Building a facial and speaker recognition application that operates on the fly for monitoring conference attendees is a challenge, but an AI-guided system is proving equal to the task.

“We are primarily trying to solve the problem of ease of use for user identification. Whether it’s speech or face, the enrollment process requires a tedious enrollment step. We can make this automatic, when the person is having a conversation with another person, or when interacting with the device in a natural way.”

- Jon Huang, Senior Staff Researcher, Intel

**Challenge**

Typical video conference tools are not very immersive, nor do they take full advantage of interactive capabilities. A solution to modernize videoconferencing recordkeeping by employing the latest artificial intelligence (AI)-guided recognition technologies would be a benefit to the industry.

**Solution**

By introducing a bot into video conferences, equipped with facial and speaker recognition tools, typical meeting activities such as recording comments and generating minutes could be fully automated, capturing attendee ideas and their positions on issues being discussed.

**Background and Project History**

Julio Zamora and Jonathan Huang combined their individual expertise—one in facial recognition and the other in voice and speech recognition technologies—to develop a unique solution that adds value and oversight to video conferences. The result of their collaborative research project is a conference-oriented, multimodal system for re-identifying participants during the event. The system creates its own database over the course of a video conference, enrolling new users automatically and learning—on the fly—how to recognize individuals using a combination of speaker recognition and facial recognition.

“We were motivated by the fact,” Julio said, “that our current videoconferencing methods are not very immersive. We wondered whether it was possible to design an application to send the meeting invitation to a virtual attendee, a bot, and then use the bot during the meeting to track faces and create the meeting minutes automatically.”

“Conventional machine-learning methods for recognizing faces assume you have the database available when you first get started,” Julio said. “Then you can train the neural network and deploy the models used to drive the inference engine. In practical applications, however, this is not always a realistic assumption. A key challenge that we addressed is learning on the fly.”

In some circumstances, however, there might be prior information about individual faces available at the start of the conference. The challenge then becomes being able to identify previous conference attendees whose clothing, hairstyles, and other characteristics may be different than before, while also detecting and classifying new attendees. “We designed the system to be resilient,” Jon explained, “and to learn faces in the conference room on a given day. The system can then associate the faces of the session with prior databases, if available. If not, the system can identify new attendees and make the association, matching faces to names over the course of the session.”
Over the history of speech research, certain obstacles have slowed progress and made it difficult to implement speaker recognition technology, primarily because of the nature of the enrollment process. “For a text-dependent speaker recognition system (such as Siri* or OK Google*),” Jon said, “it only takes a few utterances of the key phrase to build a robust model. However, for a text-independent system (speaker recognition that is not dependent on a particular phrase), it typically takes several minutes of speech input from a user to accurately capture the acoustic content of most phonemes.”

Realizing that a long, cumbersome enrollment process results in a bad user experience, Julio and Jon combined computer vision with the speaker recognition, making it possible to eliminate the explicit enrollment step. The vision system can identify and cluster faces and can also detect when faces are talking. The faces can then be associated with speech samples, which can be used to perform text-independent speaker recognition enrollment in a natural setting. With the knowledge of speaker patterns in a conference, the system can produce transcriptions using automatic speech recognition (ASR) that can be associated with individual faces and the actual user.

“As an extension to this idea,” Jon noted, “if we have models of a speaker by audio, but not by face, we can enroll using the face ID algorithm when the speech audio of a person is confidently identified. Together, these ideas fully enable self-learning of user identity by a combination of audio and vision.”

Some of the ideas behind this project originated as early as 2012, and the groundwork—speaker and face identification—has been in development for a few years. “For about one and half years,” Julio said, “we have been actively working on the creation of building blocks, and we keep improving the results over time.”

“Although the multimodal part is still being improved,” Jon commented, “we have completed the technical transfer of the speaker identification components to the business unit to become part of an Intel portfolio of speech technologies.”

The project explored recognition tasks that went beyond prior machine-learning precedents, such as discovering the number of members in a family and labeling each association, without relying on supervision to accomplish the task. Julio said, “The creation of a relationship between audio (speaker recognition) and video (facial recognition) helps us track and recognize the user when one of the modalities is missing.”

**Process of Facial and Speaker Recognition**

Both speaker and voice recognition were used in this project to positively ascertain conference attendee’s identities and attribute spoken remarks to individuals for accurately recording the minutes of a meeting.

