Article

Person Localization Model Based on a Fusion of Acoustic and Visual Inputs

Leon Koren 1, Tomislav Stipancic 1,*, Andrija Ricko 1 and Luka Orsag 1

Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Ivana Lucica 5, 10000 Zagreb, Croatia; leon.koren@fsb.hr (L.K.); ar200783@stud.fsib.hr (A.R.); orsagluka@gmail.com (L.O.)
* Correspondence: tomillav.stipancic@fsb.hr

Abstract: PLEA is an interactive, biomimetic robotic head with non-verbal communication capabilities. PLEA reasoning is based on a multimodal approach combining video and audio inputs to determine the current emotional state of a person. PLEA expresses emotions using facial expressions generated in real-time, which are projected onto a 3D face surface. In this paper, a more sophisticated computation mechanism is developed and evaluated. The model for audio-visual person separation can locate a talking person in a crowded place by combining input from the ResNet network with input from a hand-crafted algorithm. The first input is used to find human faces in the room, and the second input is used to determine the direction of the sound and to focus attention on a single person. After an information fusion procedure is performed, the face of the person speaking is matched with the corresponding sound direction. As a result of this procedure, the robot could start an interaction with the person based on non-verbal signals. The model was tested and evaluated under laboratory conditions by interaction with users. The results suggest that the methodology can be used efficiently to focus a robot’s attention on a localized person.

Keywords: spatial location; residual neural network; digital filter; person separation; cognitive robotics; multimodal signal processing; sensors; HRI

1. Introduction

Natural person-to-person interaction is highly contextual, in that a person can collect and analyze different inputs such as acquired or memorized information. During the process of reasoning, this information can shape and color the person’s current perception of the other person during the interaction, including the level of mutual understanding or attention [1].

Following similar concepts in the discipline of human-robot interaction (HRI), a robot, as a physical extension of the virtual software agent, uses different sensing modalities to gather information about the real world. Within this vision, this paper describes a model for audio-visual person separation that is used to identify and localize the talking person within a crowded place. The model is implemented for use on the interactive biomimetic robotic head PLEA [2]. Based on the multimodal approach, PLEA uses visual and voice modalities fused in a separate algorithm to determine the most appropriate hypothesis about the person’s emotional state at a given moment [3]. These decisions are changeable over time as the system receives the latest information. In this way, the robot can change the reasoning output in a timely manner, and adapt to significant changes in the environment. PLEA uses information visualization techniques to demonstrate multivariate data in the form of facial expressions or gestures. By using the light projector located in the neck, PLEA can express its emotions and communicate using non-verbal communication signals, as shown in Figure 1.
With the newly proposed model for audio-visual person separation presented in this paper, PLEA can accurately identify one person among other people and start or continue with their mutual interaction. In this approach, audio and visual techniques are combined to focus on, and to separate, the person of interest in a crowded place. The main goal of this approach is to focus on establishing the first contact and maintaining the interaction, by identifying and tracking a person to support the process of mutual understanding with the robot.

Voice perception is inherently multimodal (in this case, audio-visual (AV)). In addition to the acoustic speech signal reaching the listener’s ears, the presence of the face may be used to identify the appropriate person seeking interaction. Therefore, the model combines information about the presence of all detected faces and associates this information with the direction of the sound. The direction of the sound is precisely determined using a multi-microphone array, as shown in Figure 2.

The model relies on ResNet neural network architecture to find the person’s face [4]. As a part of the computational model, the sound is also filtered and analyzed using various parameters to identify a corresponding voice of the person and its direction.

The rest of this paper is organized as follows. Section 2 presents the current and comparable concepts used for audio-visual system perception, and the contribution of the presented methodology to date. Section 3 introduces the physical design of the system and its components including a multi-microphone array, a microcomputer, a web camera, and a workstation that is used to host the PLEA virtual agent. The same section describes the audio-visual methods, including the mathematical background used to filter and analyze the acquired sound, DL techniques used during the face detection process, and fusion techniques used to output the current reasoning hypothesis. The evaluation results are
discussed and analyzed in Section 4. At the end of the paper, a conclusion, summarizing the main concepts and the potential future research directions is provided in Section 5.

