, giving sound classification similar power as image classification. This method poses a new innovation into applying machine

learning to SSL and requires future research to see if its true effectiveness is higher than that of a regular deep neural network.

Conclusion

Future Steps

Fast and efficient Sound Source Localization poses a challenge to the scientific community wholly. A diverse range of solutions, methods, and implementations have been created in order to improve upon different elements of the problem, often at the trade-off of other parts. For example, TDOA algorithms create very accurate predictions at the costs of higher computational costs. Often microphone array size needs to be increased for more precision, which creates an exchange of improving one element for the cost of another (Mandlik, Nemec & Dolecek, 2012). Neural networks have seemingly entered the field as the newest promising solution to the problem because of its versatility and complex prediction method at a decent computational cost. As different types of Neural Networks are implemented with varying complexities, researchers hope to break the status quo and develop a system that develops all around success (X. Yue et al., 2018). Researchers will continue to tackle the problem with different methods, and future steps will be taken to progress the solutions from the research phase to have the consistency necessary for the industrial world.

Relation to the Hearing Impaired

Implementations of this technology may eventually be used in devices to benefit the hearing impaired. Because of the cost barrier to entry of many assistive technologies, a real-time SSL system could have merit in this field. The cost consists of the creation of the device itself, and with microphone arrays shrinking in size, could be utilized as wearable technology. The obstacles posed is the resolution or amount of information conveyable by an SSL system, and realistically it could only notify of the sounds location, but not of its composition or content. Although for a lower price this system, will function as a method of relieving some stress and providing aid for those who are financially unable to afford the currently available devices. Therefore, this project will be pursued in the future.

References:

A Beginner's Guide to Neural Networks and Deep Learning. (n.d.). From https://skymind.ai/wiki/neural-network.

Belloch, Jose & Cobos, Maximo & Gonzalez, Alberto & Quintana-Ortı, Enrique. (2015). Real-time Sound Source Localization on an Embedded GPU Using a Spherical Microphone Array. Procedia Computer Science. 51. 201-210. 10.1016/j.procs.2015.05.226.

Cobos, Maximo, Antonacci, Fabio, Alexandridis, Anastasios, … Bowon. (2017, August 17). A Survey of Sound Source Localization Methods in Wireless Acoustic Sensor Networks. From https://www.hindawi.com/journals/wcmc/2017/3956282/

Correll, R. (2019). Challenges That Still Exist for the Deaf Community. From https://www.verywellhealth.com/what-challenges-still-exist-for-the-deaf-community-4153447.

Deafness and hearing loss. (2019). From https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss.

How is Deafness Affecting Your Mental Health? (2016). From https://deafunity.org/article-interview/deafness-and-mental-health/

J. C. Murray, H. Erwin and S. Wermter. (2005) "A recurrent neural network for sound-source motion tracking and prediction," *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Montreal, Que., 2005, pp. 2232-2236 vol. 4.doi: 10.1109/IJCNN.2005.1556248.

J . Kim, K. Noh, J. Kim and J. Chang. (2018) "Sound Event Detection Based on Beamformed Convolutional Neural Network Using Multi-Microphones," *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, Guiyang, 2018, pp. 170-173.doi: 10.1109/ICNIDC.2018.8525597

M. Liang, L. Xi-Hai, Z. Wan-Gang and L. Dai-Zhi, (2015) "The Generalized Cross-Correlation Method for Time Delay Estimation of Infrasound Signal," *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, Qinhuangdao, 2015, pp. 1320-1323. doi: 10.1109/IMCCC.2015.283

M. Mandlik, Z. Nemec and R. Dolecek, (2012) "Real-time sound source localization," *2012 13th International Radar Symposium*, Warsaw, 2012, pp. 322-325. doi: 10.1109/IRS.2012.6233370

Risoud, M., Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., & Vincent, C. (2018). Sound source localization. From https://www.sciencedirect.com/science/article/pii/S187972961830067X.

Takeda, R & Komatani, K., (2016) "Sound source localization based on deep neural networks with directional activate

function exploiting phase information," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, 2016, pp. 405-409. doi: 10.1109/ICASSP.2016.7471706

X. Yue, G. Qu, B. Liu and A. Liu. (2018) "Detection Sound Source Direction in 3D Space Using Convolutional Neural Networks," *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, Laguna Hills, CA, USA, pp. 81-84. doi: 10.1109/AI4I.2018.8665693

Yang Geng, Jongdae Jung and Donggug Seol, (2008) "Sound-source localization system based on neural network for mobile robots," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, pp. 3126-3130.doi: 10.1109/ IJCNN.2008.4634240

Zhao, Shengkui., Ahmed, Saima., Liang, Yun., Rupnow, K., Chen, D., Jones, D. (2012). A Real-Time 3D Sound Localization System with Miniature Microphone Array for Virtual Reality. From https://www.researchgate.net/publication/258423607_A_RealTime_3D_Sound_Localization_System_with_Miniature_Microphone_Array_for_Virtual_Reality