“At a high level,” Jon said, “we perform face detection and tracking by capturing the sequence of images for each face inside a frame, to generate a preliminary database. This database is then used to train a [convolutional neural network] CNN model in which feature extraction and descriptor separation are also involved.”

![Figure 1. System for facial and speech recognition.](image)
Jon explained that the resulting trained model is then used as a real-time inference engine, assigning labels for known users—previously identified faces that appeared inside the collected frames. A label analysis over the real-time input image is executed to clean up the collected database. For example, if two faces are mistakenly assigned the same label, the system switches to a model-retraining procedure. The retraining process updates the database with new (actual) labels.

After all the current face labels have been updated, they are used to segment the audio inputs, generating a set of associated patterns per user, and equipping the system to automate enrollment based on audio and speaker profiles.

Enabling Technologies

Jonathan and Julio took advantage of several Intel technologies, including the Intel® Movidius™ Vision Processing Unit (VPU), Intel® RealSense™ technology, and the 8-microphone circular array included in the Intel® Speech Enabling Developer Kit. “These are the perfect components for this type of application,” Julio said, “because all you need is good eyes, good ears, and an efficient brain. Our algorithms are the glue that takes advantage of these components to generate a complete solution.”

“The framework we used,” Julio said, “is the Robot Operating System (ROS), which is widely used in the robotics community. ROS simplifies the multimodal data fusion and, fortunately, Intel releases ROS nodes for each component (Intel RealSense technology, the mic arrays, and so on). With these ROS nodes available, using these sophisticated hardware components is as easy as subscribing to an ROS topic.”

Another valuable ingredient for architecting solutions of this type is Multi-Task Cascaded Convolutional Networks (MTCNN), which provides effective face detection. MTCNN uses a set of face landmarks that are useful for face alignment.

Technological Milestones

As a part of the project, Jonathan and Julio developed a method for determining the number of iterations required to complete a training sequence. A paper on the resulting method, called the Dynamic learning rate approach, was presented at ICPR2018: Dynamic Learning Rate for Networks: A Fixed Time Stability Approach. The method they developed uses Lyapunov’s stability theory and implements a training rule based on this theory.

“In addition,” Julio said, “we invented a new type of adaptive convolutional neural network capable of defining its own convolutional kernels on the fly—to adapt and extract better features from the input images. The approach we used saves a substantial amount of memory for MNIST* and even more for CIFAR10*. It also lets us reduce from 16 layers used in FaceNet* recognition to 8 layers in our solution. This technology will soon be public without any patent or restriction.”

Prospective Use Cases

Combined with automatic speech recognition, this technology can provide benefits to organizations by automatically generating transcripts during meetings—keyed to individual speakers—for a real-time, accurate record of the proceedings. In enterprise markets in which collaboration is an important part of the daily workflow, the recognition tools unlock opportunities for Intel to establish technologies to power platform solutions.

“For the smart home,” Jon said, “the speech assistants (including Amazon Echo*, Google Home*, and other smart appliances) can seamlessly learn users’ voices without an explicit enrollment. If the system can figure out the identity of a person by itself, the experience can feel magical.”

The speaker recognition component has been transferred to an Intel product group. The plan is to integrate it with other technologies available in the Intel Speech Enabling Developer Kit, including wake-on-voice and far-field voice capabilities.

Forward-Looking Development Perspectives

In the development of the hardware and software infrastructure, Julio and Jonathan selected the ROS as the optimal framework for a multimodal approach (see Figure 2). “Intel hardware is 100 percent compatible with this framework,” Julio said. “Because it can effectively perform machine learning on the fly, ROS saves a lot of money and time that would otherwise be spent on the creation of databases and labeling of data.”

In implementing a multimodal approach, Julio noted that there is no single modality that is completely robust for performing user tracking. The team determined that the only viable way to tackle the user tracking operations in a home scenario would be to apply a combination of different modalities to increase the redundancy and boost the accuracy.

“We started thinking about the user re-identification problem independently for each modality (the audio team with one solution and the video team with another),” Julio said. “Now we understand this as a multimodal problem that requires architecting a solution to integrate both modalities in a new way.”