2. Current and Comparable Concepts Used for Audio-Visual System Perception and Contribution of Presented Methodology

Fusion of information about a particular phenomenon of interest using different sensing modalities can increase the perception potential of a technical system [5].

The tasks related to the fusion of visual and voice (or acoustic) modalities are usually associated with speech perception, enhancement, or separation [6]. For example, a vision modality can be used to strengthen the reasoning hypothesis of the model, as reported in [7]. A similar work has presented a computational architecture that gives the system the ability to adapt to change [8].

Classical speech separation methods have assumptions in relation to the statistical features of the signals, and aim at estimating the underlying target signal or signals according to mathematically tractable criteria, as reported in [9–12]. Modern methods for speech enhancement or separation methods are often connected to deep learning techniques [13], focusing on natural language processing (NLP) tasks where the quality of the acquired source depends on many factors, such as source distance, level of background noise, number of sources, etc., as reported in [14,15]. For example, in [16], a machine learning based method is used to combine a visual modality with an acoustic input to improve the corrupted audio source and increase the quality of information for further processing.

Neuroscience studies [17,18] and voice perception studies [19] have shown that vision, as a sensing modality, has a strong impact on a person’s ability to focus on a particular phenomenon of interest. The research presented in this paper is based on the fusion of audio and visual inputs to separate a single speaker from a crowded place. The model for audio-visual person separation can be classified under audio-visual sound separation systems (AV-SS) [6]. Such systems aim to extract one or multiple voice targets from a mixture of voice and visual information [20,21]. This information can then be used in different applications, such as focusing attention [22,23], establishing eye contact [24,25], and building a mutual understanding [26], etc.

In contrast to these state-of-the-art methods, which are mostly used in laboratory settings or in simulations, the proposed system has been designed to work in near real-time and real-world scenarios. The model can also correct wrong assumptions in a timely manner as new information is acquired, enabling the system to be more resistant to constant environmental changes and disturbances. In this way, the research fulcrum of this work is partially set on the production of the adaptation of system capabilities. Additionally, the system is fully integrated and can be used as a stand-alone module or as a part of other artificial intelligence models.

3. Materials and Methods

3.1. System Architecture

The system consisted of a LyraTD-MSC microcontroller platform, a Raspberry Pi 4B microcomputer, and a standard web camera, as shown in Figure 3. The specifications of the web camera were 78° field of view, 1080p image resolution, and an image-capturing frequency of 60 frames per second. If another camera is to be used, the parameter setting procedure should be repeated to accomplish more reliable system performance (for example, the parameter for the camera field of view should be altered during the calibration procedure).
Serial communication was used for data transfer between the microcontroller and the microcomputer. Additionally, a special audio transfer protocol (I2S) was used to transfer the audio data. From there, the data were transferred to the remote workstation via the standard network infrastructure.

Sound acquisition and direction of arrival (DOA) estimation was achieved by the LyraTD-MSC module designed by Espressif [27]. The module had 3 microphones placed at the same radius and 120° apart, as shown in Figure 4. The calculation of DOA was performed by a digital signal processing chip, located on the module itself. The module had a step resolution of 5° degrees. The same module was also responsible for filtering the sound and converting it to a digital format.

Face expression extraction was performed on the Raspberry PI microcomputer using Python library OpenCV. The camera was placed on the central axis above the microcontroller so that the DOA angle could be converted to the angle in the camera’s field of view. When properly placed and calibrated, a live camera feed shows a vertical line indicating the direction that the sound (the human voice) is coming from. The implementation of the physical model setup is shown in Figure 5.
Extraction of facial bounding boxes was performed from the live feed video based on the OpenCV library and the residual neural network model. The DOA angle was then combined with the positions of the extracted bounding boxes using a hand-crafted overlapping algorithm. In this way, the system could identify the person in front of PLEA who was speaking and interacting with it.

By focusing on a single person instead of processing the entire information stream, the proposed architecture significantly reduces the processing time on a remote workstation. Additionally, the system focuses on the reasoning attention, which is much more natural and human-like.

