Presented at the DLSU Research Congress 2016
De La Salle University, Manila, Philippines
March 7-9, 2016

Automatic Steering Platform for Mobile Device Video Telecommunication using Sound Source Localization

Ardiel Lapid, John Jerico Operio, Dan Joseph Porcioncula, Emarson Que, and Clement Ong*
College of Computer Studies – De La Salle University Manila, Philippines
**clem.ong@delasalle.ph*

Abstract: The continuing improvement of Internet bandwidth has driven the growth of VoIP, with a significant percentage of those associated with video calls. More recently, video calls using mobile devices, such as smartphones and tablets, became possible. The limited field of view (FOV) of such devices translate to a certain degree of criticality in setting up the device at a proper azimuth and elevation to ideally frame the user's face. Once set up, the user is expected to stay within the camera's FOV, making it difficult or inconvenient to continue the video chat if one has to move.

This research proposes to develop a small smartphone/tablet actuated platform based on sound-source localization technology that will allow the system to properly yaw and pitch to keep the user's face within the mobile device camera's field of view. A microphone array with four microphones is used to obtain the sound input. The Time Delay Estimation-based SSL algorithm generates the estimated location of the user and relays this information to two motors to pan and tilt the mobile device based on the generated location of the user. The platform has a tilting capability of 0 to 30 degrees and a panning capability of +/- 180 degrees.

The system requires the video conversation to be one-sided, where only one speaker/user talks and the noise that falls within the accepted voice bandwidth will be treated as another source.

Key Words: Sound Source Localization; Automatic Steering Platform; Mobile Device Platform;

1. INTRODUCTION

Video conferencing has changed the way business industry operates (Turek, 2008). It has given companies an opportunity to cut costs and save time by using video conferencing to hold meetings rather than different people meeting in one place coming from various places around the globe. Another benefit of video conferencing is that it is available on the majority, if not all, internet-enabled devices. For example, Activision used video conferencing by having programmers meet in a virtual collaboration environment and it saved them months in terms of developing a new game compared to the other companies that are competing in the market (Turek, 2008). Today, video conferencing is not available to big companies only but also to consumers who use it as another form of communication. Since its invention, many companies have integrated this method to create their own way of giving people the chance to use video calls. Skype Technologies created application allowing people to use video calling for free. Since their launch a decade ago, Skype has evolved and has given users features such as video calling, real-time speech and instant message translation and group calling capabilities (Pall, 2016). Skype's CEO Tony Bates said in 2011 that Skype users are averaging 300 million minutes per month of video calling and 50 percent of Skype's traffic is video calling (Leena, 2011).

Being unable to see the other person due to wrong camera angle or sudden movement that results in the other person being removed from the camera's field of view (FOV) defeats the purpose of video calling, however.

The problem can be solved by creating a platform that can automatically position the phone mounted on it, allowing the proper angle for the camera to capture the user in order to minimize the effect of the camera's limited field of view. The platform can be made to determine the location of person talking - such algorithms are called Sound-source Localization (SSL) algorithms and can theoretically accurately determine the user's position in 3D space, with much less resource requirements (computing complexity) compared to,

ISSN 2449-3309

for example, video/image processing approaches. The goal of this research is to produce an accurate estimation of the source position with the use of SSL algorithms.

There are two commercial products that system may be compared upon. The first would be the Polycom series created by Microsoft. The CX5000, CX5100, and CX5500 have a 360º camera and have 6 directional microphones at the base of the device. These devices are also called the Microsoft RoundTable. The differences between these three are that the CX5000 and CX5100 have 1080p video capabilities whereas the CX5500 has VoIP functionality. The drawback of these products are their price, ranging from $3000 to $4300 (Polycom CX5000 Optimized for use with Lync Data Sheet, n.d.). Another product would be the Swivl developed by Satarii (Satarii, 2014). This device has two product models, the basic and the premium. The basic model is used for photographers because the devices can only mount DSLRs. It has a 360º degree rotating capability and 20º degree up and down tilting capability. The other unit, which is the premium model, can be used to mount smartphones and tablets. This model does not have any tilting capabilities but still has a 360º rotating function. This model also has a wireless dongle with a built-in microphone for the device to be controlled and it can also be controlled by the user at the other end of the device using a mobile app (Prospero, 2012). The market price for this device is $395.