Reflecting on the lessons learned during the project, Jon had a similar perspective. “Before we integrated the audio and vision components together, the individual modalities were developed separately. With the benefit of hindsight, it would be better to conceive a solution that integrates more tightly together such fusing features at an earlier stage for improved accuracy.”

Jon noted that the speaker recognition community has an extensive history of algorithms to streamline development projects. For example, The Gaussian Mixture Model—Universal Background Model—GMM-UBM—is one of the predominant techniques for performing text-independent speaker verification.
SUCCESS STORY | AI-Powered Speaker and Facial Recognition System

Other algorithmic approaches and useful toolkits include:

- GMM-supervectors
- GMM-based i-vectors
- Deep neural network (DNN)-based speaker recognition
- Components from the Kaldi* Toolkit

**Figure2.** ROS framework for the multimodal ID implementation.

**A Quick Look at the Robot Operating System**

The Robot Operating System (ROS) provides a comprehensive set of tools, libraries, and conventions to make it easier to build robotic solutions within a consistent, but flexible framework. Sponsored by the Open Source Robotics Foundation, ROS is fully open source and works with a simulator, Gazebo*, that is also open source and includes a physics engine, graphics resources, and programmatic and graphical interfaces. The Open Robotics organization supports collaborative development projects and engages in many different types of work, including custom engineering, research and development projects, consulting, and application development. Collaborative work spans academy, government, and industry sectors.

**AI is Expanding the Boundaries of Computer Vision**

Through the design and development of specialized chips, sponsored research, educational outreach, and industry partnerships, Intel is firmly committed to advancing the state of AI to solve difficult challenges in medicine, manufacturing, agriculture, scientific research, and other industry sectors. Intel works closely with government organizations, non-government organizations, educational institutions, and corporations to uncover and advance solutions that address major challenges in the sciences.

“We’re on the cusp of computer vision and deep learning becoming standard requirements for the billions of devices surrounding us every day. Enabling devices with humanlike visual intelligence represents the next leap forward in computing. With Intel® Movidius™ Myriad™ X technology, we are redefining what a VPU means when it comes to delivering as much AI and vision compute power [as] possible, all within the unique energy and thermal constraints of modern, untethered devices.”

– Remi S. El-Ouazzane, Vice President, Chief Operating Officer, Artificial Intelligence Products Group
“Robotics is an interdisciplinary subject, requiring the integration of lots of different components to produce a working system. In the classroom, I need to be able to focus on one component at a time. For example, when talking about localization, I want my students to be able to implement and test localization algorithms without also building sensor drivers, motor controllers, and everything else that is needed to make a robot wander around its environment. ROS lets them do just that: Students can start with a full, working system on a real robot or in simulation, swap-in their localization algorithm, and quickly see the results.”

– Bill Smart, Associate Professor, Oregon State University

The Intel® AI portfolio includes:

**Intel® Xeon® Scalable processors**: Tackle AI challenges with a compute architecture optimized for a broad range of AI workloads, including deep learning.

**Framework Optimization**: Achieve faster training of deep neural networks on a robust scalable infrastructure.

**Intel® Movidius™ Myriad™ X Vision Processing Unit (VPU)**: Create and deploy on-device neural networks and computer vision applications.

**OpenVINO™ toolkit**: Make your vision a reality on Intel® platforms—from smart cameras and video surveillance to robotics, transportation, and more.

**Intel® Distribution for Python**: Supercharge applications and speed up core computational packages with this performance-oriented distribution.

**Intel® Data Analytics Acceleration Library (Intel® DAAL)**: Boost machine learning and data analytics performance with this easy-to-use library.

**Intel® Math Kernel Library (Intel® MKL)**: Accelerate math processing routines, increase application performance, and reduce development time.

**Intel® RealSense™ SDK**: Start coding projects that include Intel® RealSense™ Depth Cameras, using a cross-platform library of components.

**Intel Speech Enabling Developer Kit**: Create commercial solutions that use speech recognition through Amazon Alexa®. Kit components include an 8-mic circular array (DMIC board), dual digital signal processor with an interference engine, and connectors.

For more information, visit the portfolio page: [https://ai.intel.com/technology](https://ai.intel.com/technology).

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request. Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel, the Intel logo, Intel RealSense, Movidius, Myriad, OpenVINO, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

© 2018 Intel Corporation 0918/KP/MESH/PDF 337323-001US