3.2. Methodology

The described system was based on couple algorithms implemented on a few diverse levels and platforms. The software that controls the processing of audio signals and DOA angle was implemented in C language using the ESP IoT Development Framework. Signal fusion was performed in Python on the Raspberry Pi platform and face recognition was implemented using the OpenCV library. The input DOA angle was filtered using a discrete low-pass filter, which was also implemented on the microcontroller.

3.2.1. Sound Direction Acquisition and Filtering

A digital signal processing (DSP) chip with a predefined algorithm for calculating the DOA angle was used to determine the direction of the sound. The algorithm was based on the principle by which sound waves reach one microphone sooner than others. When positioned correctly and based on the microphone input, the algorithm used trigonometry to calculate the angle from which the sound was coming, with a high degree of certainty. The same chip also filtered signals from the microphones so that the entire system was biased towards a human voice.

The angle was refreshed every 90 ms and read from the registry of the DSP chip with an ESP32 microcontroller, after which, the algorithm in the microcontroller filtered the angle and sent the result to the Raspberry Pi. The first and second order filters were tested at three different cut-off frequencies to establish important parameters for a specific use case.

The filters were designed as continuous, and transformed to the discrete system using (bilinear) transformation. The cut-off frequencies were empirically selected from the data based on the average human interaction time, which was empirically determined.
Sample time was determined experimentally by alternating the sound between the left and right speaker. The speakers were positioned against each other with the LyraTDMSC module in the middle. The iteration time between alternations in each cycle was shortened until the point when the module could no longer determine the DOA angle. At this point, the alteration time was taken as the sample time. Both filters were implemented as recursive equations.

The first order filter in a recursive equation form is defined in (1). The input value is defined by a parameter x and the output is defined by a parameter y. The index k is a discrete step (e.g., k–1 is the value of a given variable for one step in the past).

\[ y = A_1 \cdot y_{(k-1)} + B_0 \cdot x_{(k)} + B_1 \cdot x_{(k-1)}, \]  

where:

\[ A_1 = \frac{2 - T \cdot \omega_p}{T \cdot \omega_p + 2} \]  

\[ B_0 = \frac{T \cdot \omega_p}{T \cdot \omega_p + 2} \]  

\[ B_1 = B_0 \]

The second order filter in recursive form is defined in (5).

\[ y = B_0 \cdot x_{(k)} + B_1 \cdot x_{(k-1)} + B_2 \cdot x_{(k-2)} - A_1 \cdot y_{(k-1)} - A_2 \cdot y_{(k-2)} \]  

where:

\[ B_0 = \frac{1}{4 \cdot s^2 + 2 \cdot \sqrt{T} \cdot s + 1} \]  

\[ B_1 = 2 \cdot B_0 \]  

\[ B_2 = B_0 \]

\[ A_1 = (8 \cdot s^2 - 2) \cdot B_0 \]  

\[ A_2 = (4 \cdot s^2 - 2 \cdot \sqrt{T} \cdot s + 1) \cdot B_0 \]

All equations are given as a function of sampling time, denoted by T, and cut-off frequency denoted by \( \omega_p \), in radians per second. \( A_2, A_1, B_2, B_1, B_0, s \) are used as substitution parameters in front of discrete steps in formulas (1) and (5).

### 3.2.2. Image Acquisition and Face Detection

Image acquisition was performed using a standard web camera and a Python script inside the Raspberry Pi. The OpenCV library was used to capture an image from the stream and to implement a face detection functionality. The algorithm used for face detection was a well-known single-shot detection algorithm with a residual neural network [29]. The main reason for using this approach was to make the algorithm work faster, especially for images containing multiple faces. This was combined with a deep residual neural network which achieved better results, compared with a standard convolutional neural network [30]. Residual neural networks can use a deeper architecture without increasing the training loss, resulting in better performance [31]. This is due to the direct connections between layers, so errors are not multiplied throughout layers, as shown in Figure 6. In contrast, adding additional layers to convolutional neural networks leads to an increase in losses and a decrease in accuracy. This is especially important in face detection applications, which are usually based on larger data sets that contain input features that are more easily learned by deeper networks.
From an image, the algorithm returned an array of bounding boxes. Each bounding box was a set of two coordinates in the format \((startX, startY, endX, endY)\). The relative angle from LyraTD-MSC was then converted to an X coordinate on the acquired image. To achieve this, the camera’s field of view was aligned with an arbitrarily defined angular span on the DOA angle detection module. The conversion formula is defined in (12).