Compared to the two previous commercial devices, the proposed system in this paper can be used to mount any smartphone or tablets in the current market. The prototype being developed is meant to be used to provide one-to-one teleconferencing, as such, it is limited to detecting and turning as well as tilting to only one user or sound source at a time. With advanced SSL techniques it is projected that a remote control will not be required, simplifying the use of the system.

2. METHODOLOGY

For this research, the microphone array was set up consisting of four electret condenser microphones in a pyramidal configuration (Fig. 1). The actual setup is in a room with a temperature of 24°C and can be seen in Fig. 2,3 and 4. The array input was recorded by an M-Audio Delta 1010LT sound card at 88.2 kHz and processed through MATLAB. The Delta 1010LT is a 4-channel sound card capable of recording all channels at the same time. For this experiment, the source signal used was a single clap, to lessen the chance of noise distorting the signals before time delay estimation. The possible ambient noise that could interfere with the setup is the computer's whirring noise and sound reflections.

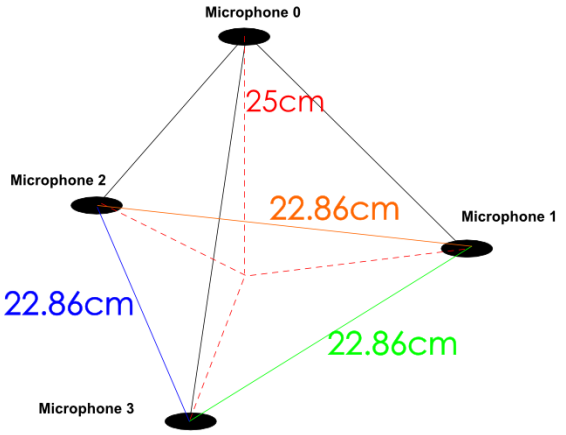

Fig. 1. Microphone Array Setup

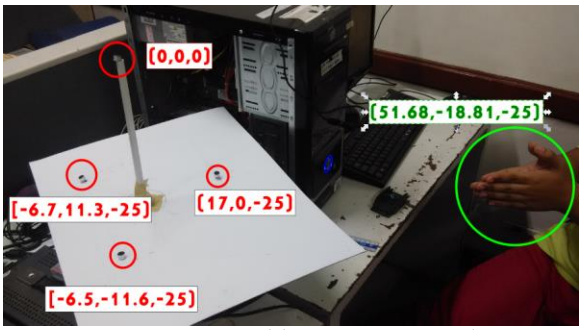

Fig. 2. Array Setup (1) with Source at (51.68, -18.81,-25) cm

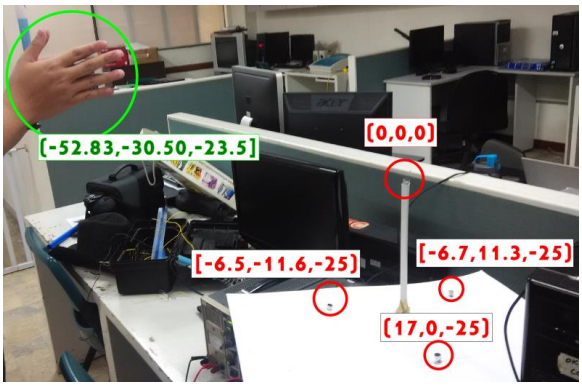

Fig. 3. Array Setup (2) with Source at (-52.83, -30.50,-23.5) cm

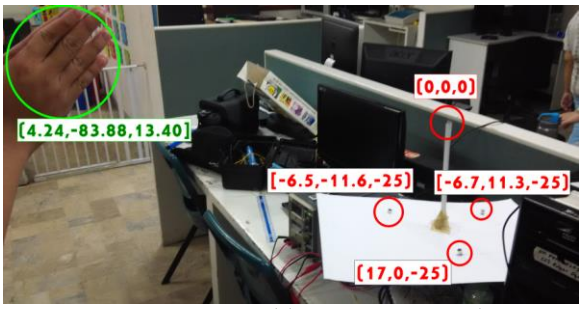

Fig. 4. Array Setup (3) with Source at (4.24, -83.88,13.40) cm

Table 1. Microphone Cartesian Coordinates

Microphone No.	Microphone Coordinates (x,y,z) in cm
Mic 0	(0, 0, 0)
Mic 1	(17,0,-25)
Mic 2	(-6.7,11.3, -25)
Mic 3	(-6.5,-11.6,-25)