\[
X = X_m \cdot \frac{\phi - \phi_s}{\phi_m - \phi_s}
\]  

where:
- \(X\) – frame coordinate in pixels
- \(X_m\) – width of frame in pixels
- \(\phi\) – DOA angle
- \(\phi_m\) – camera’s field of view maximum angle
- \(\phi_s\) – camera’s field of view minimum angle

3.2.3. Signal Fusion

The final information fusion was performed by overlapping information from bounding boxes with the angle information. When the converted angle value was placed between the minimum and maximum value of the X coordinate, the speaker marked with this bounding box was identified as the current speaker, and marked with green color, as shown in Figure 7. In this way, the right speaker was identified and localized, and the robot could start or continue the interaction process.

Systems that operate in a similar way can be sensitive to input signal synchronization problems, as described in Coviello et al. [32]. A synchronization acquisition method was used in this work to prevent all such cases. When a new direction angle was determined during the acquisition, the system could acquire the new image from a video stream, and in the following step, detect faces. The production of these two operations (which lasted less than 10 ms) was faster than the sensor sample time and eight times faster than the time required for the angle information acquisition (which lasted about 90 ms). Therefore, the system is not expected to experience synchronization problems and artifacts.
4. Results and Discussion

After assembling the system, a testing procedure was performed to determine the responsiveness of the algorithm to the detection of the angle and to determine the best filter. For this purpose, three cut-off frequencies were defined and tested, based on the average speech rates of native English speakers [28]. The system was evaluated in a controlled environment by playing the same sentences from two loudspeakers, each located on one side of the camera’s field of view. After recording, the data were processed using the RMS method and the filter with the smallest error was chosen.

For this purpose, three cut-off frequencies were defined and tested, based on the average speech rates of native English speakers [28]. The system was evaluated in a controlled environment by playing the same sentences from two loudspeakers, each located on one side of the camera’s field of view. After recording, the data were processed using the RMS method and the filter with the smallest error was chosen.

The second test was designed to evaluate the overall system in a real environment. Student volunteers were used to interact with the system. After the interaction, students were asked to evaluate the interaction and the precision of the system. The evaluations were statistically analyzed and compared with the mathematical evaluation explained in Section 3.

The speech rate of native English speakers can vary from 80 words per minute to 160 or more [28]. To analyze more cases, lower and average values were used. Additionally, data about average sentence length was used, which finally resulted in the average time it took a person to say a sentence.

The parameters for both filters and all cut-off frequencies are given in Table 1, where $f_c$ is the cut-off frequency in hertz and $A_2, A_1, B_2, B_1, B_0$ are parameters in front of discrete values for the first and second steps as described in subsections 3.1 and 3.5.

| $f_c$ (Hz) | Filter | $A_1$  | $A_2$  | $B_0$  | $B_1$  | $B_2$  |
|-----------|--------|--------|--------|--------|--------|--------|
| 0.66      | 1. Order LPF | 0.6855 | -      | 0.1573 | 0.1573 | -      |
|           | 2. Order LPF | 1.4864 | -0.5936 | 0.0268 | 0.0536 | 0.0268 |
| 0.40      | 1. Order LPF | 0.829  | -      | 0.0853 | 0.0853 | -      |
|           | 2. Order LPF | 1.6836 | -0.7273 | 0.0109 | 0.0218 | 0.0109 |
| 0.26      | 1. Order LPF | 0.883  | -      | 0.059  | 0.059  | -      |
|           | 2. Order LPF | 1.7931 | -0.8126 | 0.0049 | 0.0097 | 0.0049 |
Testing was performed in a controlled environment using a generic English-speaking audio file. The audio file was played on speakers 75° apart on the same radius. Simulation of two different DOA angles was achieved by playing the audio source alternately from one speaker to another. The angle data were recorded, and are shown in Figure 8.

![Graphs showing DOA angle data](image)

(a) First order filter response  
(b) Second order filter response  
(c) $f_c = 0.26$ Hz  
(d) $f_c = 0.4$ Hz  
(e) $f_c = 0.66$ Hz

**Figure 8.** Low pass filter responses.

In order to better evaluate different filters, the error with respect to the ideal response was also calculated. The calculation was conducted using the root mean square (RMS) algorithm for the whole recorded response. The calculated errors are listed in Table 2.
Table 2. RMS Errors.

| Cut off Frequency | Filter Type | RMS Error |
|-------------------|-------------|-----------|
| \( f_c = 0.66\text{Hz} \) | 1. Order LPF | 13.9° |
|                    | 2. Order LPF | 16.4° |
| \( f_c = 0.4\text{Hz} \) | 1. Order LPF | 25.2° |
|                    | 2. Order LPF | 25.3° |
| \( f_c = 0.26\text{Hz} \) | 1. Order LPF | 30.1° |
|                    | 2. Order LPF | 33.6° |

From the results presented, it can be concluded that the filter with the highest cut-off frequency gave the best results, which was expected because a higher cut-off frequency has a faster response. Secondly, the first order filter gave a smaller RMS error compared with the second order filter. This was because the second order filter had an overshoot on the step response. To reduce or to eliminate overshoot, the damping parameter must be lowered, which would also slow down the response time. When comparing first and second order filters (as shown in Figure 8c,d), it can be seen that this type of change in the second order filter further increases the RMS error, compared with the first order filter. Considering all this, for the given application, the first order filter was recognized as better.

The cut-off frequency depends not only on the RMS error and the values in the response graphs, but also on the subjective responses of the interacting user. Therefore, for this part, the entire system was tested with few interacting participants, as shown in Figures 9 and 10.

After testing the entire system under controlled conditions, the best filter among all variations was the first order filter with the largest cut-off frequency (\( f_c = 0.66 \text{Hz} \)). The experiment with the participants confirmed the results of the statistical tests previously performed with the filters. The main reason for choosing the first order filter over the second order filter was the lack of an overshoot characteristic, which worsened the results.

Figure 9. Speaker 1 talking.
Figure 10. Speaker 2 talking.

5. Conclusions

The model for audio-visual person separation presented in this paper was used to determine the spatial position of the interacting person through visual and acoustic signal fusion. As a part of the visual modality, ResNet neural network architecture was used to detect the faces of people in the room. Within the speech (acoustic) domain, the direction of the voice was determined and matched to the corresponding face of the person, to produce a localization.

Data obtained through testing proved that the model can be used in many situations, particularly for implementation on the social robot PLEA [2]. Based on the information obtained, the robot can turn its head or eyes towards the speaker and react to her or his presence. Additionally, if two or more users/students talk at the same time, the system will target the person who is currently the loudest. If someone else becomes louder than the person currently speaking, the system will shift its attention to the new person.

Because the system can detect angles only in steps of $5^\circ$, which is the module step resolution, the system is not efficient enough to detect people who are both too far away from the sensors and too close to each other. On the other hand, if people are positioned within close proximity, a detection is possible if the face detection algorithm can see the face and if talking does not cap microphones. Faces cannot be detected if they are too close to the system and too close to each other. In that case, the angle between them in relation to the system can be greater than $5^\circ$. The result of close proximity is that the person’s face becomes too big. Therefore, the number of detected faces is proportional to the distance between people and the camera.

This model can be used as a part of the new gaze direction algorithm for virtual agents immersed in the real environment using different interaction interfaces (e.g., large screens, TVs, PC monitors, mobile phones, etc.), which can improve the technology acceptance and interaction quality. Moreover, based on evaluation results, the proposed system outputs optimized results with a small computational overhead, so the entire system can be run on small form factor computers and further improve the distribution of computing power.

Future work will continue research on the model and associated methodology. Based on visual clues from the speaker, such as lip movements and facial expressions, the robot PLEA will focus its attention on establishing direct eye contact and following the speaker by moving its eyes. This mechanism will have an additional effect on the level of mutual understanding between the robot and the person with whom it interacts.
Moreover, the output results generated in this work will be combined with the results generated by models developed in the previous work. For example, PLEA can determine the emotional state of the person in the interaction and accordingly update and fine-tune its reactions in the form of its own facial expressions [2]. In this way, PLEA could provide responses which are much more contextually aligned to the current situation, in addition to considering other information.