To find the estimated location of the sound source, a Time Delay Estimation based localization method was used. The time delay is relative to the speed of sound and the range difference of a microphone pair as seen in Eq. 1 (Tellakula, 2007).

$$\tau_{ij} = \frac{1}{speed\ of\ sound}(d_{ij}) \quad \text{(Eq. 1)}$$

where:
d_{ij} = Range Difference of Microphone i and j
$speed\ of\ sound \approx 331 + 0.610 \times temp_{air}$

$$D_i \triangleq \|r_s - r_i\| = \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2 + (z_s - z_i)^2} \quad \text{(Eq. 2)}$$

where:
x_s, y_s, z_s = Cartesian Coordinates of Source
x_i, y_i, z_i = Cartesian Coordinates of Microphone i

$$d_{ij} \triangleq D_i - D_j, \qquad i, j = 0, \dots, N \quad \text{(Eq. 3)}$$

where:
D_i = Distance from source to Microphone i
D_j = Distance from source to Microphone j

The microphones in the array recorded audio samples simultaneously for 3 seconds to be used in acquiring the time delay estimates. For this case, the algorithms used were Cross-Correlation (CC) and Generalized Cross Correlation - Phase Transform (GCC-PHAT), implemented by using the weighting function Eq. 9, was used to compute for the time delay estimates. The Cross Correlation algorithm interprets the signal from a microphone i in an array as $x_i(t)$ in Eq. 4 (Khaddour, 2011).

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \quad \text{(Eq. 4)}$$

where:
α_i = Gain factor of microphone i
τ_i = Time delay from source signal to the microphone i
$n_i(t)$ = Noise detected by the microphone i

The algorithm gets the estimated time delay by getting the highest value of τ in Eq. 5.

$$R_{x_i x_k} = E[x_i(t) x_k(t - \tau)] \quad \text{(Eq. 5)}$$

where:
$R_{x_i x_k}$ = Cross power spectrum density function

$$\boldsymbol{S_{x_1 x_2} = E\{X_1(f) X_2^*(f)\}} \quad \text{(Eq. 6)}$$

where:
$S_{x_1 x_2}$ = Product of signals from microphone (x1) and microphone (x2)
$X_1(f)$ = Fourier Transform of the recording from microphone (x1)
$X_2^*(f)$ = Complex conjugate of the Fourier transform of the recording from microphone (x2)

$$\tau_\phi = \arg\max \psi_{x_1 x_2}(\tau) \quad \text{(Eq. 7)}$$

where:
τ_ϕ = Estimated time delay of signal arrival between microphone (x1) and microphone (x2)

$$\psi_{x_1 x_2}(\tau) = \int_{-\infty}^{\infty} \phi(f) S_{x_1 x_2}(f) \exp(j2\pi f \tau) \quad \text{(Eq. 8)}$$

where:
$\phi(f)$ = The weighting function

$$\psi_p(f) = \frac{1}{|S_{x_1 x_2}(f)|'} \quad \text{(Eq. 9)}$$

where:
$\psi_p(f)$ = The weighting function of PHAT

3. RESULTS AND DISCUSSION

Theoretically, the time difference of signal arrival between a pair of microphones can be

calculated by using Eq. 7. The speed of sound varies with the ambient temperature and during the experiment it was 25°C at sea level therefore the speed of sound was approximately 345.64 m/s. Table 2 shows the distance of each microphone from the source (Eq. 2) for the set up indicated in Fig. 1 & Fig. 2.

Table 2. Microphone Distances from Source

Microphone no.	Distance from source (Setup 1)	Distance from source (Setup 2)	Distance from source (Setup 3)
Mic 0	60.42 cm	65.37 cm	85.05 cm
Mic 1	39.46 cm	90.32 cm	93.13 cm
Mic 2	65.69 cm	78.91 cm	103.22 cm
Mic 3	59.71 cm	75.70 cm	47.06 cm

Table 3 shows the time delays for each unique microphone pair, determined at a temperature of 25°C. The algorithms used were Cross-Correlation and GCC-PHAT.

Table 3. Time Delay for each Microphone Pair (Setup 1)

	Pair 1,0	Pair 2,0	Pair 3,0
Theoretical	-606.39 us	152.63 us	-20.51 us
CC	-464.85us	-1632.65us	11.33us
GCC-PHAT	-464.85us	215.42us	56.69us

Table 4. Time Delay for each Microphone Pair (Setup 2)

	Pair 1,0	Pair 2,0	Pair 3,0
Theoretical	721.96 us	391.82 us	298.87 us
CC	419.50 us	-600.91us	79.36us
GCC-PHAT	385.49us	136.05us	11.34us

Table 5. Time Delay for each Microphone Pair (Setup 3)

	Pair 1,0	Pair 2,0	Pair 3,0
Theoretical	233.79 us	525.57 us	-1099.16 us
CC	124.72 us	-657.60 us	147.39 us
GCC-PHAT	136.05us	691.6us	260.77us

The results shows the time delay of each microphone. The values for Pair 1, 0 for the CC and GCC-PHAT has a small difference this is because the direction of the sound source is near from microphone 1. The sound source is directly being accepted by both microphones form pair 1, 0. For the Pair 2, 0, both have different values. This may be due to the reflection of the back board near microphone 2. Because of this, both algorithms interpret the reflection in a different way. For the pair 3, 0, the values have a large difference from the theoretical value and both have positive values whereas the theoretical has a negative value because microphone 0 may accept the sound source first rather than microphone 3 that is why it may have a conflict in terms of the sound source location.

4. CONCLUSIONS

The minimum distance of 22.86 cm between each microphone in the system is far enough to obtain a difference in sound source in which the approximate sound source location can be derived. This value was empirically obtained by experimenting on different distances of each microphones from one another and the 22.86cm distance gives a more readable and fine enough delays to get an approximation of sound source location. The Cross-Correlation and Generalized Cross Correlation - Phase Transform with weighting function is used by the system to compute for the time delay estimation and to shows that microphone receives the sound signal each in different time. For the future work in terms of the prototype development, the system algorithm will be finalized along with the platform. After finalizing the platform and the algorithm already installed in the platform, user testing will be implemented to know if the platform is doing what it is supposed to do and adjust the algorithm depending on the errors it gets. Also, more tests will be implemented using various smartphones and tablets. The potential area for improvement with regards to the research is the handling of the reverberations and multiple sound source localization where in it can locate the person talking and differentiate it from other sound source. The current system cannot cope with ambient noise, though GCC methods can give good results when the reverberation of the room is not very high, but when reverberation becomes important all of GCC methods will fail because they are based on a simple signal model that does not represent reality. So to address this noise problem in the system, some robust approaches with the

ISSN 2449-3309

reverberation and noise will be used in the future work. One of the approach is the adaptive eigenvalue decomposition which focuses directly on the impulse responses between the source and the microphones in order to estimate time delay. The goal of this method is not to accurately estimate the two impulse responses but rather the time delay.

5. REFERENCES

Abbasi, M., Dahlan, B., Honarbakhsh, P., Mansoor, W. (2011). Sound Source Localization for Automatic Camera Steering. Networked Computing and Advanced Information Management (NCM), 2011 7th International Conference.

Huang, Y., Elko, G. W., Benesty, J., and Mersereau, R.M. (2001), "Real-time passive source localization: A practical linear-correction least squares approach." IEEE Transactions on Speech and Audio Processing, vol. 9, no. 8, November 2001

Khaddour, H. (2011). A Comparison of Algorithms of Sound Source Localization Based on Time Delay Estimation. Elektrorevue, 2(1). Retrieved from http://elektrorevue.cz/.

Pall, G. (2015, January 14) Ten years of skype video: yesterday, today and something new. Retrieved from: http://blogs.skype.com/2016/01/12/ten-years-of-skype-video-yesterday-today-and-something-new/

Polycom® CX5000 Unified Conference Station Data Sheet. Retrieved from http://docs.polycom.com/global/documents/products/voice/conferencing_solutions/cx5000_datasheet.pdf

Prospero, M. (2012). Hands-On: Satarii Introduces Two New Swivls for Photogs and Video-Conferencing. Retrieved from http://blog.laptopmag.com/hands-on-satarii-introduces-two-new-swivls-for-photogs-and-video-conferencing

Satarii (2014). Swivl. Retrieved from http://www.swivl.com/

Tellakula, A. K. (2007). Acoustic source localization using time delay estimation. Degree Thesis. Bangalore, India: Supercomputer Education and Research Centre Indian Institute of Science.

Turek, M. (2008, May 2). Video Conferencing is Changing Businesses in Many Ways. Retrieved from http://www.informationweek.com/video-conferencing-is-changing-businesses-in-many-ways/d/d-id/1067458?