In addition to the model developed and presented in [3], this work will continue by exploring how the robot can hear all people in vicinity but “concentrate” on only one.

**Author Contributions:** Conceptualization, L.K. and T.S.; methodology, L.K. and T.S.; software, L.K. and A.R.; validation, L.K., T.S. and L.O.; formal analysis, L.K.; investigation, L.K. and A.R.; resources, T.S.; data curation, L.K., A.R.; writing—original draft preparation, L.K. and T.S.; writing—review and editing, L.K. and T.S.; visualization, L.K.; supervision, T.S.; project administration, T.S.; funding acquisition, T.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research and APC were funded by Croatian Science Foundation, grant number UIP-2020-02-7184.

**Institutional Review Board Statement:** The study was conducted in accordance with the principles of the Declaration of Helsinki, and it was approved by the Institutional Review Board. The research represented in this paper was not registered in The Clinical Trial Registration because it is purely observational and does not require registration.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Acknowledgments:** This work has been supported in part by the Croatian Science Foundation under the project “Affective Multimodal Interaction based on Constructed Robot Cognition—AMICORC (UIP-2020-02-7184).”

**Conflicts of Interest:** The authors declare no conflict of interest.

** References**

1. Barrett, L.F. *How Emotions Are Made: The Secret Life of the Brain*; Houghton Mifflin Harcourt: Boston, MA, USA, 2017. https://doi.org/10.7202/1064926ar.
2. Stipancic, T.; Koren, L.; Korade, D.; Rosenberg, D. PLEA: A social robot with teaching and interacting capabilities. *J. Pac. Rim Psychol.* 2021, 15, 18344909211037019. https://doi.org/10.1177/18344909211037019.
3. Koren, L.; Stipancic, T. Multimodal Emotion Analysis Based on Acoustic and Linguistic Features of the Voice. In Proceedings of the International Conference on Human-Computer Interaction, Virtual Event, 24–29 July 2021; Springer: Cham, Switzerland, 2021; pp. 301–311. https://doi.org/10.1007/978-3-030-77626-8_20.
4. Ranjan, R.; Patel, V.M.; Chellappa, R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 2017, 41, 121–135. https://doi.org/10.1109/tpami.2017.2781233.
5. Jiang, Y.; Li, W.; Hossain, M.S.; Chen, M.; Alelaawi, A.; Al-Hammadi, M. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf. Fusion* 2020, 53, 209–221. https://doi.org/10.1016/jинфус.2019.06.019.
6. Michelsanti, D.; Tan, Z.H.; Zhang, S.X.; Xu, Y.; Yu, M.; Yu, D.; Jensen, J. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 2021, 29, 1368–1396. https://doi.org/10.1109/taslp.2021.3066303.
7. Stipancic, T.; Jerbic, B. Self-adaptive Vision System, In Proceedings of the Doctoral Conference on Computing, Electrical and Industrial Systems, Costa de Caparica, Portugal, 22–24 February 2010. https://doi.org/10.1007/978-3-642-11628-5_21.
8. Jerbic, B.; Stipancic, T.; Tomasic, T. Robotic Bodily Aware Interaction within Human Environments. In Proceedings of the SAI Intelligent Systems Conference (IntelliSys 2015), London, UK, 10–11 November 2015. https://doi.org/10.1109/IntelliSys.2015.7361160P.
9. Lu, R.; Duan, Z.; Zhang, C. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Process. Lett.* 2018, 25, 1315–1319. https://doi.org/10.1109/lsp.2018.2853566.
10. Luo, Y.; Wang, J.; Wang, X.; Wen, L.; Wang, L. Audio-visual speech separation using i-vectors. In Proceedings of the 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), Weihai, China, 28–30 September 2019; IEEE: Piscataway, NJ, USA, 2020; pp. 276–280. https://doi.org/10.1109/ICICSP48821.2019.8958547.
